Sant'Anna
Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Published by Pearson

www.pearson.com

ISBN 9788891910776

Contents

Specialized sessions

Accounting for record linkage errors in inference (S2G-SIS) 2

Probabilistic record linkage with less than three matching variables. 3
Tiziana Tuoto and Marco Fortini

Advanced methods for measuring and communicating uncertainty in official statistics 9

A model for measuring the accuracy in spatial price statistics using scanner data. 10
Ilaria Benedetti and Federico Crescenzi

Communication of Uncertainty of Official Statistics. 16
Edwin de Jonge and Gian Luigi Mazzi

Measuring uncertainty for infra-annual macroeconomic statistics. 22
George Kapetanios, Massimiliano Marcellino and Gian Luigi Mazzi

Bayesian methods in biostatistics 27

Network Estimation of Compositional Data. 28
Nathan Osborne, Christine B. Peterson and Marina Vannucci

Using co-data to empower genomics-based prediction and variable selection. 34
Magnus M. Münch, Mirrelijn M. van Nee and Mark A. van de Wiel

Data integration versus privacy protection: a methodological challenge? 40

Statistical Disclosure Control for Integrated Data. 41
Natalie Shlomo

The Integrated System of Statistic Registers: first steps towards facing privacy issues. 47
Mauro Bruno and Roberta Radini

Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics. 53
Fabio Ricciato, Kostas Giannakouris, Albrecht Wirthmann and Martina Hahn

Designing adaptive clinical trials 59

Optimal designs for multi-arm exponential trials. 60
Rosamarie Frieri and Marco Novelli

Education: students' mobility and labour market 66

From measurement to explanatory approaches: an assessment of the attractiveness of the curricula programs supplied by Italian universities. 67
Isabella Sulis, Silvia Columbu and Mariano Porcu

Pull factors for university students' mobility: a gravity model approach. 73
Giovanni Boscaino and Vincenzo Giuseppe Genova

Spatial autoregressive gravity models to explain the university student mobility in Italy. 79
Silvia Bacci, Bruno Bertaccini and Chiara Bocci

Environmental Statistics (GRASPA-SIS) 85
A Time Clustering Model for Spatio-Temporal Data. 86
Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona

Reconstruction of sparsely sampled functional time series using frequency domain functional principal components. 93
Amira Elayouty, Marian Scott and Claire Miller

Methods for High Dimensional Compositional Data Analysis 98
Algorithms for compositional tensors of third-order. 99
Violetta Simonacci

High-dimensional regression with compositional covariates: a robust perspective. 105
Gianna Serafina Monti and Peter Filzmoser

Three-way compositional analysis of energy intensity in manufacturing. 111
Valentin Todorov and Violetta Simonacci

Modern Statistics for Physics Discoveries 117
Identification of high-energy λ-ray sources via nonparametric clustering. 118
Giovanna Menardi, Denise Costantin, and Federico Ferraccioli

Statistical Analysis of Macroseismic Data for a better Evaluation of Earthquakes Attenuation Laws. 124
Marcello Chiodi, Antonino D'Alessandro, Giada Adelfio and Nicoletta D'Angelo

Network Modelling in Biostatistics.....130
Natural direct and indirect relative risk for mediation analysis. 131
Monia Lupparelli and Alessandra Mattei

New issues on multivariate and univariate quantile regression 137
Mixtures of quantile regressions for longitudinal data: an R package. 138
Maria Francesca Marino, Maria Giovanna Ranalli and Marco Alfò

Multivariate Mixed Hidden Markov Model for joint estimation of multiple quantiles. 144
Luca Merlo, Lea Petrella and Nikos Tzavidis

Recent methodological advances in finite mixture modeling with applications (CLADAG-SIS) 150
Aggregating Gaussian mixture components. 151
Roberto Rocci

Local and overall coefficients of determination for mixtures of generalized linear models. 157
Roberto Di Mari, Salvatore Ingrassia and Antonio Punzo

Statistical Analysis of Satellite Data (SDS-SIS) 163
Functional Data Analysis for Interferometric Syntethic Aperture Radar Data Post-Processing: The case of Santa Barbara mud volcano. 164
Matteo Fontana, Alessandra Menafoglio, Francesca Cigna and Deodato Tapete

Recent Contributions to the Understanding of the Uncertainty in Upper-Air Reference Measurements. 170
Alessandro Fassò

Statistical models and methods for Business and Industry 176
Modelling and monitoring of complex 3D shapes: a novel approach for lattice structures. 177
Bianca Maria Colosimo, Marco Grasso and Federica Garghetti

Open data powered territorial planning - Case study: The Turin historical center. 183
Silvia Casagrande, Gianmaria Origgi, Alberto Pasanisi, Martina Tamburini, Pascal Terrien, Tania Cerquitelli and Alfonso Capozzoli

Process optimization in Industry 4.0: Are all data analytics models useful? 189
Alberto Ferrer

Technology and demographic behaviours (AISP-SIS) 195
Internet and the Timing of Births. 196
Maria Sironi, Osea Giuntella and Francesco C. Billari

The Internetization of Marriage: Effects of the Diffusion of High-Speed Internet on Marriage, Divorce, and Assortative Mating. 202
Francesco C. Billari, Osea Giuntella and Luca Stella

Solicited Sessions

Advanced Statistical Methods in Health Analytics 209
Assessing the impact of the intermediate event in a non-markovian illness-death model. 210
Davide Paolo Bernasconi, Elena Tassistro, Maria Grazia Valsecchi and Laura Antolini

Big data and AI: challenges and opportunities in healthcare. 216
Vieri Emiliani, Gian Luca Cattani and Fabrizio Selmi

Statistical methodology for volume-outcome studies. 222
Marta Fiocco and Floor van Oudenhoven

Advances in textual data mining 228
Distance measures for exploring pairs of novels in a large corpus of Italian literature. 229
Matilde Trevisani and Arjuna Tuzzi

Supervised vs Unsupervised Latent Dirichlet Allocation: topic detection in lyrics. 235
Mariangela Sciandra, Alessandro Albano and Irene Carola Spera

Advances in the interaction between artificial intelligence and official statistics 241
Automated Land Cover Maps from Satellite Imagery by Deep Learning. 242
Fabrizio De Fausti, Francesco Pugliese and Diego Zardetto

CROWD4SDG: Crowdsourcing for sustainable developments goals. 248
Barbara Pernici

Permanent Population Census: evaluation of the effects of regional strategies on the process efficiency. The direct experience of Tuscany. 253
Linda Porciani, Luisa Francovich, Luca Faustini and Alessandro Valentini

Capture-recapture methods 259
Bayesian Model Averaging for Latent Class Models in Capture-Recapture. 260
Davide Di Cecco

Combining “signs of life” and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts. 266
Marco Fortini, Antonella Bernardini, Marco Caputi and Nicoletta Cibella

Population size estimation with interval censored counts and external information. 272
Alessio Farcomeni

Changes in environment extremes and their impacts 278
FPCA Clustering of rainfall events. 279
Gianluca Sottile, Antonio Francipane, Leonardo Noto and Giada Adelfio

Trends in rainfall extremes in the Venice lagoon catchment. 285
Ilaria Prosdocimi and Carlo Gaetan

Copulas: models and inference 291
Analysis of district heating demand through different copula-based approaches. 292
F. Marta L. Di Lascio and Andrea Menapace

CoVaR and backtesting: a comparison between a copula approach and parametric models. 298
Michele Leonardo Bianchi, Giovanni De Luca and Giorgia Rivieccio

Estimating Asymmetric Dependence via Empirical Checkerboard Copulas. 304
Wolfgang Trutschnig and Florian Griessenberger

Strong Convergence of Multivariate Maxima. 310
Michael Falk, Simone A. Padoan and Stefano Rizzelli

Data Science: when different expertise meet 316
Bayesian stochastic modelling of the temporal evolution of seismicity. 317
Elisa Varini and Renata Rotondi

Cluster Analysis for the Characterization of Residential Personal Exposure to ELF Magnetic Field. 323
Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Isabelle Magne, Martine Souques, Serena Fiocchi, Marta Parazzini and Paolo Ravazzani

Statistical Assessment and Validation of Ship Response in High Sea State by Computational Fluid Dynamics. 328
Andrea Serani, Matteo Diez and Frederick Stern

Uncertainty Quantification for PDEs with random data using the Multi-Index Stochastic Collocation method. 334
Lorenzo Tamellini and Joakim Beck

Emerging challenges in official statistics: new data sources and methods 340
Small area poverty indicators adjusted using local spatial price indices. 341
Stefano Marchetti, Luigi Biggeri, Caterina Giusti and Monica Pratesi

Smart solutions for trusted smart statistics: the European big data hackathon experience. 347
Francesco Amato, Mauro Bruno, Tania Cappadozzi, Fabrizio De Fausti and Manuela Michelini

The ESSnet Project Smart Surveys: new data sources and tools for Surveys of Official Statistics 353

Factorial and dimensional reduction methods for the construction of indicators for evaluation (SVQS-SIS) 359
A comparison of MBC with CLV and PCovR methods for dimensional reduction of the soccer players' performance attributes. 360
Maurizio Carpita, Enrico Ciavolino and Paola Pasca

A framework of cumulated chi-squared type statistics for ordered correspondence analysis. New tools and properties. 366
Antonello D'Ambra, Pietro Amenta and Luigi D'Ambra

Exploring drug consumption via an ultrametric correlation matrix. 372
Giorgia Zaccaria and Maurizio Vichi

Ranking extraction in ordinal multi-indicator systems. 378
Marco Fattore and Alberto Arcagni

Gender statistics 384
Gender differences in Italian STEM degree courses: a discrete-time competing-risks model. 385
Marco Enea and Massimo Attanasio

Some Challenges and Results in Measuring Gender Inequality. 391
Fabio Crescenzi and Francesco Di Pede

How Deep is Your Plot? Young SIS and deep statistical learning (ySIS)..397
A modal approach for clustering matrices. 398
Federico Ferraccioli and Giovanna Menardi

A Note on Detection of Perturbations in Biological Networks. 404
Vera Djordjilović

Bayesian inference for DAG-probit models. 410
Federico Castelletti

Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process. 416
Sirio Legramanti

Measuring poverty and vulnerability 421
Choosing the vulnerability threshold using the ROC curve. 422
Chiara Gigliarano and Conchita D'Ambrosio

New advances in applications, a Bayesian nonparametric perspective 428
Bayesian Mixture Models for Latent Class Analysis. 429
Raffaele Argiento, Bruno Bodin and Maria De Iorio

Non-Parametric Inference and Forecasting of Functional and Object Data 435
An interpretable estimator for the function-on-function linear regression model with application to the Canadian weather data. 436
Fabio Centofanti and Matteo Fontana

Statistical process monitoring of multivariate profiles from ship operating conditions. 440
Christian Capezza

Prior choice in Bayesian Modelling (SISbayes) 446
Bayesian Learning of Multiple Essential Graphs. 447
Luca La Rocca, Federico Castelletti, Stefano Peluso, Francesco Claudio Stingo and Guido Consonni

Bayesian post-processing of Gibbs sampling output for variable selection. 453
Stefano Cabras

Priors on precision parameters of IGRMF models. 459
Aldo Gardini, Fedele Greco and Carlo Trivisano

Sequence Analysis: methods and applications 465
Internal migration, family formation and social stratification in Europe. A life course approach. 466
Roberto Impicciatore, Gabriele Ballarino and Nazareno Panichella

Socio economic integration of migrants 472
A study on the characteristics of spouses who intermarry in Italy. 473
Agnese Vitali and Romina Fraboni

Statistical Analysis for mobility and transportation 479
A multilevel Analysis of University attractiveness in the network flows from Bachelor to Master's degree. 480
Silvia Columbu and Ilaria Primerano

Analysis of mobility data through a novel Cheng and Church algorithm for functional data. 486
Marta Galvani, Agostino Torti and Alessandra Menafoglio

Bridge closures in a transportation network: analysis of the impacts in the region of Lombardy. 491
Agostino Torti, Marika Arena, Giovanni Azzone, and Piercesare Secchi

Statistical Methods and Applications in Social Network Analysis 496
A clustering procedure for ego-networks data: an application to Italian elders living in couple. 497
Elvira Pelle and Roberta Pappadà

Analysing the mediating role of a network: a Bayesian latent space approach. 503
Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison

Network-time autoregressive models for valued network panel. 509
Viviana Amati

University student mobility flows and network data structures. 515
Maria Prosperina Vitale, Giuseppe Giordano and Giancarlo Ragozini

Statistical Methods in Psychometrics 521
A simple probabilistic model to evaluate questionable interim analysis strategies. 522
Francesca Freuli and Luigi Lombardi

Incorporating Expert Knowledge in Structural Equation Models: Applications in Psychological Research. 528
Gianmarco Altoè, Claudio Zandonella Callegher, Enrico Toffalini and Massimiliano Pastore

Predicting social media addiction from Instagram profiles: A data mining approach. 534
Antonio Calcagnì, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale

Structural entropy based modeling for psychological measurement. 540
Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo

Statistical modelling in environmental epidemiology 546
A Time Varying Coefficient Model to Estimate the Short-Term Effects of Air Pollution on Human Health. 547
Pasquale Valentini, Luigi Ippoliti and Clara Grazian

Joint Analysis of Short and Long-Term Effects of Air Pollution. 551
Annibale Biggeri, Dolores Catelan, Giorgia Stoppa and Corrado Lagazio

Statistical Modelling of Scientific Evidence for Forensic Investigation and Interpretation 557
DNA mixtures with related contributors. 558
Peter J. Green and Julia Mortera

Forensic Statistics: How to estimate life expectancy after injury. 564
Jane L. Hutton

The additional contribution of combining genetic evidence from multiple samples in a complex case. 570
Giampietro Lago

The history of forensic inference and statistics: a thematic perspective. 576
Franco Taroni and Colin Aitken

Topological learning: interpretable representations of complex data.....581
Comparing Neural Networks via Generalized Persistence. 582
Mattia G. Bergomi and Pietro Vertechi

On the topological complexity of decision boundaries. 588
António Leitão and Giovanni Petri

Persistence-based Kernels for Data Classification. 594
Ulderico Fugacci

Topological and Mixed-type learning of Brain Activity. 600
Tullia Padellini, Pierpaolo Brutti, Riccardo Giubilei

Contributed papers and Posters

Bayesian Statistics 607

A Bayesian approach for modelling dependence among mixture densities. 608
Mario Beraha, Matteo Pegoraro, Riccardo Peli and Alessandra Guglielmi

A change of glasses strategy to solve the rare type match problem. 614
Giulia Cereda and Fabio Corradi

A new prior distribution on the simplex: the extended flexible Dirichlet. 620
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

ABC model choice via mixture weight estimation. 626
Gianmarco Caruso, Luca Tardella and Christian P. Robert

An ABC algorithm for random partitions arising from the Dirichlet process. 632
Mario Beraha and Riccardo Corradin

Bayesian Inference of Undirected Graphical Models from Count Data. 638
Pier Giovanni Bissiri, Monica Chiogna and Nguyen Thi Kim Hue

Bayesian IRT models in NIMBLE. 644
Sally Paganin, Chris Paciorek and Perry de Valpine

Bayesian modelling of Facebook communities via latent factor models. 650
Emanuele Aliverti

Bayesian nonparametric adaptive classification with robust prior information. 655
Francesco Denti, Andrea Cappozzo and Francesca Greselin

Choosing the right tool for the job: a systematic analysis of general purpose MCMC software. 661
Mario Beraha, Giulia Gualtieri, Eugenia Villa, Riccardo Vitali and Alessandra Guglielmi

Empirical Bayes estimation for mixture models. 667
Catia Scricciolo

Improving ABC via Large Deviations Theory. 673
Cecilia Viscardi, Michele Boreale and Fabio Corradi

Learning Bayesian Networks for Nonparanormal Data. 679
Flaminia Musella and Vincenzina Vitale

Measuring well-being combining different data sources: a Bayesian networks approach. 685
Federica Cugnata, Silvia Salini and Elena Siletti

Penalising the complexity of extensions of the Gaussian distribution. 691
Diego Battagliese and Brunero Liseo

Predictive discrepancy of credible intervals for the parameter of the Rayleigh distribution. 697
Fulvio De Santis and Stefania Gubbiotti

Small-area statistical estimation of claim risk. 702
Francesca Fortunato, Fedele Greco and Pierpaolo Cristaudo

Subject-specific Bayesian Hierarchical model for compositional data analysis. 708
Matteo Pedone and Francesco C. Stingo

Wasserstein consensus for Bayesian sample size determination. 714
Michele Cianfriglia, Tullia Padellini and Pierpaolo Brutti

Biostatistics 720

A comparison of the CAR and DAGAR spatial random effects models with an application to diabetics rate estimation in Belgium. 721
Vittoria La Serra, Christel Faes, Niel Hens and Pierpaolo Brutti

A functional approach to study the relationship between dynamic covariates and survival outcomes: an application to a randomized clinical trial on osteosarcoma. 727
Marta Spreafico, Francesca Ieva and Marta Fiocco

A Statistical Approach to the Alignment of fMRI Data. 733
Angela Andreella, Ma Feilong, Yaroslav Halchenko, James Haxby and Livio Finos

Adaptive clinical trials: Bayesian decision-theoretic and frequentist approaches for cost-effectiveness analysis. 739
Martin Forster and Marco Novelli

Bootstrap corrected Propensity Score: Application for Anticoagulant Therapy in Haemodialysis Patients. 745
Maeregu W. Arisido, Fulvia Mecatti and Paola Rebora

Combining multiple sources to overcome misclassification bias in epidemiological database studies. 751
Francesca Beraldi, Rosa Gini, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini

Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics. 756
Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Nicola Rares Franco, Tiziana Rancati and Catharine West

Graphical models for count data: an application to single-cell RNA sequencing. 762
Nguyen Thi Kim Hue, Monica Chiogna and Davide Risso

Interregional mobility, socio-economic inequality and mortality among cancer patients. 768
Claudio Rubino, Mauro Ferrante, Antonino Abbruzzo, Giovanna Fantaci and Salvatore Scondotto

PET radiomics-based lesions representation in Hodgkin lymphoma patients. 774
Lara Cavinato, Martina Sollini, Margarita Kirienko, Matteo Biroli, Francesca Ricci, Letizia Calderoni, Elena Tabacchi, Cristina Nanni, Pier Luigi Zinzani, Stefano Fanti, Anna Guidetti, Alessandra Alessi, Paolo Corradini, Ettore Seregni, Carmelo Carlo-Stella, Arturo Chiti and Francesca Ieva

Prediction of late radiotherapy toxicity in prostate cancer patients via joint analysis of SNPs sequences. 780
Nicola Rares Franco, Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Tiziana Rancati and Catharine West

Predictive versus posterior probabilities for phase II trial monitoring. 785
Valeria Sambucini

Profile networks for precision medicine. 791
Andrea Lazzerini, Monia Lupparelli and Francesco C. Stingo

Proton-Pump Inhibitor Provider Profiling via Funnel Plots and Poisson Regression. 797
Dario Delle Vedove, Francesca Ieva and Anna Maria Paganoni

Selecting optimal thresholds in ROC analysis with clustered data. 803
Duc Khanh To, Gianfranco Adimari and Monica Chiogna

Environment, Physics and Engineering 809

A hidden semi-Markov model for segmenting environmental toroidal data. 810
Francesco Lagona and Antonello Maruotti

An experimental analysis on quality and security about green communication. 816
Vito Santarcangelo, Emilio Massa, Davide Scintu, Michele Di Lecce and Massimiliano Giacalone

An improved sensitivity-data based method for probabilistic ecological risk assessment. 822
Sonia Migliorati and Gianna Serafina Monti

Comparing predictive distributions in EMOS. 828
Giummolè Federica and Mameli Valentina

Compositional analysis of fish communities in a fast changing marine ecosystem. 834
Pierfrancesco Alaimo Di Loro, Marco Mingione, Giovanna Jona Lasinio, Sara Martino and Francesco Colloca

FDA dimension reduction techniques and components separation in Fourier-transform infrared spectroscopy. 840
Francesca Di Salvo, Elena Piacenza and Delia Francesca Chillura Martino

Functional Data Analysis for Spectroscopy Data. 846
Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Diego Perugini, Alessandro Pisello, Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis and Simone Vantini

Functional graphical model for spectrometric data analysis. 852
Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini

Local LGCP estimation for spatial seismic processes. 857
Nicoletta D'Angelo, Marianna Siino, Antonino D'Alessandro and Giada Adelfio

Observation-driven models for storm counts. 863
Mirko Armillotta, Alessandra Luati and Monia Lupparelli

Statistical control of complex geometries, with application to Additive Manufacturing. 869
Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio, Piercesare Secchi

Tree attributes map by 3P sampling in a design-based framework. 875
Lorenzo Fattorini and Sara Franceschi

Unsupervised classification of texture images by gray-level spatial dependence matrices and genetic algorithms. 880
Roberto Baragona and Laura Bocci

Finance, business and official statistics 886

A discrete choice approach to analyze contractual attributes in the durum wheat sector in Italy. 887
Stefano Ciliberti, Simone Del Sarto, Giulia Pastorelli, Angelo Frascarelli and Gaetano Martino

A fuzzy approach to the measurement of the employment rate. 893
Bruno Cheli, Alessandra Coli and Andrea Regoli

A proposal to model credit risk contagion using network count-based models. 898
Arianna Agosto and Daniel Felix Ahelegbey

A similarity matrix approach to empower ESCO interfaces for testing, debugging and in support of users' experience. 904
Adham Kahlawi, Cristina Martelli, Lucia Buzzigoli, Laura Grassini

Adding MIDAS terms to Linear ARCH models in a Quantile Regression framework. 910
Vincenzo Candila and Lea Petrella

Company requirements in Italian tourism sector: an analysis for profiles. 916
Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga

Determinants of Firms' Default Risk after the 2008 and 2011 Economic Crises: a Latent Growth Models Approach. 921
Lucio Masserini, Matilde Bini and Alessandro Zeli

Double Asymmetric GARCH-MIDAS model - new insights and results. 927
Alessandra Amendola, Vincenzo Candila and Giampiero M. Gallo

European SMEs and Circular Economy Activities: Evaluating the Advantage on Firm Performance through the Estimation of Average Treatment Effects. 933
Luca Secondi

Financial Spillover Measures to Assess the Stability of Basket-based Stablecoins. 939
Paolo Pagnottoni

Forecasting Banknote Flows in BdI Branches: Speed-up with Machine Learning. 945
Marco Brandi, Monica Fusaro, Tiziana Laureti and Giorgia Rocco

Fully reconciled GDP forecasts from Income and Expenditure sides. 951
Luisa Bisaglia, Tommaso Di Fonzo and Daniele Girolimetto

GLASSO Estimation of Commodity Risks. 957
Beatrice Foroni, Saverio Mazza, Giacomo Morelli and Lea Petrella

Measuring the Effect of Unconventional Policies on Stock Market Volatility. 963
Giampiero M. Gallo, Demetrio Lacava and Edoardo Otranto

Multidimensional versus unidimensional poverty measurement. 969
Michele Costa

Multiple outcome analysis of European Agriculture in 2000-2016: a latent class multivariate trajectory approach. 975
Alessandro Magrini

Nowcasting GDP using mixed-frequency based composite confidence indicators. 981
Maria Carannante, Raffaele Mattera, Michelangelo Misuraca, Germana Scepi and Maria Spano

On the tangible and intangible assets of Initial Coin Offerings. 987
Paola Cerchiello and Anca Mirela Toma

Seasonality variation of electricity demand: decompositions and tests. 993
Luigi Grossi and Mauro Mussini

SMEs circular economy practices in the European Union: Implications for sustainability. 999
Nunzio Tritto, Josè G. Dias and Francesca Bassi

Tax Incentives' Effect on the Provision of Occupational Welfare in Italian Enterprises. 1005
Alessandra Righi

The determinants of eco-innovation: a country comparison using the community innovation survey. 1011
Ida D'Attoma and Silvia Pacei

World ranking of urban sustainability through composite indicators. 1017
Elena Grimaccia, Alessia Naccarato and Silvia Terzi

Machine Learning and Data Science.....1023

A novel approach for Artificial Intelligence through Lorenz zonoids and Shapley Values. 1024
Paolo Giudici and Emanuela Raffinetti

A warning signal for variable importance interpretation in tree-based algorithms. 1030
Anna Gottard and Giulia Vannucci

Assessment of the effectiveness of digital flyers: analysis of viewing behavior using eye tracking. 1036
Gianpaolo Zammarchi, Claudio Conversano and Francesco Mola

At risk mental status analysis: a comparison of model selection methods for ordinal target variable. 1042
Elena Ballante, Silvia Molteni, Martina Mensi and Silvia Figini

Categorical Encoding for Machine Learning. 1048
Agostino Di Ciaccio

Dynamic Quantile Regression Forest. 1054
Mila Andreani and Lea Petrella

Estimating the UK Sentiment Using Twitter. 1059
Stephan Schlosser, Daniele Toninelli and Michela Cameletti

Forecasting local rice prices from crowdsourced data in Nigeria. 1065
Ilaria Lucrezia Amerise and Gloria Solano Hermosilla

Generalized Mixed Effects Random Forest: does Machine Learning help in predicting university student dropout? 1071
Massimo Pellagatti, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

HateViz: a textual dashboard Twitter data-driven. 1077
Emma Zavarrone, Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Nicola Canestrari

How to perform cyber risk assessment via cumulative logit models. 1083
Silvia Facchinetti, Silvia Angela Osmetti and Claudia Tarantola

Machine learning prediction for accounting system. 1087
Chiara Bardelli and Silvia Figini

Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory. 1093
Cristina Davino, Rosa Fabbricatore, Carla Galluccio, Daniela Pacella, Domenico Vistocco, Francesco Palumbo

The weight of words: textual data versus sentiment analysis in stock returns prediction. 1099
Riccardo Ferretti and Andrea Sciandra

Unsupervised Energy Trees: clustering with complex and mixed-type variables. 1105
Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti

Using anchoring vignettes to adjust self-reported life satisfaction: a nonparametric approach leading to a Semantic Differential scale. 1111
Sara Garbin, Serena Berretta, Maria Iannario and Omar Paccagnella

Variable selection for robust model-based learning from contaminated data. 1117
Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy

Variable Selection in Text Regressions: Back to Lasso? 1123
Marzia Freo and Alessandra Luati

Web Usage Mining and Website Effectiveness. 1129
Maria Francesca Cracolici and Furio Urso

Models and methods - Categorical, Ordinal, Rank Data 1135

Aberration for the analysis of two-way contingency tables. 1136
Roberto Fontana and Fabio Rapallo

An investigation of the paradoxical behaviour of **κ**-type inter-rater agreement coefficients for nominal data. 1142
Amalia Vanacore and Maria Sole Pellegrino

Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach. 1148
Luigi Lombardi and Antonio Calcagnì

BOD – min range: A Robustness Analysis Method for Composite Indicators. 1154
Emiliano Seri, Leonardo Salvatore Alaimo and Vittoria Carolina Malpassuti

Comparing classifiers for ordinal variables. 1160
Silvia Golia and Maurizio Carpita

Discovering Interaction Effects Between Subject-Specific Covariates: A New Probabilistic Approach For Preference Data. 1166
Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio, Mark De Rooij and Elise Dusseldorp

Hybrid random forests for ordinal data. 1171
Rosaria Simone and Gerhard Tutz

Model-based approach to biclustering ordinal data. 1177
Monia Ranalli and Francesca Martella

New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution. 1183
Cristina Mollica and Luca Tardella

Non-metric unfolding on augmented data matrix: a copula-based approach. 1189
Marta Nai Ruscone and Antonio D'Ambrosio

Ordinal probability effect measures for dyadic analysis in cumulative models. 1194
Maria Iannario and Domenico Vistocco

Simulated annealing for maximum rater agreement. 1200
Fabio Rapallo and Maria Piera Rogantin

Models and methods – Regression.....1206

A Clusterwise regression method for Distributional-valued Data. 1207
Rosanna Verde, Francisco de A. T. de Carvalho and Antonio Balzanella

A nonparametric approach for nonlinear variable screening in high-dimensions. 1213
Francesco Giordano, Sara Milito and Lucia Maria Parrella

Adjusted scores for inference in negative binomial regression. 1219
Euloge C. Kenne Pagui, Alessandra Salvan and Nicola Sartori

Estimation of the treatment effect variance in a difference-in-differences framework. 1224
Marco Doretti and Giorgio E. Montanari

Exploring multicollinearity in quantile regression. 1230
Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco

Generalized M-quantile random effects model. 1236
Francesco Schirripa Spagnolo and Vincenzo Mauro

Goodness-of-fit assessment in linear quantile regression. 1242
Ilaria Lucrezia Amerise and Agostino Tarsitano

Joint Redundancy Analysis by a multivariate linear predictor. 1248
Laura Marcis and Renato Salvatore

M-quantile regression shrinkage and selection via the lasso. 1254
M. Giovanna Ranalli, Lea Petrella and Francesco Pantalone

New insights into the Conditioning and Gain Score approaches in multilevel analysis. 1260
Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini

Simultaneous confidence regions and curvature measures in nonlinear models. 1265
Claudia Furlan and Cinzia Mortarino

Models and methods – Sampling 1271

Design-based consistency of the Horvitz-Thompson estimator for spatial populations. 1272
Lorenzo Fattorini, Marzia Marcheselli, Caterina Pisani and Luca Pratelli

Empirical likelihood in the statistical matching for informative samples. 1278
Daniela Marella and Danny Pfeffermann

Evaluating a Hybrid One-Staged Snowball Sampling through Bootstrap Method on a Simulated Population. 1284
Venera Tomaselli and Giulio Giacomo Cantone

How optimal subsampling depends on guessed parameter values. 1290
Laura Deldossi and Chiara Tommasi

Indicators for risk of selection bias in non-probability samples. 1296
Emilia Rocco and Alessandra Petrucci

On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design. 1302
Caterina May and Chiara Tommasi

Pseudo-population based resamplings for two-stage design. 1308
Pier Luigi Conti, Daniela Marella and Vincenzina Vitale

Models and methods - Theoretical Issues in Statistical Inference 1314

A new mixture model for three-way data. 1315
Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato

A Sequential Test for the Cpmk Index. 1320
Michele Scagliarini

Probability Interpretations and the Selection of the Most Effective Statistics Method. 1326
Paolo Rocchi

Robust Composite Inference. 1332
Valentina Mameli, Monica Musio, Erlis Ruli and Laura Ventura

Statistical hypothesis testing within the Generalized Error Distribution: Comparing the behavior of some nonparametric techniques. 1338
Massimiliano Giacalone and Demetrio Panarello

Stochastic dependence with discrete copulas. 1344
Fabrizio Durante and Elisa Perrone

Models and methods - Time Series and Longitudinal Data.....1350

Bootstrap test in Poisson–INAR models. 1351
Lucio Palazzo and Riccardo Ievoli

Continuous Time-Interaction Processes for Population Size Estimation. 1357
Linda Altieri, Alessio Farcomeni, Danilo Alunni Fegatelli and Francesco Palini

Longitudinal data analysis using PLS-PM approach. 1363
Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia and Marina Marino

Long-memory models for count time series. 1369
Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto

Combining multiple frequencies in Realized GARCH models. 1375
Antonio Naimoli and Giuseppe Storti

Models with Time-Varying Parameters for Realized Covariance. 1381
Luc Bauwens and Edoardo Otranto

Pitman-Yor mixture models for survival data stratification. 1387
Riccardo Corradin, Luis Enrique Nieto Barajas and Bernardo Nipoti

Prediction is not everything, but everything is prediction. 1393
Leonardo Egidi

The Generalized Dynamic Mixtures of Factor Analyzers for clustering multivariate longitudinal data. 1399
Francesca Martella, Antonello Maruotti and Francesco Tursini

Trends and long-run relations in cointegrated time series observed with noise. 1405
Angelica Gianfreda, Paolo Maranzano, Lucia Parisio and Matteo Pelagatti

Population and society 1411

A dimensionality assessment of refugees' vulnerability through an Item Response Theory approach. 1412
Simone Del Sarto, Michela Gnaldi, Yara Maasri and Edouard Legoupil

Accounting for Interdependent Risks in Vulnerability Assessment of Refugees. 1418
Daria Mendola, Anna Maria Parroco and Paolo Li Donni

Active ageing in China: What are the domains that most affect life satisfaction in the elderly? 1424
Ilaria Rocco

Analyzing the waiting time of academic publications: a survival model. 1430
Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi and Francesco Porro

Clustering of food choices in a large sample of students using university canteen. 1436
Valentina Lorenzoni, Isotta Triulzi, Irene Martinucci, Letizia Toncelli, Michela Natilli and Roberto Barale, Giuseppe Turchetti

Cruise passengers' expenditure at destinations: Review of survey techniques and data collection. 1442
Caterina Sciortino, Stefano De Cantis, Mauro Ferrante and Szilvia Gyimóthy

Educational integration of foreign citizen children in Italy: a synthetic indicator. 1448
Alessio Buonomo, Stefania Capecchi and Rosaria Simone

Estimating the Change in Housework Time of the Italian Woman after the Retirement of the Male Partner: An Approach Based on a Two-Regime Model Estimated by ML. 1454
Giorgio Calzolari, Maria Gabriella Campolo, Antonino Di Pino and Laura Magazzini

First and Second Year Careers of STEM Students in Italy: A Geographical Perspective. 1460
Antonella D'Agostino, Giulio Ghellini and Gabriele Lombardi

Future Scenarios and Support Interventions for the Family: Involving Experts' Participation through a Mixed-Method Research Study. 1466
Mario Bolzan, Simone Di Zio, Manuela Scioni and Morena Tartari

Gender and Monetary Policy Preferences: a Diff-in-Diff Approach. 1472
Donata Favaro, Anna Giraldo and Ina Golikja

Headcount based indicators and functions to evaluate the effectiveness of Italian university education. 1478
Silvia Terzi and Francesca Petrarca

Identify the speech code through statistics: a data-driven approach. 1484
Andrea Briglia, Massimo Mucciardi and Jérémi Sauvage

Inspecting cause-specific mortality curves by simplicial functional data analysis. 1490
Marco Stefanucci and Stefano Mazzuco

Intertemporal decision making and childless couples. 1495
Daniela Bellani, Bruno Arpino and Daniele Vignoli

Italian Households' Material Deprivation: Multi-Objective Genetic Algorithm approach for categorical variables. 1501
Laura Bocci and Isabella Mingo

LI-CoD Model. From Lifespan Inequality to Causes of Death. 1507
Andrea Nigri and Susanna Levantesi

Modeling Well-Being through PLS-SEM and K-M. 1513
Venera Tomaselli, Mario Fordellone and Maurizio Vichi

News life-cycle: a multiblock approach to the study of information. 1519
Rosanna Cataldo, Marco Del Mastro, Maria Gabriella Grassia, Marina Marino and Rocco Mazza

Short-term rentals in a tourist town. 1525
Silvia Bacci, Bruno Bertaccini, Gianni Dugheri, Paolo Galli, Antonio Giusti and Veronica Sula

Sportlstat: a playful activity to developing statistical literacy. 1531
Alessandro Valentini and Francesca Paradisi

Statistical modeling for some features of Airbnb activity. 1537
Giulia Contu and Luca Frigau

Tertiary students with migrant background: evidence from a cohort enrolled at Sapienza University. 1543
Cristina Giudici, Donatella Vicar and Eleonora Trappolini

The Causal Effect of Immigraton Policies on Income Inequality. 1549
Irene Crimaldi, Laura Forastiere, Fabrizia Mealli and Costanza Tortù

The job condition of academic graduates: a joint longitudinal analysis of AlmaLaurea and Mandatory Notices of the Ministry of Labour. 1557
Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci

The joint effect of childcare services and flexible female employment on fertility rate in Europe. 1565
Viviana Cocuccio and Massimo Mucciardi

The Left Behind Generation: How the current Early School Leavers affect tomorrow's NEETs? 1571
Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano and Antonella Rocca

The probability to be employed of young adults of foreign origin. 1577
Alessio Buonomo, Francesca Di Iorio and Salvatore Strozza

The risk of inappropriateness in geriatric wards: a comparison among the Italian regions. 1583
Paolo Mariani, Andrea Marletta, Marcella Mazzoleni and Mariangela Zenga

The role of the accumulation of poverty and unemployment for health disadvantages. 1589
Annalisa Busetta, Daria Mendola, Emanuela Struffolino and Zachary Van Winkle

Unemployment and fertility in Italy. A regional level data panel analysis. 1595
Gabriele Ruiu and Marco Breschi

University drop out and mobility in Italy. First evidences on first level degrees. 1601
Nicola Tedesco and Luisa Salaris

Worthiness-based Scale Quantifying. 1607
Giulio D'Epifanio

Young people in Southern Italy and the phenomenon of immigration: what is their perception? 1613
Nunziata Ribecco, Angela Maria D'Uggento and Angela Labarile

Preface

The COVID-19 pandemic is putting our society under incredible health, emotional, and economic stress. Facing its harmful effects and their uncertainty, the Executive Board of the Italian Statistical Society (SIS) and the Local Organizing Committee, to ensure the highest level of safety for members and delegates, deliberated to cancel the 50th Meeting of the Italian Statistical Society originally planned to be held in Pisa in June 2020 and to postpone the conference to June 2021. The Executive Board and the Local Organizing Committee continue to monitor closely the pandemic evolving situation, and keep the members of SIS and the researchers informed about the potential new dates for the next meeting. To give value to the work of those who prepared their presentation for the conference, the Program Committee decided to publish the volume *Book of short papers - SIS 2020* despite the conference cancellation.

The conference program included 4 plenary sessions, 16 specialized sessions, 24 solicited sessions, 32 contributed sessions and the poster exhibition. Plenary sessions concerned with robust statistics, human longevity, statistical models for climate changes and small area estimation for educational poverty. The meeting had to host also 2 round tables on data privacy and innovation in statistics. Activities focused on topics of interest for a wider audience included two round tables on Teaching Statistics and on the SIS journal Statistical Methods & Applications, and the Stats Under the Stars (SUS6) competition for young statisticians. The SUS6 event attracted many sponsors from statistical, financial and editorial firms as well as numerous students. The conference committee had registered 345 accepted submissions, including 143 to be presented in invited plenary, specialized and solicited sessions, and 202 spontaneously submitted for oral and poster sessions.

This book includes most of the scientific contributions that had to be presented at the 50th Meeting of the Italian Statistical Society. It is organized into 49 chapters corresponding to 15 specialized, 23 solicited sessions, and to 11 general topics for contributed papers and posters. All 268 contributions provide a wide overview of the state-of-the-art of the subjects, from methodological and theoretical contributions, to applied works and case studies. The result is a very lively picture of the Italian statistical community with its international connections.

We would like to thank all contributors for having submitted their work to the conference, the members of the Program Committee and the extra reviewers for their efforts in this difficult period. Although the Conference did not take place, the organization went on until cancellation was decided for safety reasons. It would have been impossible without the joint effort of Università di Pisa, Scuola Superiore Sant'Anna and National Research Council of Pisa. Members these three institutions took part actively in the Local Organizing Committee. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

This book is our contribution to encourage the scientific community and the network of the Italian Statistical Society to go on and transform this difficult period into an opportunity of scientific debate for better statistics in a better world.

Alessio Pollice
Università degli Studi di Bari Aldo Moro
Chair of the Program Committee

Nicola Salvati
Università di Pisa
Chair of the Local Organizing Committee

Francesco Schirripa Spagnolo
Università di Pisa

Program Committee: Alessio Pollice (Chair), Serena Arima, Marilena Barbieri, Alessandra Brazzale, Eugenio Brentari, Alessia Caponera, Antonio Lepore, Antonella Plaia, Tommaso Proietti, MarcoRiani, Nicola Salvati, Pasquale Sarnacchiaro, Mauro Scanu, Manuela Stranges, Valentina Tocchioni, Simone Vantini, Massimo Ventrucci, Paola Vicard, Donatella Vicari.

Local Organizing Committee: Nicola Salvati (Chair), Gaia Bertarelli, Bruno Cheli, Alessandra Coli, Paolo Frumento, Fosca Giannotti, Caterina Giusti, Piero Manfredi, Stefano Marchetti, Lucio Masserini, Vincenzo Mauro, Barbara Pacini, Dino Pedreschi, Francesco Schirripa Spagnolo, Chiara Seghieri.

Organizers of Specialized and Solicited Sessions: Giada Adelfio, Bruno Arpino, Emanuele Aliverti, Nicoletta Balbo, Mara Bernardi, Silvia Bozza, Pierpaolo Brutti. Annalisa Busetta, Michela Cameletti, Carlo Cavicchia, Fabrizio Durante, Leonardo Egidi, Pietro D. Falorsi, Francesco Finazzi, Livio Finos, Stefania Galimberti, Michele Gallo, Caterina Giusti, Francesca Greselin, Alessandra Guglielmi, Francesca Ieva, Tiziana Laureti, Achille Lemmi, Brunero Liseo, Fabio Massimo Lo Verde, Daria Mendola, Roberta Pappadà, Lea Petrella, Alessandra Petrucci, Alessia Pini, Sabrina Prati, Maria Giovanna Ranalli, Davide Risso, Fabrizio Ruggeri, Silvana Salvini, Monica Scannapieco, Francesco Stingo, Luca Tardella, Grazia Vicario, Susanna Zaccarin, Maroussa Zagoraiou.

Specialized sessions

Accounting for record linkage errors in inference (S2G-SIS)

Probabilistic record linkage with less than three matching variables

Record linkage probabilistico con meno di tre variabili di confronto

Tiziana Tuoto and Marco Fortini

Abstract Probabilistic record linkage based on Fellegi-Sunter theory is a methodology for integrating data collected in different sources when a unique common identifier is not available. It requires at least three matching variables are available to identify the probability model. In official statistics, it is emerging the need to join archives even with less than three common variables, this is the case for instance of addresses and business archives of poor quality. For this problem, we compare available common variables by means of string comparators and propose mixtures of continuous and categorical distributions rather than usual the latent class models to estimate linkage probabilities.

Abstract *Il record linkage probabilistico basato sulla teoria di Fellegi-Sunter è una metodologia per integrare dati raccolti in fonti diverse quando non è disponibile un codice identificativo comune univoco. In questo caso, sono necessarie almeno tre variabili di confronto per identificare il modello di probabilità. Nella statistica ufficiale, emerge la necessità di abbinare archivi anche quando sono disponibili meno di tre variabili in comune, come ad esempio nel caso di archivi di indirizzi o di imprese di scarsa qualità. Per questo problema, confrontiamo le variaibli disponibili mediante comparatori di stringhe e proponiamo misture di distribuzioni continue e categoriche piuttosto che i modelli a classi latenti solitamente utilizzati per la stima delle probabilità di abbinamento.*

Key words: Fellegi-Sunter record linkage, mixture models, string metrics

1 Introduction

[1] Marco Fortini, Istat; email: fortini@istat.it
Tiziana Tuoto, Istat; email: tuoto@istat.it

Data linkage is a common practice in National Statistical Offices and in many other Institutions to enlarge and enrich the availability of information without incurring the costs of new surveys and burden to respondents. Nowadays in many National Statistical Institutes a new statistical production system has been established, mainly based on integrated datasets, including both administrative archives and traditional sample surveys. This new statistical production system has been by the large availability of administrative data, often with unique identifiers for the units of interest, and almost unlimited computing and storage capacity. Integrated datasets provide complementary variables for the same units; they make it possible to discover relationships between different types of units (e.g. households and enterprises, households and schools) and to study the changes over time of the units and the variables.

When units unique identifiers are not available or corrupted, e.g. for privacy reasons or for lack of quality in the data sources, the data linkage is a not trivial task. The most widespread methodology to face linkage issues is the probabilistic record linkage. Probabilistic Record Linkage, according to the theory by Fellegi and Sunter (1969) and the implementation proposed by Jaro (1989) requires at least three matching variable to identify the probability model. This minimum number of common variables is easily available when the reference units are people (e.g. names, surnames, date and place of birth, gender). Unfortunately, this is not the case when integration is needed between other reference units, such as addresses or businesses. In this paper, we propose a model for probabilistic record linkage that uses less than three matching variables. The method is applied to link data from the Palestinian Business Census and an administrative business register. The behaviour of the proposed model is discussed by means of a simulation study.

2 Mixture models for probabilistic record linkage

In this section we shortly recall the well-known probability model for record linkage as proposed by Jaro (1989), that requires at least three matching variables are available, to move to the description of our proposal, that allows probabilistic record linkage with only two matching variables. Both models rely on mixture models.

2.1 Probabilistic record linkage

The goal of record linkage is to recognize records referred to the same unit even when this is differently represented in different sources. To fix the idea, let us consider two data sources, file A and file B, of size N_A and N_B, respectively. The whole comparisons between records *(a,b)* from A and B generate a comparison pairs space Ω of size $N_\Omega = N_A \times N_B$. The goal of linkage procedure is to identify in Ω two disjoint sets M and U such that $\Omega = M \cup U$ and $M \cap U = \emptyset$, where M is the set of Matches, i.e. $\omega = (a,b)$ represents the same unit, $a=b$; while U is the set of Non-matches, i.e.

$\omega = (a,b)$ refer to two different units. $a \neq b$. The pairs assignment to the sets M and U is determined on the basis of K common matching variables, and the comparison vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$ reporting the agreement/disagreement between the matching variables. When the comparison on the matching variables admits dichotomous outcome, the vector $\boldsymbol{\gamma}$ assumes 2^K possible patterns. Jaro (1989) firstly approaches the record linkage problem by means of mixture models with latent variables. The not observed variable representing the real matching status is the latent one, to be predicted by observing the results of the comparisons vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$ on the K observed matching variables.

Probabilistic record linkage models the 2^K observed frequencies of pairs of the comparison vector $\boldsymbol{\gamma}$ as a mixture coming from two different distributions, $p_{\boldsymbol{\gamma}} = p \cdot m_{\boldsymbol{\gamma}} + (1-p) \cdot u_{\boldsymbol{\gamma}}$, the distribution of the comparisons $\boldsymbol{\gamma}$ given that the pairs belong to the population of Matches, i.e. $m_{\boldsymbol{\gamma}} = P(\boldsymbol{\gamma} \mid M)$, and the distribution of comparisons $\boldsymbol{\gamma}$ given that the pairs belong to the population of Non-Matches, $u_{\boldsymbol{\gamma}} = P(\boldsymbol{\gamma} \mid U)$. The probability of the set Matches over the comparison space Ω, $p = P(M)$, is the weight of the mixture.

We aim to estimate of the probability of a pair to be a match among those showing the pattern $\boldsymbol{\gamma}$: $\pi_{\boldsymbol{\gamma}} = P(\omega \in \mathrm{M} \mid \boldsymbol{\gamma}), \forall \boldsymbol{\gamma}$. To solve this problem, Jaro (1989) suggests applying the EM algorithm: with at least three variables and under the conditional independence assumption, we obtain estimated distributions $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$. Hence, applying the Bayes rule we can evaluate the probability of a pair to be a match given its pattern $\boldsymbol{\gamma}$:

$$\pi_{\boldsymbol{\gamma}} = (p \cdot m_{\boldsymbol{\gamma}}) / (p \cdot m_{\boldsymbol{\gamma}} + (1-p) \cdot u_{\boldsymbol{\gamma}}).$$

The most likely pairs to be Matches are those with values of $\pi_{\boldsymbol{\gamma}}$ close to 1.

To be identifiable, the model needs at least three matching variables, in this way we have $2^3 = 8$ observed frequencies and need to estimate 7 parameters $p, m_1, m_2, m_3, u_1, u_2, u_3$

2.2 Mixture of Beta and Bernoulli distribution for record linkage

In some real cases, some of them described in the following paragraph, the matching variables are less than three, preventing the application of the previous models. However, it is quite common the matching variables are strings of characters and are compared via string comparators that result in [0, 1] intervals rather than dichotomous outcomes. The most common and widespread string comparators for names are Levenstein, Jaro, Jaro-Winkler, qgram, Jaccard. In principle, to obtain comparison outcome in the range [0,1] we don't need to constrain ourselves to string variables, but the same reasoning can be applied to numeric variables, adopting the most convenient comparison function.

When only two matching variables are available, we propose the following model for record linkage: for each pair $\omega \in \Omega$ we consider a first variable with outcome $\delta_\omega \in [0,1]$ and the other variable with dichotomous outcome $\gamma_\omega \in \{0,1\}$.

The joint probability of observing δ_ω and φ_ω can be still modeled as a mixture of Matches and Non-matches distributions:

$$P(\delta_\omega, \varphi_\omega) = P(\omega \in M)P(\delta_\omega, \varphi_\omega|\omega \in M) + \big(1 - P(\omega \in M)\big)P(\delta_\omega, \varphi_\omega|\omega \in U).$$

Under the usual conditional independence assumption, we can factorize the joint observed probability:

$$P(\delta_\omega, \varphi_\omega) = P(\omega \in M)P(\delta_\omega|\omega \in M)P(\varphi_\omega|\omega \in M)$$

$$+ \big(1 - P(\omega \in M)\big)P(\delta_\omega|\omega \in U)P(\varphi_\omega|\omega \in M).$$

where $P(\omega \in M) = p; \quad P(\delta_\omega|M) = m_\delta(\omega); \quad P(\varphi_\omega|M) = m_\varphi(\omega);$

$$P(\delta_\omega|U) = u_\delta(\omega); \quad P(\varphi_\omega|U) = u_\varphi(\omega).$$

Let us assume the following probability distributions:

- $m_\delta(\omega) = Beta(\delta_\omega; \alpha_M, \beta_M), \quad \alpha_M, \beta_M > 0;$
- $u_\delta(\omega) = Beta(\delta_\omega; \alpha_U, \beta_U), \quad \alpha_U, \beta_U > 0;$
- $m_\varphi(\omega) = Bernoulli(\ \ _\omega; m_\varphi), \quad m_\varphi \in [0,1];$
- $u_\varphi(\omega) = Bernoulli(\varphi_\omega; u_\varphi), \quad u_\varphi \in [0,1].$

This allow us to write the compete likelihood, which includes also the latent variable, which can be factorized for the parameters $p, m_\varphi, u_\varphi, \alpha_M, \beta_M, \alpha_U, \beta_U$ and solved via an EM algorithm. Dealing with Beta distributions, the EM algorithm requires the solution of a system of partial derivatives involving non linear equations, its description and the related algebra can be provided in Appendix.

3 An application to real data and a simulation

The method proposed above might be applied when less than three matching variables are available, as it is the case, e.g. in the linkage of some residual addresses in the Italian Statistical Addresses Register. In this section, we propose a real-case application to the first version of the Palestinian Statistical Business Register, built through the linkage of several statistical and administrative registers. Moreover, to understand the behaviour of the proposed modelling, we show the results of a simulation with fictitious data where the linkage status is known.

Among others, the Palestinian Statistical Business Register links the 2017 Palestinian Business Census, managed by the Palestinian Central Bureau of Statistics PCBS and the Municipality Business Archive, managed by each Municipality for administrative reasons, as e.g. the delivery of services such as water, electricity, etc. For the municipality of Salfit, the census counts 662 establishments and the Salfit municipality archive reports 394 establishments. The data sets have some common

variables, i.e. owner name, kind of activity, address; unfortunately the kind of activity is not coded in the same way, preventing the comparison of the reported information, and the address are often missing or registered in an incomparable way, referring to rural areas. This implies the only available matching variable is the owner name. This info from the Salfit municipality archive can be compared to both the owner name and the commercial name in Census. In this exercise, we model the linkage process as follows: the Jaccard distance between owner names in both sources is modelled as a mixture of two Beta for M and U, the Jaro-Winker distance between owner name and commercial name is dichotomised and modelled as a Bernoulli. The proposed model identifies 198 matches, with $\pi_{\gamma} > 0.5$, whilst the deterministic linkage performed at PCBS mainly via manual inspection identifies 171 matches. Manual check to evaluate the goodness of the additional matches is not trivial, due to the Arabic language of the reported information.

To facilitate the performance evaluation of the proposed method a simulation is performed, using public synthetic data for which the true match status is known, i.e the dataset created for the ESSnet Data Integration project (Essnet DI, 2011). The database consists of over 26000 records, with matching variables such as names, dates of birth and addresses. The matching variables contain simulated missing values and typos, mimicking those encountered in reality. From this database, 100 samples of size 1000 are independently selected, each by simple random sampling without replacement. From each sample of 1000 units, two files A and B are independently created by Bernoulli sampling with probability of selection $p_A = 0.93$ and $p_B = 0.92$, respectively, i.e. the two files to be linked are of sizes $N_A = 930$ and $N_B = 920$ on average over the 100 replications, and the number of true matches between them is 858 on average.

Three linkage models are compared: the first (*mod1*) models the Jaccard distance on *Surname* with Beta distribution and Bernoulli distribution for the matching variable *Year of Birth*. In the second model (*mod2*) we create a new variable pasting "*Surname*" and "*Name*" and model the Jaccard distance on the *SurnameName* variable with Beta distribution and again Bernoulli distribution for the matching variable *Year of Birth*. This second model aims at introducing more information into the linkage, so to apply a standard procedure in the third model (*mod3*) where we consider the traditional Fellegi-Sunter model based on the three variables, *Name*, *Surname* and *Year of Birth*. It's worthwhile noting that we introduce a constraint on the Beta distribution for *mod2*, i.e. we fix the parameters of the Beta distribution for the M set to be $\alpha_M = \beta_M = 0.5$, in order to evaluate the modelling adequacy of the Beta distribution for linkage purpose, comparing results from *mod1* and *mod2*. Some results from the simulation are shown in table 1, where match rate and false match rate are reported, averaging over the 100 simulation. The simulation results allow for deeper analysis, presented in Appendix.

Table 1: *Results from the simulation*

Linkage Model	***Match rate***	***False Match rate***
Mod1	0.77	0.28
Mod2	0.82	0.05

Mod3	0.88	0.02

4 Concluding remarks

The proposed method seems a valid alternative to judgmental deterministic record linkage in situations with little information, when less than three matching variables are available.

The methodology can be extended in several ways, based on the specific features of the real data. A possible extension is presented in the simulation, where we compare the results of modelling Beta distributions where all the parameters vary in the parameter space, with Beta distributions whit the parameters constrained to fixed values. Other possible extensions include modelling a mixture of two Beta distributions, as well as extending the mixture of Beta and Bernoulli distributions to the comparison of three (or more) matching variables. To this extent, the simulation provides some insights yet. In the proposed setting, the standard Fellegi-Sunter approach seems exploiting the identification power of the matching variables in a way that over-performing the Beta-Bernoulli mixture, in terms of both match rate and false match rate. Obviously, to some extent, this is due to the use of three variables instead of two. Actually, whether the proposed modelling might substitute the standard approach based on multinomial distribution is still an open question and need further analysis. The intent of this contribution is not to provide an alternative of the Jaro implementation of the Fellegi Sunter model, but rather to provide a probabilistic solution in cases where the standard model is not applicable at all.

References

1. Fellegi, I.P., Sunter, A.B. (1969), A Theory for Record Linkage, Journal of the American Statistical Association, 64, pp. 1183-1210
2. Jaro M.A. (1989) “Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida”, Journal of the American Statistical Association, 89, 414-420.
3. Essnet DI - McLeod, Heasman and Forbes, (2011) Simulated data for the on the job training, http://www.cros-portal.eu/content/job-training

Accounting for record linkage errors in inference (S2G-SIS)

A model for measuring the accuracy in spatial price statistics using scanner data

Un modello per la misurazione dell'accuratezza nell'indice spaziale dei prezzi tramite l'uso degli scanner data

Benedetti Ilaria and Crescenzi Federico

Abstract Scanner data coming from the retail trade outlets of modern distribution has the potential to significantly change how to compile spatial price indexes. Over the last years, the availability of new sources of data have stimulated various research studies for adopting more developed statistical techniques for constructing price indexes and assessing their accuracy. In order to evaluate the accuracy associated to point estimates of sub-national PPPs, the Jackknife replication technique is suggested and an empirical application is provided by ISTAT 2018 Scanner Data for the Tuscany region for the basic heading "mineral water", "coffee" and "dried pasta".

Abstract *L'uso degli scanner data provenienti dai punti vendita al dettaglio della grande distribuzione organizzata rappresentano una sfida cruciale per la compilazione di indici di prezzo spaziali. La disponibilità di questi nuovi dati ha stimolato i ricercatori nell'introduzione di tecniche statistiche al fine di valutare l'accuratezza degli indici dei prezzi. Con l'obiettivo di quantificare l'accuratezza associata alle stime puntuali delle stime sub-nazionali delle PPP, il metodo di replicazione Jackknife è stato suggerito e un applicazione empirica è stata realizzata da ISTAT tramite l'uso dei dati scanner del 2018. Questo studio è limitato alla regione Toscana per gli aggregati di consumo: acqua minerale, caffè e pasta secca.*

Key words: Scanner data, uncertainty, purchasing power parities, jackknife replication method

Benedetti Ilaria
University of Tuscia, Department of Economic, Engineering Business and Society, e-mail: i.benedetti@unitus.it

Crescenzi Federico
University of Florence, Department of Statistics, Computer Science, Applications "G. Parenti" e-mail: federico.crescenzi@unifi.it

1 Introduction

Spatial price indexes (SPIs), that measure the differences in price levels across geographical areas, are essential for comparing real income, standards of living and consumer expenditure patterns. Therefore, it is extremely important to provide accurate point estimates of price differences across space with information about the presence and magnitude of uncertainty. Yet, despite the importance attached to these indicators, no standard errors or reliability measures are computed and published both at an international and national level. Little research has been done for measuring and communicating errors inherent in official spatial price indices (Manski, 2015). Nevertheless, during the last years there has been a growing interest for a more explicit evaluation of the accuracy of official economics statistics (Spiegelhalter et al., 2011) including the problems related to the uncertainty of the Purchasing Power Parities PPPs (Deaton, 2012; Deaton and Aten, 2017; Rao and Hajargasht, 2016). Uncertainty in the PPPs comes, not only from the choice of index number formula, but also from the dispersion of relative prices. A first attempt to solve the accuracy issue in PPPs is provided whithin the stochastic approach (Hajargasht and Rao, 2010) based on the country-product dummy method which can be viewed as a signal extraction problem. This approach derive the index numbers using different formulae used in the computation of PPPs which in turn could be used in the computation of standard errors (Rao and Hajargasht, 2016). Scanner Data (SD) has the potential to improve the accuracy of temporal and spatial price indices thus increasing NSOs' credibility and reputation. The aim of this paper is to contribute to the advancement of spatial price index literature by exploring the issue of evaluating the uncertainty associate to point estimates of sub-national PPPs using SD and the Jackknife replication technique (JRR). The paper is structured as follows. Features of SD in term of accuracy in the computation of spatial price index are presented in Section 2. The replication methodology used for estimating variance in sub-national PPPs is described in Section 3 while in Section 4 some of the results obtained are presented and discussed for a set of product aggregates (called in the international terminology Basic Headings-BHs).

2 The use of SD to improve the accuracy of spatial price indexes

In order to classify the several possible errors which can arise when estimating price indexes, we followed the mean-square error model suggested by Biggeri and Giommi (1987) in which the total error is defined as:

$$MSE = \mathrm{E}[P - \mathrm{E(P)}]^2 + [\mathrm{E(P)} - P'']^2 + [P'' - P^*]^2 + 2[\mathrm{E(P)} - P''][P'' - P^*] \quad (1)$$

Where P is the estimated survey index, E(P) is the expected value of the index, P'' is the defined goal of the index, P^* is the ideal goal of the index. In this model, $E[P - \mathrm{E(P)}]^2$ represents total variance including both sampling and measurement variance.

These errors are due to samples of representative products and point of sales, $[\mathrm{E(P)} - P'']^2$ includes measurement errors resulting from surveys for collecting prices and defining weights, $[P'' - P^*]$ indicates errors arising when the computational formula or the definition of the index are not adequate with the respect to the ideal goal of the index.

The availability of SD, which contain transactions of all goods that have been sold, the prices actually paid by consumers, and the quantities sold for each item code or GTIN[1] may significantly improve the accuracy of SPIs (Laureti and Polidoro, 2018). In this contex, SD has proved to be extremely useful for constructing official price index for temporal comparisons given its extensive coverange of transaction, information on weights, price and characteristics of items.

Unfortunately, a few of theoretical research has been conducted on SPIs in particular no standard errors or reliability measures are computed and published for PPPs. To this aim, the mean square error model, generally applied to temporal price index, should be extended to SPIs in order to evaluate their accuracy.

By using SD it is possible to calculate indices based on a variety of "superlative" index number formulae, including the Fisher ideal index, thus reducing the amount of $[P'' - P^*]$. In this framework, several studies evaluated the effect of using different price index formulae based on SD (Imai, Diewert e Shimizu, 2015; Laureti and Polidoro, 2018).

Measurement errors $[\mathrm{E(P)} - P'']^2$ are also reduced when using SD for replacing on-field collected prices thanks to the increased number of products priced, the improved territorial and population coverage (prices may be collected in each city across the province and not only in the provincial capital). In contrast, the traditional basket is a relatively small sample of the complete universe of goods and quantities sold are not available. Moreover, the use of unit value as "price", calculated as the total expenditure for that item code divided by the total quantities sold, represents a more accurate measure of transaction price than an isolated price quotation (Diewert 1995) moreover, it is a representative price paid by consumers over the reference period, thus reducing conceptual uncertainty.

The accuracy of a price index depends also on the selection of representative items: with SD it is possible to obtain probabilistic sample for products sold in modern retail chains. Empirical approaches for measuring variance for temporal price index are based on the use of replication technique repeated sampling techniques from a model population (Heravi and Morgan, 2014).

The jackknife method of variance estimation could provide a way to get index variability information by resampling from a single sample. This is the aim of our paper, the innovative contribution of our paper is to estimate the accuracy of PPP obtained by using ISTAT SD.

Since January 2018 Italy has been using SD for compilying official CPI. However, standard error are not computed even if a measure of transitory uncertainty is published in new releases. lmost a third of EU countries are using SD for compiling

[1] Global Trade Item Number is the current name of the barcode and it identifies a unique product over time and space.

CPIs even if using different methos. However, little is done to used SD for spatial price comparison.

3 Jackknife replications for computing PPP uncertainty

Originally introduced as a technique of bias reduction, the Jackknife method has by now been widely tested and used for variance estimation (Durbin, 1959). Efron and Stein (1981) and Efron (1982) provide a discussion of the Jackknife methodology. Verma (1993) and Verma and Betti (2011) provide a general description of JRR and other practical variance estimation methods in large-scale surveys. Like other resampling procedures, the JRR method estimates the sampling error from comparisons among sample replications which are generated through repeated resampling of the same parent sample. Each replication needs to be a representative sample in itself and to reflect the full complexity of the parent sample. The JRR variance estimates take into account the effect on variance of aspects of the estimation process which are allowed to vary from one replication to another. In principle, these can include complex effects such as those of imputation and weighting. The basic JRR model which shall be adopted in this work can be summarized as follows. Consider a design in which two or more primary units have been selected independently from each stratum in the population. As in the case of the linearization approach, subsampling of any complexity may be involved within each PSU, this does not affect the variance computation formulae. In the standard 'delete one-PSU at a time Jackknife' version, each JRR replication is formed by eliminating one sample PSU from a particular stratum at a time and increasing the weight of the remaining sample PSU's in that stratum appropriately so as to obtain an alternative but equally valid estimate to that obtained from the full sample. This procedure involves creating as many replications as the number of primary units in the sample.

Let j be a subscript to indicate a sample PSU and let k indicate its stratum; moreover, let $a_k \geq 2$ be the number of PSU in stratum k, assumed to be selected independently. Let λ be a full sample estimate of any complexity, and $\lambda_{(kj)}$ the estimate obtained after eliminating primary unit j in stratum k and increasing the weight of the remaining $a_k - 1$ units in that stratum Also, let $\lambda_{(k)}$ be the simple average of the $\lambda_{(kj)}$ over the a_k values of j in k. The variance of λ is then estimated as follows:

$$var(\lambda) = \sum_k \left[(1 - f_k) \frac{a_k - 1}{a_k} \sum_j (\lambda_{(kj)} - \lambda_{(k)})^2 \right] \tag{2}$$

Where $(1 - f_k)$ is the finite population correction which in typical social surveys is approximately equal to 1.

Under quite general conditions for the application of the procedure, the same and relatively simple variance estimation formula (**??**) holds for λ of any complexity. This in fact is the major attraction of the JRR method for practical application.

In this framework we use the rationale behind JRR to estimate the uncertainty due to the selection of the retail trade outlets as well as the sampling error component in (1).

4 Data and results

The analysis has been carried out by ISTAT using 2018 SD for all outlets of the Tuscany region. Annual provincial average prices which are obtained by aggregating the weekly price of each GTIN code by considering outlet-type (hypermarket or supermarket of a specific chain) and modern distribution chains for the Tuscany region. The dataset consists in 85,345 annual price quotes from the ten Tuscany provincial capitals concerning the eight most important modern distribution chains. In order to illlustrate the potential of the suggested methodology, in this analysis the BHs for Mineral water, Coffee and Pasta are used. The Eurostat-OECD methodology for computing international PPPs has been used, in which it does not consider real weights for items at BH level. Laspeyres and Paasche indexes has been computed by using expenditure share for each product sold in both provinces in the comparison.

$$P_{jk}^{L} = \frac{\sum_{i \in N_{jk}} p_{ik} q_{ij}}{\sum_{i \in N_{jk}} p_{ik} q_{ij}}, \quad P_{jk}^{P} = \frac{\sum_{i \in N_{jk}} p_{ik} q_{ik}}{\sum_{i \in N_{jk}} p_{ij} q_{ik}}, \quad P_{jk}^{F} = \sqrt{P_{jk}^{L} \times P_{jk}^{P}} \tag{3}$$

Then Fisher index are computed. In order to garantee the transitivity property we apply the Elteto-Köves-Szulc (EKS) method:

$$P_{jk}^{GEKS-FISHER} = \prod_{l=i}^{M} \left[P_{jl}^{F} \times P_{lk}^{F} \right]^{1/M} \tag{4}$$

With the aim to obtaining variance estimation of the calculated PPP, the jackknife replications are performed by setting the outlet as PSU and strata to provinces. Results in Table 1 show little variation among provinces.

Table 1 Estimated PPPs for Provinces of Tuscany(FI=1.00). JRR estimates of coefficient of variation in parenthesis.

	Provinces								
BHs	AR	GR	LI	LU	MS	PI	PO	PT	SI
Mineral water	0.9884	0.9990	1.0140	1.0124	1.0234	0.9997	0.9950	1.0005	1.0016
	(*0.36%*)	(*0.65%*)	(*0.71%*)	(*0.80%*)	(*0.51%*)	(*0.37%*)	(*0.41%*)	(*0.52%*)	(*0.72%*)
Coffee	0.9984	1.0313	1.0367	1.0117	1.0380	1.0006	0.9996	0.9985	1.0073
	(*0.67%*)	(*0.73%*)	(*0.58%*)	(*0.54%*)	(*1.19%*)	(*0.57%*)	(*0.72%*)	(*0.48%*)	(*1.67%*)
Dried pasta	1.0027	0.9926	1.0092	1.0060	1.0299	1.0013	1.0018	1.0008	1.0152
	(*0.57%*)	(*0.61%*)	(*0.35%*)	(*0.38%*)	(*0.78%*)	(*0.31%*)	(*0.54%*)	(*0.32%*)	(*0.83%*)

Interestingly, the province of Siena (SI) is the one showing highest variation for two out of three BHs, namely coffee and pasta. Regarding mineral water, the province of Lucca (LU) shows the highest variation followed again by the province of Siena. On the contrary, for each BH the lowest variations are to be found in the provinces of Arezzo (AR), Pistoia (PT) and Pisa (PI).

Acknowledgements.

The authors thank Dott. Antonella Simone from ISTAT for the availability of data.

References

1. Biggeri, L., Giommi, A. On the accuracy and precision of the consumer price indices. Methods and applications to evaluate the influence of the sampling of households. Bulletin of the International Statistical Institute, 52, 137-154 (1987)
2. Deaton, A. Calibrating measurement uncertainty in purchasing power parity exchange rates. International Comparison Program (ICP) Technical Advisory Group, Washington, DC, September, 17-18 (2012)
3. Deaton, A., Aten, B. Trying to Understand the PPPs in ICP 2011: Why are the Results so Different?. American Economic Journal: Macroeconomics, 9(1), 243-64 (2017)
4. Diewert, W. E. On the stochastic approach to index numbers . Department of Economics, University of British Columbia, Vol. 1, p. 995 (1995)
5. Durbin, J. A note on the application of Quenouilles method of bias reduction to the estimation of ratios, Biometrika, 46 m pp. 477–480 (1959)
6. Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, (1982)
7. Efron, B. and Stein C., The Jackknife estimate of variance, Ann. Stat. 9 , pp. 586–596 (1981)
8. Hajargasht, G., Rao, D.S.Prasada. Stochastic approach to index numbers for multilateral price comparisons and their standard errors. Rev. Income Wealth **56 (1)**, S32–S58 (2010)
9. Heravi, S., Morgan, P. . A comparison of six sampling schemes for price index construction in a COICOP food group. Applied Economics, 46(22), 2685-2699 (2014)
10. Imai, S., Diewert, E., Shimizu, C. Consumer Price Index Biases (2015)
11. Laureti, T., Polidoro, F. Big data and spatial price comparisons of consumer prices. In 49th Scientific meeting of the Italian Statistical Society (2018)
12. Manski, C. F. Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern, Journal of Economic Literature **53(3)**, 631-653 (2011)
13. Rao D.S. Prasada, Hajargasht G. "Stochastic approach to computation of purchasing power parities in the International Comparison Program (ICP)". Journal of Econometrics, 191(2), pp. 414-425 (2016)
14. Spiegelhalter, D., Pearson, M., Short, I. Visualizing uncertainty about the future. Science (New York, N.Y.), 333(6048):1393–400 (2011)
15. Verma, Sampling Errors in Household Surveys: A Technical Study, INT-92-P80-15E, United Nations Department for Economic and Social Information and Policy Analysis, Statistical Division, New York, (1993)
16. Verma, V., Betti, G. Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. Journal of Applied Statistics 38(8), 1549–1576 (2011)

Communication of Uncertainty of Official Statistics

La comunicazione dell'incertezza delle statistiche ufficiali

Edwin de Jonge and Gian Luigi Mazzi

Abstract Communication of uncertainty of official statistics is of fundamental importance, but is not a common practice. We advocate that uncertainty measures in official statistics should be calculated and communicated and discuss issues in the communication of uncertainty. Best practices in verbal, numerical and visual communication of uncertainty measures in statistics are presented and discussed.
Abstract *La comunicazione dell'incertezza delle statistiche ufficiali ha una importanza fondamentale ma non è pratica comune. Noi riconosciamo che le misure di incertezza nella statistica ufficiale dorebbero essere calcolate e comunicate e discutiamo i problemi di comunicazione dell'incertezza. Inoltre presentiamo e discutiamo le "best practices" nella comunicazione verbale, numerica e visuale delle misure di incertezza in statistica*

Key words: Uncertainty measures, Visualisation, Official Statistics

1 Introduction

Statistics has the aim of producing accurate, precise and reliable measures for a given social, economic and environmental phenomenon. All statistical data should ideally be accompanied by indications of its accuracy, precision and reliability. Usually official statistics are presented and communicated as a singe value without any indication of accuracy, precision and reliability. Official statistics are inherently uncertain since they always include a measurement error which is given by the difference between the true and often latent and unobservable value of the target phe-

Edwin de Jonge
Statistics Netherlands, Henri Faasdreef 312, The Hague, The Netherlands, e-mail: e.dejonge@cbs.nl

Gian Luigi Mazzi
senior consultant e-mail: glmazzi@pt.lu

nomenon and its estimate produced by statisticians. The European level the Code of Practice Principle 12 explicitly mentions that "Sampling errors and non-sampling errors are measured and systematically documented according to the European standards". Despite this, errors have received little attention and little was done to communicate uncertainty.

This has been noted by [2], [6], [3] and [9], who emphasise as headline statistical estimates, are often presented as point estimates, arguably conveying a misleading degree of reliability, without explicitly expressing underlying and inherent uncertainties. It is important to try to provide an answer to a legitimate question: why are official statisticians ignoring the inherent uncertainty associated to the statistics they are producing? Providing an exhaustive answer is out of scope of this paper but we give two main reasons explaining this conservative approach. The first one reflects that official statisticians are worried that, explicitly acknowledging the presence of uncertainty in their statistics, could lower the reputation and credibility of their institution. The second is the assumption of statisticians that their users do not understand uncertainty and to properly deal with uncertainty without being confused. When dealing with uncertainty in official statistics there are three main aspects which need to be carefully considered and analysed. The first one is represented by the identification of the causes of uncertainty. Manski distinguishes three main source of uncertainty: temporary, permanent and conceptual. The second aspect is the measurement of uncertainty, which can be a challenging endeavour especially when trying to measure independently various types of uncertainty. The third challenging aspect is the communication of uncertainty and to this specific topic the rest of this paper is devoted.

2 Various types of communication

The communication of uncertainty in economic and social statistics is of fundamental importance. Not communicating uncertainty of statistics leads to wrong conclusions about their certitude, which leads to subsequent political, financial, social or economic decisions which are not correctly factoring in those uncertainties. Even more so, it would raise questions *why was a certain statistic revised after some time if it was communicated with certitude in the first instance?* and therefore causes distrust in future publications. Communicating uncertainty in an unclear manner however, can also create negative perceptions in users. Consequently uncertainty measures need to be communicated in a clear, transparent, easily understandable, unambiguous and meaningful manner. Communication tools should guide users to have a correct and positive perception of the uncertainty phenomenon helping them to integrate it in their own analysis or decision process. The way of communicating uncertainty also depends on what kind of uncertainty one is concerned about: *direct* and *indirect*, as [9] argues. Direct uncertainty refers to the uncertainty about a fact, number or scientific hypothesis that can be communicated either in quantitative terms , for instance a probability distribution or confidence interval, likelihood ratios

or verbal indication of its probability. On the other hand, indirect uncertainty refers to the quality of the underlying knowledge that forms a basis for the measured number. This will generally be communicated as a list of caveats about the underlying sources of evidence in a qualitative scale. In this paper we are focusing on communicating direct uncertainty. Literature considers three types of communication: verbal, numerical and visual communication which are not necessarily substitutes but often complementary each other. Following [9] we can say that the appropriate format also depends on the medium of communication.

2.1 Verbal communication

Verbal communication aims at providing users with short, standardised messages giving information on the uncertainty level associated to a specific statistical estimates. Verbal communication is widely used for forecasting, because the verbal terms describe probabilities. In policy when a forecasting estimate is used, it is important to describe the certainty of the estimate with a verbal indication. A well known example of forecasts are the climate analyses and global temperature forecasts of the The Intergovernmental Panel on Climate Change (IPCC). The following terms are used in the communication of the IPCC [8] (table 1):

Virtually certain	>99%
Very likely	90% – 99%
Likely	66% – 90%
About as likely as not	33% – 66%
Unlikely	10% – 33%
Very unlikely	1% – 10%
Exceptionally unlikely	< 1%

Table 1 IPCC verbal uncertainty assessments

The main difficulties limiting the use of verbal communication in official statistics is the constitute by the identification of a clear glossary, consistent across languages, as well as of a standardised set of messages.

2.2 Numerical communication

The most common communication method for uncertainty in official statistics is numerical communication. Often they are sampling error indications, but can also include other sources of uncertainty. Some statistical output is pseudo-accurate: the numerical precision of the number is higher then its statistical precision, e.g. the number of farm chicken in 2019 the Netherlands according to Statistics Netherlands was 100 992 944, which seems overly precise. Rounding a statistical estimate to match its statistical precision is therefore a good practice.

Sampling errors are often expressed in standard errors (SE), which is the standard deviation σ of the sampling distribution. When this approximates the normal distribution the true value x of statistical estimate $\hat{x}$ lies within $[\hat{x} - 1.96 \cdot \mathrm{SE}, \hat{x} + 1.96 \cdot \mathrm{SE}]$ with 95% confidence. The 95% confidence interval is very common in many sciences and official statistics. Other common, but currently less used, uncertainty measures include a Bayesian credible interval and a prediction interval. A credible interval is a summary statistic of a Bayesian data analysis in which the posterior probability distribution of a statistic is estimated. A prediction interval expresses the uncertainty of a prediction, which is conceptually different because there is no true value (yet).

When numerical uncertainty measures are communicated by statistical institutes they are most often not presented together with the statistic but described in a footnote or methodological appendix.

While a confidence interval expressed in percentages gives an indication of the size of the uncertainty interval and therefore its error, it does not communicate the probable range in which the true value lies. A good practice for numerical communication of uncertainty is therefore to present the user an interval with lower and upper bound in the same scale as the statistic it self, e.g., 2.3 ($[2.2, 2.4]$) in stead of $2.3 \pm 5\%$, which also allows for asymmetric interval resulting from an advanced statistical method. Which interval measure is used is for most users less important, for them interval indicates the certainty that the producer of the statistics has found for this statistic.

2.3 Visual communication

Visualisation tools have proven to be a powerful tool for displaying and communicating statistics [14, 10, 12, 11, 13]. The visual perception channel of users of statistics allows for detection of data patterns and abnormalities, such as outliers or missing data with ease [15]. Often visualisation of statistics provides a dense data summary, summarising and compacting many data into one single picture. A properly constructed visualisation shows the main message of the data, trends or other data patterns, but also reveals subtle details. Visualisation techniques are a promising method, allowing to show at the same time statistical data together with the associated uncertainty. this can be achieved by incorporating the uncertainty measures into the visualisation tools. In practice there are two main ways to measure and consequently communicating uncertainty: the use of an interval measure (e.g. confidence interval) and probability distribution (e.g. density distribution of errors).

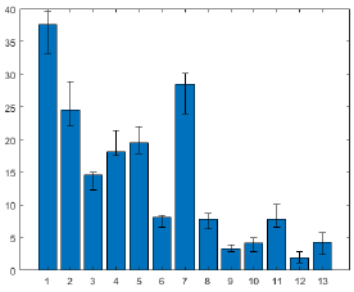


Fig. 1 Bar chart with error bars

Figure 1 shows the most used method in scientific papers of plotting statistical uncertainty: error bar charts.

Each point estimate is visualised with a bar, and the confidence interval is plotted as an interval on top each bar. This is a direct encoding the interval and has the benefits that the visual focus is on the point estimate and that the confidence interval is a detail. It follows the same good practice as for numeric communication, that it shows that range of probable values. The error bar makes it also possible to compare the sizes of the intervals for the different estimates. A disadvantage is that the error bar is visually asymmetric, the lower bound is less visible than the upper bound.

When data are expressed as time series, such as for many macroeconomic indicators, The line chart is the most appropriate and used visualisation tool.

Figure 2 shows the different options to visually add uncertainty to a line chart from the user study in [16].

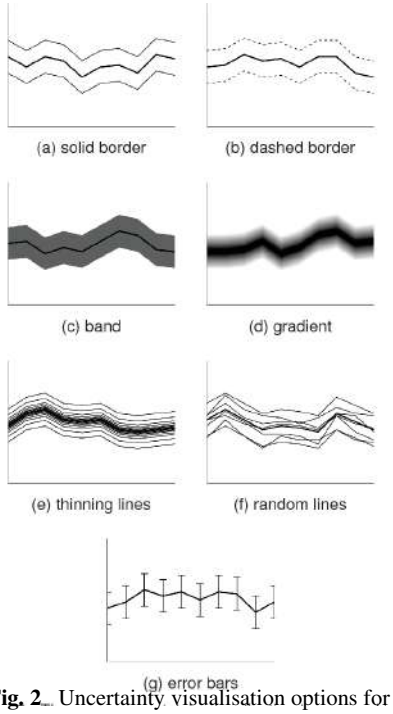


Fig. 2. Uncertainty visualisation options for line charts from [16]

Options a, b, c and g are confidence interval options. All these options ”work” for users, but [9] showed that c) and g) are slightly better in reading off the overall trend of a time series. The paper indicates that the error band is most natural to users and can be used as a good practice for showing time series. When dealing with probability distributions, several methods can use such as diamonds (often used in medical statistics), box and violin plots, giving a more accurate idea of the underlying distribution but not of the uncertainty since the point estimate is missing. We can also use a density plot, which shows the probability of a value or a gradient approach (especially appropriate for time series data) where probability is encoded in transparency, the more likely, the less transparent and more. In the recent year one of the most popularised approach is the fan plot which is designed to show the bounds of several different confidence intervals (often coloured to emphasise the changing probability density going further from the point) and are used by the Bank of England when communicating past and forecasts of future GDP estimates.

3 Conclusions

Uncertainty of official statistics should be communicated in a clear, transparent and understandable manner. Using the best practices for verbal, numerical and visual communication as presented in this paper, should lead to increased and improved publication of uncertainty in official statistics. Contrary to common belief, perception research indicates that common users of statistics have an understanding that measured official statistics have uncertainty and are able to handle communicated uncertainty measures. Not only that uncertainty of statistics is communicated is important, but also how they are communicated.

References

1. A. Cunningham, J. Eklund, C. Jeery, G. Kapetanios, and V. Labhard. A state space approach to extracting the signal from uncertain data. Journal of Business and Economic Statistics, 30:173-180, 2012.
2. Manski Charles F. Communicating uncertainty in official economic statistics: an appraisal fifty years after Morgenstern. Journal of Economic Literature 53:631-653, 2015
3. van der Bles, Anne Marthe and van der Linden, Sander and Freeman, Alexandra LJ and Mitchell, James and Galvao, Ana B and Zaval, Lisa and Spiegelhalter, David J Communicating uncertainty about facts, numbers and science. Royal Society Open Science 6, 2015
4. Morgenstern, Oskar and others On the accuracy of economic observations. Princeton University Press, 1963
5. Simon Kuznets National income: a new version. Review of Economics and Statistics 30:151-179, 1948
6. Charles F Manski. Communicating uncertainty in policy analysis. Proceedings of the National Academy of Sciences, 116(16):7634-7641, 2019.
7. G. Kapetanios, M. Marcellino, Felix Kempfs, g. L. mazzi, Jana Eklund Vincent Labhard Measuring and communicating uncertainty: status of the art and perspectives. Statistical working paper collection Eurostat 2020 forthcoming
8. Mastrandrea, Michael D and Field, Christopher B and Stocker, Thomas F and Edenhofer, Ottmar and Ebi, Kristie L and Frame, David J and Held, Hermann and Kriegler, Elmar and Mach, Katharine J and Matschoss, Patrick R and others Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental panel on climat change 2010
9. van der Laan, D and de Jonge, Edwin and Solcer, Jessica Effect of Displaying Uncertainty in Line and Bar Charts-Presentation and Interpretation. proceedings of the International Conference on Information Visualization Theory and Applications 2:225-232 Citepress, 2015
10. Yau, Nathan Data points: Visualization that means something. John Wiley & Sons, 2013
11. cairo Alberto The Functional Art: An introduction to information graphics and visualization. New Riders, 2012
12. Meirelles, Isabel Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations. Rockport publishers, 2013
13. Robbins, Naomi B Creating more effective graphs. Wiley, 2013
14. Wilke, C.O. Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. O'Reilly Media, 2019
15. Cleveland, William S. Research in statistical graphics. Journal of the American Statistical Association 82:419-423, 1987
16. Susanne Tak and Alexander Toet and Jan van Erp The perception of visual uncertainty representation by non-experts. 2013

Measuring uncertainty for infra-annual macroeconomic statistics

La misura dell'incertezza nelle statistiche macroeconomiche infra-annuali

George Kapetanios *, Massimiliano Marcellino †, Gian Luigi Mazzi ‡

Abstract The Comunikos project launched by Eurostat aims at investigating new methods and tools for measuring and communicating uncertainty in official statistics. This paper, which originates from the Comunikos outcome, describes an approach for measuring and displaying uncertainty in macroeconomic infra-annual statistics together with an application

Abstract *Comunikos e' un progetto lanciato da Eurostat che si propone di investigare nuovi metodi e strumenti per misurare e comunicare l'incertezza nelle statistiche ufficiali. Questo paper, originato dai risultati del progetto Comunikos, presenta una metodologia per misurare l'incertezza associata alle statistiche macroeconomiche infra-annuali assieme ad una applicazione*

Key words: official statistics, uncertainty, signal extraction

1 Introduction

Statistical offices and other public agencies producing statistics usually communicate a variety of official economic and social indicators in general as single values (normally corresponding to the central point estimate), without explicitly mentioning the associated inherent and unavoidable uncertainty. While the technical documentation associated with official data often acknowledges the possible presence of errors, little is done to communicate widely such features. While it is difficult to derive a valid scientific or professional explanation for this circumstance, Manski (2019) argues that one possible reason for this status quo lies in the partly political nature of official statistics. He argues that policy makers or other public agencies may be incentivised to express strong certitude in their communication rather than providing further information about the underlying and inherent uncertainty. However, conveying strong certitude about data or economic analysis can be harmful for the development of public policies in multiple ways. If policy makers incorrectly believe that existing statistical analyses provide an errorless description of the current state of the

* Kings College London, george.kapetanios@kcl.ac.uk

† Bocconi University Milan, massimiliano.marcellino@unibocconi.it

‡ Senior Consultant, glmazzi@pt.lu

economy, they will not take into proper account the underlying uncertainty when taking their decisions. Moreover, communicating official statistics with strong certitude leads to further difficulties because of the way that third parties, especially media, distribute this information to a wider audience, namely by largely taking them at face value, which may lead to further miscommunication. On the other hand official statisticians are worried by the possibility that showing that statistics are affected by uncertainty could lower their credibility. Furthermore, they consider that uncertainty, especially when it is relatively high, could confuse or even mislead policy makers and analysts. This explains, even if it does not justify completely, the traditional conservative position taken by official statistical agencies. Nowadays things are starting to move, even if slowly, and the attention to all aspects related to the uncertainty in official statistics is progressively growing up within statistical institutions. In such a promising context, also Eurostat has decided in 2019 to play an active role contributing to the measurement and communication of uncertainty in official statistics by launching a new research project within its methodological framework contract. This Eurostat project labeled "COMmunicating UNcertainty In Key Official Statistics" (Comunikos) aims at:

- review and categorise the various sources of uncertainty in official statistics and identify their impact on the disseminated data;
- review the methods and metrics used in official statistics as well as in other disciplines to measure uncertainties;
- propose new methods and metrics to present uncertainties based on the review of existing sources and methods;
- develop case studies and empirical applications related to both time series and cross-sectional and survey based statistics;
- provide enhanced recommendations and guidance on measuring and communicating uncertainty in official statistics;

The rest of this paper (which is based on the outcome of the Comunikos project) is devoted to the presentation of the method identified to measure uncertainty in time-series based statistics and of a short description of an empirical application.

2 The Methodology

The methodology is based on state space modelling. Such models provide a natural avenue since they permit the presence of unobserved variables that can proxy for the true process that statistical agencies and other policy making bodies are trying to measure. Such models have an added benefit of allowing for the consideration of particular economic structures that can inform the quantification of conceptual uncertainties. The proposed model is a state space representation of the signal extraction problem following the work of Cunningham et al (2012). Using business surveys and other indirect measures, the model allows for an array of measures of each macroeconomic variable of interest. Then, for each variable of interest, the model comprises alternative indicators, a transition law and separate measurement equations describing the latest official estimates. The model is presented in a vector notation, assuming m variables of interest. However, we simplify estimation by assuming block-diagonal structure throughout the model so that the model can be estimated on a variable-by-variable basis for each of the m elements in turn. Let the m dimensional vector of variables of interest that are subject to data uncertainty at time t be denoted by $\mathbf{y}_t$, $t = 1, \ldots, T$. The vector $\mathbf{y}_t$ contains the unobserved true value of the economic concept of interest. The model for the true data $\mathbf{y}_t$ is given by

$$\mathbf{y}_t = \mu + \sum_{i=1}^{q} \mathbf{A}_i \mathbf{y}_{t-i} + \varepsilon_t, \tag{1}$$

$\mathbf{A}_1, \ldots, \mathbf{A}_q$ are $m \times m$ matrices, $\mathbf{A}(L) = \mathbf{I}_m - \mathbf{A}_1 L - \ldots - \mathbf{A}_q L^q$ is a lag polynomial whose roots are outside the unit circle, μ is a vector of constants, $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{mt})'$ and $\mathrm{E}(\varepsilon_t \varepsilon_t') = \Sigma_\varepsilon$, where we denote the main diagonal of Σ_ε by $\sigma_\varepsilon^2 = (\sigma_{\varepsilon_1}^2, \ldots, \sigma_{\varepsilon_m}^2)'$. We further assume that $\mathbf{A}_1, \ldots, \mathbf{A}_q$ are diagonal. Let $\mathbf{y}_t^{t+n}$ denote a noisy estimate of $\mathbf{y}_t$ published by the statistical agency at time $t+n$, where $n = 1, \ldots T-t$. The model for these published data is

$$\mathbf{y}_t^{t+n} = \mathbf{y}_t + \mathbf{c}^n + \mathbf{v}_t^{t+n} \tag{2}$$

where $\mathbf{c}^n$ is the bias in published data of maturity n and $\mathbf{v}_t^{t+n}$ the measurement error associated with the published estimate of $\mathbf{y}_t$ made at maturity n. One of the main building blocks of the model is the assumption that revisions improve estimates so that official published data become more accurate as they become more mature. Reflecting this assumption, both the bias in the published estimates and the variance of measurement errors are allowed to vary with the maturity of the estimate - as denoted by the n superscript. The constant term $\mathbf{c}^n$ is included in equation (2) to permit consideration of biases in the statistical agency's data set. Specifically, $\mathbf{c}^n$ can be modelled as

$$\mathbf{c}^n = \mathbf{c}^1 (1+\lambda)^{n-1}, \tag{3}$$

where $\mathbf{c}^1$ is the bias in published data of maturity $n = 1$ and λ describes the rate at which the bias decays as estimates become more mature $(-1 < \lambda < 0)$. This representation assumes that the bias tends monotonically to zero as the estimates become more mature. The measurement errors, $\mathbf{v}_t^{t+n}$, are assumed to be distributed normally with finite variance. Serial correlation in $\mathbf{v}_t^{t+n}$ is allowed. Concerning the estimation strategy, we mainly use maximum likelihood via the Kalman filter.

3 Application

Our application is based on a dataset containing data for 5 key macroeconomic indicators: Harmonized Index of Consumer Prices (HICP), Industrial Production index (IPI) Retail Trade Volume Index (RTI), Unemployment Rate (UR) and Gross Domestic Product in volume (GDP) for the Euroarea and its 4 major countries: Germany, France, Italy and Spain. Data start in 2000 with vintages available from March 2018 to March 2019 (13 vintages or releases). Then, we have created a quarterly dataset taking for each monthly variable the average of the 3 months of a given quarter. Data were further transformed in quarterly growth rates except for UR which was transformed into a quarterly difference. As a first consideration, the HICP is almost never revised while the RTI is the most often revised indicator among those considered. Over the 2000-2018 sample, the data for GDP show the mean growth rate to have been highest for Spain and lowest for Italy, with the growth rate for the Euroarea, Germany and France being in between, and quite similar to each other. The data for France have the lowest volatility over the period, with the data for the other countries up to twice as volatile. The data for all countries show a skew to the downside, reflecting the great recession, with the downside skew greatest for Germany, and smallest for Spain. Out of the monthly series, IP and RTI are shown to be somewhat more volatile than GDP, even when considered at the quarterly frequency. In broad terms, the patterns across the countries are similar to what is observed for GDP, with the exception of the volatility of RTI growth which turns out to be higher in the case of Spain than for the other countries. In contrast to GDP growth, which has been positive on average over the 2000-2018 sample for all countries, the growth rates of IP and RTI have not been positive on average for all the countries. IP

growth has been negative on average over the sample for Italy, Spain and France, and RTI growth for Italy and Spain, what might be an indication of a shift in activity from industry and retail trade to other parts of the service sector. The UR, finally, has declined on average over the sample in the cases of Germany and the Euroarea, increasing for the cases of Italy and Spain, and unchanged for France. As GDP growth, the changes in the UR are shown to have the lowest volatility over the sample in the case of France. The highest volatility is recorded in the case of Spain, even though the volatility of real GDP growth for that country has been the lowest across the countries. In terms of revisions, the data suggest that the average revision is of the same order of magnitude for the (quarterly) growth rate of real GDP and RTI, while revisions to the IPI growth rate have tended to be somewhat larger on average, at least for the Euroarea, Germany and France. The volatility of the revisions is quite similar across the countries, with exceptions to the upside for IPI for Italy and Spain, and RTI for Germany. When applying our modelling strategy, due to the limited number of releases we set the initial error variance to 1, the parameter beta in serial correlation to -0.2 implying an AR(1) process for v(t+1). Initial bias in the statistical agency's data set and bias decay were both set to zero, as based on an earlier study by Cunningham et al (2012) for UK data. We have experimented with the decay parameter delta, setting it to -0.01 and -0.05 and with the correlation of the measurement errors with the underlying state of the economy. The correlation was set to -0.5; 0 and 0.5 in turn. The results show that delta=-0.05 and rho=-0.5 perform best in describing the revisions in inflation, GDP for all countries and retail trade for all countries but Spain as measured by Mean Squared Error of the final release and the filtered estimate. For industrial production the same combination of variables performs best for Germany and Italy, while for the remaining three countries moving to a slower decay of delta=-0.01 provides better results. In the case of unemployment the no correlation case for delta=-0.05 works best for all countries but Spain for which delta=-0.01. What is interesting to note about those results is that with respect to the correlation of the measurement error, the best-performing specification is the same (rho=-0.5) for the variables expressed as growth rates - GDP, IP and RTI - and another (rho=0.0) for the UR, expressed as a simple difference, and irrespective of the decay parameter (whether delta=-0.05 or -0.01). This supports the notion that variables tend to display different patterns of data uncertainty, and therefore require different treatment for data uncertainty, depending on whether they are stationary or trending.

4 Conclusions

In this short paper, which summarizes an extensive work carried out within the Comunikos project of Eurostat, Kapetanios et al (2020) we considered a state of the art framework for modelling real-time data and quantifying the uncertainty surrounding them. We applied this framework to a post-crisis sample (2000-2018) for the Euroarea and its four largest countries to provide an extensive set of empirical results. The results indicate that the model-based estimate of the true process in terms of mean squared error (MSE) outperforms the most recent published estimate for all the variables considered. The results further illustrate the sensitivity of the performance of the model in terms of the mean squared error (MSE) with respect to the two model parameters, the decay parameter δ, and the parameter capturing the correlation in measurement error, ρ. Overall, the results suggest that the treatment for data uncertainty is both feasible and important, and that the appropriate choice of parameters is quite relevant to achieve the best empirical performance. One aspect that must be given some consideration in applying the methodology is the nature of the variables - trending or stationary - that are being estimated.

References

1. A. Cunningham, J. Eklund, C. Jeffery, G. Kapetanios, and V. Labhard. A state space approach to extracting the signal from uncertain data. Journal of Business and Economic Statistics, 30:173–180, 2012.
2. Charles F Manski. Communicating uncertainty in policy analysis. Proceedings of the National Academy of Sciences, 116(16):7634–7641, 2019.
3. G. Kapetanios, M. Marcellino, Felix Kempfs, g. L. mazzi, Jana Eklund Vincent Labhard Measuring and communicating uncertainty: status of the art and perspectives. Statistical working paper collection Eurostat 2020 forthcoming

Bayesian methods in biostatistics

Network Estimation of Compositional Data

Stima di un Grafo di Dati Composizionali

Nathan Osborne, Christine B. Peterson, Marina Vannucci

Abstract Network estimation for Gaussian data has been extensively studied in the statistical literature. In this paper, we seek to develop a novel method for compositional count data. We use a hierarchical Bayesian model with latent layers and employ spike-and-slab priors for edge selection. For posterior inference, we utilize the expectation maximization algorithm to enable efficient estimation. Through simulation studies, we demonstrate that the proposed model outperforms existing methods in its accuracy of network recovery. We show the practical utility of our model via an application to microbiome data from the Human Microbiome Project.
Abstract *La stima di grafi per dati Gaussiani è stata ampiamente studiata nella letteratura statistica. In questo contributo, proponiamo un nuovo metodo per i dati di conteggio composizionali. Usiamo un modello gerarchico Bayesiano con strati latenti e a priori spike-and-slab per la selezione degli archi. Per l'inferenza a posteriori, si utilizza l'algoritmo EM che consente una stima efficiente. Attraverso studi di simulazione, dimostriamo che il modello proposto supera i metodi esistenti nell'accuratezza della stima del grafo. Mostriamo l'utilità pratica del nostro modello tramite un'applicazione a dati di microbioma dello Human Microbiome Project.*

Key words: graphical model, EM algorithm, count data, Bayesian hierarchical model, microbiome data

Nathan Osborne & Marina Vannucci
Dept of Statistics, Rice University, Houston, Texas, USA. e-mail: Nathan.Osborne@rice.edu; marina@rice.edu

Christine B. Peterson
Dept of Biostatistics, UT MD Anderson Cancer Center, Houston, Texas, USA. e-mail: CBPeterson@mdanderson.org

1 Introduction

In this paper, we propose a Bayesian hierarchical model for compositional data that allows for estimation of network interactions. In the Gaussian setting, the problem of selecting edges in the graph reduces to the estimation of a sparse inverse covariance matrix, since exact zeros in this matrix, which is also known as the precision matrix, correspond to conditional independence relations. In frequentist settings, penalized likelihood methods, such as neighborhood selection and the graphical LASSO, have been proposed. These methods have also been extended to count data by using data transformations or penalized log-likelihood methods. In Bayesian inference, the G-Wishart prior, which is the conjugate prior that imposes exact zeros in the precision matrix, has been explored by several authors for inference of Gaussian graphical models, but poses significant computational challenges and it is not easily scalable. Alternative shrinkage constructions that employ continuous priors on the off-diagonal elements of the precision matrix have been proposed, including the Bayesian graphical lasso, which relies on double exponential priors, or mixture *spike-and-slab* priors [5], for which efficient expectation conditional maximization methods have been recently proposed [4], that allow scalability.

In our approach, we consider multivariate count data, and specifically compositional data that have a fixed sum constraint. We model the data using a Dirichlet-Multinomial likelihood and then introduce a latent layer by modeling the log concentration parameters via a Gaussian distribution. We allow additional covariates to influence the variable counts and account for this through the mean function. We also capture the dependence relationships of the concentration parameters by estimating the inverse covariance matrix via the shrinkage prior of [5]. For posterior inference, we implement an expectation-minimization (EM) algorithm to estimate the model. This allows us to gain flexibility by using a Bayesian model, while still remaining computationally efficient. We test our model on a simulated dataset and see that it outperforms competing models. We apply the model on data from the gut microbiome from the Human Microbiome Project.

2 Model

Suppose we have observed multivariate counts arranged in an $n \times p$ matrix, $\mathbf{X}$, where p is the number of observed variables measured across n samples. We then let the p-vector $\mathbf{X}_i$ correspond to the measurements for observation i, and the matrix entry $x_{i,j}$ correspond to the j^{th} variable measurement for the i^{th} observation. We also observe q covariate measurements for each of the n observations, with these q additional factors possibly influencing the measured counts for each observation. We arrange this covariate data in an $n \times q$ matrix, $\mathbf{M}$.

We are interested in understanding the conditional dependence relationships among the p variables. We adopt a hierarchical model formulation with a latent Gaussian layer, similarly to [6], as

$$\begin{aligned} \mathbf{Z}_i \mid \mathbf{B_0}, \mathbf{M}_i, \mathbf{B}, \Omega &\sim \text{MVNorm}(\mathbf{B_0} + \mathbf{M}_i\mathbf{B}, \Omega^{-1}) \\ \alpha_i &= \exp\{\mathbf{Z}_i\} \\ \mathbf{h_i} \mid \alpha_i &\sim \text{Dirichlet}(\alpha_i) \\ \mathbf{X_i} \mid \mathbf{h}_i &\sim \text{Multinomial}(\mathbf{h_i}). \end{aligned} \tag{1}$$

In this hierarchical formulation, we introduce a latent normal variable $\mathbf{Z}_i$, which is a direct transformation of the concentration parameter α_i and therefore controls the observed counts $\mathbf{X}_i$. This model has several important features: the Dirichlet-Multinomial likelihood for count data, $\mathbf{X}_i$, allows us to account for overdispersion as well as the compositional nature of the data. The dependence among the $\mathbf{Z}_i$ is captured by the inverse covariance matrix, also known as the precision matrix, Ω. We are also able to control for the influence of additionally observed covariates and the abundance of a covariate with the mean term, $\mathbf{B_0} + \mathbf{M}_i\mathbf{B}$.

Prior on $\mathbf{B_0}$ and B We first introduce the priors on the mean elements of $\mathbf{Z}$, which will allow us to control for additional covariates. We put a non-informative prior on each element of $\mathbf{B_0}$, specifically $B_{0_j} \propto 1$. We then consider a horseshoe-like prior on the elements of $\mathbf{B}$, which will encourage the effects of non-influential covariates to be shrunk to zero. Following the work of [1], we say

$$B_{k,j} \mid \mu_{k,j}, \theta_j \sim N(0, \frac{\theta_j}{2\mu_{k,j}}) \text{ , } p(\mu_{k,j}) = \frac{1 - exp(-\mu_{k,j})}{2\pi_\mu^{1/2}\mu_{k,j}^{3/2}}, 0 < \mu_{k,j} < \infty, \theta_j > 0,$$

where $\mu_{k,j}$ is a shrinkage parameter for each individual regression coefficient, and θ_j a global shrinkage parameter for each coefficient related to the j^{th} variable in $\mathbf{Z}$.

Prior on Inverse Covariance Matrix Next we introduce the prior on the precision matrix Ω, which allows us to learn a sparse association network. We consider the prior introduced by [5] as a prior on the precision matrix entries:

$$\begin{aligned} \pi(\Omega \mid \delta, \nu_1, \nu_0, \lambda) \propto \prod_{i<j} \big\{ (1 - \delta_{i,j}) \text{Normal}(\omega_{i,j} \mid 0, \nu_0^2) + \delta_{i,j} \text{Normal}(\omega_{i,j} \mid 0, \nu_1^2) \big\} \cdot \\ \prod_i \text{Exp}(\omega_{i,i} \mid \lambda/2) \mathbf{1}_{\Omega \in M^+}, \end{aligned} \tag{2}$$

where ν_0 and ν_1 are fixed standard deviations, that assume small and large values respectively, and $\delta_{i,j}$ is a latent variable indicating whether or not an edge is present between nodes i and j. The mixture of normals on the off-diagonal precision matrix entries enables the selection of interactions, represented by edges in a network, since non-zero precision matrix entries reflect conditional dependence relationships. Here, entries reflecting conditional independence relations do not equal exactly zero, but get shrunk to close to zero. The diagonal entries are drawn from a

common exponential prior. The final term in equation (2) expresses a constraint to the space of positive definite matrices M^+. This prior is particularly advantageous in our model, as it allows for efficient estimation via the EM algorithm and leads to less bias in graph estimation than the graphical LASSO, as shown by [4].

We complete the modeling by setting the prior on the graph structure, assuming independent Bernoulli distributions on the inclusion of each edge as follows:

$$p(\delta_{i,j} \mid \pi) \propto \pi^{\delta_{i,j}}(1-\pi)^{1-\delta_{i,j}} \text{ , } \pi \mid a_\pi, b_\pi \sim \text{beta}(a_\pi, b_\pi). \tag{3}$$

Model Estimation via EM We define the objective function to maximize as

$$\begin{aligned} F(\Omega, \delta, \pi, \mathbf{Z}, \mathbf{B}, \mathbf{B}_0) =& p(\mathbf{Z} \mid \mathbf{M}, \Omega, \mathbf{B}) p(\Omega \mid \delta, \nu_1, \nu_0, \lambda) p(\mathbf{B} \mid \mu, \theta) p(\mu) \\ & p(\delta \mid \pi) p(\pi \mid a_\pi, b_\pi) p(\mathbf{X} \mid \alpha). \end{aligned}$$

Following the work of [1, 4], we take the expectation of the objective function in terms of δ and $\mu_{k,j}$. This gives an E step of updating

$$\begin{aligned} E_{\delta \mid \Omega, \pi, \mathbf{Z}, \mathbf{B}, \mathbf{B}_0}[\delta_{i,j}] &= p^*_{ij} \equiv \frac{a_{ij}}{a_{ij} + b_{ij}} \\ E_{\delta \mid \Omega, \pi, \mathbf{Z}, \mathbf{B}, \mathbf{B}_0}[\frac{1}{\nu_0^2(1-\delta_{i,j}) + \nu_1^2 \delta_{i,j}}] &= d^*_{ij} \equiv \frac{1 - p^*_{ij}}{\nu_0} + \frac{p^*_{ij}}{\nu_1} \\ E_{\mu_{j,k} \mid .}[\mu_{j,k} \mid \mathbf{B}, \mathbf{Z}, \theta] &= \tilde{\mu}_{j,k} = \frac{1}{2\pi\theta_j^{1/2}}(\frac{\theta_j}{B^2_{j,k}} - \frac{\theta_j}{B^2_{j,k} - \theta_j}) \end{aligned}$$

where $a_{ij} = p(\omega_{i,j} \mid \delta_{i,j} = 1)\pi$ and $b_{ij} = p(\omega_{i,j} \mid \delta_{i,j} = 0)(1-\pi)$. We can think of the parameter p^*_{ij} as an estimate of the posterior probability that the edge i,j is selected in the network. We use these updated values in the M step to update the precision matrix, Ω and standard deviation parameter θ_j.

In the M step, we first update the centering parameters, $B_{0_j} = \frac{\sum_{i=1}^N Z_{i,j} - \mathbf{M}_i \mathbf{B}_j}{N}$. We next update $\mathbf{B}$, updating each column, $\mathbf{B}_j$, and corresponding θ_j independent of each other column. These updates are

$$B_{j,k} = \Big(\frac{M^T M}{\sigma_j} + \text{diag}(\frac{2\tilde{\mu}_{j,k}}{\theta_j})\Big)^{-1}(\frac{M^T Z_j}{\sigma_j}) \text{ , } \theta_j = \frac{1}{q}\sum_{k=1}^{q} 2\tilde{\mu}_{j,k} B^2_{j,k}$$

Note that σ_j is the standard deviation of the j^{th} column of $\mathbf{Z}$, found by using the properties of the multivariate normal distribution shown in equation (1).

We then perform a column-wise update of the precision matrix, Ω. Consider the following notation and block structures:

$$\Omega = \begin{pmatrix} \Omega_{11} & \omega_{12} \\ \omega_{21}^T & \omega_{22} \end{pmatrix}, (\mathbf{Z} - (\mathbf{MB} + \mathbf{B_0}))^T(\mathbf{Z} - (\mathbf{MB} + \mathbf{B_0})) = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{21} & s_{22} \end{pmatrix}.$$

We can then do a column-by-column update as

$$\omega_{12}^{k+1} = -((s_{22}+\lambda)(\Omega^{k+1})^{-1} + \text{diag}(d_{ij}^*))^{-1})\mathbf{s}_{12}\ ,\ \omega_{22}^{k+1} = \omega_{12}^{k+1}\Omega_{11}^{k+1}\omega_{12}^{k+1} + \frac{n}{\lambda + s_{22}}.$$

The point estimates of π is also updated as $\pi = \frac{(a_\pi + \sum_{i<j} p_{ij}^* - 1)}{(a_\pi + b_\pi + \frac{p(p-1)}{2} - 2)}$.

Finally, the matrix of latent variables can be estimated by finding a point estimate for each entry. This is done by updating each row of the matrix independently. As shown in [6], the objective function to optimize with respect to $\mathbf{Z}$ is

$$\log P(\mathbf{Z} \mid \mathbf{X}, \mathbf{M}, \mathbf{B_0}, \mathbf{B}, \Omega) = -\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\tilde{\Gamma}(\alpha_{ij} + x_{ij}) - \tilde{\Gamma}(s(\alpha_i) + s(\mathbf{X}_i)) -$$

$$\sum_{j=1}^{p}\tilde{\Gamma}(\alpha_{ij}) + \tilde{\Gamma}(s(\alpha_i)) - \frac{1}{2}\log \mid \Omega \mid + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{Z}_i - (\mathbf{B_0} + \mathbf{M}_i\mathbf{B}))\Omega(\mathbf{Z}_i - (\mathbf{B_0} + \mathbf{M}_i\mathbf{B})),$$

where $\tilde{\Gamma}$ is the log-gamma function, and $s(x_i) = \sum_{j=1}^{p}(x_{ij})$. To accomplish optimization of each $\mathbf{Z}_i$ we use the limited-memory quasi-Newton (L-BFGS) algorithm, which is a quasi-newton gradient descent method that makes use of the inverse gradient to direct where to search through the variable space.
The parameters are updated by alternating between the E and M steps, updating each parameter in their respective step. Since Ω is updated via a column wise update, in each M step the updates of Ω are repeated until the estimate of Ω for given iteration has converged. For a more detailed derivation of each E and M step refer to [1, 4].

3 Simulated Data

We simulate data to be similar to microbiome data, and follow a simulation set up similar to [6] but use dimensions n = 350, p = 100, and q = 5. We compare the proposed models performance to SparCC, SpiecEasi, and mLDM [2, 3, 6] which are specifically designed for network discovery of microbiome data.

Results in Table 1 show that in all cases the proposed model performs best in terms of TPR and AUC. In all network structures other than a hub structure, the proposed model also does best in F1 and MCC scores. SpiecEasi and mLDM both did well in regards to FPR, but also selected more sparse networks.

4 Human Microbiome Data

Figure 1 shows an estimated network using microbiome data from the Human Microbiome Project. We find that OTUs of the same phyolgenetic family are more likely to be conditionally dependent on each other and also that OTUs of different families have more negative partial correlations, indicating that those are more likely to be in competiton with each other, as also observed by [3, 6]

Table 1 Simulation results shown as averages across 25 different simulated data sets. Band, Hub, Block, and Random refer to the underlying network structure used to generate the data.

p = 100, n = 350	TPR	FPR	F1	MCC	AUC	TPR	FPR	F1	MCC	AUC
	Band					Hub				
SparCC	0.576	0.097	0.371	0.343	0.645	0.841	0.112	0.224	0.303	0.781
SpiecEasi	0.173	**0.020**	0.233	0.217	0.878	0.371	0.021	0.309	0.298	0.934
mLDM	0.364	0.022	0.430	0.419	0.838	0.719	**0.017**	**0.609**	**0.618**	0.914
Proposed Model	**0.768**	0.049	**0.604**	**0.589**	**0.933**	**0.883**	0.094	0.267	0.350	**0.956**
	Block					Random				
SparCC	0.611	0.092	0.391	0.369	0.687	0.713	0.094	0.413	0.412	0.741
SpiecEasi	0.200	**0.020**	0.261	0.245	0.892	0.193	0.016	0.256	0.245	0.925
mLDM	0.395	0.017	0.477	0.470	0.864	0.435	**0.013**	0.525	0.529	0.910
Proposed Model	**0.810**	0.048	**0.622**	**0.613**	**0.946**	**0.852**	0.067	**0.551**	**0.560**	**0.954**

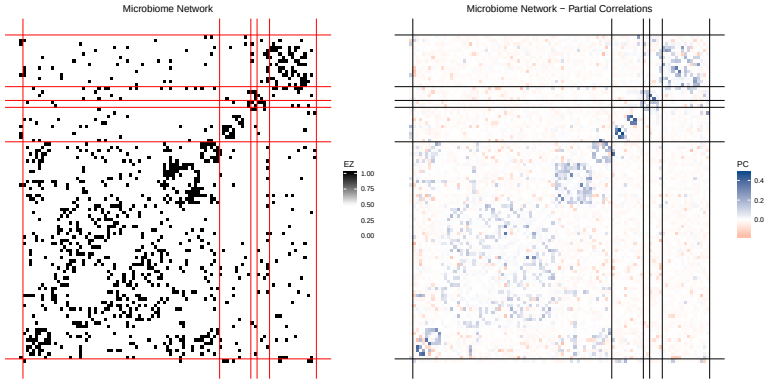


Fig. 1 Results when using data from the Human Microbiome Project. The figure on the left shows the network, with conditional dependence of OTUs indicated with a black point and red lines separating the phyogentic families. The figure on the right shows results from the same data but looks at the partial correlations of each OTU obtained from the estimated precision matrix.

References

1. Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. The horseshoe-like regularization for feature subset selection. *Sankhya B*, 2019.
2. Jonathan Friedman and Eric J. Alm. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):1–11, 2012.
3. Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Computational Biology*, 11(5):1–25, 05 2015.
4. Zehang Richard Li and Tyler H McCormick. An expectation conditional maximization approach for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, To appear, 2019.
5. H. Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.
6. Yuqing Yang, Ning Chen, and Ting Chen. Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical Bayesian statistical model. *Cell Systems*, 4(1):129 – 137, 2017.

Using co-data to empower genomics-based prediction and variable selection

L'uso dei dati complementari per migliorare le previsioni basate su dati genetici e la selezione delle variabili

Magnus M. Münch, Mirrelijn M. van Nee and Mark A. van de Wiel

Abstract Genomics-based prediction is cursed by the dimensionality of the covariates and their collinearity. Fortunately, several sources of complementary data (co-data) are available in the public domain, such as pathway information, p-values from similar studies and genomic annotation. We discuss several types of co-data and how such co-data can be combined to inform the regularization (i.e. the priors) for subsets of the genomics variables in (Bayesian) regression settings. Several priors (and corresponding penalties) are considered, and estimation techniques facilitating efficient computation, such as variational and empirical Bayes, are discussed.
Abstract *La qualità delle previsioni dei modelli per dati genetici è affetta dal problema del dimensionalità e della collinearità presente tra covariate. Fortunatamente, esistono diverse fonti pubbliche di dati complementari (co-data), come pathway data, p-values da studi simili e annotazioni genetiche. In questo lavoro, noi discutiamo diversi tipi di co-data e come tali dati possono essere formalizzati, ad esempio mediante la specificazione di una distribuzione a priori in una regressione Bayesiana. Si studieranno inoltre diverse distribuzioni a priori e le rispettive performance in termini di efficienza computazionale per vari approcci inferenziali quali il variational e empirical Bayes.*

Key words: Co-data, empirical Bayes, variational Bayes, penalized regression

1 Introduction

Clinical research studies the wealth of genomics to predict disease-related outcomes and discover biomarkers of diseases that may be driven by few or many variables. Samples of high-dimensional genomics provide a limited amount of information on the covariates, making prediction and covariate selection difficult. Performance may

Dep. Epidemiology & Data Science, Amsterdam University medical centers, Amsterdam, Netherlands, e-mail: mark.vdwiel@amsterdamumc.nl

be improved when additional information on the covariates, termed *co-data*, is incorporated [8], accessible, for example, from external studies or public repositories.

Ideally, one would like to extract and exploit all relevant information from multiple and various co-data sources. Group-lasso type of methods can handle grouped co-data, penalising covariates in groups to favor group-sparse solutions. Group- and latent overlapping group-lasso [3, 13] form the basis for structured group-lasso penalties like grouped trees [5] and hierarchical groups [12]. As these penalties are parameterised by only one hyperparameter, their flexibility to adapt to the main data is insufficient: non-informative or contradictory grouping information may lead to sub-optimally performing prediction models [8, 2]. Alternative methods like the adaptive lasso [14] exist, but are prone to overfitting, because of the covariate-specific weight that is typically estimated from the same data.

Our focus is on methods that allow co-data-adaptive regularization, in particular methods with a(n) (empirical) Bayesian flavor. We focus on clinical prediction, and therefore on methods that can deal with dichotomous and/or survival response.

2 Co-data

Co-data stands for *co*mplementary *data*, which contains information on the *covariates*; it may originate from any source of information, including data from different or the same samples. In the latter case, it should not include the response labels of those samples, as this may lead to overfitting when training the learner on the main data using co-data adaptive penalties. Below we describe several structures of co-data. Such structures determine the type of penalization induced by the co-data.

i) Non-overlapping groups: the covariates are grouped in non-overlapping groups. An example in clinical genomics are the published gene signatures: a set of genes that is known to play an important role in the development of the disease. This set defines two groups: each gene either belongs to a signature or not; **ii) Overlapping groups:** Pathways are a well-known example of overlapping groups: genes can belong to multiple pathways, which induces overlap between the group defined by the pathways; **iii) Hierarchical or graph-structured groups:** In molecular biology many relations are represented by a graph or hierarchy. If one has exon-level data, a simple hierarchy is: chromosome $\rightarrow$ gene $\rightarrow$ exon. Another example is gene ontology, which represents groups of genes in a directed acyclic graph (DAG). At the top of the hierarchy nodes represent general functions, which are refined in more specific biological functions downwards the DAG; **iv) Continuous co-data:** Some co-data may present itself on a continuous scale. Examples are p-values from external data or correlations with other omics covariates than those used by the learner.

The co-data functionality differs substantially between the methods discussed below. All methods can handle non-overlapping groups, but only some can handle the other co-data structures. In addition, the methods differ in how these are handled, and in the ability to handle *multiple* co-data sources.

3 Regularized regression

Let $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{X} = (X_{ij})_{i=1,j=1}^{n,p}$ be the response vector and high-dimensional covariate matrix, respectively. We consider regularized regression:

$$\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) - f_{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \tag{1}$$

where $\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$ is an appropriate log-likelihood for modeling $\mathbf{y}$ and $f_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is a penalty function, to be specified for the methodologies below. Alternatively, a Bayesian formulation is pursued, in which case a prior $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ plays the role of regularizer. All methods below allow a vector of penalties or hyperparameters $\boldsymbol{\lambda}$, instead of a scalar, where the specific penalty for β_j is modeled by the use of co-data.

4 Co-data informed regression

Group-adaptive ridge regression employs a penalty:

$$f_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \lambda_{g(j)} \beta_j^2, \tag{2}$$

where $g(j) = g$ if $j \in \mathscr{G}_g$, a group based on the co-data, or alternatively $\beta_j \sim N(0, \tau^2_{g(j)})$, with $\tau^2_{g(j)} = C\lambda^{-1}_{g(j)}$. For fixed hyperparameters and suitable constant C, the Bayesian MAP estimate equals the maximum penalized likelihood estimate. Below we discuss two methodologies that employ ridge with empirical Bayes penalty parameter estimation: `GRridge` [8] and `ecpc` [10]. The fully Bayesian method `graper` [11] also handles ridge, and is discussed further on. The crux behind both `GRridge` and `ecpc` is the moment-based empirical Bayes estimation of $\boldsymbol{\lambda}$. Both methods first conventionally estimate a global penalty λ by cross-validation. Then, group penalties are estimated by solving the equations

$$\frac{1}{|\mathscr{G}_g|} \sum_{j \in \mathscr{G}_g} (\hat{\beta}_j^{\lambda})^2 = \frac{1}{|\mathscr{G}_g|} \sum_{j \in \mathscr{G}_g} E[(\hat{\beta}_j^{\lambda})^2],$$

with expectation taken with respect to $\mathbf{y}$ and the prior of β_j, $N(0, \tau^2_{g(j)})$. This renders a linear system of G equations with G unknowns, $(\tau_1^2, \ldots, \tau_G^2)$, denoted by $A\boldsymbol{\tau^2} = \boldsymbol{b}$. Solving this system renders $\boldsymbol{\tau^2}$, so by equivalence, $\boldsymbol{\lambda}$. Overlapping groups are modeled by averaging variances of all groups of which a covariate is a member.

The extension `ecpc` [10] introduces an extra layer of shrinkage on group level. It applies penalised moment-based empirical Bayes estimation for $\boldsymbol{\tau^2}$:

$$\tilde{\boldsymbol{\tau}}^2 = \underset{\boldsymbol{\tau^2}}{\operatorname{argmin}} \, ||A\boldsymbol{\tau^2} - \boldsymbol{b}||_2^2 + f_{pen}\left(\boldsymbol{\tau^2}; \hat{\lambda}_{\tau}\right), \quad \hat{\tau}_g^2 = \max(0, \tilde{\tau}_g^2). \tag{3}$$

The estimate for the hyperpenalty $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\tau}}$ parameterising the penalty function f_{pen} is obtained via a data-driven way of randomly splitting the groups of covariates. The extra level of shrinkage renders a flexible framework as any penalty can be used on the group level, e.g. to shrink group estimates to counter overfitting in the number of groups or to include group structure by using structured penalties on group level. Different co-data sources may demand different penalisation strategies suitable for the type of co-data. The model first estimates group weights independently for each co-data set, using a hypershrinkage strategy suitable for that specific co-data set. Co-data weights are then estimated to integrate multiple co-data by a weighted average.

The following hypershrinkage strategies are allowed: 1) the default L_2 penalty. Such a smooth penalty counters overfitting for an increasing number of groups and handles overlapping, correlated groups. 2) a lasso penalty combined with the default penalty, for group selection; 3) a hierarchical lasso penalty [3, 12], which selects groups under a given hierarchical structure. After selection, the default penalty is used to obtain group weight estimates of the selected groups. This strategy is used to find a well-fitting, data-driven adaptive discretisation of continuous co-data.

Group-regularized elastic net (`gren`) [6] consideres the penalty:

$$f_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \sqrt{\lambda'_{g(j)}}|\beta_j| + \lambda_2 \sum_{j=1}^{p} \lambda'_{g(j)}\beta_j^2, \tag{4}$$

where the global L_1- and L_2-norm penalization parameters λ_1 and λ_2 determine overall penalization of the features, while the group-specific penalty multipliers $\lambda'_{g(j)}$ adaptively shrink the features based on the prior feature group structure. We use the square root in the L_1-norm such that the feature-specific penalty weights have the same effect on the β_j scale: $\sqrt{\lambda'_g}|\beta_j| = |w_g\beta_j|$ and $\lambda'_g\beta_j^2 = (w_g\beta_j)^2$. As in `GRridge` and `ecpc`, the global penalty parameters λ_1, λ_2 are estimated by cross-validation. The λ'_g, are estimated by maximum marginal likelihood empirical Bayes:

$$\hat{\boldsymbol{\lambda}}' = \underset{\boldsymbol{\lambda}'}{\operatorname{argmax}} \operatorname{ML}(\boldsymbol{\lambda}';\mathbf{y}), \text{ with } \operatorname{ML}(\boldsymbol{\lambda}';\mathbf{y}) = \int_{\boldsymbol{\beta}} \exp[\ell(\boldsymbol{\beta};\mathbf{y},\mathbf{X})]\pi_{\boldsymbol{\lambda}'}(\boldsymbol{\beta})d\boldsymbol{\beta}. \tag{5}$$

The Bayesian prior version of the penalty in (4) is obtained by reparameterising the ridge prior as $\tau_j^2 = (\gamma_j^2 - 1)/(\lambda'_{g(j)}\lambda_2\gamma_j^2)$ and adding an extra layer to the prior [4]: $\gamma_j^2 \sim \text{tr-}\Gamma(1/2, 8\lambda_2/\lambda_1^2, (1,\infty))$, where tr-$\Gamma(k,\theta,(x_l,x_u))$ denotes the truncated gamma distribution with shape k, scale θ, and domain (x_l,x_u). Direct maximisation of (5) requires the repeated numerical calculation of a $p-$dimensional integral, which is unfeasible or computationally prohibitive for many modern problems. Indirect maximisation is achieved by repeated expectation-maximization (EM) updates:

$$\boldsymbol{\lambda}'^{(l+1)} = \underset{\boldsymbol{\lambda}'}{\operatorname{argmax}} E_{\boldsymbol{\gamma}^2|\mathbf{y}}[\log \pi_{\boldsymbol{\lambda}'}(\boldsymbol{\gamma}^2)|\boldsymbol{\lambda}'^{(l)}], \tag{6}$$

where the expectation is with respect to the posterior $p(\boldsymbol{\gamma}^2|\mathbf{y})$. This expectation is difficult to compute due to the unavailability of a closed-form posterior. The posterior is therefore approximated with variational Bayes (VB): $Q(\boldsymbol{\beta},\boldsymbol{\gamma}^2) = q(\boldsymbol{\beta})q(\boldsymbol{\gamma}^2) \approx p(\beta,\boldsymbol{\gamma}^2|\mathbf{y})$. The approximate expectation $E_Q \approx E_{\boldsymbol{\gamma}^2|\mathbf{y}}$ is available in closed-form, and leads to an update (6):

$$\boldsymbol{\lambda}'^{(l+1)} = \underset{\boldsymbol{\lambda}'}{\operatorname{argmax}} \frac{1}{2}\sum_{g=1}^{G} |\mathscr{G}_g| \log \lambda'_g - \frac{\lambda_2}{2}\sum_{g=1}^{G} \lambda'_g d_g^{(l)}, \quad \text{s.t.} \prod_{g=1}^{G} (\lambda'_g)^{|\mathscr{G}_g|} = 1, \quad (7)$$

with $d_g^{(k)} = \sum_{j\in\mathscr{G}_g} E_Q[\gamma_j^2\beta_j^2/(\gamma_j^2-1)|\lambda_g'^{(l)}]$. The constraints ensures that overall penalization is determined by the global penalty parameters λ_1, λ_2. The Bayesian elastic net does not give point estimates of the regression coefficients. To produce such point estimates, we refit the regular, frequentist elastic net with fixed λ_1, λ_2, and λ'_g, as estimated by cross-validation and emprical Bayes, respectively. Similarly to `ecpc`, the `gren` penalty in (5) may be extended to include multiple sources of feature groupings by simply modeling the penalty multiplier as a product within both the L_1- and L_2-norm [6].

Group-adaptive spike-and-slab, `graper` [11], is an hierarchical, full Bayes method that allows differential priors based on co-data groups. In short, the likelihood is either Gaussian or Bernoulli (for dichotomous outcome) and the prior for each regression coefficient β_j is a spike-and-slab:

$$\pi(\beta_j) = \pi_{g(j)}\delta_0 + (1-\pi_{g(j)})N(0,\tau^2_{g(j)}). \quad (8)$$

Hence, both the inclusion probability $\pi_{g(j)}$ and the Gaussian prior variance $\tau^2_{g(j)}$ may depend on the group $g(j)$. Then, π_g and τ_g^2 are endowed with their 'default' priors: a uniform and vague inverse-Gamma distribution, respectively. Much of the elegance of `graper` lies in the variational Bayes estimation strategy. The authors supply two versions: one where the posterior of $\boldsymbol{\beta}$ is approximated by a full factorization, and one where a more realistic, but computationally more demanding, multivariate normal is used. The authors show that the latter is more accurate in several applications. Moreover, they show superior performance to group-agnostic methods like (adaptive) lasso and elastic net, as well as the group-lasso. Finally, they report competitive predictive performances of `GRridge` and `graper` for dense settings (applying `graper` without spike and a Gaussian slab).

5 Discussion

We have discussed several co-data adaptive methods. Of these methods `graper` is the only full Bayes method, which may therefore be superior in terms of uncertainty propagation. This may be particulary attractive when uncertainty of predictions is of interest. In [9] a hybrid alternative to empirical Bayes ridge and full Bayes

ridge is discussed: the overall penalty parameter is endowed with a prior, while co-data-based penalty multipliers are estimated by empirical Bayes. A small simulation study shows that the coverage of the prediction intervals is as least as good as that of the empirical Bayes and full Bayes counterparts.

If the underlying data generating mechanism is truly sparse, `graper` (sparse setting) and `gren` likely outperform the ridge-based methods. Although posterior selection methods for the latter exist [1], `graper` and `gren` may be preferred when a (very) sparse model is desired. By far the most flexible method is `ecpc`: it can seemingly deal with multiple co-data sources of various kinds, and addresses hyper-parameter shrinkage. While `graper` and `gren` can only handle linear and logistic regression, `ecpc` and `GRridge` can also handle penalized Cox regression.

Besides `graper`, other hierarchical Bayes methods, such as [7], may provide an alternative solution to include external information for specific settings. These methods, however, typically do not scale well computationally, and often offer limited flexibility in terms of implementation.

References

1. H.D. Bondell and B.J. Reich. Consistent high-dimensional bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.*, 107:1610–1624, 2012.
2. A.-L. Boulesteix, R. De Bin, X Jiang, and M. Fuchs. IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Comp. Math. Meth. Med.*, 2017, 2017.
3. Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.
4. Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.
5. Jun Liu and Jieping Ye. Moreau-yosida regularization for grouped tree structure learning. In *Advances in neural information processing systems*, pages 1459–1467, 2010.
6. M. M. Münch, C. F. W. Peeters, A. W. van der Vaart, and M. A. van de Wiel. Adaptive group-regularized logistic elastic net regression. *Biostatistics*, 2019.
7. M. A. Quintana and D. V. Conti. Integrative variable selection via Bayesian model uncertainty. *Statist. Med.*, 32(28):4938–4953, 2013.
8. M. A. van de Wiel, T. G. Lien, W. Verlaat, W. N. van Wieringen, and S. M. Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statist. Med.*, 35(3):368–381, 2016.
9. M. A. van de Wiel, D. E. te Beest, and M. Münch. Learning from a lot: Empirical Bayes in high-dimensional prediction settings. *Scand. J. Stat.*, pages 1–24, 2018.
10. M. M. van Nee, L. F. A. Wessels, and M. A. van de Wiel. Flexible co-data learning for high-dimensional prediction. *arXiv preprint arXiv:2005.04010*, 2020.
11. B. Velten and W. Huber. Adaptive penalization in high-dimensional regression and classification with external covariates using variational bayes. *arXiv preprint arXiv:1811.02962*, 2018.
12. X. Yan, J. Bien, et al. Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560, 2017.
13. M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. B*, 68(1):49–67, 2006.
14. H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

Data integration versus privacy protection: a methodological challenge?

Statistical Disclosure Control for Integrated Data

Controllo statistico della riservatezza di dati integrati

Natalie Shlomo[1]

Abstract The digitalization of all aspects of our society leads to new and integrated data sources offering unprecedented opportunities for research and evidence-based policies. On the other hand, detailed personal information is more easily accessible leading to increased disclosure risks. Traditional statistical disclosure control (SDC) risk scenarios of identity and attribute disclosures and methods for confidentializng statistical data may no longer be sufficient. This has led to statistical agencies placing more control on the release of the data by restricting and licensing access. Guided by government initiatives for open unrestricted access to statistical data, we explore whether the computer science definition of a differentially private perturbation mechanism can be included in the SDC tool-kit. This implies that users of open-access data will have to work with perturbed data.

Abstract *La digitalizzazione in corso nella società offre l'opportunità per la ricerca di utilizzare nuove fonti di dati ottenute tramite integrazione. Ciò porta a un maggior rischio di identificazione. Gli scenari tradizionali di controllo statistico del rischio (SDC) di identificazione personale o di un suo attributo e i corrispondenti metodi per la creazione di dati statistici anonimizzati possono non essere più sufficienti. Ciò ha indotto gli istituti di statistica a porre sotto maggior controllo il rilascio dei loro dati restringendone l'accesso. Le iniziative governative sull'accesso aperto ai dati statistici ci ha indotto a investigare se il meccanismo di pertubazione noto come differential private perturbation tipico nella comunità computer science possa essere incluso fra gli strumenti di SDC. In questo contesto, gli utenti di dati open-access dovranno usare dati perturbati.*

Key words: Differential Privacy, Table Generation, Remote Analysis, Synthetic Data

1. Introduction

Statistical disclosure control (SDC) for traditional forms of statistical data is well established in the literature. These traditional forms of statistical data include microdata from social surveys, magnitude tables from business statistics and frequency tables for whole-population register/census counts (see Hundepool, et al. 2012 for more details). In this paper, we focus on microdata. SDC methods for protecting microdata arising from social surveys include both 'safe data' and 'safe access' approaches. In terms of 'safe data', survey microdata is generally protected by coarsening the quasi-identifiers, deleting sensitive variables, such as low-level geographies, and top-coding sensitive variables such as the size of the household, income and expenditures. Since social surveys typically have very small sample fractions, statistical agencies generally assume that a potential attacker would not have response knowledge, meaning that the attacker would not know if an individual is included in the survey microdata or not. Sampling therefore provides an inherent

[1] Natalie Shlomo, Social Statistics Department, University of Manchester; email: Natalie.Shlomo@manchester.ac.uk

level of protection and is considered an SDC method in itself. In terms of 'safe access', the survey microdata is generally released into data archives or under special licenses so that users need to undergo an application process and state the purpose of their request prior to gaining access to the data. The disclosure risk scenario for survey microdata is the ability of an attacker to link the microdata to a population database on a set of quasi-identifiers, i.e. identity disclosure, to assess whether a data subject in the microdata is unique in the population. A data subject with rare and unique identifying variables is more likely to be a population unique. Therefore, disclosure risk measures for quantifying identity disclosure are typically matching probabilities based on the quasi-identifying variables. Given the large number of variables for each data subject in the microdata, an identity disclosure leads to attribute disclosure where many new variables can be learnt about the data subject. Denoting F_k the population size in cell k of a table defined by identifying variables having K cells and f_k the sample size and $\sum_{k=1}^{K} F_k = N$ and $\sum_{k=1}^{K} f_k = n$, the set of sample uniques, is defined as: $SU = \{k: f_k = 1\}$. The sample uniques are potential high-risk records since they may be population uniques. Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures as follows (where I is the indicator function): Number of sample uniques that are population uniques: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$; Expected number of correct matches for sample uniques (i.e., a matching probability) $\tau_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$).

When the population is unknown, Elamir and Skinner (2006) propose using a Poisson Distribution to estimate the disclosure risk measures with log-linear modelling to estimate population parameters inferred from the observed sample counts. Skinner and Shlomo (2008) developed goodness of fit criteria for determining the optimal log-linear model which produces unbiased estimates of the disclosure risk measures. Shlomo and Skinner (2010) adapt the estimation of risk measures to take into account measurement and perturbation errors. An extension of the probabilistic modelling by Reiter (2005a) accounts for the probability of re-identification weighted by suppositions on attacker knowledge regarding the methods of perturbation. More recently, Manrique-Vallier and Reiter (2012) used mixed membership models to estimate the probability of re-identification.

With the digitalization of all aspects of our society, statistical agencies are able to develop new, integrated and linked data sources offering unprecedented opportunities for research and evidence-based policies. New data sources can contain whole population sectors, such as in big data, and can be combined and integrated statistically through correlation structures. However, with detailed personal information easily accessible from the internet, traditional SDC disclosure risk scenarios are no longer sufficient and statistical agencies are relying more on restricting and licensing data. One disclosure risk scenario that is growing in prominence with new forms of integrated data is inferential disclosure as it encompasses the traditional disclosure risk scenarios of identity and attribute disclosures. Inferential disclosure risk is the ability to learn new attributes with high probability from combining and manipulating datasets. Inferential disclosure can arise if there is a proportion of some characteristic that is very high within a subgroup or a regression model with very high predictive power. In fact, an individual does not even have to be in the dataset in order to disclose information.

Driven by demand from governments to provide open access data alongside the restricted access data, this increased disclosure risk has led to statisticians investigating more rigorous definitions and approaches for protecting the confidentiality of data subjects in microdata. The computer science definition of differential privacy (Dwork, et al. 2006) is currently under research and development on its use in the SDC tool-kit. Differential privacy is a definition of an output perturbation mechanism that states that no one individual can be revealed and all possible outputs must be indistinguishable up to a degree of epsilon with and without the data of one individual. For data to be differentially private, a stochastic perturbation mechanism is applied under certain parametrizations. Traditional approaches of deterministic mechanisms such as sampling and coarsening are not differentially private (Shlomo and Skinner 2012). In Section 2 we define differential privacy and then provide an overview of some open-access data dissemination methods based on protecting outputs in Section 3. We conclude in Section 4.

2. Differential Privacy

Differential privacy was developed by computer scientists as a standard for a perturbation mechanism for protecting outputs in a remote query-based system with the aim to specifically protect against inferential disclosure. See Dinur and Nissim (2003) and an overview book by Dwork and Roth (2014) for more details on differential privacy. In differential privacy, a 'worst case' scenario is allowed for, in which the potential attacker has complete information about all the units in the database except for one data subject of interest. The definition of a perturbation mechanism M satisfies ε-differential privacy if for all queries on neighbouring databases $a, a'\epsilon A$ differing by one individual and for all possible outputs defined as sub-sets $S\epsilon Range(M)$ we have:

$$P(M(a) \in S) \leq e^{\varepsilon}\ P(M(a') \in S) \quad (1)$$

This means that when observing a perturbed output S, little can be learnt (up to a degree of e^{ε}) and the attacker is unable to decipher whether the output was generated from database a or a'. In other words, the ratio $\frac{P(M(a)\epsilon S)}{P(M(a')\epsilon S)}$ is bounded (the probability in the denominator cannot be zero). The solution to guarantee differential privacy in the computer science literature is by adding noise/perturbation to the outputs of the queries under specific parameterizations. The noise is generated from the Laplace Distribution. In addition, a relaxation was proposed leading to the definition of (ε, δ) differential privacy where δ is a small probability of obtaining an unbounded ratio. This extension allows more utility, for example by limiting the extent of the perturbation.

3. Open-access Dissemination Strategies

Since access to microdata, particularly those arising from integrated and linked data sources, is generally licenced within safe data enclaves, open-access options will rely on generating outputs from the microdata through structured web-based internet platforms that return protected (perturbed) outputs in the form of tables and statistics. Another option for open-access data is to produce synthetic data where data is generated from a model based on the sufficient statistics obtained through the original data. We discuss both approaches here.

3.1 Web-based Dissemination

Online flexible table generating servers allow users to define, generate and download their own tables. A good example is the Australian Bureau of Statistics (ABS) TableBuilder for disseminating census tables. In flexible table generating, users access the servers via the internet and define their own table of interest from a set of pre-defined variables and categories typically from drop-down lists. The generated table undergoes a series of ad-hoc SDC checks. If it passes the criteria, it is confidentialized and then downloaded to the user's PC without the need for human intervention. These SDC checks may include, for example, limiting the number of dimensions in the table, minimum population and sparsity thresholds, ensuring consistent and nested categories of variables to avoid disclosure by differencing. One characteristic of the ABS TableBuilder is that for any cell that is generated from census microdata, the perturbation of the cell value will always be the same. Fraser and Wooton (2005) describe the 'same cell-same perturbation' approach where each individual in the microdata is assigned a random number. Any time data subjects are aggregated to form a cell in a table, their random numbers are also aggregated and this becomes the seed for the perturbation. Therefore, the same cell will always have the same perturbation. This reduces the chance of identifying the true cell value through multiple requests of the table and averaging out the perturbations. According to this setting, all possible tables and all possible cells that can be generated in the flexible table generating server are essentially known in advance and hence can be protected under a given privacy budget ε in (1). This is known as a non-interactive mechanism in the theory of differential privacy. Any post-processing of a differentially private output will still be differentially private.

To produce a table for discrete variables (continuous variables may be discretized for the purpose of producing tables), Rinott et al (2018) propose using a differentially private exponential mechanism (McSherry, et al. 2007) based on a utility function: $u(a, b)$ described as follows:

Given a list of all possible cells $k = 1, \dots, K$: $a = (a_1, \dots, a_K) \epsilon A$ choose output $b = (b_1, \dots, b_K) \epsilon B$ with probability proportional to $\exp\left(\frac{\frac{\varepsilon}{2} u(a,b)}{\Delta u}\right)$ where ε is the privacy budget and the scale is defined as: $\Delta \mathrm{u} = \max_{b \in B} \max_{a, a' \in A} |u(a, b) - u(a', b)|$where a and a' are neighboring databases that differ by removing one individual. Δu is known as the sensitivity in the differential privacy mechanism. The utility function is defined through a loss function: $u(a, b) = -l_1 = \sum_{k=1}^{K} |a_k - b_k|$. Under this definition, for a list of internal cells where a data subject appears only once, the sensitivity Δu is 1. If marginal totals are also included and an individual appears several times in the list then Δu will increase. This mechanism is essentially a discretized Laplace distribution and is optimal for the case of perturbing count data with respect to preserving utility. Furthermore, bounding the perturbations such that $|a_k - b_k| \leq m$ for all k leads to (ε, δ)- differential privacy where δ is the probability at the cap m. Shlomo et al. (2019) compared standard SDC methods and differential privacy for a flexible table builder containing survey-weighted counts. The differential privacy mechanism consisted of perturbing the sample counts and then adjusting the survey-weighted cell count according to the average survey weight in the cell. They showed that for the case of internal cells of tables and relatively large sample counts there was less perturbation required under differential privacy and higher utility compared to the SDC approaches.

It is now recognized that differential privacy for protecting frequency tables can be a viable technique in the SDC tool-kit at statistical agencies. Open questions remain and are subject to further research. Whilst the use of the non-interactive differential privacy mechanism will avoid depleting a privacy budget under multiple generation of tables, how to set this budget and determine the sensitivity of the mechanism given the large-scale dissemination of tables containing internal and marginal cells need careful consideration. In addition, policy makers need to understand the consequences of the privacy parameters ε and δ. Moreover, there is a need to go beyond the flexible table generating server towards a remote analysis server where users can query a dataset for exploratory analysis, measures of association and regression modelling and in return receive a confidentialized output under the differentially private mechanism, smoothed histograms and sequential boxplots for scatterplots and residual analyses.

3.2 Synthetic Data

In recent years, there have been initiatives to produce synthetic microdata as public-use files that preserve some of the statistical properties of the original microdata. This type of open data allows researchers to plan their research questions and data analysis and prepare their code prior to gaining access to the licenced restricted data. In addition, these data can be used for teaching purposes. The data elements are replaced with synthetic values generated from an appropriate probability model that is developed through the sufficient statistics of the original data. Similar to imputation techniques, the synthetic data is generated through a posterior predictive distribution. See Reiter (2005b) and references therein for more details of generating synthetic data. The synthetic data can be implemented on parts of data so that a mixture of real and synthetic data is released although this means that a thorough disclosure risk assessment is needed prior to releasing such data. In practice, it is very difficult to capture all conditional relationships between variables and within sub-populations. If models used in a statistical analysis are sub-models of the model used to generate data, then the analysis of multiple synthetic samples should give valid inferences. The subject of using differential privacy in the production of synthetic data is still undergoing research. One early application for generating synthetic data using a differential privacy mechanism that is embedded in the Bayesian Multinomial-Dirichlet model is the US Census Bureau 'On the Map' available at http://onthemap.ces.census.gov/. It is a web-based mapping and reporting application that shows where workers are employed and where they live according to the Origin-Destination Employment Statistics from the US Census 2001. More information is given in Abowd and Vilhuber (2008). New research is ongoing where differentially private noise is added to the estimating equations used to generate synthetic data through the sequential regression modelling approach. In addition, the regression models are adapted to ridge regression which bounds the parameter space. Early results show that adapting the sequential regression modelling approach to the differentially private standard does not have a significant utility loss compared to the standard regression approaches.

4. Discussion

In recent years, with increasing opportunities for linking and integrating data, statistical agencies are relying more on restricting access to such data due to their inability to ensure the confidentiality of statistical units. However, with government

initiatives for more open and accessible data, statistical agencies are exploring alternative means for disseminating statistical data through web-based applications and synthetic data. Given the rising concerns of inferential disclosure under these new dissemination strategies, this has led to fruitful collaborations between statisticians and computer scientists and initial research on whether the formal 'by-design' privacy guarantee of differential privacy can be embedded in the SDC tool-kit. However, perturbative methods of SDC come at a cost in that researchers will have to cope with the perturbation when carrying out statistical analysis which may require more training. Under differential privacy, the parameters of the privacy mechanism are not secret and can be used to correct statistical analysis of perturbed data and this provides a large incentive for including differential privacy into the SDC tool-kit.

References

1. Abowd, J.M. and Vilhuber, L., (2008). How Protective Are Synthetic Data? In PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 239-246.
2. Dinur, I. and Nissim, K. (2003). Revealing Information While Preserving Privacy. PODS 2003, 202-210.
3. Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.
4. Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, 9, 211-407.
5. Elamir, E. and Skinner, C.J. (2006). Record-Level Measures of Disclosure Risk for Survey Micro-data. Journal of Official Statistics, 22, 525–539.
6. Fraser, B. and Wooton, J. (2005). A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 November.
7. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S.,Spicer, K. and de Wolf, P. P. (2012). Statistical Disclosure Control. John Wiley & Sons, United Kingdom.
8. Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating Identification Disclosure Risk Using Mixed Membership Models. Journal of the American Statistical Association, 107, 1385-1394.
9. McSherry, F. and Talwar, K. (2007). Mechanism Design via Differential Privacy. Foundations of Computer Science, 2007, FOCS'07, 48th Annual IEEE Symposium on 94-103. IEEE, New York.
10. Reiter, J.P. (2005a). Estimating Risks of Identification Disclosure in Microdata. Journal of the American Statistical Association 100, 1103-1112.
11. Reiter, J.P. (2005b), Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. Journal of the Royal Statistical Society, A, Vol.168, No.1, 185-205.
12. Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018). Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statistical Sciences, 33, No. 3 (2018), 358-385.
13. Shlomo, N., Krenzke, T. and Li, J. (2019). Confidentiality Protection Approaches for Survey Weighted Frequency Tables. Transaction on data privacy
14. Shlomo, N. and Skinner. C.J. (2012). Privacy Protection from Sampling and Perturbation in Survey Microdata. Journal of Privacy and Confidentiality, Vol. 4, Issue 1.
15. Shlomo, N. and Skinner, C.J. (2010). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Annals of Applied Statistics, Vol. 4, No. 3, 1291-1310.
16. Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. Journal of American Statistical Association, Vol. 103, Number 483, 989–1001.

The Integrated System of Statistic Registers: first steps towards facing privacy issues

Il Sistema Integrato dei Registri: primi passi per affrontare i problemi di privacy

Mauro Bruno and Roberta Radini[1]

Abstract The Italian Integrated System of Statistical Registers (ISSR) provides coherent and high quality data that support all the steps of statistical production, ranging from the extraction of frames to aggregated data for dissemination. The ISSR processes personal data on a large scale, such data can refer to populations with strong privacy and confidentiality restrictions. In this paper, we will briefly describe how we can enrich the data architecture of ISSR with privacy preservation solutions. In particular, the system ISSR must be equipped with a privacy-preserving statistical analysis system that: (i) performs privacy-preserving access on the input parties by controlling elaborations that can be performed by the computing parties, and (ii) decides which results can be published to which result parties.

Abstract *Il Sistema Integrato dei Registri italiano (SIR) fornisce dati coerenti e di alta qualità che supportano tutte le fasi della produzione statistica, che vanno dall'estrazione di frame a dati aggregati per la diffusione. Il SIR elabora i dati personali su larga scala, tali dati possono fare riferimento a popolazioni comportando rischi di forti restrizioni sulla privacy e sulla riservatezza. In questo documento, descriveremo brevemente come possiamo arricchire l'architettura dei dati del SIR con soluzioni di tutela della privacy. In particolare, il SIR deve essere dotato di un sistema di analisi statistica che preservi la privacy: (i) eseguendo l'accesso in modalità "privacy-preserving" alle parti di input e controllando le elaborazioni che possono essere eseguite dalle strutture di calcolo e (ii) decidendo quali risultati possono essere pubblicati e a quali parti fanno riferimento.*

Key words: Integrated System of Statistical Registers, Privacy, Ontologies

[1] Mauro Bruno, Istat; mbruno@istat.it
Roberta Radini, Istat; radini@istat.it:

1 Background

In 2016, Istat launched a modernization program with the aim to redesign the statistical production chain, moving from a silos-based model to a process-oriented approach. Within this transformation, a major choice was the centralization and organization of the statistical information from administrative sources and surveys in an integrated system of registers, called the Italian Integrated System of Statistical Registers (ISSR).

The main purpose of the ISSR is to provide to the production processes coherent and high quality data to be used in all the production steps, from extraction of frames for the selection of samples to aggregated data for dissemination and information for statistical research on new issues.

The integrated system of registers is obtained from processing of personal data on a large scale, also concerning particular categories such as minors, disabled people or in general populations with particular privacy and confidentiality restrictions. In view of the unquestionable value in obtaining important statistical outputs, the integrated system of registers must be equipped with particular policies for the protection of the privacy, defined as the right of an individual or business to keep information about them from being disclosed. Personal privacy is recognised and upheld by the European Union as a fundamental human right [1].

In this paper, we will briefly describe how we can enrich the data architecture of ISSR with privacy preservation solutions. In particular, the system ISSR must be equipped with a privacy-preserving statistical analysis system that: (i) performs privacy-preserving access on the input parties by controlling elaborations that can be performed by the computing parties, and (ii) decides which results can be published to which result parties.

1.1 Integrated System of Statistical Registers (ISSR)

The Italian Integrated System of Statistical Registers (ISSR) is a 'micro data' repository that guarantees: i) homogeneous management process of domain specific registers, ii) statistical, conceptual and physical integration of the statistical units contained in each register. The registers that compose the ISSR refer to the core themes of statistical production, i.e. Individuals, Families and Cohabitations, Economic/Institutional Units and Places. Moreover, the relationship among the themes allows it to be called "Integrated System" [2, 3, 4].

The design of ISSR data architecture is compliant with modern semantic integration approaches, with a strong emphasis on active metadata guiding the access to the system. More precisely the whole data architecture (see **Figure 1**) can be split into three layers: all registers, ISSR and data consumption, defined by all type of publication or analysis outputs.

ISSR offers flexibility and agility to retrieve on-the-fly data from all registers, overcoming the limitations of older approaches such as ETL and data replication. Currently, the System is composed by four sub-layers: 1) Physical Integration that

combines data, using the unit identification code, 2) Integrated Micro Layer that selects, integrates and transforms data, then exposing them as views, 3) Macro-data Layer provides aggregated data to users to conduct analyses or business intelligence activities, 4) Semantic level provides access to data via Ontology Based Management System (ODBM) to micro and macro data [5].

The Ontologies is a formal, shared and explicit representation of the conceptualization of the domain of interest expressed through the formal language, which makes it "machine-actionable", allow representing metadata “coupled” with data. Therefore, in this system the metadata are not limited to a “documentation” role but allow governing the data integration step by ensuring the high quality of integrated data. Moreover, ontologies allow the access to the data and the accesses management. This functionality can be a component to handle some Output privacy issues as described below in *Use Case 2*.

In the 2th sub-layers, Integrated Micro Layer, we need to implement solutions that solve or mitigate input privacy issues, as described in the *Use Case 1*.

1.2 Input and Output privacy: Techniques

In this section, we introduce input and output privacy core concepts. In literature, the term Input Privacy refers to a set of methods and technologies that allow to perform processing without accessing or deriving any input value provided, nor accessing intermediate values or statistical results during the processing of data other than output established for processing [6].

Input Privacy techniques allow processing registers linked data without accessing directly to the data of each register, thus reducing liability and simplifying compliance with data protection regulations. Secure computation techniques, which process data while it remains encrypted or obfuscated from regular access, allow achieving this goal. Input privacy techniques are generally used when a party wants to share data with other parties, without providing access or sharing its information. Concerning the ISSR, its input sources are the statistical registers and privacy measures have to be put in place in order to enforce privacy on computations that involve multiple registers.

Another class of techniques to preserve privacy is Output Privacy. Such class guarantees that the published results do not contain identifiable input data beyond what is allowable by Input Parties.

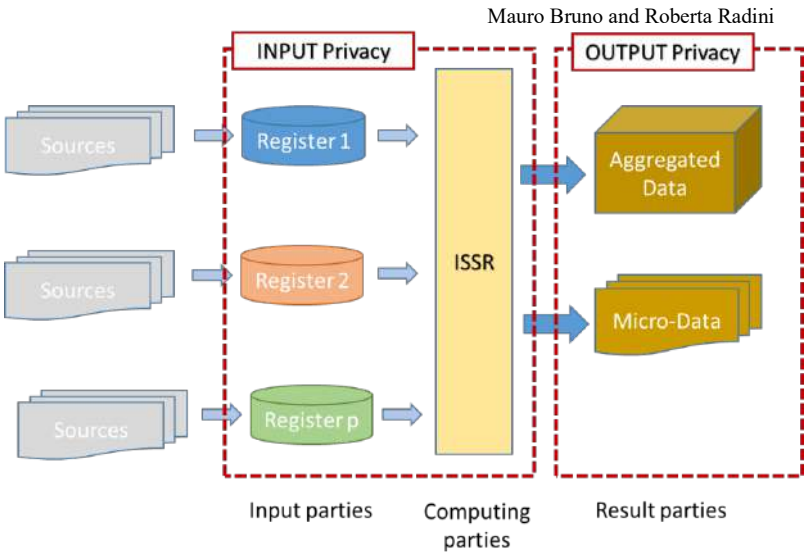


Figure 1: *ISSR privacy*

Following the general description provided above in **Figure 1** we identify the parties involved in the ISSR and the 'boundaries' of input and output privacy.

In the following subsections, we will describe two applications of input and output privacy techniques in the context of ISSR. 'Use Case 1' describes how input privacy techniques allow merging two or more registers in a single archive, in compliance with data protection regulations. 'Use Case 2' describes the access the ISSR to obtain a non-aggregated or aggregated output combining differential privacy with ontology-based access to information.

Use Case 1: Input Privacy

Suppose we want to perform analyse the employment conditions of individuals with disabilities with respect to the municipality of residence. This analysis requires linking the basic *Register of individuals* with the selection of the disabled, enriched with the information of the municipality of residence and the *Register of Labor*, which records the job positions of the selected individuals.

In this scenario, access to register data could be ontology-driven. Ontologies allows defining concepts and relationships between concepts, therefore allow identifying queries - or more general analysis - that can have strong impact on privacy of individuals. In addition, these analyses allow defining possible harm to individuals.

The use of input privacy techniques such as (Full) Homomorphic Encryption (HE, FHE) [7] are extremely onerous from a computational perspective, thus limiting for the number of operations that can be performed. Therefore, ontology-driven accesses could determine in which cases processing according to input privacy.

We can define, through ontologies, confidential and sensitive the information about disabilities and potentially discriminating if combined with information on their employment.

The link of two registers using privacy-preserving record linkage techniques is described in [8]. This type integration is classified as Private Set Intersection with Analytics (PSI-A) [9]. Each register has an owner internal at Istat; the analysis can be executed by one of owners or a third party (internal at Istat). In order to avoid the disclosure of sensitive information during the linkage process, we need sophisticated techniques. We will examine HE and FHE, a family of encryption schemes with a special algebraic structure that allows computations to be performed directly on encrypted data without requiring a decryption key. In particular, Full Homomorphic Encryption allows both additions and multiplications to be performed on encrypted data.

Use Case 2: Access Privacy

Suppose we want to perform analyses relating to the residential individuals in small areas (sub municipality). This analysis requires linking the basic *Register of individuals* with the selection of residential people, enriched with the information of the addresses or the sub municipality areas of residence and the *Register of Places,* which records the geographic identification.

As described in previous case, the use of the ontologies can guide in the definition of risky concepts and dangerous with respect to privacy relationships between multiple concepts.

We can define through the ontologies the relationship between individuals and the residential address or area as a risky information and that can determine the damage of location of the people and therefore define required the use of privacy input techniques to the access to this information.

A second level of use of ontologies can be to intercept people with a set of characteristics that can make them re-identifiable.

Besides the use of ontologies, we are investigating the use of output privacy techniques belonging to Differential Privacy. Differential Privacy is an approach used to quantify and limit the number of information about individuals in a database that leak out releasing the result of an aggregate calculation on that database.

References

1. Privacy and Data Protection by Design, European Union Agency for Network and Information Security (2015) https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design
2. Wallgren, A., Wallgren, B. (2014): Register-Based Statistics. Statistical Methods for Administrative Data. Wiley
3. G. Garofalo, R. Radini, M. Scannapieco: The Italian System of Statistical Registers: On the Design of an Ontology-based Data Integration Architecture. In proceeding of NTTS Conference on New Tecniques and Tecnologies for Statistics (2017)

4. P.D. Falorsi, R. Radini, M. Scannapieco, L. Tosco: *Modernizing Data Integration Systems at Istat*. European Conference on Quality in Official Statistics, Q2018, Krakov, Poland (2018)
5. R. Radini, M. Scannapieco, L. Tosco: The Italian Integrated System of Statistical Registers: Design and Implementation of an Ontology-based Data Integration Architecture. (2018) https://www.istat.it/it/files//2018/11/Scannapieco_original-paper.pdf
6. UN Handbook on Privacy-Preserving Computation Techniques. (2019) http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf
7. Gentry, C., Boneh, D. (2009): A fully homomorphic encryption scheme. Vol. 20. No. 09. Stanford: Stanford University
8. Vatsalan, D., Christen, P., O'Keefe, C. M. & Verykios, V. S., An Evaluation Framework for Privacy-preserving Record Linkage. Journal of Privacy and Confidentiality, 6(1), 35-75 (2014)
9. Falorsi, P.D., Liseo, B., Scannapieco, M. (2019): Dealing with Privacy Issues in Data Integration Systems, versione estesa dei proceedings di 'Law via the Internet 2018', Knowledge of the Law in the Big Data Era, Series Frontiers in Artificial Intelligence and Applications, ISBN 978-1-61499-984-3, IOS Press

Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics

Trusted Smart Surveys: una possibile applicazione delle tecnologie di miglioramento della privacy nella statistica ufficiale

Fabio Ricciato, Kostas Giannakouris, Albrecht Wirthmann, Martina Hahn

Sommario In this discussion paper we outline the general concept of Trusted Smart Surveys as a possible use-case for the application of Privacy Enhancing Technologies (PET). Central to the Trusted Smart Survey paradigm is the notion of *extracting output information without revealing the input data*. This can achieved by leveraging PET to compute the final desired information — a statistical indicator or and aggregate table — without centralising the individual data from the respondents. Together with additional mechanisms for ensuring transparency, auditability and public control over the whole process these technologies can help to scale up public trust in new survey models, hence promoting participation, and in this way contribute to improve accuracy and relevance of official statistics.

Sommario *In questo contributo viene delineato il concetto delle* Trusted Smart Surveys *intesa come una possible applicazione delle moderne tecnologie di miglioramento della privacy (PET). Un elemento centrale di questo paradigma è il concetto che il risultato di una computazione possa essere estratto senza rivelare i dati in ingresso. In questo modo, le tecnologie PET consentono di ottenere l'informazione desiderata — ad esempio un indicatore statistico oppure una tavola di valori aggregati — senza centralizzare i dati individuali dei partecipanti. Insieme ad altri meccanismi tesi a garantire la trasparenza, la verificabilità e il controllo sull'inte-*

Fabio Ricciato
Eurostat, 5 rue A. Weicker, Luxembourg, e-mail: Fabio.Ricciato@ec.europa.eu

Kostas Giannakouris
Eurostat, 5 rue A. Weicker, Luxembourg, e-mail: Konstantinos.Giannakouris@ec.europe.eu

Albrecht Wirthmann
Eurostat, 5 rue A. Weicker, Luxembourg, e-mail: Albrecht.Wirthmann@ec.europaa.eu

Martina Hahn
Eurostat, 5 rue A. Weicker, Luxembourg, e-mail: Martina.Hahn@ec.europaa.eu

ro processo, queste tecnologie possono aiutare a rafforzare la fiducia da parte del pubblico nei nuovi modelli d'indagine, promuovendo una maggiore partecipazione, contribuendo cosi a migliorare l'accuratezza e la rilevanza delle statistiche ufficiali.

Key words: Trusted Smart Survey, Trusted Smart Statistics, Privacy Enhancing Technologies, Official Statistics.

1 From traditional surveys to Smart Surveys

The traditional survey model based on paper questionnaire is based on a batch of questions (questionnaire) presented to the respondent, by a human interviewer, during a single survey session. Such questionnaire-based interaction model has not changed with telephone-based and computer-assisted interviews. Instead, leveraging the smartphone (via a mobile app) allows for several important changes to the survey model. First, the so-called *response burden* can be diluted over much longer intervals, even months or years, with a continuous flow of questions presented at very low rate in the background of the respondent's normal activities. Second, the *active data* generated in response to explicit queries may be complemented by and combined with *passive data* collected automatically by sensors onboard the smartphone (e.g. GPS, accelerometer, etc.) or acquired from other devices within the personal sphere of the respondent (e.g., home appliances, Internet-of-Things devices), again contributing to reduce the response burden (e.g., automatic diary). Third, the human interviewer can be replaced by an Artificial Intelligence (AI) assistant. The combination of these elements enable the design of *context-aware* survey models where *timing and content* of the presented questions are decided dynamically based on *context* information as inferred based on the sensor data. For example, a question of the kind "*It seems you are travelling today: is this trip mainly intended for work or for leisure?*" might be posed during the trip instead of the more traditional "*How many business trips did you perform in the previous 6 months?*" In other words, the combined use of AI and passive data allow to decide dynamically *when* and *what* question to ask. Furthermore, questions can request explicit action and non-textual response, for example taking photos or recording sound ("*Can you please take a photo of your surrounding environment?*").

What we have outlined insofar is basically the concept of *Smart Surveys*. Through the smartphone app, Statistical Offices (SO) have the possibility to establish a continuous, long-term dialogue with each respondent. Furthermore, they have the opportunity to feed back individualised reports to the respondent, to disseminate selected figures and statistics of special interest for particular groups, and to better communicate the importance of official statistics. In this way, the bi-directional dialogue between the SO and the respondent may become even more interactive.

Such fundamental change of paradigm is enabled not only by the availability of mobile technologies, but also by the fact that continuous interaction with mobile apps is considered rather normal nowadays. It is not only a matter of exploiting

new digital technologies, but also leveraging the *new digital behaviours* and *new digital attitudes* that citizens have developed through the use of such technologies. In the private sector, marketing strategies are increasingly relying on continuous low-intensity interaction with customers via mobile apps, and in principle similar practices (with appropriate adaptations) may be adopted by public institutions for public interest purposes, including by statistical offices to produce better official statistics. That does not mean that SOs are called to emulate marketing practices from the commercial sector. Rather on the contrary, SOs have the opportunity to pioneer new models of data use based on their constituent principles of transparency, openness, independence and democratic control, and to *show the way* to other public institutions (but also to the private sector) as to how the collectivity can (re)gain democratic control over its data. In so doing, SO should remain critical and vigilant of the possible negative consequences and unintended effects, not least the risk of distorting — or anyway influencing excessively — the social phenomena they aim at measuring. Taking a critical design approach, SO can design new survey models that ripe the potential benefits of the new (data, technologies, behaviours, attitudes) while minimising the potential risks. In the remainder of this short contribution we propose an initial reflection on one of several aspects related to the design of a suitable Smart Survey model, namely the issue of *trust*.

2 From Smart Surveys to *Trusted* Smart Surveys

The *smart survey* model is considerably more intrusive in the life of the participant than the traditional survey model: it would allow SO to plug into a much richer, wider and deeper set of data about the individual respondent and his/her activities, with much greater level of detail and for longer times. The rising awareness by the public about the dystopian risks of mass surveillance would make such model difficult to accept (if at all desirable) unless a set of convincing safeguards are put in place to offset the risks of any possible data misuse — against the individuals, groups and the whole collectivity — and to guarantee a level of transparency and *trustworthiness* that is commensurate to the risks.

Trust is a complex issue where aspects of *trust in the process*, *trust in the input data* and *trust in the output statistics* are closely intermingled and reciprocally inter-dependent. If the respondents fear negative consequences resulting from participating in Smart Surveys, they will not provide their truthful data (or provide no data at all) and the resulting final figures will be poorly representative of reality. In this way, lack of trust in the process causes lack of trust in the input data and therefore in the final statistics. To stir a virtuous cycle of trust, SOs must provide strong trust guarantees on the whole survey process in the first place.

As the survey model evolves from traditional to smart, additional guarantees must be provided, also based on new technological means. The success of a well designed Smart Survey model requires SOs to put in place (and convincingly communicate to the respondents and to the general public) a set of strong mechanisms to ensure

by design public control on *what information is extracted* from such data, *how* (methodology), *why* (purpose) and *by whom*. This requires the deployment of a coherent combination of “soft" measures (i.e. legal provisions, organisational processes, practices) and “hard" measures (i.e. technological solutions) that are adequate to the new scenario[1]. Putting in place a coherent set of strong safeguards is what qualifies the *Trusted* Smart Survey model, i.e., an augmentation of the Smart Survey concept by technological and non-technological elements aimed at increasing the degree of *trustworthiness of data use*.

3 Design principles: an initial proposal

Designing a Trusted Smart Survey model is a task that involves multiple dimensions: it is matter of *trustworthiness engineering* where security, privacy, organisational and communication elements must be blended together, each of them entailing hardware, software and *humanware* aspects. To this aim, it is important to establish a common terminology and a clear set of design principles and requirements in the first place. A key aspect of the Trusted Smart Survey paradigm, and more in general of the whole Trusted Smart Statistics concept [1], is a shift of focus from *data collection* to *data use*. This implies that the central pivot of the whole engineering process is the *combination of DATA and CODE* — not merely DATA.

Generally speaking, given a collection of input data $\boldsymbol{x}$ and pre-defined function $\boldsymbol{f}$, statistical production entails the computation of the output $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. The *function* $\boldsymbol{f}$ denotes here any generic algorithmic workflow (or statistical methodology) described by a pre-defined sequence of mathematical operations of arbitrary complexity. The most complete and non-ambiguous description of a complex function $\boldsymbol{f}$ is by means of a computer program, i.e., software code. A computation instance, or *query*, is therefore associated to the application of CODE $\boldsymbol{f}$ to DATA $\boldsymbol{x}$, hence to the DATA-CODE pair $\langle \boldsymbol{x}, \boldsymbol{f} \rangle$. Both elements may be centralised at a single machine administered by a single entity, or distributed among multiple machines possibly administered by different entities. In the envisioned model of Trusted Smart Survey the individual data of each participant remain local to the participant's private space (e.g., her personal device or private cloud space). The code execution is split between the participant's machine, for the local pre-processing of individual data (“micro" processing), and an institutional or inter-institutional infrastructure supporting secure private computation of aggregate values (“macro" processing) through PET. Such PET infrastructure might possibly comprise multiple machines administered by different organisations, as e.g. with Secure Multi-Party Computation (SMPC), or it may reduce to a single machine controlled jointly by multiple institutions, e.g., based on Trusted Execution Environment (TEE) combined with

[1] While a set of measures is already in place to ensure trustworthiness of statistical production and protection of personal data collected from traditional sources, such measures need to be scaled up and strengthened in order to cope with new, more intrusive non-traditional kinds of data sources.

mechanisms for multi-party authorisation. Whatever PET solution is adopted, it is important to avoid concentrating control into a *single point of trust*.

Paraphrasing a famous quote: "*DATA is neither good nor bad; nor is it neutral*" (the original version had "technology" instead of "data" [2]). In fact, the discourse about what is ethically desirable vs. undesirable — hence operationally allowed or disallowed, technically enabled vs. disabled — would be more clearly scoped if centered specifically on the DATA-CODE pair $\langle \boldsymbol{x}, \boldsymbol{f} \rangle$, rather than solely on DATA $\boldsymbol{x}$. It is possible to envision data governance models wherein the legal right *and the technical possibility* of applying a particular CODE on a particular set of DATA is decoupled from who holds the data. Keeping data holding and data usage as distinct issues is indeed a pivotal aspect in the discourse.

Based on the above considerations, we list below a possible set of design principles for the future Trusted Smart Survey model. Our goal here is to present an initial proposal serving as starting point for a critical discussion within the statistical community. With the caveat above, we propose to consider the following design principles:

1. For a generic respondent i, the individual data element $\boldsymbol{x}_i$ remains private to the respondent and is never disclosed as such to any single other entity — DATA are closed and remain stored at their respective sources.
2. The source code of $\boldsymbol{f}$ is always made openly available for public scrutiny *before* and *after* performing the computation, in order to allow for ex-ante and ex-post checks — CODE is open.
3. The final result of the computation $\boldsymbol{y}$ (query output) is made available to the SO but not automatically released to the public — access to the query output remains restricted to authorised SO.
4. Before execution, the source code of $\boldsymbol{f}$ must be approved by a pre-determined set of ex-ante controllers.
5. The execution of every query must be logged in a secure way and made accessible to ex-post controllers. Each log must contain all appropriate meta-data, including an unmodifiable pointer to the exact code version that was executed and detailed information as to who had access to the computation result.
6. Technical and organisation means are adopted to ensure that the binary (executable) code run on the data corresponds to the (source) code $\boldsymbol{f}$ that was preliminarily approved by *all* ex-ante controllers (e.g., code authentication).
7. Technical and organisation means are adopted to ensure that the information stored in every log corresponds truthfully to the query execution.

The role of ex-ante or ex-post controller may be taken by any entity (institution, organisation, association, etc.) bearing a legally recognised interest related to the query. To illustrate, consider the example that entity A is mandated to ensure methodological quality (the code $\boldsymbol{f}$ is appropriate and fit for the purpose), entity B cares that the prospective output to be computed by $\boldsymbol{f}$ is ethically acceptable, and entity C is concerned that the same output does not cause harm to some legitimate private business. One possible way to reassure and gain support by all entities is to assign to all such entities a *shared, non-exclusive* ex-ante controller role. Similar conside-

rations apply for the assignment of ex-post controller roles. We do not claim here that each and every entity with some interest in the computation has automatically the right to act as controller. However, the PET infrastructures must allow to flexibly configure ex-ante and ex-post controller roles to different entities as needed in each specific use-case. In other words, having such feature in place at the software and hardware level allows to flexibly engineer the process and governance model at the *humanware* level, and to assign shared, non-exclusive controls to the relevant and legitimate stakeholders.

Note that Principle (3) foresees that access to the computation output $\boldsymbol{y}$ may still be restricted to the SO. This provision allows to execution of computation function $\boldsymbol{f}$ for which the output $\boldsymbol{y}$ still yields some residual risk of personal re-identification. In this case, it is responsibility of the SO to apply the usual Statistical Disclosure Control (SDC) checks in post-processing before releasing the final statistics to the public. Alternatively, Differential Privacy (DP) mechanisms may be embedded in the computation function $\boldsymbol{f}$. Depending on the specific use-case, both variants can be considered. In summary, referring to the notions of Input Privacy and Output Privacy (see e.g. to [3] for a discussion), the whole chain from raw input data until final statistics must compose an Input Privacy solution (e.g., SMPC or TEE) with a solution for Output Privacy (SDC or DP).

The above principles imply that *(i)* the computation function $\boldsymbol{f}$ is defined beforehand, and *(ii)* the whole statistical production process is fully automatised and represented in machine-readable code. These conditions are not problematic in the statistical *production* stage, but may be somewhat critical for *research* and *methodological development* where some degree of manual data exploration is unavoidably needed. To cope with such situation, a slightly modified set of design principles could be formulated, where some items are loosened (e.g., ex-ante checks) while others are strengthened (e.g., ex-post checks) in order to maintain adequate control while allowing for the necessary degree of flexibility during manual exploration.

Riferimenti bibliografici

1. F. Ricciato, A. Wirthmann, K. Giannakouris, F. Reis, and M. Skaliotis. Trusted smart statistics: motivations and principles. *Statistical Journal of the IAOS*, 35, 2019. https://ec.europa.eu/eurostat/cros/system/files/sji190584.pdf.
2. M. Kranzberg. Technology and history: “kranzberg’s laws". *Technology and Culture*, 27(3), 1986. doi:10.2307/3105385.
3. F. Ricciato, A. Bujnowska, A. Wirthmann, M. Hahn, and E. Barredo-Capelot. A reflection on privacy and data confidentiality in official statistics. In *ISI World Statistics Congress*, 2019. https://www.bis.org/ifc/events/isi_wsc_62/ips177_paper3.pdf.

Designing adaptive clinical trials

Optimal designs for multi-arm exponential trials

Disegni ottimi per prove cliniche a risposte esponenziali

Rosamarie Frieri and Marco Novelli

Abstract Most of the randomized clinical trials for treatment comparisons have been designed to obtain a balanced allocation among the treatments. This is mostly due to the so-called universal optimality of balance. However, with several treatments the balanced allocation may not be efficient and could be strongly ethically inappropriate, in particular for phase III-trials. In [3], taking into account the exponential model, the target allocation maximizing the power of Wald test under a suitable ethical constraint has been derived. In this paper, we further explore the operating characteristics of such allocation through a comparison with other targets proposed in the literature, showing that the constrained optimal target exhibits good performances in terms of inferential precision and ethical demands.

Abstract *La maggior parte degli studi clinici randomizzati per il confronto tra trattamenti sono stati disegnati per ottenere un'allocazione bilanciata tra i gruppi. Questo è dovuto principalmente alla cosiddetta ottimalità universale del bilanciamento. Tuttavia, in presenza di molti trattamenti, l'allocazione bilanciata potrebbe risultare inefficiente e fortemente non etica, in particolare nelle prove cliniche di Fase III. In [3], considerando risposte esponenziali, è stata derivata l'allocazione ottimale che massimizza la potenza del test di Wald basato sui contrasti, sotto un appropriato vincolo etico. In questo articolo, verranno approfondite le caratteristiche operative di tale allocazione anche attraverso un confronto con le altre allocazioni proposte in letteratura, allo scopo di valutare la sua efficienza rispetto a criteri di natura sia etica che inferenziale.*

Key words: unbalanced allocations, efficiency, power of the Wald test, ethics

---

Rosamarie Frieri
Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, Bologna
e-mail: rosamarie.frieri2@unibo.it

Marco Novelli
Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, Bologna
e-mail: marco.novelli4@unibo.it

1 Introduction

In randomized clinical trials two competing goals, i.e. individual vs. collective ethics, have to be balanced. Indeed, the need to allocate more patients to the best available treatment (individual ethics) usually conflicts with the rigourous pursuit of scientific knowledge obtained with high inferential precision (collective ethics). So far, the vast majority of randomized clinical trials have been designed to achieve a balanced allocation among treatment groups, thanks to the optimal properties of balance especially in terms of estimation accuracy (see [6]). However, adopting equal allocation in the presence of several treatments could be neither efficient - since it is different from the optimal design for hypothesis testing (see [2, 3]) - nor ethical, especially in the context of phase-III trials, where the need to care for the well-being of the subjects involved in the trial is of primary importance. To overcome this trade-off, target allocations depending on a metric that accounts for treatment effects and/or their variabilities have been proposed, in order to obtain a valid compromise between ethical demand and inferential precision (see [1, 5]). Generally, these allocations depend on the unknown model parameters and can be targeted by suitable response adaptive randomization procedures, namely sequential allocation rules that, making use of the information accrued along the trial, change the assignment probabilities at each step in order to skew allocations toward the superior treatment. Recently, taking into account the problem of testing statistical hypothesis in normal homoscedastic trials, Baldi Antognini et. al. in [2] proposed an optimal target which maximizes the power of the Wald test of homogeneity, subject to an ethical constraint reflecting the effectiveness of the treatments. Moreover, Frieri and Zagoraiou in [3] derived a constrained target for exponential outcomes that are particularly relevant for oncological trials with survival endpoints. In this paper, we explore in depth the operating characteristic of the allocation derived in [3] through a comparison with other targets proposed in the literature. Our results show that the constrained optimal target guarantees very good performance in terms of statistical power, estimation precision, and ethical demands.

2 Framework and notation

In this work, clinical trials in which each subject is sequentially allocated to one of $K \geq 2$ available treatments are considered. Let δ_{kj} be the treatment assignment indicator such that $\delta_{kj} = 1$ when patients j is assigned to treatment $k\,(k = 1, \ldots, K)$ and 0 otherwise, with $\sum_{k=1}^{K} \delta_{kj} = 1$. The experimental outcome of the corresponding patient, Y_j, is assumed to be exponentially distributed with $E(Y_j|\delta_{kj} = 1) = \mu_k$, the treatment effect, and $V(Y_j|\delta_{kj} = 1) = \mu_k^2$ its variance. At each stage n, let $\boldsymbol{\pi}_n = (\pi_{1n}, \ldots, \pi_{Kn})^\top$, with $\pi_{kn} = n^{-1} \sum_{j=1}^{n} \delta_{kj}$ and $\sum_{k=1}^{K} \pi_{kn} = 1$, be the vector collecting the treatment assignment proportions up to that point, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^\top$ the vectors of treatment effects. In what follows, without loss of generality, we adopt

the-larger-the-better scenario, that is an higher response is more desirable for the patient's care, and we will work under the non-restrictive assumption that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$, i.e. the best performing treatment is labelled as the first one and the worst as the Kth one.

Here, the inferential focus is on the treatment contrasts, so letting $\mathbf{A}^\top = [\mathbf{1}_{K-1}| -\mathbf{I}_{K-1}]$, where $\mathbf{1}_p$ and $\mathbf{I}_p$ represent the p-dim vector of ones and the identity matrix, we denote by $\boldsymbol{\mu}_c = \mathbf{A}^\top \boldsymbol{\mu} = (\mu_1 - \mu_2, \ldots, \mu_1 - \mu_K)^\top$ and $\hat{\boldsymbol{\mu}}_{cn} = (\hat{\mu}_{1n} - \hat{\mu}_{2n}, \ldots, \hat{\mu}_{1n} - \hat{\mu}_{Kn})^\top$ the vector of contrasts wrt the first treatment and their MLEs, respectively. Under well-known regularity conditions, $\hat{\boldsymbol{\mu}}_{cn}$ is strongly consistent and asymptotically normal, i.e. $\hat{\boldsymbol{\mu}}_{cn} \xrightarrow{a.s.} \boldsymbol{\mu}_c$ and $\sqrt{n}(\hat{\boldsymbol{\mu}}_{cn} - \boldsymbol{\mu}_c) \xrightarrow{d} \mathrm{N}(\mathbf{0}_{K-1}, \mathbf{A}^\top \mathbf{M}^{-1} \mathbf{A})$, where $\mathbf{M} = \mathbf{M}(\boldsymbol{\mu}|\boldsymbol{\pi}) = \mathrm{diag}\left(\mu_i^{-2}\pi_i\right)_{i=1,\ldots,K}$ is the Fisher information matrix associated with $\boldsymbol{\mu}$. Finally, let us define by $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)^\top$, with $\rho_k \geq 0$ and $\sum_{k=1}^K \rho_k = 1$, the desired target allocation proportion, that can be obtained through suitable optimization problems.

The experimental strategy adopted to obtain the optimal target depends on the objective of the trial. When the aim is to maximize the inferential precision in the estimation of the treatment contrasts, Sverdlov and Rosenberger in [7] derived the tr_A optimal target by minimizing $\mathrm{tr}[\mathbf{A}^\top \mathbf{M}^{-1} \mathbf{A}]$, i.e.

$$\rho_1^A = \frac{\mu_1 \sqrt{K-1}}{\mu_1 \sqrt{K-1} + \sum_{i=2}^K \mu_i} \quad \text{and} \quad \rho_k^A = \frac{\mu_k}{\mu_1 \sqrt{K-1} + \sum_{i=2}^K \mu_i} \text{ for } k = 2, \ldots, K. \quad (1)$$

It is easy to show that $\rho_i^A \geq \rho_{i+1}^A \iff \mu_i \geq \mu_{i+1}$ for $i = 1, \ldots, K-1$ so this target, for exponential outcomes and under the-larger-the-better scenario, is ethical. On the side of hypothesis testing instead, a typical problem in multi-arm trials is to test the null-hypothesis of equality of treatment effects, i.e. $\boldsymbol{\mu}_c = \mathbf{0}_{K-1}$, where $\mathbf{0}_{K-1}$ is the $(K-1)$-dimensional vector of zeros. The target allocation maximizing the power of the Wald test of homogeneity is $\boldsymbol{\rho}^* = (\mu_1/(\mu_1+\mu_K), 0, \ldots, 0, \mu_K/(\mu_1+\mu_K))^\top$ (see [3]), which is clearly inappropriate for both statistical and ethical reasons. To avoid empty treatment arms, Zhu and Hu in [9], adopting the same framework in [8], set an optimization problem in which the power of the test is maximized subject to a constraint on the lower bound of the minimum number of subject assigned to each treatment. More specifically, the ensuing target $\boldsymbol{\rho}^Z$ should satisfy $\rho_i^Z \geq T$ for $i = 1, \ldots, K$, where $T \in [0, 1/K]$ is selected by the user. The target $\boldsymbol{\rho}^Z$ is available in closed form (which is not reported here for brevity, see [9]), however, it is only defined when $\mu_1 = \cdots = \mu_s > \mu_{s+1} \geq \ldots \geq \mu_{K-g} > \mu_{K-g+1} = \cdots = \mu_K$, for some positive integers s and g such that $s + g < K$. Notice that this framework does not include the configurations of the parameters in which $s + g = K$, i.e. $\mu_1 = \cdots = \mu_j > \mu_{j+1} = \cdots = \mu_K$ for $j = 2, \ldots, K-1$.

Finally, Frieri and Zagoraiou in [3], by adopting a multipurpose design methodology, derived the optimal allocation maximizing the power of Wald test subject to an ethical constraint reflecting the efficacy of the treatments. The ensuing constrained optimal target maximizing the non centrality parameter $\phi(\boldsymbol{\rho}) = n \cdot \mu_c^\top [\mathbf{A}^\top \mathbf{M}^{-1} \mathbf{A}]^{-1} \mu_c$ of the multivariate Wald test subject to $\rho_i \geq \rho_{i+1}$ for $i =$

$1,\dots,K-1$ is

$$\boldsymbol{\rho}^C = \begin{cases} (1-(K-1)x, x, \dots, x)^\top & \text{if } x < K^{-1}, \\ \boldsymbol{\rho}^B & \text{if } x \geq K^{-1}, \end{cases} \tag{2}$$

where $\boldsymbol{\rho}^B$ is the balanced allocation and

$$x = \frac{\frac{1}{\mu_1}\sum_{k=1}^{K}\left(\frac{1}{\mu_k}-\frac{1}{\mu_1}\right)^2}{\sum_{k=1}^{K}\left(\frac{1}{\mu_k}-\frac{1}{\mu_1}\right)\sum_{k=1}^{K}\left(\frac{1}{\mu_k^2}-\frac{1}{\mu_1^2}\right)}.$$

Note that in the presence of a cluster of superior treatments $\mu_1 = \dots = \mu_s > \mu_{s+1} \geq \dots \geq \mu_{K-g} > \mu_{K-g+1} = \dots = \mu_K$ all targets $\boldsymbol{\rho}^C = (\rho_1^C, \dots, \rho_s^C, x, \dots, x)^\top$ such that $\sum_{i=1}^{s}\rho_i^C = 1-(K-s)x$ and $s+g<K$ are optimal. Instead, when $s+g=K$ every allocation such that $\sum_{i=1}^{s}\rho_i^C = 1-(K-s)x = \mu_1/(\mu_1+\mu_K) = 1-\sum_{i=s+1}^{K}\rho_i^C$ is optimal.

3 Comparisons of optimal allocations for the exponential model

In this section, we compare the statistical performances of the previously introduced designs, that is $\boldsymbol{\rho}^A, \boldsymbol{\rho}^B, \boldsymbol{\rho}^C$ and $\boldsymbol{\rho}^Z$, in terms of the normalized power, $E_P(\boldsymbol{\rho}) = \phi(\boldsymbol{\rho})/\phi(\boldsymbol{\rho}^*)$ and the tr_A efficiency, $E_{tr_A}(\boldsymbol{\rho}) = \frac{\mathrm{tr}[\mathbf{A}^\top\mathbf{M}^{-1}(\boldsymbol{\rho}^A)\mathbf{A}]}{\mathrm{tr}[\mathbf{A}^\top\mathbf{M}^{-1}(\boldsymbol{\rho})\mathbf{A}]}$. As a measure of ethics, we consider the ratio between the total expected outcome and its maximum value, that is $E_E(\boldsymbol{\rho}) = \mu_1^{-1}\sum_{i=1}^{K}\mu_i\rho_i$. Figure 1 summarizes the operating characteristics of the targets for $\mu_2 = 10, \mu_3 = 9$ and $\mu_4 = 8$, as μ_1 varies from 15 to 35. As far as the statistical power is concerned, for values of μ_1 close to μ_2, $\boldsymbol{\rho}^Z$ with $T = 0.1$ shows the highest power efficiency while, as μ_1 increases (greater than 25) the best performance in terms of $E_P(\cdot)$ is achieved by $\boldsymbol{\rho}^C$ in (2), whose power is always increasing wrt μ_1. Note that, this property is not shared by all the targets. The $\boldsymbol{\rho}^A$ target in (1), the balanced one and $\boldsymbol{\rho}^Z$ for $T = 0.15$ and $T = 0.2$ always present lower power than $\boldsymbol{\rho}^C$. In terms of ethical demand, $\boldsymbol{\rho}^C$ outperforms all the competitors with a gain wrt the second best ($\boldsymbol{\rho}^Z$, $T = 0.1$) up to 7%. The ethical efficiency of $\boldsymbol{\rho}^C$ - as also confirmed by other studies omitted here for brevity - is slightly decreasing for values of μ_1 close to μ_2 and tends to increase as μ_1 grows. A similar behaviour is retrieved only for $\boldsymbol{\rho}^A$, whereas all the remaining targets have decreasing ethical efficiency as μ_1 increases. In terms of estimation precision, the second best is $\boldsymbol{\rho}^Z$ with $T = 0.15$ for $\mu_1 < 26$, while the same target with $T = 0.1$ shows a value of E_{tr_A} approaching 1 as μ_1 grows. The constrained target $\boldsymbol{\rho}^C$ exhibits an estimation efficiency almost constant wrt μ_1 with a value always greater than 0.93. In general, the balanced design shows the lowest efficiency in all the measures considered, while the performances of $\boldsymbol{\rho}^Z$ strongly depends on the subjective choice of T.

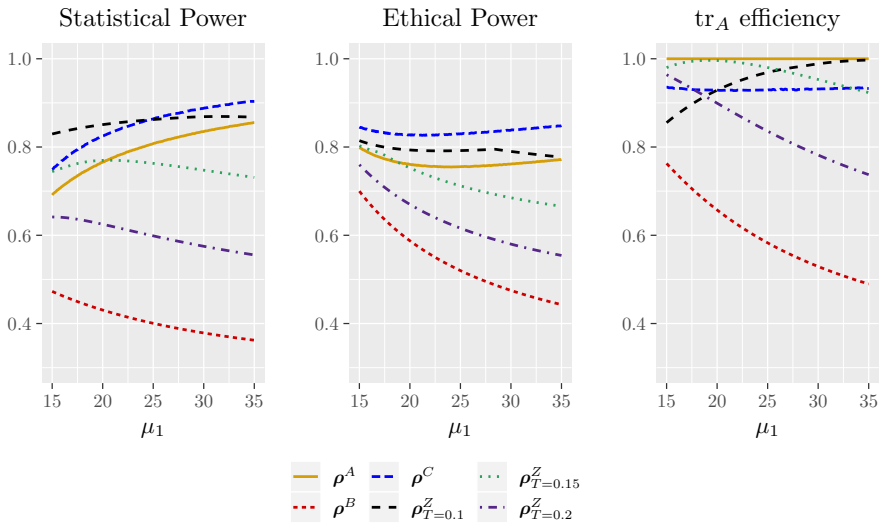


Fig. 1 Efficiency measures for $\boldsymbol{\rho}^A$, $\boldsymbol{\rho}^C$, $\boldsymbol{\rho}^B = (0.25, 0.25, 0.25, 0.25)^\top$, and $\boldsymbol{\rho}^Z$ for $T = 0.1, 0.15$ and 0.2 where $\mu_1 \in [15, 35]$, $\mu_2 = 10$, $\mu_3 = 9$ and $\mu_4 = 8$.

In Table 1, we present some examples in which groups of treatments with the same efficacy are considered . We can notice that $\boldsymbol{\rho}^C$, which coincides with the balanced design in scenario *(a)*, skews the allocations to the best performing treatment as the differences between μ_1 and the other treatment effects increases. In all the configurations considered, $\boldsymbol{\rho}^C$ leads to the highest power while keeping the ethical and the estimation efficiency always greater than 92.3% and 91.5%, respectively.

Scenario	$\boldsymbol{\mu}$	$\boldsymbol{\rho}$	$E_P(\boldsymbol{\rho})$	$E_E(\boldsymbol{\rho})$	$E_{tr_A}(\boldsymbol{\rho})$
(a)	$(12, 12, 12, 10)^\top$	$\boldsymbol{\rho}^A = (0.379, 0.219, 0.219, 0.183)^\top$	0.670	0.970	1
		$\boldsymbol{\rho}^C = \boldsymbol{\rho}^B$	0.818	0.958	0.915
		$\boldsymbol{\rho}^Z$	-	-	-
(b)	$(12, 12, 10, 10)^\top$	$\boldsymbol{\rho}^A = (0.394, 0.227, 0.189, 0.189)^\top$	0.975	0.936	1
		$\boldsymbol{\rho}^C = (0.318, 0.227, 0.227, 0.227)^\top$	1	0.923	0.969
		$\boldsymbol{\rho}^B$	0.992	0.917	0.898
		$\boldsymbol{\rho}^Z$	-	-	-
(c)	$(12, 10, 10, 10)^\top$	$\boldsymbol{\rho}^A = (0.409, 0.197, 0.197, 0.197)^\top$	0.928	0.902	1
		$\boldsymbol{\rho}^C = (0.545, 0.152, 0.152, 0.152)^\top$	1	0.925	0.932
		$\boldsymbol{\rho}^B$	0.682	0.875	0.881
		$\boldsymbol{\rho}^Z$	-	-	-

Table 1 Behaviour of $\boldsymbol{\rho}^A$, $\boldsymbol{\rho}^C$ and $\boldsymbol{\rho}^B = (0.25, 0.25, 0.25, 0.25)^\top$ in presence of groups of treatments with the same efficacy.

As discussed in Section 1, it is worth noticing that the target proposed by Zhu and Hu [9] cannot be computed for some parameters configurations, e.g. scenarios *(a), (b), (c)* of Table 1. This drawback can strongly affect its applicability.

Notice also that in clinical trials comparing $K > 2$ treatments the definition of ethics is not unequivocally determined and the requirement of being ethical by skewing more patients to the superior treatment may sometimes be misleading (see [3]). The structure of the ethical constraint $\rho_i \geq \rho_{i+1}$ for $i = 1, \ldots, K-1$ in $\boldsymbol{\rho}^C$ ensures that the target has its components ordered accordingly to the magnitude of the treatment effects. In general, this property is not shared by the considered targets: for example the $\boldsymbol{\rho}^A$ allocation in (1) assigns more patients to the reference treatment, which in our set-up coincides with the best treatment. However, if for example we consider $\boldsymbol{\mu} = (10, 12, 12, 12)^\top$, then $\boldsymbol{\rho}^A = (0.325, 0.225, 0.225, 0.225)^\top$. Moreover, if we adopt $\boldsymbol{\rho}^Z$ with $T = 0.1$ in a configuration close to scenario *(a)*, e.g. $\boldsymbol{\mu} = (12.1, 12, 11.9, 10)^\top$ then the ensuing target is $\boldsymbol{\rho}^Z = (0.357, 0.1, 0.1, 0.443)^\top$. In both examples the highest proportion of patients is receiving the less effective drug showing that these targets may be inappropriate from an ethical viewpoint.

The results of Figure 1, Table 1 and the discussion above, show that the constrained optimal target $\boldsymbol{\rho}^C$ represents a valid trade-off between statistical power, inferential precision and ethical demand.

Acknowledgements Marco Novelli was supported by the Italian Ministry of Education, University and Research under PRIN 2015 "Environmental processes and human activities: capturing their interactions via statistical methods (EphaStat)".

References

1. Baldi Antognini, A., Giovagnoli, A.: Adaptive designs for sequential treatment allocation. Chapman and Hall/CRC (2015)
2. Baldi Antognini, A., Novelli, M., Zagoraiou, M.: Optimal designs for testing hypothesis in multiarm clinical trials. *Stat. Meth. Med. Res.*, **28**, 3242-3259 (2019)
3. Frieri, R., Zagoraiou, M. : Optimum and ethical designs for hypothesis testing in multi-arm exponential trials. Submitted (2020)
4. Hu, F., Zhang, L. X.: Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Ann. Stat.*, *32*, 268-301 (2004)
5. Rosenberger, W. F., Lachin, J. M.: Randomization in clinical trials: theory and practice. John Wiley and Sons (2015)
6. Silvey, S.D.: Optimal Designs. Chapman & Hall, London (1980)
7. Sverdlov, O., Rosenberger, W. F.: On recent advances in optimal allocation designs in clinical trials. *J. Stat. Theory Prac.*, **7**, 753-773 (2013)
8. Tymofyeyev, Y., Rosenberger, W. F., Hu, F.: Implementing optimal allocation in sequential binary response experiments. *J. Am. Stat. Assoc.*, **102**, 224-234 (2007)
9. Zhu, H., Hu, F.: Implementing optimal allocation for sequential continuous responses with multiple treatments. *J. Stat. Plan. Infer.*, **139**, 2420-2430 (2009)

Education: students' mobility and labour market

From measurement to explanatory approaches: an assessment of the attractiveness of the curricula programs supplied by Italian universities

Dalla misurazione all'analisi delle determinanti: una valutazione dell'attrattività dei Corsi di Studio in Italia

Isabella Sulis and Silvia Columbu and Mariano Porcu

Abstract The aim of this study is to define the attractiveness of degree programs supplied by Italian Universities using data provided by the National Student Archive (NSA) on a cohort of students enrolled at the universities in a.y. 2017-18. For this sake, micro data on students university mobility choices have been summarized at field of study level in each of the Italian University by defining several indicators of degree programs attractiveness. A classification of universities based on the attractiveness of the fields of study they offer have been advanced which relies on a robust definition of mover and which takes into account multiple aspects of the investigated phenomena. The resulting data set has been investigated using Multilevel Latent Profile Analysis to identify profiles of universities and degree programs.

Abstract *Il lavoro analizza l'attrattiva dei corsi di laurea attivati nelle Università italiane utilizzando i dati forniti dall' Anagrafe Nazionale Studenti relativi alla coorte di immatricolati nell'a.a. 2017-18. I micro dati sulle scelte di mobilità degli studenti sono stati aggregati a livello di corso di studio delle singole università attraverso la definizione di diversi indicatori di attrattività. La scelta di usare indicatori multipli per la definizione di un indice di attrattività ha consentito di avanzare una classifica delle università, in termini di attrattività dei corsi di studio attivati nell'offerta curriculare, che si basa su una definizione robusta di studente in mobilità e che tiene conto di molteplici aspetti complementari. L'utilizzo della Multilevel Latent Profile Analysis ha permesso di identificare profili di università in base all'attrattività dei rispettivi corsi di laurea.*

Key words: university attractiveness, Multilevel Analysis, Latent Profile Analysis, Field of study

---

Isabella Sulis and Mariano Porcu
University of Cagliari, Department of Political and Social Sciences, e-mail: isulis@unica.it, mrporcu@unica.it

Silvia Columbu
University of Cagliari, Department of Mathematics and Computer Science, Italy, e-mail: silvia.columbu@unica.it

1 Introduction

The study of the capability of universities of attracting students from other geographical areas has become a topic of great relevance in the last years, since it is strictly linked with the reputation of the universities and their perceived quality in the public opinion [2, 3, 9], with the depopulation of the youth and educated people from the less developed areas of the country and the related brain drain phenomenon [6, 7, 5, 1, 8, 9, 10, 4]. This study aims to measure the attractiveness of Italian universities focusing on degree programs they supply and addressing the following research questions: (i) How Italian degree programs can be classified according to their capability to attract students? (ii) Which are the main profiles that arise of universities and degree programs in terms of attractiveness? We operationalize attractiveness using multiple indicators and we adopted a multilevel latent profile analysis for shaping profiles of universities and degree programs characterized by different levels of attractiveness. For this sake micro data provided by the National Student Archive (NSA) on a cohort of students enrolled in # 1540 bachelor degrees of # 80 Italian Universities in a.y. 2017-18 has been adopted and individual student mobility choices have been summarized for each university at field of study level (degree program) by defining six indicators of attractiveness which are function of the number of incoming movers, the distance from their city of residence and their status of free or forced movers.

2 Data

We use the micro data provided by the NSA related to a cohort of freshman enrolled in 2017/18 and the information on the two surveys AlmaLaurea on graduated in 2015 in order to build up a data set of Italian bachelor degree programs. The micro data provided by NSA have been summarized (averaged) at degree program level for each university to sketch profiles in terms of socio-demographic background of their students. The distribution of students and degree programs by disciplinary field is listed in Table 1. We use a robust definition of mover by considering as student in mobility only those students who attend a university outside their region (for regional mobility) or the province (for provincial mobility) of residence and who take more than 90 minutes to reach their municipality of residence from the university.

Moreover, we identified free and forced mover students. Namely we classified as free movers only those students who had the possibility to attend the degree programs where they enrolled in their province of residence and chose to move. In this way we overcome issues related to considering as movers, students coming from bordering areas or universities which offer a limited number of degree programs. Six indicators have been considered in order to operationalize the attractiveness dimension of bachelor degree programs offered by Italian universities: 1) the number of incoming robust regional mover students in each degree program over the num-

Table 1 Number of students and degree Programs by disciplinary field

Disciplinary field	N. degrees	N. students
Agrarian Sciences	64	7,979
Other Scienfic	78	7,797
Architercture and Ingeneering	180	45,026
ArteeDesign	63	10,350
Biology and Chemestry	70	12,808
Economics	134	41,759
Pharmacy	48	8,190
Education	63	12,613
Computer Sciences and ICT	65	9,145
Law	91	19,894
Literature	99	13,342
Languges	63	20,229
Mathematics ad Phisic	67	6,447
Politics and Social Sciences	153	25,249
Health Profession	147	15,402
Psychology	32	7,023
Physical education	33	6,192
Total	1450	269,445

ber of enrolled students in each degree program (`INCIDENCER`); 2) the number of incoming robust provincial mover students in each degree program over the number of enrolled students in each degree program (`INCIDENCEP`); 3) the number of incoming robust regional movers attracted by a degree program of a given university on the total number of regional movers in the specific study field (classe di laurea) of the degree program in Italy (`QUOTEMOVERR`); 4) the number of incoming robust provincial movers attracted by a degree program of a given university over the total number of regional movers in the specific study field (classe di laurea) of the degree program in Italy (`QUOTEMOVERP`); 5) the number of free movers coming from other regions (without applying the condition of the 90 minutes travel time) over the total number of enrolled students (`FREEMOVERR`); 6) the average travel time of students coming from other regions using the travelling time between municipalities estimated by ISTAT (`AVERAGETRAVEL`)[1]. The average values of the indexes and their standard deviations (see Table 2) show a remarkable variability of the six indicators across the four geographical macro-areas and within the macro-areas. This clearly highlight the divergences between Centre-North and South universities not only in the capability to attract students but also in the type of mover students attracted (free versus forced movers) and in the extension of their catching area in the national framework. The dimensionality of the selected attractiveness indicators has been assessed using Principal Component Factors Analysis on the standardized indicators. Results clearly suggest the presence of a dominant factor which explains about 72% of the common variance.

[1] For this last indicator we had to impute values for incoming and outgoing mover students in Sardinia and Sicily (less than 3% of the population) as the Istat database provides for the two islands only the travel time for between municipalities within region movements.

Table 2 Descriptive Statistics attractiveness of indicators

	INCIDENCER		INCIDENCEP		QUOTEMOVERR	
MACRO-AREA	mean	sd	mean	sd	mean	sd
CENTRE	0.18	0.15	0.20	0.16	0.04	0.06
ISLANDS	0.02	0.04	0.12	0.11	0.003	0.01
NORTH	0.15	0.14	0.16	0.14	0.05	0.09
SOUTH	0.06	0.11	0.09	0.12	0.02	0.05
	QUOTEMOVERP		FREEMOVERR		AVERAGETRAVEL	
MACRO-AREA	mean	sd	mean	sd	mean	sd
CENTRE	0.04	0.06	0.20	0.15	65.83	39.85
ISLANDS	0.02	0.02	0.04	0.09	40.61	13.78
NORTH	0.05	0.08	0.23	0.17	64.87	36.20
SOUTH	0.02	0.05	0.10	0.15	40.38	22.89

3 Modelling Approach

Multilevel Latent Profile Analysis [12] has been used to jointly classify universities and degree programs in Level-2 and Level-1 latent classes and to sketch profiles for both. The number of lower and higher level classes has been defined on the basis of the improvement in the goodness of fit measures, heterogeneity in the classes and interpretability of the results [11]. By indicating with $j = 1,\ldots,n_k$ the 1540 degree programs on which the analysis refers to (Level-1 units) and with $k = 1,\ldots,K$ the 80 universities (Level-2 units) in which degree programs are clustered, the vector $\boldsymbol{y}_{kj}$ of observed values for units kj is explained by defyning two discrete latent variables U (defined by $h = 1,\ldots,H$ latent classes) and V (defined by $m = 1,\ldots,M$ classes), respectively at Level-1 and Level-2 units. A 2-level latent profile model is described by two equations representing respectively the level-2 and the level-1 membership. That is at second level

$$f(y_k) = \sum_{h=1}^{H} P(v_k = h) \prod_{j=1}^{n_k} f(\boldsymbol{y}_{kj}|v_k = h),$$

with $P(v_k = h)$ membership probability of Level-2 units to profile h, and at first level

$$f(\boldsymbol{y}_{kj}|v_k = h) = \sum_{m=1}^{M} P(u_{kj} = m|v_k = h) \prod_{i=1}^{I} f(y_{kji}|u_{kj} = m, v_k = h),$$

where $P(u_{kj} = m|v_k = h)$ is the conditional membership probability of first level units to profile m, given the belonging in level-2 class h, and y_{kji} is the value of the indicator i ($i = 1,\ldots,I$) for unit kj. We assume that $f(y_{kji}|u_{kj} = m, v_k = h)$ is normally distributed with mean μ_{jmkh} and variance σ^2_{jmkh}).

4 Results

On the basis of these aspects, four Latent Classes have been identified to profile degree programs attractiveness and three Latent Classes to differentiate between universities, as Table 3 shows. The profiles of the latent classes at Level-1 (Table 3 (a)) have been drawn looking at the differences in the expected values of the indicator items (expressed in z-score) between degree programs belonging to different classes. Namely the degree programs in LC2 have on average a value of the z-score equal to -0.96, thus it detects *Low attractive degree programs*, whereas degree programs in LC4 have an average z-score equal to 1.53, thus it identifies *High attractive degree programs*. LC3 and LC2 with values equal to -0.61 and 0.05 detect respectively *Medium-Low attractive* and *Medium attractive* degree programs. The value of the Wald test statistics show that the average values of the indicators significantly differ between LCs. At Level-2 three profiles of Universities arise on the basis of the proportion of degree programs classified in each Level-1 category (Table 3 (b)). *Low* attractive universities have about 50% of the degree programs classified in the *Low* attractive LC at Level-1, *Medium* attractive universities have about 68% of degree programs classified as *Medium-Low and Medium* attractive, whereas *High* attractive universities have 53% of degree programs classified as *High attractive*.

Table 3 Results Multilevel Latent Profile Analysis: L-1 Degree Programs, L-2 Universities

a) Latent Classes Profiles: Level 1							
Indicators	Medium-Low	Low	Medium	High	Wald	p-value	R^2
	LC1	LC2	LC3	LC4			
INCIDENCER	-0.72	-1.06	0.07	1.71	2072.67	0.00	0.79
rank	2	1	3	4			
INCIDENCEP	-0.62	-1.15	0.10	1.67	2698.00	0.00	0.80
rank	2	1	3 4				
QUOTEMOVERR	-0.59	-0.74	-0.13	1.46	882.73	0.00	0.54
rank	2	1	3 4				
QUOTEMOVERP	-0.54	-0.76	-0.10	1.40	1128.68	0.00	0.51
rank	2	1	3	4			
FREEMOVERR	-0.65	-1.11	0.12	1.63	2084.44	0.00	0.77
rank	2	1	3	4			
AVERAGETRAVEL	-0.58	-0.95	0.24	1.29	1166.73	0.00	0.58
rank	2	1	3	4			
mean	**-0.61**	**-0.96**	**0.05**	**1.53**			
average rank	2	1	3	4			
Labels LC	Medium-Low	Low	Medium	High			
b) Latent Classes Profiles: Level 2							
Level 2	Level-1 Latent Classes						
Latent Classes	Medium-Low	Low	Medium	High			
Low	0.378	**0.501**	0.103	0.018			
Medium	**0.323**	0.173	**0.361**	0.142			
High	0.089	0.001	0.381	**0.530**			

5 Conclusion

Main findings clearly suggest the clustering of universities and degree programs in lower and higher level latent classes with different intensity of the underlying latent trait. Latent classes at Level-1 and Level-2 are well differentiated at both levels of analysis. Further analyses focused on the e determinants of class assignment are still in progress.

References

1. Attanasio, M., Enea M.: La mobilità degli studenti universitari nell'ultimo decennio in Italia. In: De Santis G., Pirani E., Porcu, M. (eds) Rapporto sulla popolazione. L'istruzione in Italia, pp 43 – 58. Il Mulino, Bologna (2019).
2. Bratti, M. and Verzillo, S.: The gravity of quality: research quality and the attractiveness of universities in Italy. Regional Studies. **53** (10), 1385–1396 (2019).
3. Ciriaci, D.: Does university quality influence the interregional mobility of students and graduates? The case of Italy. Regional Studies. **48**(10):1592-1608 (2014).
4. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.: Exploring the Italian student mobility flows in higher education. In: Bini M., Amenta P., D'Ambra A., Camminatiello I. (eds). Statistical Methods for Service Quality Evaluation. Book of short papers of IES 2019 Rome, Italy, July 4-5, pag. 46-49. Cuzzolin Editore, Napoli (2019).
5. D'Agostino A., Ghellini G., Longobardi S.: Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. Electronic Journal of Applied Statistical Analysis. **12**(4): 826-245 (2019).
6. Dotti, N., Fratesi, U., Lenzi, C., and Percoco, M.: Local labour markets and the interregional mobility of Italian university students. Spatial Economic Analysis. **8**(4):443-468 (2013).
7. Fratesi, U., Percoco, M. (2014): Selective Migration, Regional Growth and Convergence: Evidence from Italy. Regional Studies **48**(10): 1650–1668.
8. Genova V. G., Tumminello M., Enea M., Aiello F.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electronic Journal of Applied Statistical Analysis. **12**(4): 774-800 (2019).
9. Giambona, F., Porcu, M., and Sulis, I.: Students mobility: assessing the determinants of attractiveness across competing territorial areas. Social Indicators Research, **133**(3):1105-1132 (2017).
10. Lombardi, G., Ghellini, G.: The effect of grading policies on Italian Universities' attractiveness: A Conditional Multinomial Logit approach. Electronic Journal of Applied Statistical Analysis. **12**(4): 801-825 (2019).
11. Lukočienė, O., Varriale, R, Vermunt J.K.: The Simultaneous Decision(S) About The Number Of Lower- And Higher-Level Classes In Multilevel Latent Class Analysis. Sociological Methodology. **40**(1):247 - 283 (2010).
12. Vermunt J. K.: Multilevel Latent Class Models. Sociological Methodology. **33**(1): 213-239 (2003).

Pull factors for university students' mobility: a gravity model approach

Il gravity model per la individuazione dei fattori di attrazione per la mobilità studentesca

Giovanni Boscaino and Vincenzo Giuseppe Genova

Abstract Migration phenomena are characterised by flows that are typically multilateral. Often an outgoing flow corresponds to one or more incoming flows, which counterbalances it. When the balance is negative, socio-economic problems can arise. Italy has been afflicted for years by a particular type of unilateral migration: student mobility. Thousands of students leave the South to study in the Centre-North, driven by the better socio-economic conditions of those areas, and by more prosperous job prospects. Since this flow is not followed by a return one, nor by other incoming flows, the historical, socio-economic gap between North and South is widened. Taking advantage of the big dataset concerning the cohorts of students enrolling in Italian universities, made available by the Ministry of Education, we adopted the viewpoint of gravity models to investigate the social, economic and educational aspects of universities and destination areas that can make mobility attractive, studying them also over time.

Abstract *I fenomeni migratori sono caratterizzati da flussi tipicamente multilaterali: ad un flusso in uscita ne corrisponde uno o più in entrata che ne bilancia il saldo. Ma accade a volte si assiste a flussi sbilanciati, con saldo molto negativo conducendo a problemi socio-economici. L'italia da anni è afflitta da un particolare tipo di migrazione unilaterale: la mobilità studentesca. Migliaia di studenti abbandonano il Sud per studiare al Centro-Nord, attratti dalle migliori condizioni socio-economiche. Poiché tale flusso non è poi seguito da uno di ritorno, né da un altro in entrata, lo storico divario socio-economico tra il Nord e il Sud è amplificato. Sfruttando il dataset delle coorti degli studenti che si iscrivono alle università italiane, fornito dal MUR, per indagare gli aspetti socali, economici e formativi delle università e delle aree di destinazione che possono rendere appetibile la mobilità, studiandoli anche nel tempo, secondo l'approccio del gravity model.*

Key words: Student mobility, gravity model, mobility pull factors

---

Giovanni Boscaino and Vincenzo Giuseppe Genova
Dipartimento di Scienze Economiche, Aziendali e Statistiche, viale delle Scienze, Edificio 13, Palermo (Italia), e-mail: giovanni.boscaino@unipa.it, vincenzogiuseppe.genova@unipa.it

1 Introduction

The critical role of a university for the development of a territory is widely accepted in the literature. Its importance is due mostly to two factors. First, universities play a core role in the level of education in the area in which they are situated. This importance is amplified where there is a knowledge-based economy, for which human capital is of primary importance and the labour market requires highly qualified people [9]. Second, Perry & Wiewel [12] introduced the notion of *university as urban developer and regenerator*, having positive, direct and indirect effects to the environment around. If it can attract students, then income will be created from them, due to the increase in both population and demand of consumer goods. In Italy, since the 90's students' mobility has brought a significant national concern for the socio-economic future of the Country. Specifically, students from the southern regions tend to move toward the northern ones, and this phenomenon is seen as a perpetual mechanism of the regional disparities and the existing dualism between the southern and northern regions of Italy [1]. From the origin region perspective, students' migration may severely hamper the regional potential in many ways. First of all, when students move from one region to another one, they have to be financially supported by their families, creating a transfer of capital to the destination ones. Second, the origin regions could lose their best students and so the best human capital for their future growth, since students are more likely to find a job in that region. If graduates do not come back to their home regions, these regions will have less human capital, losing the ability to attract external investments, that together with their economy, will be negatively affected. In the academic year 2016-2017, compared to the 685,000 southerners enrolled at the University, 25.6% study at a University in the Centre-North. The share of residents in the Centre-North studying in the South is only 1.9%. The net migration balance is about 157,000 units, and is constantly increasing. Students who emigrate for study purposes make up about 0.7% of the resident southern population [5]. The empirical evidence shows how, when we consider the entire range of working life, the rate of Return On Investments in human capital in the southern regions is about half of that observed in the "richest" regions of the Center-North (Piedmot, Lombardy and Emilia-Romagna, [6]). Besides, students' choice regarding where to study often depends on the (perceived) "quality" of the training offered by the university. In Italy, the largest regions usually offer the broadest range of topics, becoming an attractive pole for the migratory flow of students. It emerges rather clearly the presence of a market failure [8]. The economic theory suggests how there should be market equilibrium characterised by investments smaller than the optimal ones, or sub-optimal choices in terms of "quality" of universities or courses that fit better the students' needs. The study carried out by Dotti et al. [9] shows that, although the quality of a university and its characteristics play a central role in the decision process of a student, the characteristics of the labour market in a specific region must be taken into account as well. They point out that in the southern regions there are not universities which attract students living more than 200 kilometres of distance, and the number of registrations of a university is linked to the job prospects of students, so to the expected number of job offers.

Amendola & Restaino [7] focus the attention to the determinants and characteristics of the mobility process from and toward Italy. They show that the number of students interested in spending part of their academic education abroad is steadily increasing. This is not a surprise, given that students who choose to study abroad earn much in terms of life experience, intercultural competencies and quality of studies, being more and more willing to learn and/or to strengthen their competencies and knowledge. The study carried out by Camillo et al. [11] on Italian graduates who work abroad, shows that more than 80% of graduates interviewed five years after getting their degree would make the same choice because they were very satisfied and would not go back in Italy. So, it seems that the Italian labour market is not able to increase the value of human capital generated by Italian universities. Students' enrolment has decreased significantly, especially after the economic crisis of 2008, while the students' migration from the South to the other regions of the Country has increased. This phenomenon has created further inequalities within the Country and a cultural and socio-economic loss for the South of Italy. The study presented in this paper aims to explore the determinants of students' mobility in Italy. In particular, the focus is on the pull factors of the destination universities and areas. We have considered information coming from institutional statistics about universities' and city areas characteristics, seeking for some aspects that can motivate students to abandon the South.

2 Methodology

In literature, the Gravity Models approach is one of the most adopted to study migration flows from an origin to a destination place. The main idea behind the Gravity Model is the Newton gravity law: a gravity model assumes that flow (the attraction force) is proportional to the sizes of origin and destination (of bodies' masses) but inversely proportional to the distance of these two places (bodies). In its general form a gravity model can be expressed as $N_{ij} = G\frac{Y_i Y_j}{d_{ij}}$ where, considering mobility flows, N_{ij} is the number of people move from area i to area j; G is a constant; Y_i is the number of people in area i; Y_j is the number of people in area j; d_{ij} is the distance between countries i and j. Under a modelling approach previous equation can be expressed by logarithmic transformation:

$$\ln(N_{ij}) = \beta_0 + \beta_1 \ln(Y_i) + \beta_2 \ln(Y_j) + \beta_3 \ln(d_{ij}) + \varepsilon_{ij} \quad (1)$$

In literature many reformulation of Eq. 1 were proposed, in particular authors proposed to substitute the d_{ij} with variables that act as detractors, *e.g.* the travel costs. For example, concerning students' mobility, Beine *et al.* [2] argued that students' mobility could be affected by migration costs and the size of the network at the destination. According to them, the network at the destination is defined as migrants from the origin city living at the destination and this network can facilitate migration flows. Indeed previous migrants are likely to provide assistance and information

to those students decide to migrate in the city j, reducing migration costs [2, 3]. The goal of our analysis is to study the determinants of students' mobility and the possible effects of the migration chain. When students move because they are advised and invited by friends or relatives living in the destination area, we are in the presence of a migratory chain [4]. According to Beine *et al.* [2], we decided to analyse students' mobility following a Gravity Model approach with mixed effects on clusters of Sicilian municipalities (clusters of origin), and on the provinces of "out of Sicily" universities (provinces of destination). Clusters come from [10]: 38 clusters grouping the 390 Sicilian municipalities, based on economic, commercial and geographical proximity aspects, and the spoken language (Italian vs local dialects). In our analysis we assume the response variable is $N_{ij|k} \sim NB(\mu_{ij|k}, \sigma_{ij|k})$, where $N_{ij|k}$ is the number of outgoing students from cluster i to university's province j conditioned to the $k-th$ covariate profile. Under these assumptions the model is:

$$\ln(\mu_{ij|k}) = \ln(N_{i|k}) + \alpha_{ij} + V_i\phi + X_j\beta + Z_u\gamma + \theta_{ij}\ln(M_{iju}) + \varepsilon_i + \omega_j \qquad (2)$$

where $N_{i|k}$ is the total number of freshmen of the i cluster of Sicilian municipalities, α_{ij} is the intercept, V_i is the vector of student's characteristics in the cluster i, X_j are socio-economic covariates of the provinces of destination, Z_u are covariates related to the characteristics of university u, and M_{iju} are students of previous years from cluster i that study at university u in the province j. ϕ, β, and γ are vectors of unknown regression parameters and θ_{ij} is the network effect. $\varepsilon_i \sim N(0, \sigma^2)$ is the cluster of origin random effect, and $\omega_j \sim N(0, \sigma^2)$ is the province of destination level random effect. As the aim of this analysis is to evaluate pull factors in students' mobility, we decided to include in the model socio-economic covariates and universities covariates that can act as pull factors. Another goal of this analysis is to measure the migration chain effect in students' mobility. In this first approach, we decided to use as migration chain effect the network measure proposed in [3], that is expressed by θ in eq. 2. Our idea is if a chain migration effect in Italian university students there exists, the parameter θ should be positive and significant.

3 Results and comments

The model in Eq. 2 was estimated for data coming from the Italian Ministry of Education. In particular, data regards individual information about two cohorts (2014 and 2017) of the population of the Italian freshmen followed up to graduation. Many variables are available: socio-demographic (Gender, Area of study, High School Finale Grade, and Diploma Type), life quality of the province of destination (Unemployment rate, House Price, Free time indicator, Public order indicator), and university characteristics (Scholarship, Indicators of Services, Structures, and Communication). The baseline profile is a student with high school grade $[60-70]$, from scientific high school, that enrol at a Degree Course of the humanistic area, with quantitative covariates. For sake of simplicity and pages limit, the estimates of the

four models are reported in Table 1. Concerning the baseline profile, the expected number of outgoing students, for the cohort 2014, is grater for those students choose a non-humanistic area, and in particular the expected number of outgoing students for male seems to be greater in the social area and for female in the scientific area. Looking at the cohort 2017 (last two columns) we notice that outgoing male students prefer the scientific area and females the health one, *ceteris paribus*. In the 2017 cohort, instead, male students with lower mobility are those choose the social area and females the scientific area, maybe due to the labour market in the destination provinces which could have determined the degree courses choice.

Table 1 Model estimates by gender and cohort, standard errors in parenthesis.

	2014		2017	
Parameters	Males (*s.e.*)	Females (*s.e.*)	Males (*s.e.*)	Females (*s.e.*)
Intercept	-3.396 (1.11)	-4.904 (0.39)	-7.997 (2.30)	-8.877 (1.49)
Health area	0.090 (0.06)	0.040 (0.04)	0.071 (0.07)	0.069 (0.04)
Scientific area	0.118 (0.07)	0.093 (0.04)	0.080 (0.06)	-0.051 (0.04)
Social area	0.169 (0.06)	0.068 (0.04)	0.042 (0.06)	0.019 (0.03)
Grade (70,80]	0.128 (0.04)	0.122 (0.04)	-0.070 (0.04)	0.024 (0.04)
Grade (80,90]	0.101 (0.04)	0.119 (0.04)	-0.043 (0.04)	0.053 (0.04)
Grade (90,100]	0.288 (0.04)	0.255 (0.04)	0.233 (0.04)	0.146 (0.04)
Classical	-0.299 (0.04)	-0.103 (0.03)	-0.283 (0.04)	-0.058 (0.03)
Professional	-0.528 (0.09)	-0.343 (0.08)	-0.308 (0.07)	-0.121 (0.06)
Technical	-0.304 (0.04)	-0.296 (0.05)	-0.213 (0.04)	-0.126 (0.04)
Other diploma	-0.394 (0.07)	-0.207 (0.04)	-0.372 (0.06)	-0.083 (0.03)
Unemployment rate	-0.012 (0.01)	-0.005 (0.01)	-0.010 (0.01)	-0.004 (0.01)
House price	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)
Free time	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)	0.001 (0.00)
Public order	-0.001 (0.00)	0.000 (0.00)	-0.002 (0.00)	0.000 (0.00)
M_{iju}	0.004 (0.00)	0.003 (0.00)	0.004 (0.00)	0.004 (0.00)
Services	-0.010 (0.01)	-0.001 (0.00)	-0.018 (0.01)	-0.011 (0.01)
Scholarships	0.011 (0.00)	0.001 (0.00)	0.015 (0.00)	0.005 (0.00)
Structures	0.002 (0.00)	0.000 (0.00)	0.015 (0.01)	0.006 (0.01)
WEB	-0.011 (0.01)	-0.002 (0.00)	0.030 (0.01)	0.000 (0.01)

Furthermore, regardless the gender and the cohorts, looking at the high school grade, the higher the grade, the bigger the expected number of outgoing students and this result could be in according to the idea that families let move their children if they are very talented and motivated. Instead, we can see differences due to the high school diploma. Indeed, the expected number of outgoing students is lower for students have not a Scientific lyceum diploma, and this result suggests the students with a higher mobility profile come from the Scientific and Classical Lyceum. Looking at pull factors related to the areas and the universities of destination, Table

1 highlights that the unemployment rate at the destination seems to play a role in the student mobility: indeed, regardless the cohort and the gender the higher the rate, the lower the expected number of outgoing students. In addition, such an effect is higher for male students than females, and it does not show great differences between the two years. Moreover, the more accessible the scholarships are at the enrolment, the higher the number of students that chose an "out of Sicily" university is. Also, the effect related to the scholarship seems to change according to gender. With respect to the migration chain effect in students' mobility, the explanatory variable M_{iju} (defined as students that from cluster of origin i move to province of destination j to study at university u in previous years) highlights a migration chain if it will be positive and significant (as explained in section 2). Looking at Table 1, such effect seems exist and it's highly significant, and as expected this result could suggest migration chain plays a role in student mobility. Also, this result seems to be similar and stable with respect to the gender and over time.

Acknowledgements This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P

References

1. M. Attanasio, M. Enea: La mobilità degli studenti universitari nell'ultimo decennio in italia. Rapporto sulla Popolazione. L'istruzione in Italia. Il Mulino (2019).
2. M. Beine, R. Noel, L. Ragot: Determinants of the international mobility of students. Economics of Education Review (2014) doi:10.1016/j.econedurev.2014.03.003
3. M. Beine, F. Docquier, C. Ozden: Diasporas. Journal of Development Economics (2011) issn:0304-3878
4. Haug, S. (2008). Migration networks and migration decision-making. Journal of Ethnic and Migration Studies, 34(4):585–605.
5. Svimez (2018). *Rapporto Svimez 2018*. Bologna, Il Mulino.
6. Guadagnini, M., Mussida, C. (2009). *Il rendimento dell'istruzione nelle regioni italiane*. Prometeia, mimeo.
7. Amendola A., Restaino M. (2016). *An evaluation study on students' international mobility experience*. Quality and Quantity. Pag.1-20.
8. Ciriaci, D., Nuzzi, A., (2012). *Qualità dell'Università e mobilità dei laureati italiani: evidenze empiriche e proposte di policy*. Istituzioni del Federalismo
9. Dotti, N. F., Fratesi, U., Lenzi, C. & Percoco, M. (2013). *Local Labour Markets and Interregional Mobility of Italian University Students*, Spatial Economic Analysis, Vol. 8 No. 4, 443-468.
10. Genova, G.V., Tumminello, M., Enea, M., Aiello, F., and Attanasio, M. (2019). *Student mobility in higher education: Sicilian otflow network and chain migration*, Electronic Journal of Applied Statistical Analysis, Vol. 12, No. 4, 774-800.
11. Camillo, F., Vittadini, G., Binassi, S. (2016). *International migration of Italian graduates*.
12. Perry, D. C., Wiewel, W. (2005). *The University as Urban Developer. Case Studies and Analysis.* Cities and Contemporary Society.

Spatial autoregressive gravity models to explain the university student mobility in Italy

Modelli gravitazionali spaziali autoregressivi per spiegare la mobilità studentesca in Italia

Silvia Bacci, Bruno Bertaccini, Chiara Bocci

Abstract We investigate the mobility of Italian academic students among geographical areas (i.e., provinces) to attend university. The study relies on data collected by the Italian National Student Registry and concerns students enrolled in the academic year 2011–2012 in a bachelor degree program or a five-years degree program of any Italian university. The methodological approach we adopt is based on the analysis of the flows of students among provinces through spatial autoregressive gravity models. The gravity component of this type of models accounts for the deterrence effect due to the distance among province of origin and province of destination. Instead, the spatial autoregressive component is introduced to capture homogenous behaviours among contiguous geographical areas. In particular, we focus on alternative ways to specify the spatial weight matrix that characterises the spatial autoregressive component of the models at issue.

Abstract *Oggetto di questo contributo è l'analisi della mobilità degli studenti universitari italiani tra aree geografiche (province) per frequentare l'università. Lo studio si basa su dati raccolti dall'Anagrafe Nazionale Studenti e riguarda gli immatricolati ad un corso di laurea triennale o a ciclo unico nell'anno accademico 2011-2012 in un qualsiasi ateneo italiano. L'approccio metodologico adottato è basato sull'analisi dei flussi di studenti tra province tramite un'idonea specificazione di un modello gravitazionale spaziale autoregressivo. La componente gravitazionale del modello tiene conto dell'effetto deterrente dovuto alla distanza tra provincia di origine e provincia di destinazione. La componente autoregressiva spaziale viene, invece, introdotta per catturare comportamenti omogenei tra aree geografiche contigue. Una particolare attenzione è posta sulle possibili specificazioni della matrice dei pesi spaziali, caratterizzante la componente autoregressiva spaziale dei modelli in oggetto.*

Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence
e-mail: silvia.bacci@unifi.it, bruno.bertaccini@unifi.it, chiara.bocci@unifi.it

Key words: Anagrafe Nazionale Studenti, Gravity model, Origin destination flows, Spatial correlation, Student mobility.

1 Introduction

Mobility of academic students across a country (inter-regional mobility) to attend university has important consequences on the economic development of the geographical areas involved, as the migration for study reasons is often just a precursor of the migration for job reasons and students that move away from home to take a degree tend not to return at the end of the academic studies. This phenomenon is particularly remarkable in Italy where many students living in the South are attracted by university courses offered in the North of Italy. Some recent contributions in such a setting are [6], [7], [14].

In the present contribution we are interested in detecting homogenous geographical areas having a similar behaviour in terms of student mobility. More in detail, we investigate the possibility of aggregation of: (i) geographical areas of origin on the basis of their tendency to "send out" young people living there to attend university and (ii) geographical areas of destination where academic campus are located on the basis of their tendency to attract students from outside. The study deals with the first-level mobility, that is the transition from the high-school to the bachelor (or five-years) degree. Hence, the focus is on the flows of students from the place of residence to a university located elsewhere, with provinces considered as the territorial unit of reference.

The statistical methodology we adopt is based on the Spatial Auto-Regressive (SAR) gravity models.

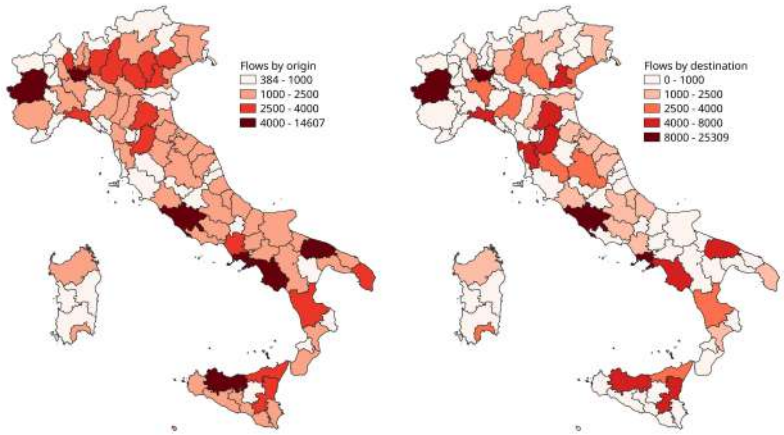
The class of gravity models (for a review see [1] and [13]) is typically adopted to explain incoming and outgoing flows (of migrants, goods, workers, and so on) between geographical areas (e.g. countries). In the higher education field, gravity models are useful to explain the flows of students across territories and/or universities; examples of gravity models applied to the university context are given by [12], [3] and [6]. In addition, some spatial components may be introduced to take into account the spatial correlation between flows [10], after having controlled for observable characteristics of the origins and destinations.

The data used in the study comes from the Italian National Student Registry (in Italian, Anagrafe Nazionale Studenti - ANS), which is an administrative database that monthly registers and monitors the careers of university students enrolled in any degree program in a public or private university located in Italy.

2 Data

Analyses are carried on freshmen enrolled in the academic year 2011–2012 in a bachelor degree program or a five-years degree program (i.e., degree program that usually lasts five or six years), defined according to Law 509 of 1999 and Law 270 of 2004, of any Italian higher education institution, with the exclusion of on-line universities. The set of students amounts to 207,388 distributed in 71 universities and 102 provinces (note that a certain number of universities have campus dislocated on several provinces). The biggest universities are the University of Rome "La Sapienza" (5.3%), Bologna (5.0% of students) and Neaples "Federico II" (5.0%). The 28.9% of universities is located in the North-West of Italy, the 21.5% in the North-East, the 23.1% in the Centre, the 19.5% in the South of Italy and the remaining 7.0% in the Islands (i.e., Sicily and Sardegna). A more detailed description of the data at issue together with some assessments of the capability of attraction and retention of students by universities is provided by [2].

Around one-half (48.6%) of freshmen enrolled in a bachelor or five-years degree program comes from a province different from the one where he/she lives. Obviously, in certain cases the migration is "forced" because the higher education offer of several provinces is completely absent or strongly limited (see Fig. 1). For this reason, the study of flows needs to take into account the distinction between forced migrant students and free migrant students [9] or, similarly, the distinction between provinces that supply a complete and varied offer of bachelor degree programs and provinces with a limited or absent offer.



(a) (b)

Fig. 1 Number of outgoing students by province of residence (panel a) and number of incoming students by province of destination (panel b).

3 The SAR gravity model

The gravity model we rely on is formulated as [8]

$$\log Y_{hi} = \beta_0 + \gamma_t \log d_{hi} + \log \vec{x}'_{Oh}\vec{\beta}_O + \log \vec{x}'_{Di}\vec{\beta}_D + e_{hi}, \tag{1}$$

with Y_{hi} denoting the flow of students from the origin h to the destination i, with h representing the province of residence and i the province where is located the university attended by the student. Moreover, d_{hi} is the road distance in kilometers between h and i: as the distance between h and i increases, we should observe a reduction in the flows towards i. Vectors $\vec{x}_{Oh}$ and $\vec{x}_{Di}$ collect characteristics of the origin h and destination i, respectively, that affect outgoing and incoming flows, such as the type of higher educational offer supplied (e.g., complete and varied, limited, absent). Moreover, β_0 is the constant term and γ_t, $\vec{\beta}_O$, and $\vec{\beta}_D$ are the regression coefficients associated with the distance and the other characteristics of the origin and the destination; e_{hi} is the error term. Considering n provinces, the model analyses all flows for n^2 origin-destination pairs of provinces. The $n \times n$ flow matrix contains intra-provincial flows in its main diagonal and inter-provincial flows in its off-diagonal elements; it is vectorised by stacking the columns to form an $n^2 \times 1$ vector of log-flows contained in $\vec{y}$. Hence, model (1) may be written in matrix notation as

$$\vec{y} = \vec{\beta}_0 + \vec{\gamma}\vec{d} + \vec{X}_O\vec{\beta}_O + \vec{X}_D\vec{\beta}_D + \vec{e}, \tag{2}$$

where $\vec{d}$ is the $n^2 \times 1$ vector of log-distances $\log d_{hi}$, $\vec{X}_O = \vec{X} \bigotimes \vec{1}_n$ and $\vec{X}_D = \vec{1}_n \bigotimes \vec{X}$, with $\vec{X}$ being the $n \times p$ matrix of the provinces' p characteristics. To take simple the notation we assume the same covariates for origin and destination provinces, but this assumption is not necessary. In practice, the Kronecker product $\bigotimes$ repeats the same values of the n provinces to create a matrix of covariates associated with each origin ($\vec{X}_O$) and each destination ($\vec{X}_D$).

The gravity model (2) assumes independence among the origin-destination flows. It is possible to relax this assumption to account for different types of spatial dependence between geographical areas [10, 11]. In detail, the SAR gravity model is specified as follows

$$\vec{y} = \vec{\beta}_0 + \rho_O\vec{W}_O\vec{y} + \rho_D\vec{W}_D\vec{y} + \rho_{OD}\vec{W}_{OD}\vec{y} + \vec{\gamma}\vec{d} + \vec{X}_O\vec{\beta}_O + \vec{X}_D\vec{\beta}_D + \vec{e}, \tag{3}$$

with $\vec{W}_O = \vec{W} \bigotimes \vec{1}_n$, $\vec{W}_D = \vec{1}_n \bigotimes \vec{W}$, and $\vec{W}_{OD} = \vec{W} \bigotimes \vec{W}$, where $\vec{W}$ is the $n \times n$ spatial weight matrix that accounts for the neighbour relationships between the n provinces. The coefficients ρ_O, ρ_D, and ρ_{OD} measure the strength of the origin-based, destination-based and origin-to-destination-based dependence, respectively. In other words, these correlation coefficients denote how larger flows from origin h to destination i are accompanied by: larger flows from neighbours of h to i (ρ_O), larger flows from h to neighbours of i (ρ_D), larger flows from neighbours of h to neighbours of i (ρ_{OD}). Note that the model specification could include one or more of the spatial components.

4 Specification of the spatial weight matrix $\vec{W}$

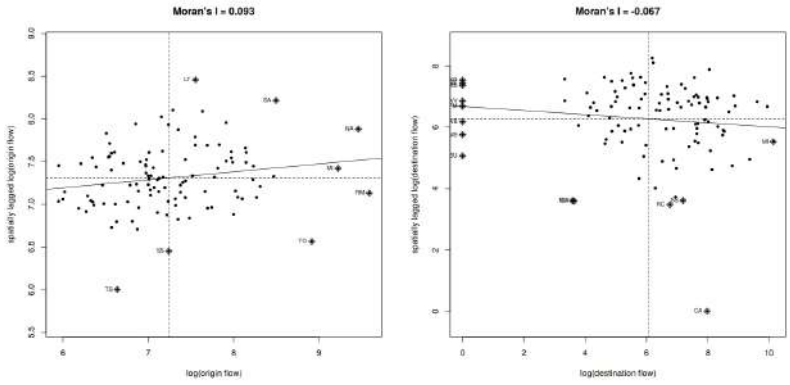
The specification of a SAR gravity model requires a suitable definition of the spatial weight matrix $\vec{W}$. We can adopt two main approaches to define the elements w_{hi} of $\vec{W}$: (i) taking into account actual geographical relations or (ii) considering socio-economic similarities. Note that, whatever the definition of $\vec{W}$ matrix is, elements on the principal diagonal are set to 0 by definition.

Point (i) represents the traditional approach to the definition of $\vec{W}$. Typical choices are given by:

- contiguity matrices of order 1 (i.e., we consider only areas that share a border) or higher (i.e., we also consider neighbours of neighbours): $w_{hi} = 1$ if area h and area i are contiguous (of order 1 or higher), and 0 otherwise;
- k-nearest neighbours: $w_{hi} = 1$ if i is one of the k-nearest areas to h, and 0 otherwise;
- distance-based neighbours: $w_{hi} = 1$ if i is within a specified distance from h, and 0 otherwise;
- inverse distance: $w_{hi} = 1/d_{hi}$.

A different perspective is adopted under point (ii), where the elements of $\vec{W}$ are defined according to measures of similarity alternative or complementary to the geographical ones. The recent literature proposes the use of socio-economic variables that characterise the substantial (not necessarily geographical) neighbourhood among areas; for instance, [5] use a distance between areas based on a quality of life indicator.

Preliminary results obtained using contiguity matrices of order 1 are displayed in Figure 2, where the spatial autocorrelation Moran's index is plotted both for outgoing flows (panel a) and for the incoming flows (panel b).



(a) (b)

Fig. 2 Moran's index plot for the outgoing flows (panel a) and the incoming flows (panel b).

The Moran's index outlines weak (although significant) spatial correlation between flows of students from different provinces, with a few provinces having a leverage effect. In order to better investigate the spatial correlation structure of data at issue, the future work will be focused on the evaluation of alternative specifications of the spatial weight matrix $\vec{W}$, mainly based on socio-economic similarities (point (ii) above) in order to detect groups of provinces sharing a common behaviour in terms of propensity to send out students to attend university elsewhere and/or to attract students from other provinces.

Acknowledgements The authors acknowledge the financial support provided by the "Dipartimenti Eccellenti 2018-2022" ministerial funds.

References

1. Anderson, J. E., The gravity model. Working Paper 16576, National Bureau of Economic Research, Cambridge, MA (2010)
2. Bacci, S. and Bertaccini, B., Assessment of the university reputation through the analysis of the student mobility, Social Indicators Research, DOI: 10.1007/s11205-020-02322-x (2020)
3. Bruno, G. and Genovese, A., A spatial interaction model for the representation of the mobility of university students on the Italian territory, Networks and Spatial Economics, 12, 41–57 (2012)
4. Columbo, S., Porcu, M., Primerano, I., Sulis, I., and Vitale, M. P. Exploring the Italian student mobility flows in higher education. In M. Bini, P. Amenta, A. D'Ambra, and I. Camminatiello (Eds.), Statistical methods for service quality evaluation, 46–49. Cuzzolin (2019)
5. Delgado, F. J. and Lago-Peñas, S. and Mayor, M.: Local tax interaction and endogenous spatial weights based on quality of life, Spatial Economic Analysis, 13, 296–318 (2018)
6. Dotti, N. F., Fratesi, U., Lenzi, C., and Percoco, M., Local labour market conditions and the spatial mobility of science and technology university students: evidence from Italy, Review of Regional Research, 34, 119–137 (2014)
7. Enea, M. From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree. In P. C., M. Pratesi, and A. Ruiz-Gazen (Eds.), Studies in theoretical and applied statistics, 239–249. Springer (2018)
8. Fotheringham, A. S. and O'Kelly, M. E., Spatial interaction models: Formulations and applications. Kluwer Academic, Dordrecht (1989)
9. Giambona, F., Porcu, M., and Sulis, I., Students mobility: assessing the determinants of attractiveness across competing territorial areas, Social Indicator Research, 133, 1105–1132 (2017)
10. LeSage, J. P. and Pace, R. K.: Spatial econometric modeling of origin-destination flows, Journal of Regional Science, 48, 941–967 (2008).
11. LeSage, J. P. and Pace, R. K.: Introduction to Spatial Econometrics. CRC Press, Boca Raton (2009)
12. Sà, C., Florax, R. J. G. M., and Rietveld, P., Determinants of the regional demand for higher education in The Netherlands: a gravity model approach, Regional Studies, 38, 375–392 (2004)
13. Simini, F., Gonzàlez, M. C., Maritan, A., and Barabàsi, A.-L., A universal model for mobility and migration patterns, Nature, 484, 96–100 (2012)
14. Tosi, F., Impacciatore, R. and Rettaroli, R. Individual skills and student mobility in Italy: a regional perspective. Regional Studies, 53, 1099–1111 (2018)

Environmental Statistics (GRASPA-SIS)

A Time Clustering Model for Spatio-Temporal Data

Un modello di clustering temporale per dati spazio-temporali

Clara Grazian and Gianluca Mastrantonio and Enrico Bibbona

Abstract Clustering methods are ubiquitous in many areas of science: ecology, environmental sciences, microbiology, linguistics, transport models, etc. Model-based methods, like mixture models, allow to quantify the uncertainty on the clustering, however extensions to data showing specific types of dependence, as spatio-temporal data, present some hidden difficulties. In this work, we propose an approach to identify clusters in spatio-temporal data and study their evolution over time; the method is invariant with respect to the ordering of the clusters and to the choice of the reference cluster.

Abstract *Metodi di clustering possono essere trovati in ogni area della science: ecologia, science climatiche, microbiologia, linguistica, trasporti, etc. Metodi basati su modelli, come i modelli mistura, permettono di quantificare l'incertezza sulla struttura di clustering, ma l'introduzione di strutture di dipendenza particolari, come quelle caratteristiche dei dati spazio-temporali, presentano alcune difficoltà nascoste. In questo lavoro, proponiamo un approccio per identificare cluster in dati spazio-temporali e studiare la loro evoluazione nel tempo; il metodo è invariante rispetto alla scelta dell'ordine dei gruppi e del grouppo di riferimento.*

Key words: logistic-normal process, hidden Markov models, coregionalization, invariance

Clara Grazian
School of Mathematics and Statistics, University of New South Wales, Sydney, Australia, e-mail: c.grazian@unsw.edu.au

Gianluca Mastrantonio
Dipartimento di Scienze Matematiche, Politecnico di Torino, Torino, Italy, e-mail: gianluca.mastrantonio@polito.it

Enrico Bibbona
Dipartimento di Scienze Matematiche, Politecnico di Torino, Torino, Italy, e-mail: enrico.bibbona@polito.it

1 Introduction

Clustering analysis is one of the main tasks in statistics and applied statistics; some examples can be find in the task of characterising the number and the features of animal behaviours from tracking trajectories, identifying the genetic characteristics of regional flora which increase the probability of survival in presence of pollution, or defining levels of resistance to antibiotics of bacteria. In this respect, mixture models represents a flexible tool to define a probability distribution for the problem of clustering. The natural temporal extension to mixture models is defined by the class of hidden Markov models (HMMs) [7].

Although HMMs are very popular and applied in various settings, the Markov structure can be too restrictive in some situations. One typical example is the dependence of the allocation to each group to external variables or processes. A possible solution would be to introduce covariate information that models the probabilities, however this type of information is not always available and the strucutre of this dependence can be too complex to be easily defined in a model. Another problem of HMMs is that the probabilities to belong to each cluster are time-independent: this is a strong assumption that is rarely satisfied. Finally, the full set of clusters should be present in all the temporal window of observation, while it could be reasonable to assume that some clusters can appear and disappear in specific moments, in particular if the time framework is large. The main reason why the HMM is so widely used lies on its efficiency and ease of implementation.

We propose a model where the clustering probabilities are time-dependent and depend on the particular locations of the data. These is achieved by assuming that the vector of probabilities is marginally distributed as a logistic-normal model (*LogitN*) [1] and the structured temporal dependence is induces by a Gaussian process obtained through the coregionalization method [2]. The logitN distribution was proposed by [1] as a distribution for compositional data, i.e. vectors of positive probabilities with the sum-to-one constraint. With respect to the Dirichlet distribution, which is usually introduced to model compositional data, the logitN distribution has the advantage that relates the vector of probabilities to a latent Gaussian random variable. The structure of the simplex, where the probabilities are defined and which is imposed by the sum-to-one constraint, imposes that in a vector of K elements, only $(K-1)$ are free to vary, while the K-th is deterministically determined; this element is usually defined the *reference element*.

Since there is no reason to choose any group as the reference one, inference should not depend on this choice; however, the introduction of a coregionalization in the Gaussian process involved in the Gaussian latent variable can change the marginal distribution of the elements of the probability vector; a solution to this problem is usually reached by introducing simplifying assumptions.[5] assumes independence between the $K-1$ process and spatio-temporal correlation functions, while [3] and [6] assume dependence between the vectors but only a common functional form for the spatial correlation is used for all regimes.

In this work, we propose a method which is invariant with respect to the choice of the reference element and to the choice of the ordering of the groups.

2 A LogitN representation of the temporal clustering

Let $\{t_1, t_2, \ldots, t_T\} \equiv \mathscr{T}$ be the set of observed time points and let $\mathbf{x}(\mathbf{s}) = \{\mathbf{x}_t(\mathbf{s})\}_{t\in\mathscr{T}}$ be the data observed at locations $\mathbf{s} \in \mathbb{R}^2$. Let $\mathbf{z}(\mathbf{s}) = \{z_t(\mathbf{s})\}_{t\in\mathscr{T}}$ with $z_t(\mathbf{s}) \in \{1,2,\ldots,K\} \equiv \mathscr{K}$, and $f(\cdot)$ be a generic density, with parameter of location ξ_k and scale Ω_k. We assume that the data come from a mixture-type model based on (possibly multivariate) normal densities:

$$f(\mathbf{x}(\mathbf{s})|\mathbf{z}\{\xi_k, \Omega_k\}) = \prod_{t\in\mathscr{T}} f(\mathbf{x}_t(\mathbf{s})|\xi_{z_t(\mathbf{s})}, \Omega_{z_t(\mathbf{s})}), \tag{1}$$

$$\mathbf{x}_t(\mathbf{s})|\xi_{z_t(\mathbf{s})}, \Omega_{z_t(\mathbf{s})} \sim N_2(\xi_{z_t(\mathbf{s})}, \Omega_{z_t(\mathbf{s})}). \tag{2}$$

It is necessary to define the distribution of the allocation variable $\{z_t(\mathbf{s})\}_{t\in\mathscr{T}}$ in such a way that it can describe the dynamic evolution of the clustering probabilities. The most popular model used in this setting is the hidden Markov model, which assume that

$$z_{t_i}(\mathbf{s})|z_{t_{i-1}}(\mathbf{s}), \{\pi_{k,k'}\}_{k,k'\in\mathscr{K}} \sim \sum_{k\in\mathscr{K}} \pi_{z_{t_{i-1}},k}(\mathbf{s})\delta_k, \tag{3}$$

with $z_{t_0} = 1$ and where δ_k is the Kronecker delta function, i.e. the evolution is governed by the Markov property. The mixing probability $\pi_{t,k}(\mathbf{s})$ is the probability that the location $\mathbf{s}$ belongs to component k at time t and are defined in such a way that $\pi_{t,k}(\mathbf{s}) > 0$ and $\sum_{k=1}^{K} \pi_{t,k}(\mathbf{s}) = 1$ for each $\mathbf{s}$ and t.

The cluster probabilities $\{\pi_{t,k}(\mathbf{s})\}_t$ are allowed to vary across space and time. In particular, we want to introduce dependence among them such that observations close in space or time are more likely to be allocated to the same cluster than observations which are more distant from each other. In order to work on the dependence structure, the standard approach involves to work on a transformation of the probabilities

$$\pi_{t,k}(\mathbf{s}) = \frac{e^{\omega_{t,k}(\mathbf{s})}}{1+\sum_{j=1}^{K-1} e^{\omega_{t,j}(\mathbf{s})}}, \quad k \neq K. \tag{4}$$

From (4), it is evident that there is an identification problem, since the vector of probabilities is invariant by adding a constant c to all the $\omega_{t,k}(\mathbf{s}) \in \mathbb{R}$, then an identification constraint is needed; it is usual to set one of the $\omega_{t,k}(\mathbf{s})$ to zero, e.g. $\omega_{t,K}(\mathbf{s}) = 0$ for all $t \in \mathscr{T}$ and $\mathbf{s} \in \mathbb{R}^2$.

The $(K-1)$-dimensional vector $\omega_t(\mathbf{s}) = (\omega_{t,1}(\mathbf{s}), \ldots, \omega_{t,K-1}(\mathbf{s}))'$ is assumed to be a realization of a time structured Gaussian process

$$\omega_t(\mathbf{s}) = (\mathbf{I}_{K-1} \otimes \mathbf{X}_t(\mathbf{s}))\,\beta + \mathbf{A}\eta_t(\mathbf{s}) \tag{5}$$

where $\mathbf{X}_t(\mathbf{s})$ is a vector of p (possible) time-dependent covariates, β are $(K-1)p$ regressors, $\mathbf{A}$ is an $(K-1)\times(K-1)$ matrix and $\eta_t(\mathbf{s}) = (\eta_{t,1}(\mathbf{s}), \ldots, \eta_{t,(K-1)}(\mathbf{s}))'$, where $\eta_{\cdot,k} = \{\eta_{t,k}\}_{t\in\mathscr{T}}$ is a realization of a Gaussian processes with zero mean and isotropic and stationary correlation functions $C_k(|t-t'|;\psi_k)$. Moreover, $\eta_t(\mathbf{s})$ as a function of space are independent-in-time spatially correlated errors distributed

accordingly to a Gaussian process with zero mean and with spatial covariance function, given by a isotropic correlation function $C_k(|\mathbf{s}-\mathbf{s}'|;\theta_k)$, where θ_k describes the decay rate of the correlation as a function of the distance between locations. The dependence structure of the Gaussian process induces a dependence on the compositional vectors. Since marginally ω_t is normally distributed with mean $(\mathbf{I}_{K-1}\otimes\mathbf{X}_t(\mathbf{s}))\,\beta$ and covariance matrix $\Sigma=\mathbf{AA}$, then $\pi_t(\mathbf{s})=(\pi_{t,1}(\mathbf{s}),\dots\pi_{t,K-1}(\mathbf{s}))'$ is *LogitN* distributed:

$$\pi_t(\mathbf{s}) \sim LogitN((\mathbf{I}_{K-1}\otimes\mathbf{X}_t(\mathbf{s}))\,\beta,\Sigma).$$

The term $(\mathbf{I}_{K-1}\otimes\mathbf{X}_t(\mathbf{s}))\,\beta$ rules the mean of the processes, while $\mathbf{A}$ introduce dependence between the $\omega_{\cdot,k}(\mathbf{s})=\{\omega_{t,k}(\mathbf{s})\}_{t\in\mathscr{T}}$, i.e. if $\mathbf{A}$ is diagonal the $\omega_{\cdot,k}(\mathbf{s})$ are independent; finally, $\eta_t(\mathbf{s})$ models the structured temporal dependence. The covariance between $\omega_t(\mathbf{s})$ and $\omega_{t'}(\mathbf{s})$ is given by

$$\Sigma_{t,t'}=\mathbf{AC}_{\eta,|t-t'|}\mathbf{A}'$$

where $\mathbf{C}_{|t-t'|}$ is a $(K-1)\times(K-1)$ diagonal matrices with i-th diagonal elements equal to $C_i(|t-t'||;\psi_i)$.

The sum-to-a-constant constranint of compositional vectors introduces some difficulties in the modelling of the clustering probabilities, since

$$\mathrm{Cov}(\pi_{t,k}(\mathbf{s}),\pi_{t,1}(\mathbf{s})+\cdots+\pi_{t,k}(\mathbf{s})+\cdots+\pi_{t,K}(\mathbf{s}))=0$$

and, therefore,

$$-\mathrm{Var}(\pi_{t,k}(\mathbf{s}))=\sum_{\substack{j=1\\k\neq j}}^{K}\mathrm{Cov}(\pi_{t,k}(\mathbf{s}),\pi_{t,j}(\mathbf{s})).$$

The left hand side is negative except for the trivial case when $\pi_{t,k}(\mathbf{s})$ is a constant.

[1] and following works pointed out that a more consistent measure of dependence between compositional elements can be measure as

$$\tau_{ij,kl}(t,t')=\mathrm{Cov}\left(\log\frac{\pi_{t,i}(\mathbf{s})}{\pi_{t,k}(\mathbf{s})},\log\frac{\pi_{t',j}(\mathbf{s})}{\pi_{t',l}(\mathbf{s})}\right),\ i,j,k,l\in 1,\dots,K,$$

that are the covariance between all possible combinations of log-ratios; it has to be noticed that the elements of compositional vectors provide information about the relative value of the components (i.e. ratios) instead of their absolute values. Moreover there is a one-to-one correspondence between the composition data and the log-ratios then the covariance may be expressed in terms of log-ratios without modifying the statistical problem or losing information.

The measure of dependence $\tau_{ij,kl}(t,t')$ is related to the covariances of the Gaussian variables ω_t through the fact that

$$\log\frac{\pi_{ti}(\mathbf{s})}{\pi_{tk}(\mathbf{s})}=\log\frac{\exp(\omega_{ti}(\mathbf{s}))}{\exp(\omega_{tk}(\mathbf{s}))}=\omega_{ti}(\mathbf{s})-\omega_{tk}(\mathbf{s})\qquad\forall i,k\in\mathscr{K},$$

and then

$$\begin{aligned}\tau_{ij,kl}(t,t') = \mathrm{Cov}(\omega_{ti}(\mathbf{s}),\omega_{t'j}(\mathbf{s})) + \mathrm{Cov}(\omega_{tk}(\mathbf{s}),\omega_{t'\ell})(\mathbf{s}) \\ - \mathrm{Cov}(\omega_{ti}(\mathbf{s}),\omega_{t'\ell})(\mathbf{s}) - \mathrm{Cov}(\omega_{tk}(\mathbf{s}),\omega_{t'j}(\mathbf{s}))\end{aligned} \tag{6}$$

Equation (6) highlights an important interpretation problem: the structure of $\tau_{ij,kl}(t,t')$ changes if they involves the last components of the compositiona vector.

It is easy to see [1] that a *LogitN* process has independent components in term of log-ratios at time lag $|t-t'|$, i.e. $\tau_{ij,kl}(t,t') = 0$ for arbitrary i, j, k and l, only if the covariance matrix between $\omega_t(\mathbf{s})$ and $\omega_{t'}(\mathbf{s})$ can be written as

$$\Sigma_{t,t'} = \begin{pmatrix} a_1 + a_K & a_K & \dots & a_K \\ a_K & a_2 + a_K & \dots & a_K \\ \dots & \dots & \dots & \dots \\ a_K & a_K & \dots & a_{K-1} + a_K \end{pmatrix}.$$

It is then easy to see that independent ω's, i.e. diagonals $\Sigma_{t,t'}$, does not imply independence between the elements of π. In general, the dependence structured assumed for $\omega_{t,k}(\mathbf{s})$ is not automaticcaly transferred to the vector of probabilities $\pi_{t,k}(\mathbf{s})$.

2.1 A new parametrization

A new parametrization which allows an easier interpretation of the parameters is now proposed.

It is possible to model the Gaussian variable defined without imposing the identifiability constraint:

$$\gamma_t(\mathbf{s}) = (\mathbf{I}_K \otimes \mathbf{X}_t(\mathbf{s}))(\beta_1,\beta_2) + \mathbf{A}^*\eta_t(\mathbf{s}) \tag{7}$$

where $\gamma_t(\mathbf{s})$ is a K dimensional process, aand $\mathbf{A}^*$ is a $K \times K$ matrix, β_1 is a $p(K-1)$-dimensional vector, β_2 is a p dimensional vector and $\eta_{.k}(\mathbf{s}) \sim GP(0, C_k(|t-t'|;\psi_k))$, where $C_k(|t-t'|;\psi_k)$ is a covariance function that dependes on parameters vector ψ_k. At time t, the covariance matrix of $\gamma_t(\mathbf{s})$ is $\Sigma^* = \mathbf{A}^*(\mathbf{A}^*)'$. Moreover, $\eta_{.k}(\mathbf{s}) \sim GP(0, C_k(|\mathbf{s}-\mathbf{s}'|;\theta_k))$, where $C_k(|\mathbf{s}-\mathbf{s}'|;\theta_k)$ is a covariance function that dependes on parameters vector θ_k.

Similarly to (5), the compositional vector is defined by using the $\gamma_t(\mathbf{s})$

$$\pi^*_{t,k}(\mathbf{s}) = \frac{e^{\gamma_{t,k}(\mathbf{s})}}{\sum_{j=1}^K e^{\gamma_{t,j}(\mathbf{s})}} = \frac{e^{\gamma_{t,k}(\mathbf{s})-\gamma_{t,K}(\mathbf{s})}}{\sum_{j=1}^K e^{\gamma_{t,j}(\mathbf{s})-\gamma_{t,K}(\mathbf{s})}}, \quad k = 1,\dots,K.$$

In order to deal with the identifiability problem, we subtract the value $\gamma_{t,K}(\mathbf{s})$ to all the components of $\gamma_t(\mathbf{s})$, instead of setting to zero one of the them:

$$\pi^*_{t,k}(\mathbf{s}) = \frac{e^{\gamma_{t,k}(\mathbf{s})-\gamma_{t,K}(\mathbf{s})}}{\sum_{j=1}^{K} e^{\gamma_{t,j}(\mathbf{s})-\gamma_{t,K}(\mathbf{s})}}, \quad k = 1,\dots,K.$$

We have then created a link between the two parametrizations, and the parameters of (5) can be derived by the ones of (7):

$$\mathbf{A} = [\mathbf{A}^*]_{1:(K-1),1:K} - [\mathbf{A}^*]_{K,1:K},$$

where with notation $[\cdot]_{\cdot,\cdot}$ we select the sub-matrix with given indexes; simlarly,

$$\beta = \beta_1 - \mathbf{1}_{K-1} \otimes \beta_2,$$

and, finally,

$$[\Sigma^*]_{1:(K-1),1:(K-1)} + \mathbf{1}_{K-1}[\Sigma^*]_{K,K}\mathbf{1}'_{K-1} - [\Sigma^*]_{1:(K-1),K}\mathbf{1}'_{K-1} - \mathbf{1}_{K-1}[\Sigma^*]'_{1:(K-1),K},$$

where $\mathbf{1}_c$ is a c-dimensional vector of elements equal to one. Given this link, it follows that $\pi_{t,k}(\mathbf{s}) = \pi^*_{t,k}(\mathbf{s})$ for any $k = 1,\cdots,K$ and $t = 1,\cdots,T$.

The process $\gamma_t(\mathbf{s})$ provides more interpretable results: the set of log-ratio variances are easy to compute since

$$\tau_{ij,kl}(t,t') = \mathrm{Cov}\left(\gamma_{t,i} - \gamma_{t,k}, \gamma_{t',j} - \gamma_{t',l}\right);$$

now all the $\tau_{ij,kl}(t,t')$ have the same structure, solving the inconsistencies shown in equation (6).

3 Conclusion

We have introduced a new model to describe and predict spatio-temporal trajectories characterised by a clusters. This approach introduces a latent representation which allows to maintain the assumption on the dependence structure among the clustering probabilities. It represents a generalisation of the approach proposed in [4] in order to introduce the possibility of spatial dependence in the clusters.

Areas of application are hugely variable: environmental sciences, animal movements, epidemiology, economics, genomics, among many others.

References

1. Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK, UK: Chapman & Hall, Ltd.
2. Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman and Hall.
3. Martins, A. B. T., Bonat, W. H., and Jr., P. J. R. (2016). Likelihood analysis for a class of spatial geostatistical compositional models. *Spatial Statistics*, 17: 121 – 130.

4. Mastrantonio, G., Grazian, C., Mancinelli, S. and Bibbona, E. New formulation of the logistic-Gaussian process to analyze trajectory tracking data. *The Annals of Applied Statistics*, 13(4): 2483–2508.
5. Paci, L. and Finazzi, F. (2017). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, 1–16.
6. Tjelmeland, H. and Lund, K. V. (2003). Bayesian modelling of spatial compositional data. *Journal of Applied Statistics*, 30(1): 87–100.
7. Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Reconstruction of sparsely sampled functional time series using frequency domain functional principal components

Ricostruzione di serie temporali funzionali scarsamente campionate utilizzando componenti principali funzionali nel dominio delle frequenze

Amira Elayouty, Marian Scott and Claire Miller

Abstract In this paper, we present a useful application of the recently developed time-varying Functional Principal Components (FPCs) to the analysis of non-stationary Functional Time Series (FTS) that commonly occurs in high-frequency environmental data. The time-varying FPCs (Elayouty, 2017) adapt to the changes in the auto-covariance structure and vary smoothly over frequency and time by taking into account (i) the temporal dependence between the functional observations and (ii) the changes in the covariance/variability structure over time. This makes them an attractive option for a number of analyses including the assessment of stationarity, the clustering of functional curves as well as the reconstruction of sparsely sampled functional data in a dynamic FTS context. The performance of the time-varying FPCs in reconstructing sparsely sampled FTS has been assessed here by a simulation study.

Abstract *In questo paper, presentiamo un'applicazione delle Componenti Principali Funzionali (CPF) tempo-varianti, di recente sviluppo, all'analisi delle serie temporali funzionali non stazionarie che tipicamente si osservano su dati ambientali raccolti ad alta frequenza. Le CPF tempo-varianti si adattano a cambiamenti nella struttura dell'auto-covarianza e variano in modo smooth nel tempo e nello spettro delle frequenze prendendo in considerazione (i) la dipendenza temporale tra le osservazioni funzionali e (ii) le variazioni nel tempo nella struttura di varianza-covarianza. Questo rende le CPF un'opzione interessante per una serie di analisi*

Amira Elayouty
School of Mathematics and Statistics, University of Glasgow, UK, e-mail: amira.elayouty@glasgow.ac.uk
Faculty of Economics and Political Science, Cairo University, Egypt, e-mail: a.ayouti@feps.edu.eg

Marian Scott
School of Mathematics and Statistics, University of Glasgow, UK, e-mail: Marian.Scott@glasgow.ac.uk

Claire Miller
School of Mathematics and Statistics, University of Glasgow, UK, e-mail: Claire.Miller@glasgow.ac.uk

che includono la valutazione di stazionariet, il raggruppamento di curve funzionali e la ricostruzione di dati funzionali scarsamente campionati in un contesto di serie temporali funzionali. Le prestazioni delle CPF tempo-varianti nel ricostruire serie temporali funzionali scarsamente campionate sono state valutate tramite uno studio di simulazione.

Key words: Functional Time Series, Frequency Domain, Smoothing, Sparsely-sampled, Principal Components; Non-stationarity, Functional Spectral Density

1 Introduction

Current sensor technology enables environmental monitoring programs to record measurements at high-temporal resolutions over long time periods, for processes which are in reality continuous in time. Examples include water quality measurements from automatic monitoring buoys/sensors recorded every 2 minutes over time or temperature recordings from thermistor chains at different depths in a lake over time. Unfortunately, statistical modelling and analysis as well as feature extraction from these environmental High-Frequency Data (HFD) are challenging due to the persistent and dynamic dependence structure over the different timescales (Elayouty et al. 2016) that can even become more problematic with the presence of missing measurements. Functional Time Series (FTS) analysis and its recent developments (Hormann and Kokoszka, 2012; Hormann et al. 2015; Elayouty, 2017) provides an appropriate framework for the analysis and feature extraction from these HFD, taking into consideration these challenges.

Frequency domain Functional Principal Components (FPCs) that vary smoothly over time, taking into account both the temporal correlation and the non-stationarity in the series, have been recently developed by Elayouty (2017). These time-varying FPCs proved useful to the exploration and analysis of dynamic FTS that commonly manifest in environmental data. Testing for second-order stationarity and clustering of individual curves in a FTS are two typical applications of these timevarying FPCs (Elayouty, 2017). This paper presents an application of these FPCs to the reconstruction of sparsely sampled dynamic FTS, assessed through their performance in a simulation study.

2 Frequency domain time-varying FPCs

In FTS, data are viewed as realizations of a functional stochastic process $\{X_k(t) : k \in \mathbb{Z}, t \in \mathscr{T}\}$, with k denoting the discrete time parameter e.g. day and t being the continuous time parameter defined on $\mathscr{T}$ e.g. intra-day. The time-varying FPCs evaluate the Spectral Density (SD) of the FTS process at each time point k. An eigen-decomposition of the SD is then performed at each time point k to extract

the frequency domain eigenfunctions, under the assumption that the process varies smoothly over time. The time-varying FPCs accommodate the varying autocorrelation structure in a FTS through the decomposition of the SD which, unlike the lag-0 covariance, contains information on the whole family of lag-h covariances (Hormann et al., 2015).

As only a limited number of replicates are available at each time point k, the local lag-h covariances and spectral densities are estimated by smoothing the sample lag-h covariances over time using a weight kernel $w_s(.)$ with smoothing parameter s,

$$\hat{V}_{k,h} = \frac{1}{\sum_{k' \in \mathbb{Z}} w_s(|k-k'|)} \sum_{k \in \mathbb{Z}} w_s(|k-k'|) X_{k'} \otimes X_{k'+h}. \tag{1}$$

where $w_s(.)$ is a monotonically decreasing weight function of the distance $|k-k'|$ regardless of the lag h, ensuring that the highest weights are assigned to the pairs $(X_k, X_{k'})$ near the target point k. The neighbourhood contributing to the covariance estimation is determined by the choice of the kernel and smoothing parameter.

After estimating $\hat{V}_{k,h}$, the local SD is estimated at each time point k by:

$$\hat{F}_{k,\theta} = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \hat{V}_{k,h} \exp(-\mathrm{i}h\theta), \quad \theta \in [-\pi, \pi], \tag{2}$$

and the local eigenvalues $\hat{\lambda}_{k,m}(\theta)$ and eigenvectors $\hat{\varphi}_{k,m}(\theta)$ of $\hat{F}_{k,\theta}$ are computed. The local frequency domain FPCs $\{\hat{\phi}_{kml}(t) : l \in \mathbb{Z}\}$ are then obtained via the inverse Fourier transform of $\hat{\varphi}_{k,m}(\theta)$.

3 Reconstruction of sparsely sampled FTS

Although sensor data often come in large volumes they may involve periods of missing data as a result of temporary breakdown in equipment or in transmission of data. Imputing or reconstructing these missing periods with a better accuracy is of interest to develop a complete picture of the phenomenon. Obtaining the frequency domain functional PCs at each time point k requires a full data set. For this reason a naive simple imputation method, using the average for instance, can be used to provide initial values for the missing periods. After that the time-varying FPCs $\{\hat{\phi}_{kml}(t) : l \in \mathbb{Z}\}$ are obtained and used subsequently to filter the original FTS across a number of lags and leads l to obtain the m^{th} local dynamic FPC scores at k by: $\hat{Y}_{m,k}^{(k)} = \sum_{l=-L}^{L} \int_{t \in \mathscr{T}} X_{k-l}(t) \hat{\phi}_{kml}(t) dt$. All (complete or sparsely sampled) curves can thus be approximately reconstructed based on these scores, using q-term (smooth) time-varying FPCs, $q < \infty$, as follows:

$$\hat{X}_k(t) \approx \sum_{m=1}^{q} \sum_{l=-L}^{L} \hat{Y}_{m,k+l}^{(k)} \hat{\phi}_{kml}(t), \quad \forall k. \tag{3}$$

This procedure can then be re-iterated to achieve better results.

4 Results and Discussion

An extensive simulation study was conducted to assess the performance of the proposed time-varying FPCs versus the stationary dynamic FPCs (Hormann et al, 2015) in reconstructing sparsely sampled FTS, under a variety of non-stationary data-generating processes. For each of the scenarios 200 FTS were simulated and for each FTS a random collection of curves constituting $\nu \in (10,35)\%$ of the total number of curves in the series are made incomplete. For each randomly chosen incomplete curve, a block of $\tau \in (35,60,100)\%$ of the discrete data, representing the realisations of the curve at a set of finite time points, are made missing. Naive imputations of the missing values are used as initial values before applying the FPCA.

The results of the simulation study indicated that, in basically all missing data patterns, the missing data were better recovered using the time-varying FPCs (see Fig 1). The performance of FPCs has been measured in terms of the normalised mean squared errors (NMSE) between the initially simulated complete curves and the reconstructed ones using the stationary dynamic FPCs and the time-varying FPCs with different values of the smoothing parameter allowing for different levels of smoothness. As the smoothing parameter decreases the NMSE decreases, reflecting the necessity of using local FPCs in case of non-stationary FTS. It can also be noted that the variation in NMSE between the 200 replications is systematically smaller for the time-varying FPCs.

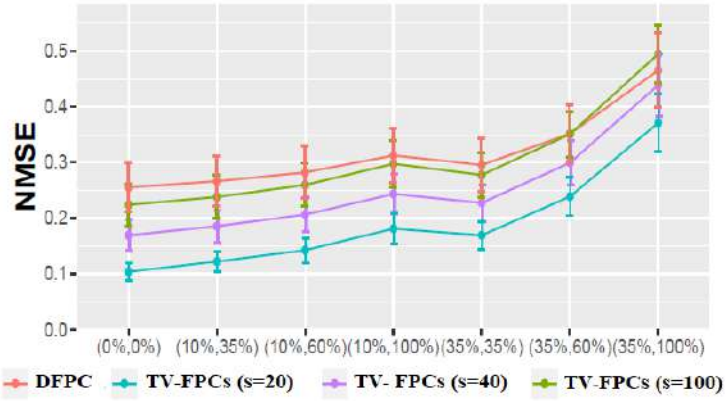


Fig. 1 Average NMSE between the originally simulated curves and their recovered versions using the first dynamic (red) and the first smooth dynamic FPCs with a smoothing parameter $s = 20$ (turquoise), 40 (purple), 100 (olive) based on the simulation results under the different levels of sparsity specified by $(\nu,\tau)\%$. The vertical bars represent one standard deviation from the mean NMSE.

5 Conclusion

Time-varying FPCA proved useful for reconstructing sparsely sampled curves in dynamic FTS with an increased accuracy compared to the dynamic FPCA by Hormann et al. (2015). This accuracy can even be improved through iteratively repeating the same procedure.

Acknowledgements A.E. is grateful to the University of Glasgow sensor studentship for funding part of the work.

References

1. Elayouty, A.: Time and frequency domain statistical methods for high-frequency time series. Ph.D. Thesis: University of Glasgow (2017)
2. Elayouty, A., Scott, M., Miller, C., Waldron, S. and Franco-Villoria, M.: Challenges in modeling detailed and complex environmental data sets: a case study modeling the excess partial pressure of fluvial CO_2. Journal of Environmental and Ecological Statistics. **2**3(1), 65–87 (2016)
3. Hormann, S., Kidzinski, L. and Hallin, M.: Dynamic functional principal components. Journal of the Royal Statistical Society - Series B. **7**7(2), 319–348 (2015)
4. Hormann, S. and Kokoszka, P.: Functional time series. In: Handbook of Statistics: Time series analysis - Methods and Applications. Elsevier: Amsterdam, pp. 15-186 (2012)

Methods for High Dimensional Compositional Data Analysis

Algorithms for compositional tensors of third-order

Algoritmi per tensori composizionali di terzo ordine

Violetta Simonacci

Abstract The PARAFAC-ALS procedure for estimating CP parameters on tridimensional tensors is sensitive to data collinearity. This inefficiency is especially problematic if collinearity is paired with other issues such as data of large dimensions and difficulties in establishing correct model rank. When dealing with compositional data, i.e. positive values with a covariance bias, multicollinearity is inherent by definition, and it is preserved also if the data is transformed in log-ratios by means of the *clr* function. For this reason, alternative estimating procedures may be considered, such as INT and INT-2. These dual-step methods use the properties of the SWATLD and ATLD algorithms during initialization to overcome ALS inefficiency while still providing least squares results. Their comparative performance is tested in an extensive simulation study on collinear data.

Abstract *La procedura PARAFAC-ALS per la stima dei parametri CP su tensori tridimensionali è sensibile alla collinearità dei dati. Questa inefficienza è particolarmente problematica se la collinearità è accoppiata ad altre difficoltà come dati di grandi dimensioni e difficoltà nello stabilire il corretto rango del modello. Quando si trattano dati composizionali, valori positivi con covarianza vincolata, la multicollinearità è data per definizione e viene preservata anche se i dati sono trasformati in logaritmi dei rapporti attraverso la trasformazione clr. Per questo motivo, è possibile prendere in considerazione procedure di stima alternative, come INT e INT-2. Questi metodi a doppia fase utilizzano rispettivamente le proprietà delle procedure SWATLD e ATLD durante l'inizializzazione per sopperire alle inefficienze del PARAFAC-ALS pur fornendo risultati in termini di minimi quadrati. Le loro prestazioni comparative sono testate in un ampio studio di simulazione su dati collineari.*

Key words: centered logratios, CP model, collinearity, PARAFAC-ALS, three-way data

---

Violetta Simonacci
Università degli Studi di Napoli - "L'Orientale", DISUS, Largo S. Giovanni Maggiore, 30, Napoli,
e-mail: vsimonacci@unior.it

1 Introduction

Third order tensors of compositional data are positive data charachterized by spurious correlations and organized along three indexes (individuals, variables, occasions). These tensors, especially when large, can be quite difficult to model. Compositional data are generally treated as log-ratio coordinates in order to get to rid of their biased structure [2]. The preferred log-ratio transformation for both two and three-way exploratory analysis [1, 9, 6] is the centered log-ratio transformation (*clr*) which preserves full multicollinearity of the data, i.e. one redundant dimension in the variable space.

Multicollinearity does not present a major issue when bilinear SVD-based methods are applied, however it can become problematic for the estimation of the CANDECOMP/PARAFAC (CP) model [3, 7]. This decomposition is an adaptation of bilinear SVD to higher orders which guarantees a unique solution under mild condition by imposing the restriction that the same latent structure is found throughout samples. For this reason, it represents the preferred modeling tool for data with a true trilinear structure. In this sense it can be differentiated from the Tucker3 procedure [12], an unrestricted high order version of bilinear models which does not provide unique outputs, thus, can prove difficult to interpret.

CP uniqueness comes at the price of estimating issues, which become more likely in case of large data, such as slow convergence and degeneracy. In particular both permanent or temporary degenerate solutions may occur. This latter case is detected when an iterative estimating algorithm encounters bottlenecks or swamps: it starts converging towards a degenerate solution, slows down excessively but eventually emerges from the swamp and finds the real optimal point. Degenerate solutions are generally recognizable by the fact that two factors appear highly correlated but with opposite signs. Abnormal convergence can be caused by bad-initialization, wrong rank selection (the dimensionality of the model does not match the real underlying one), data and factor collinearity.

Many algorithms have been proposed in the literature to fit the CP model, all with different performances with respect to these difficulties. In general, unless specific data conditions require it, the preferred option is PARAFAC-ALS (ALS), the procedure originally proposed in [7]. This is because of a key advantage: its Loss of Fit (LoF) function monotonically decreases, thus convergence is stable and results are found in the least squares sense.

Nonetheless, this algorithm is quite slow at converging on large data sets and it is sensitive to bad initial values, wrong rank and data collinearity. This last weakness, pointed out in [8], is particularly relevant for compositions because it makes the algorithm underperform on *clr*-coordinates especially for large tensors.

In particular it was observed in [5] that multicollinearity appears to slow down the algorithm more than usual, and when it is paired with over-specification it may also make it quite difficult for the procedure to find the correct results. As a solution it was proposed to use an alternative estimation process, an integrated procedure called INT, which adds a Self-Weighted Alternating Tri-Linear Decomposition (SWATLD) initialization step to ALS in order to overcome its deficiencies. This

is because SWATLD, proposed by [4], is somewhat complementary to ALS in the sense that it is robust in presence of wrong model specifications, collinearity and it is fast at converging. However, it is less stable and precise, so it constitute only an ideal initialization step.

INT was introduced for compositions but it was thoroughly tested only on non-collinear data [11].

Starting from INT another procedure was also implemented later on in [10], called INT-2, where the initialization step is no longer carried out by SWATLD but by means of the Alternating TriLinear Decomposition (ATLD) proposed by [13]. This second version was not developed for dealing with collinearity but for modeling large data-sets, as ATLD is a more efficient version of SWATLD albeit less precise. It was never tested for compositions, however it has good potential to work well because it is fast and robust to collinearity.

In this perspective the purpose of this work is to find the best algorithmic alternative for large tensors of *clr*-transformed data between INT, INT-2 and ALS by evaluating which one is preferable in terms of accuracy and efficiency. Multicollinearity, especially when associated with large dimensions and wrong factorization can severely impair estimation of the CP model and recognizing the best way to address this issue can be of great interest, especially with the increasing dimensionality of data sets in all fields of research.

2 Methods

2.1 Third-order tensors of compositions

A third order tensor $\mathscr{T} \in \mathbb{R}^{n\times m\times p}$ with generic element t_{ijk} is a data structure arranged along three dimensions or modes identified as first-mode with index $i = 1,\ldots,n$, second-mode with index $j = 1,\ldots,m$ and third mode with index $k = 1,\ldots,p$. Usually it is shaped like a cube containing the measurement of n individuals over m variables at p occasions along the vertical, horizontal and depth dimensions respectively.

If one of the three indices is fixed while the other ones are free to vary, $\mathscr{T}$ can be partitioned in three different sets of second-order tensors, referred to as slabs: frontal slabs $\mathbf{T}_{(k)}^{(n\times m)}$ with $k = 1,\ldots,p$, vertical slabs $\mathbf{T}_{(j)}^{(n\times p)}$ with $j = 1,\ldots,m$ and horizontal slabs $\mathbf{T}_{(i)}^{(p\times m)}$ with $i = 1,\ldots,n$. Slabs can be juxtaposed so that $\mathscr{T}$ is unfolded along only two dimensions and treated as a matrix. Specifically we have three unfolded tensors which are usually considered: $\mathbf{T}_{A}^{(n\times mp)}$, $\mathbf{T}_{B}^{(m\times np)}$ and $\mathbf{T}_{C}^{(p\times mn)}$.

If two of the indexes are fixed, the third order tensor can also be subdivided in rank-one tensors, known as fibers. Specifically there are $n\times p$ horizontal fibers or rows $\mathbf{t}_{ik}(1\times m)$, $m\times p$ vertical fibers or columns $\mathbf{t}_{jk}(1\times n)$ and $n\times m$ "depth" fibers or tubes $\mathbf{t}_{ij}(1\times p)$.

A third order tensor presents a compositional structure if its generic row $\mathbf{t}_{ik} = [t_{i1k},\ldots,t_{ijk},\ldots,t_{imk}]$ consists of positive values characterized by a covariance bias: $cov(t_{i1k},t_{i2k}) + cov(t_{i1k},t_{i3k}) + \ldots + cov(t_{i1k},t_{imk}) = -var(t_{i1k})$. The m elements of a compositional row are denoted as parts because they describe the portions of a whole and do not vary independently from each other, i.e. $\sum_{j=1}^{m} t_{ijk} = \kappa$ where κ is a constant representing a generic total.

The variability of such row-vectors can be fully explained only in relative terms, by referring to ratios amongst them, thus the value of κ is irrelevant and can be scaled to one. When compositions are analyzed in absolute scores frontal slices are perfectly collinear and their covariance matrix is singular.

The compositional problem can also be explained form a geometric stand point: the biased structure translates into one dimension of the row-vector being redundant, thus the sample space of the generic row composition $\mathbf{t}_{ik}$ is not $\mathbb{R}_{+}^{m}$ but rather the unit-simplex $\mathbb{S}^{m-1}$. This vector space is characterized by its own geometric rules called Aitchison geometry. In order to project compositions onto real space and apply standard multilinear tools, they can be transformed in the log-ratios amongst vector parts.

In exploratory analysis the most used transformation is the centered log-ratio (*clr*). The *clr* function provides a correspondence between $\mathbb{S}^{m-1}$ and $\mathbb{R}^m$ and considers the logarithms of the ratios between each part and the geometric mean of the full composition. Each frontal slab $\mathbf{T}_{(k)}^{(n\times m)}$ is transformed in $\mathbf{Z}_{(k)}^{(n\times m)}$ with generic element $z_{ijk} = t_{ijk}/g(\mathbf{t}_{ik})$, with $g(\mathbf{t}_{ik}) = \prod_{j=1}^{m} t_{ijk}$.

This transformation is symmetric and isometric, however, it maintains a redundant dimension, thus the condition of perfect multicollinearity of frontal slabs is not eliminated.

2.2 Compositional CP algorithms

A third-order tensor of *clr*-coordinates $\mathscr{Z} \in \mathbb{R}^{n\times m\times p}$ can be decomposed by means of the CP model, which is based on the concept of polyadic decomposition.

A tensor is in polyadic form when it is represented as the linear combination of $1,\ldots,f,\ldots$F rank-one tensors. For a tridimensional tensor of *clr*-coordinates we have:

$$\mathscr{Z}^{n,m,p} = \sum_{f=1}^{F} \mathscr{D}_f = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \quad \text{with} \sum_{j=1}^{m} b_{jf} = 0 \quad \forall \quad 1,\ldots,f,\ldots\text{F} \tag{1}$$

The generic rank-one tensor $\mathscr{D}_f$ is the result of a triad, namely the outer product of the corresponding generic vectors $\mathbf{a}_f \in \mathbb{R}^n$, $\mathbf{b}_f \in \mathbb{R}^m$ and $\mathbf{c}_f \in \mathbb{R}^p$, which indicate the factors of the first, second and third mode respectively. Each factor in every mode is related to only one factor in the other modes, i.e. it belongs to only one triad. These factors can be arranged as columns of loading matrices $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_f,\ldots,\mathbf{a}_F] \in \mathbb{R}^{n\times F}$, $\mathbf{B} = [\mathbf{b}_1,\ldots,\mathbf{b}_f,\ldots,\mathbf{b}_F] \in \mathbb{R}^{m\times F}$ and $\mathbf{C} = [\mathbf{c}_1,\ldots,\mathbf{c}_f,\ldots,\mathbf{c}_F] \in \mathbb{R}^{p\times F}$.

If the total number of factors F corresponds to the tensorial rank of $\mathscr{Z}$, that it is the minimal number of real underlying constructs which fully describe the information in the array, this decomposition is known as rank decomposition.

The CP model aims to model a tensor with noise contamination $\mathscr{Z} = \hat{\mathscr{Z}} + \mathscr{E}$, where $\mathscr{E}$ is the tensor of errors, by providing its rank decomposition without modeling excessive noise. The CP model is generally written using a frontal slab notation in the following manner:

$$\mathbf{Z}_{(k)} = \hat{\mathbf{Z}}_{(k)} + \mathbf{E}_{(k)} = \mathbf{A}\mathbf{D}^{(k)}\mathbf{B}^t + \mathbf{E}_{(k)} \quad k = 1, \ldots, p \tag{2}$$

Here the diagonal matrices $\mathbf{D}^{(i)}$, $\mathbf{D}^{(j)}$ and $\mathbf{D}^{(k)}$ contain the ith, jth and kth rows of the first-, second- and third-mode factor matrices respectively while $\mathbf{E}_{(i)}$, $\mathbf{E}_{(j)}$ and $\mathbf{E}_{(k)}$ are the horizontal, lateral and frontal slices of $\mathscr{E}$.

One of the problems connected with this procedure, however, is that the rank of a tensor is hard to establish in advance, thus, when estimating parameters, F is often chosen so that $F > R$ and a rank decomposition is not reached. This case is called over-factoring and results in the CP model being over-specified, which may create problems in the estimation process, leading to degenerate solutions.

The first and most used algorithm for fitting the CP model is ALS. which carries out least-squares optimization steps based on the following loss function:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} = \sum_{k=1}^{p} \|\mathbf{Z}_{(k)} - \hat{\mathbf{Z}}_{(k)}\|^2 = \sum_{k=1}^{p} \|\mathbf{Z}_{(k)} - \mathbf{A}\mathbf{D}^{(j)}\mathbf{B}^t\|^2 \tag{3}$$

The symbol $\|\cdot\|$ is used to represent the Frobenius norm.

Since this algorithm is sensitive to data collinearity, intrinsic to *clr*-coordinates, and it is quite slow, thus hardly preferable for large data-sets, other procedures are considered for comparison: i) INT developed to address collinearity issues but not fully tested in this framework and ii) INT-2 developed for efficiency but could also perform well on compositions.

INT is articulated in two optimization steps: an initialization phase in which SWATLD is used to overcome collinearity and over-factoring issues, and a second step in which ALS is used to refine the solution to obtain least-squares results.

SWATLD properties which make it robust to these difficulties are directly determined by the fact that instead of one, it uses three loss function and that the focus is on extracting the trilinear information.

For a full presentation of INT and of its estimation steps refer to [11].

In the other procedure considered, INT-2 [10], ATLD is used rather than SWATLD to initialize estimation, for the rest it is analogous to INT. ATLD has similar properties to SWATLD when it comes to overcoming deficiencies and it is also based on three loss functions. ATLD is however more efficient than SWATLD but way less stable in its final results.

3 Experimental design and results

The aim of this study is to understand which algorithm is best suitable for large tensors of compositional data. For this purpose the performance of the status quo algorithm, standard ALS, is compared with the two alternative procedures, INT and INT-2, developed for multicollinear and for large tensors respectively. A Monte Carlo design is thus implemented in which the algorithms are tested on large tensors of artificial data sets with three dimensionality: $n = m = p = 50$, $n = m = p = 100$, $n = m = p = 150$ and varied degrees of factor congruence and noise contamination. Performance is tested in terms of efficiency (CPU time, number of iterations employed, number of bottlenecks encountered) and of accuracy (congruence with real solution and MSE). Both the rank decomposition and the over-specification case are considered

Preliminary results show that ALS clearly undeperformes with respect to both INT and INT-2. In particular this latter procedure appears to be the best performing option for efficiency and accuracy in over-factoring.

References

1. Aitchison J., Greenacre M.: Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 51(4):375–92 (2002)
2. Aitchison J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):139–160 (1982)
3. Carroll J.D.,Chang J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35(3):283–319 (1970)
4. Chen Z.P., Wu H.L., Jiang J.H., Li Y., Yu R.Q.: A novel trilinear decomposition algorithm for second-order linear calibration. Chemometrics and Intelligent Laboratory Systems 52(1):75–86 (2000)
5. Gallo M., Simonacci V., Di Palma M.A.: An integrated algorithm for three-way compositional data. Quality & Quantity 53(5):2353–2370 (2019)
6. Graffelman J., Pawlowsky-Glahn V., Egozcue J.J., Buccianti, A.: Compositional Canonical Correlation Analysis. bioRxiv: 144584, Cold Spring Harbor Laboratory (2017)
7. Harshman R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. UCLA Working Papers in Phonetics, 16:1–84 (1970)
8. Kiers H.A.: A three–step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. Journal of Chemometrics: A Journal of the Chemometrics Society 12(3):155–171 (1998)
9. Pawlowsky-Glahn V., Egozcue J.J., Tolosana-Delgado R.: Modeling and analysis of compositional data. John Wiley & Sons (2015)
10. Simonacci V., Gallo M.: An ATLD–ALS method for the trilinear decomposition of large third-order tensors. Soft Computing: 1–12 (2019)
11. Simonacci, V. and Gallo, M.: Improving PARAFAC-ALS estimates with a double optimization procedure. Chemometrics and Intelligent Laboratory Systems, 192:103822 (2019)
12. Tucker L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3):279–311 (1966)
13. Wu H.L., Shibukawa M., Oguma K.: An alternating trilinear decomposition algorithm with application to calibration of hplc–dad for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. Journal of Chemometrics 12(1):1–26 (1998)

High-dimensional regression with compositional covariates: a robust perspective

Metodi di regressione per high-dimensional data con covariate composizionali: una prospettiva robusta

Gianna Serafina Monti and Peter Filzmoser

Abstract In this contribution we propose a robust approach to high-dimensional regression with compositional covariates. The proposed procedure is based on a class of shrinkage estimators for least trimmed squares regression in combination with elastic-net penalty. We illustrate the effectiveness of our methods by an application to a human microbiome data set with the aim to robustly predict the body mass index as a function of the gut microbiome composition.

Abstract *In questo contributo presentiamo un approccio robusto ai metodi di regressione penalizzata per la selezione automatica delle variabili nel caso in cui la matrice del disegno abbia una natura composizionale. La procedura proposta é basata su una regressione di tipo LTS (least trimmed squares) in combinazione con una penalizzazione di tipo elastic-net. Al fine di illustrare l'efficacia del metodo proposto, viene presentata un'applicazione all'analisi del microbiota intestinale umano, strettamente connesso alle abitudini alimentari e agli stili di vita, al fine di stimare in modo robusto l'indice di massa corporea .*

Key words: Lasso; elastic net models; Log-contrast model; Model selection; Regularization; Sparsity

1 Introduction

Compositional data carry relative information, and often they are expressed in proportions of percentages (Filzmoser et al., 2018). If there is a sum constraint, standard

Gianna Serafina Monti
Department of Economics, Management and Statistics, University of Milano Bicocca, e-mail: gianna.monti@unimib.it

Peter Filzmoser
Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology
e-mail: P.Filzmoser@tuwien.ac.at

linear regression with compositional covariates cannot be applied; but even without the constraint, standard regression models on the original compositions are not meaningful.

In the seminal work of Aitchson and Bacon-Shone (1984), a regression model for compositional data was proposed, known as log-contrast model. In this model the response variable is a linear combination of log-ratios of the original compositions. Let y be a response vector of dimension $n \times 1$ and $\mathbf{X}_{n\times p}$ a design matrix of compositional covariates, w.l.o.g. expressed with constant sum 1, thus each row lies in the unit simplex $\mathcal{S}^p = \{x_{ij} : x_{ij} > 0 \text{ and } \sum_{j=1}^{p} x_{ij} = 1\}$. The linear log-contrast model has the following structure

$$y = \mathbf{Z}^p \beta_{\setminus p} + \varepsilon, \tag{1}$$

where $\mathbf{Z}^p = \{\log(x_{ij}/x_{ip})\}$, $j = 1, \ldots, p-1$, is the additive log ratio transform matrix of dimension $n \times (p-1)$, with the pth component as reference, and $\beta_{\setminus p} = (\beta_1, \ldots, \beta_{p-1})$ is the regression coefficient vector, and ε is the error component, usually assumed normally distributed around zero, with constant variance σ^2. The model (1) can be expressed in a symmetric form (Lin et al., 2014) introducing a constraint on the coefficient vector:

$$y = \mathbf{Z}\beta + \varepsilon, \quad \sum_{j=1}^{p} \beta_j = 0, \tag{2}$$

where $\mathbf{Z} = \{\log(x_{ij})\} \in \mathbb{R}^{n\times p}$, and $\beta = (\beta_1, \ldots, \beta_p)$.

Considering a high-dimensional setting, Lin et al. (2014) proposed a variable selection procedure and estimation for model (2):

$$\hat{\beta} = \arg\min_{\beta} \left(\frac{1}{2n} \|y - \mathbf{Z}\beta\|_2^2 + \lambda \|\beta\|_1 \right), \quad \text{subject to } \sum_{j=1}^{p} \beta_j = 0, \tag{3}$$

where $\lambda > 0$, is the regularization parameter and $\| \|_2$ and $\| \|_1$ indicate the ℓ_2 and ℓ_1 norm, respectively.

Altenbuchinger et al. (2017) combined the variable selection problem and estimation for model (2) with the elastic-net regularization (Friedman et al., 2010):

$$\hat{\beta}_{zeroSum} = \arg\min_{\beta} \left(\frac{1}{2n} \|y - \mathbf{Z}\beta\|_2^2 + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right), \quad \text{subject to } \sum_{j=1}^{p} \beta_j = 0, \tag{4}$$

where α is a tuning parameter for the compromise between the ℓ_2 and ℓ_1 penalty. Model (4) is known as *zeroSum elastic-net regression*, to emphasize the constraint of the regression coefficients in conjunction with the elastic-net regularization.

2 Robust and sparse regression models for compositional data

With the aim to explore sparse relationships between a response variable and a set of covariates in a high-dimensional setting, Alfons et al. (2013) proposed a robust estimator by adding an ℓ_1 penalty on the coefficient estimates to the least trimmed squares (LTS) estimator (Huber and Ronchetti, 2009), leading to the sparse LTS estimator:

$$\hat{\beta}_{sparseLTS} = \arg\min_{\beta} \left\{ \sum_{i=1}^{h} r^2_{(i)}(\beta) + h\lambda \|\beta\|_1 \right\}, \tag{5}$$

where $r^2_{(i)}$ are the order statistics of the squared residuals, and $\lfloor (n+p+1)/2 \rfloor \leq h \leq n$, and $\lfloor . \rfloor$ means rounding down to the nearest integer. Sparse LTS regression is equivalent to detecting the subset of h observations whose least squares fit produces the smallest (penalized) sum of squared residuals. The authors demonstrated the robustness of the estimator (5) to multiple regression outliers, as well as to leverage points.

Kurnaz et al. (2018) extended the work of Alfons et al. (2013), they combined the sparsity of the elastic-net (EN) procedure with the robustness of LTS regression. They proposed the trimmed (EN)LTS estimator, based on the idea of repeatedly applying the non-robust classical estimators to data subsets only. The (EN)LTS estimator is defined by:

$$\hat{\beta}_{(\mathrm{EN})LTS} = \arg\min_{\beta} \arg\min_{H \subseteq \{1,2,\ldots,n\}: |H|=h} \left(\frac{1}{2h} \|y - \mathbf{X}\beta\|_2^2 + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right), \tag{6}$$

A constrained penalized LTS regression estimator for (EN)-type penalized regression is proposed here as follows,

$$\hat{\beta}_{RobzeroSum} = \arg\min_{\beta} \arg\min_{H \subseteq \{1,2,\ldots,n\}: |H|=h} \left(\frac{1}{2h} \|y - \mathbf{Z}\beta\|_2^2 + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right)$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j = 0\,, \tag{7}$$

which conveys the zero-sum constraint typical of compositional covariates. We refer to model (7) as *RobzeroSum elastic-net regression*, to combine the constraint of the regression coefficients with the elastic-net regularization in a robust way. In the following we will only consider the special case with $\alpha = 1$. An extensive simulation study, not reported here for space constraints, demonstrates its robustness in presence of outliers in the data.

2.1 Parameter selection

We suggest to perform a (10 fold) cross-validation procedure, over an appropriate grid of values for λ, to determine the optimal value of the elastic-net regularization in (7). Thus the regression coefficient estimates are calculated in correspondence of λ which minimizes the CV-error.

3 An application to human gut microbiome

The study of human microbiome, i.e. the collection of all microorganisms found in the human body, including bacteria, viruses, and fungi, is crucial in the understanding of the roles that symbionts play on human health, and their impacts on a number of diseases. We applied the proposed model to a cross-sectional study of the association between diet and gut microbiome composition (Wu et al., 2011). In this study, fecal samples from 98 healthy individuals were collected and the DNA samples were analyzed by the 454/Roche pyrosequencing of 16S rRNA gene segments of the V1-V2 region. The resulting microbiome database, produced by high-throughput sequencing of 16S rRNA, is a matrix of counts of clustered sequences, known as operational taxonomic units (OTUs), that depict bacteria types. Only the relative abundances have a meaning, as the number of sequencing reads varied a lot across samples (Weiss et al., 2017), thus microbiome data have essentially a compositional nature. Due to the high proportion of zero counts in the OTU table, we considered, among the 78 taxa at genus level, only those with an abundance $\geq 0.2\%$ in at least one sample and those which appeared in more than 10% of the samples, resulting in a matrix of dimension $n \times p = 98 \times 37$. Zero counts were replaced by the minimum non-negative observed count. We fitted RobzeroSum elastic-net regression to predict BMI response as a function of a subset of the most important taxa.

To measure the prediction accuracy of the RobzeroSum model and the zero-Sum model, we calculated the trimmed-squared prediction error $||y - \mathbf{Z}\hat{\beta}(\hat{\lambda}_{CV})||_2^2/h$ (trimming level equal to 0.1). The RobzeroSum model achieved a value of 6.91 and performs clearly better compared to the zeroSum model with a value of 11.37. This can also be seen from the two scatter plots of the measured versus predicted BMI values reported in Figure 1.

To determine the statistical strength of the selected taxa, we performed a classical nonparametric bootstrap of the units (Hastie et al., 2015). Figure 2 (left) reports the estimated regression coefficients, $\hat{\beta}_j(\hat{\lambda}_{CV})$, for the selected taxa. Thick lines and thin lines represent $\hat{\beta}_j(\hat{\lambda}_{CV}) \pm SE$, and $\hat{\beta}_j(\hat{\lambda}_{CV}) \pm 2SE$ respectively, where the SE is estimated through the trimmed standard deviation of the 1000 bootstrap realizations. Figure 2 (right) reports the proportion of times each coefficient is zero in the 1000 bootstrap realizations. We found that Clostridium is significantly negatively correlated with BMI, as obtained by Lin et al. (2014). Furthermore the major part of the selected taxa belongs to the Firmicutes phylum, confirming the strictly dependence of the BMI with this phylum.

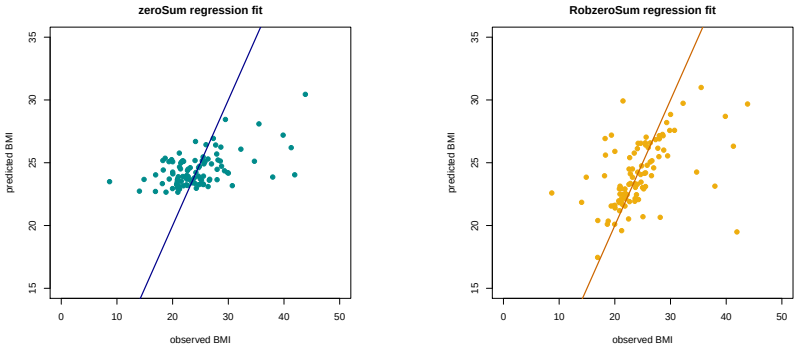


Fig. 1 Left: Scatter plot of the fitted values for the BMI response variable according to the zeroSum regression model estimates. Right: Scatter plot of the fitted values for the BMI response variable according to the RobzeroSum regression model estimates.

References

Aitchson, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330.

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.*, 7(1):226–248.

Altenbuchinger, M., Rehberg, T., Zacharias, H. U., Stämmler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P. J., and Spang, R. (2017). Reference point insensitive molecular data analysis. *Bioinformatics*, 33 2:219–226.

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied Compositional Data Analysis. With Worked Examples in R*. Springer Series in Statistics, Springer, Cham, Switzerland.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. Wiley, New York, 2nd edition edition.

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172:211 – 222.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, Jesse R.and Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and

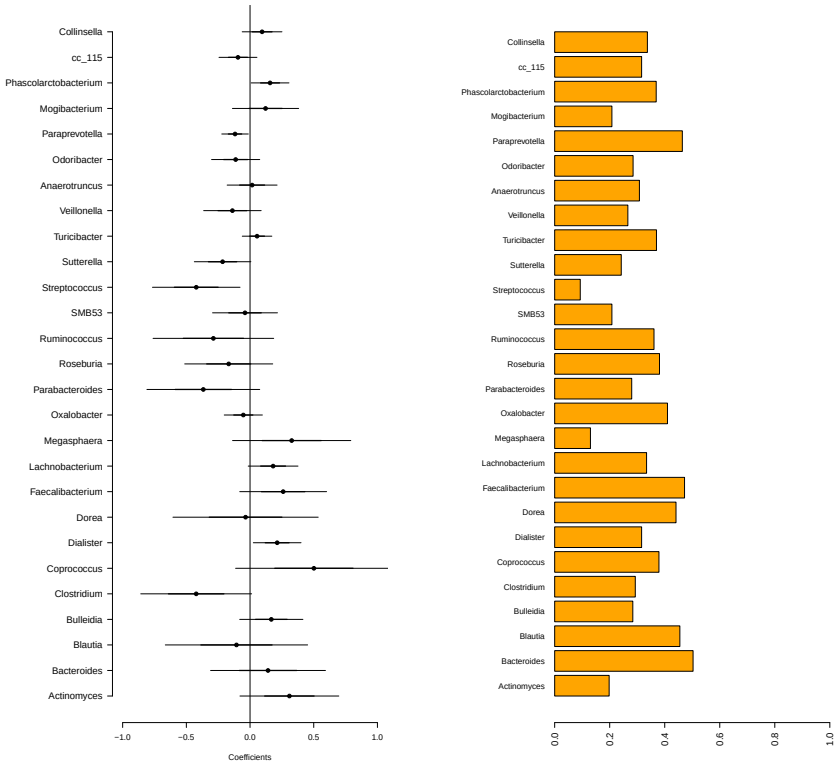


Fig. 2 Left: The dots, thick lines, and thin lines represent coefficient estimates via the RobzeroSum penalized regression model, $\hat{\beta}_j(\hat{\lambda}_{CV}) \pm SE$, and $\hat{\beta}_j(\hat{\lambda}_{CV}) \pm 2SE$, obtained by the nonparametric bootstrap. Right: Proportion of times each coefficient is zero in the bootstrap realizations. Only the coefficients related to the selected taxa from fitted RobzeroSum model are considered in both figures.

Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(27 (1)).

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.

Three-way compositional analysis of energy intensity in manufacturing

Analisi composizionale a tre vie dell'intensità energetica nel settore manifatturiero

Valentin Todorov and Violetta Simonacci

Abstract Both the scientific and political communities agree that significant reductions in CO2 emissions are necessary to limit the magnitude and extent of climate change and of course the energy efficiency is one of the most interesting issues analyzed by economists and policy makers within this debate. Different measures of energy efficiency in manufacturing can be defined but broadly this is the ratio of the production output to the energy input, usually disaggregated by industry. We create a global data set of energy intensity in manufacturing and analyze its structure by country, time and industry applying parallel factor analysis (CP). Since we are interested in the structure of the energy intensity, the absolute values are no more relevant for the analysis and the nature of this data set is compositional which requires specific adaptation of the methodology and suitable software.

Abstract *La comunità scientifica e il mondo politico concordano sul fatto che la riduzione dell'emissione di CO2 è necessaria per limitare il cambiamento climatico, quindi un efficiente uso dell'energia è tra gli aspetti più trattati dagli economisti e dai politici nell'affrontare tale tema. Nella produzione manifatturiera sono definite diverse misure di efficienza energetica, ma in generale il rapporto tra consumo energetico e output, disaggregato per settore, è tra i più usati. Considerando i paesi, gli anni e i settori merceologici, un dataset sull'intensità energetica nel settore manifatturiero è stato costruito e studiato con l'analisi dei fattori paralleli (CP). Tuttavia, l'analisi dell'intensità energetica in valore assoluto non è l'aspetto più rile-*

Valentin Todorov
United Nations Industrial Development Organization (UNIDO), VIC, Vienna, e-mail: v.todorov@unido.org

Violetta Simonacci
University of Naples-L'Orientale, Naples, 80134, Italy e-mail: vsimonacci@unior.it

vante da studiare, inoltre, la natura di tali dati è composizionale e quindi richiede un'adeguata metodologia e software.

Key words: energy intensity, manufacturing value added, PARAFAC, compositional data

1 Introduction

Energy efficiency is one of the key emissions reduction policy tools and is becoming a top priority in energy policies. If energy is used more efficiently, this can contribute to steadier and potentially higher economic growth since the amount of energy required per unit of output can be reduced. This reduction will subsequently lead to a reduced product price. Energy intensity is measured by the amount of energy used to produce one unit of economic output. It is the inverse of energy efficiency: less energy intensity means more energy efficiency. To calculate the energy intensity the amount of energy used (in physical terms, kilotons of oil equivalent, or *ktoe*) is divided by a measure of the produced output in monetary terms. Energy intensity in manufacturing is measured by dividing the energy consumed by the manufacturing value added. Many studies of energy efficiency are related to the analysis in one or few countries and do not consider the issue globally. Kepplinger et al [7] analyzes the energy efficiency in manufacturing industry using mixed-effect models, but is limited to the aggregated level. Cantore [2] studies the energy efficiency in developing countries for the manufacturing sector using the Fisher Ideal Index but covers only around 20 countries.
Data on manufacturing value added at annual frequency are available from the UNIDO INDSTAT database and data on energy consumption can be obtained from the International Energy Agency (IEA). Using the data from these data sources in the present paper we will analyze the structure of energy intensity of the manufacturing industry across countries, time and industries. The outline of the paper is as follows. The next Section 2 briefly introduces the concept of compositional data and its relevance for parallel factor analysis (CP). Section 3 describes the data and presents the results of the analysis. The last section concludes and outlines topics for further research.

2 CP and compositional data

Compositional data as defined by Aitchison [1] are strictly positive multivariate observations that carry only relative information, i.e. the only relevant information is contained in ratios between parts of a composition. Compositional data with a given constant sum constraint are represented in the simplex sample space, S^D, which consists of D-part compositions

$$S^D = \{x = (x_1, \ldots, x_D)^\top; x_j > 0, j = 1, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\} \tag{1}$$

where κ is an arbitrary constant (κ equals 100 for the case of percentages and 1 for proportions). On the simplex the compositional data follow a specific geometry, called Aitchison geometry. This means that standard statistical methods relying on the Euclidean geometry in real space cannot be applied to capture the multivariate structure. Therefore, to enable the application of multivariate statistical methods on compositional data is necessary first to transform the data to the usual Euclidean geometry. Several transformation for this purpose were introduced and it depends on the analysis which of them to use. A simple and popular transformation is the centered log-ratio transformation defined as follows:

$$clr(x) = \left\{ ln\frac{x_1}{g(x)}, \ldots, ln\frac{x_D}{g(x)} \right\}^\top \tag{2}$$

where $g(x)$ is the geometric mean of all parts in the composition. The most important property of the *clr* transformation is that it preserves the distances among data, i.e. it presents an isometry between the Aitchison geometry and the Euclidean geometry.

The CP model [see 3, 5] decomposes the 3-way data array $\underline{X}$ into three loading matrices A $(I \times F)$, B $(J \times F)$, C $(K \times F)$ with F components (using the same number of components for each mode). The CP model can be written formally as

$$X_A = AI_A(C \otimes B)^\top + E_A, \tag{3}$$

where X_A, I_A $(F \times FF)$ and E_A $(I \times JK)$ are the original array, the superdiagonal three-way identity array and the error array, all matricized with respect to the mode A. The symbol $\otimes$ represents the Kronecker product between two matrices. To estimate the optimal component matrices the residual sum of squares

$$||E_A||^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2 = \sum_{i=1}^{I}||x_i - \hat{x}_i||^2 = \sum_{i=1}^{I} RD_i^2$$

is minimized. Three-way models are usually fitted by an iterative procedure based on *PARAFAC-ALS (ALS)*. In the case of compositional data the rotational invariance of the ALS algorithm is important since it allows the application of any logratio transformation with the isometry property [see 4, where CP was considered for the first time in compositional context].

The computations in this paper were carried out using the R package **rrcov3way** [8]. It provides functions for classical and robust three-way modeling using Tucker3 and CP. These functions can be applied on both compositional and non-compositional data selecting the appropriate transformation.

3 Energy intensity in manufacturing

The data on energy consumption is obtained from the world energy balances database compiled by IEA [6]. It comprises data for more than 180 countries, 68 products and 98 flows in the period 1960-2017 in ktoe. From the products we select only the total and from the flows—the 11 flows pertaining to manufacturing (see Table 3). Manufacturing value added at annual frequency, disaggregated by industry, is available from the UNIDO INDSTAT database (`http://stat.unido.org`) for a large number of countries for the period 1963 to 2017 (at different level of detail, but we will use the 22 ISIC Revision 3 divisions). The data can be downloaded in current USD. Combining these two data sets in order to calculate the energy intensity by country poses a number of challenges. The first issue that we observe is that many countries, at least for a given period, do not disaggregate the data properly by flow but put everything into the category INONSPEC, therefore we exclude from the analysis data where the value in INONSPEC is 90% or more of the total or there are less than 5 flows with nonzero value. The discrepancy between the 11 IEA flows and the 23 ISIC Revision 3 divisions is obvious from Table 3—it was necessary to combine several ISIC divisions into one flow but also several flows together (BASICMET=NONFERR+IRONSTL). When considering only the relative nature of data, we do not need to convert the current US dollars into constant US dollars. Both data sets are compositional (sum up to a total energy consumption in manufacturing or total manufacturing value added, respectively). We can take their ratios for every flow (to obtain the energy intensity by flow), and further continue with the statistical processing in *clr* coordinates. After linking the two data sets together and removing countries where either energy consumption or value added is missing as well as removing obvious inconsistencies, we remain with 40 countries, 10 aggregated flows for the period 2000 to 2015 (16 years). This results in a three-dimensional array with dimensions $40 \times 10 \times 16$. We could start the analysis by applying principle component analysis (PCA) to the data year by year, and then present the results in a sequence of compositional biplots, but will be skipping this step for the sake of saving space. Also, if we want to get the complete picture about the development in a larger time frame, we should apply a CP model. It is

Flow	ISIC R3	ISIC Description
FOODPRO	15, 16	Manufacture of food, beverages and tobacco
TEXTILES	17, 18, 19	Manufacture of textiles, wearing apparel; dressing and dyeing of fur Tanning and dressing of leather
WOODPRO	20	Manufacture of wood and of products of wood and cork
PAPERPRO	21, 22	Manufacture of paper and paper products; Publishing
CHEMICAL	24	Manufacture of chemicals and chemical products
NONMET	26	Manufacture of other non-metallic mineral products
BASICMET	27	Manufacture of basic metals
MACHINE	28, 29	Manufacture of fabricated metal products; Machinery and eq.
	30, 31, 32	Manufacture of office eq.; Electrical machinery and apparatus n.e.c.; Radio
TRANSEQ	34, 35	Manufacture of motor vehicles; Other transport equipment
INONSPEC	25, 33, 36	Any manufacturing industry not included above

Table 1 Correspondence of IEA flows and ISIC Revision 3 descriptions

possible to apply CP to the raw data but it would be hard to recognize any structure. Therefore we conduct CP on the *clr* transformed data. The number of dimensions is chosen as F=2 using a convex hull procedure. The paired components plots of the first two modes A and B are shown in Figure 1. Three clusters of industries are seen in the mode B plot: the high tech MACHINE and TRANSPEQ, the energy intensive industries CHEMICAL, METALS and NONMETALS and the rest. Also there are clusters of countries in the mode A plot. The all-component plot is useful in our case since the third mode has a natural (time) ordering. The two components from the CP model of the clr transformed energy intensity data, shown in the left panel of Figure 2 have completely different trend. While component 1 runs smoothly throughout the period, component 2 grows steeply up to 2008, jumps abruptly up in 2009 and then continues smoothly. The per-component plot shown in the right panel of Figure 2 presents the tree modes on the first component. The flows mode (B) shows, similarly as the pair component plot of mode B the clustering of the high tech MACHINE and TRANSEQ flows against the energy intensive METALS, CHEMICAL and NONMET with the rest in the middle. The third mode does not show visible trend throughout the years but in the per-component plot on the second component (not shown here) three periods are clearly identified.

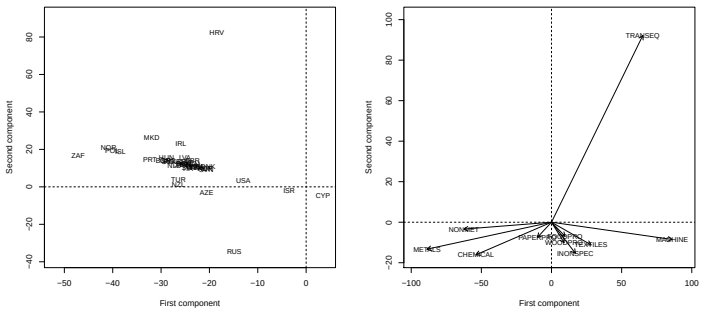


Fig. 1 Paired component plots of the CP model for *clr* transformed energy intensity (mode A and mode B, in the left, resp. right, panel)

4 Summary and conclusions

Only illustrative results of the novel approach to analysis of the structure of the extremely important nowadays energy efficiency issue is presented in this short contribution. It clearly demonstrates that the compositional character of the data must be taken into account in order to obtain meaningful results. The importance of creating a high quality, coherent data set based on internationally comparable data sources is emphasized. Future research should improve the quality of the presented data set.

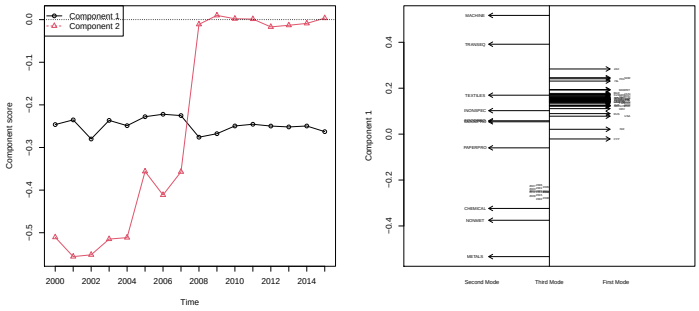


Fig. 2 All components plot and per-component plot of the CP model for *clr* transformed energy intensity (in the left, resp. right, panel)

and link the global energy intensity structure to other variables applying supervised methods. This will go hand in hand with the future development of the software for three-way data analysis implemented in the package **rrcov3way**.

References

[1] Aitchison J (1986) The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability, Chapman & Hall Ltd., London (UK), London (UK)

[2] Cantore N (2011) Energy efficiency in developing countries for the manufacturing sector. UNIDO Staff Working Paper, Vienna

[3] Carroll J, Chang J (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. Psychometrica 35(3):283–319

[4] Gallo M (2013) Log-ratio and parallel factor analysis: An approach to analyze three-way compositional data. In: Prat A (ed) Advanced Dynamic Modeling of Economic and Social Systems. Studies in Computational Intelligence, vol 448, Springer, pp 209–221

[5] Harshman RA (1970) Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. Tech. Rep. 10

[6] IEA (2019) Extended world energy balances `https://doi.org/10.1787/data-00513-en`

[7] Kepplinger D, Templ M, Upadhyaya S (2013) Analysis of energy intensity in manufacturing industry using mixed-effects models. Energy 59:754–763

[8] Todorov V, Simonacci V, Di Palma MA, Gallo M (2020) **rrcov3way**: Robust methods for multiway data analysis, applicable also for compositional data. `http://CRAN.R-project.org/package=rrcov3way`, R package version 0.2

Modern Statistics for Physics Discoveries

Identification of high-energy $\gamma-$ray sources via nonparametric clustering

Individuazione di sorgenti di raggi γ ad elevata energia mediante clustering non parametrico

Giovanna Menardi, Denise Costantin, and Federico Ferraccioli

Abstract High-energy $\gamma-$ray sources exhibit as energy flares in the sky map, produced by variously concentrated photon emissions. The identification of these sources is a fundamental task to better understand the mechanisms that both create and accelerate particles emitted by celestial objects. We discuss the application of nonparametric clustering for $\gamma-$ray source detection and provide an algorithm specific for this task. The procedure accounts for the intrinsic uncertainty associated to the available data, arising as an effect of the instrument-pitch and multiple scattering.

Abstract Le sorgenti di raggi γ ad elevata energia si manifestano come bagliori nella mappa celeste, prodotti da emissioni di fotoni molto concentrati. L'identificazione di tali sorgenti è fondamentale per una migliore comprensione dei meccanismi che determinano la creazione e l'accelerazione delle particelle emesse dai corpi celesti. In questo lavoro si discute l'applicazione dell'approccio non parametrico al *clustering* per l'identificazione di sorgenti di raggi γ ad elevata energia e si sviluppa una procedura specifica per tale obiettivo. La procedura sfrutta l'incertezza insita nei dati a disposizione, dovuta al livello di precisione dello strumento di rilevazione dei fotoni e al fenomeno dello *scattering* multiplo.

Key words: directional data, kernel estimator, $\gamma-$ray sources

Giovanna Menardi
Department of Statistical Sciences, University of Padova e-mail: menardi@stat.unipd.it

Denise Costantin
Center for Astrophysics, Guangzhou University e-mail: denise.costantin@gmail.com

Federico Ferraccioli
Department of Statistical Sciences, University of Padova e-mail: ferraccioli@stat.unipd.it

1 Introduction

The Large Area Telescope (LAT) is a wide field-of-view pair-conversion telescope onboard the Fermi spacecraft. It performs an all-sky survey generally aimed to a better understanding of the mechanisms that both create and accelerate particles emitted by celestial objects. Discovering and locating high-energy $\gamma-$ray sources in the whole sky map is one of main purposes of the survey, and a declared target of the Fermi LAT collaboration. High-energy $\gamma-$ray sources exhibit as flares in the sky map, produced by variously concentrated photon emissions.

The standard procedure of the Fermi LAT collaboration for source detection relies on so-called *single-source* models, where the presence of a possible new source is evaluated after splitting the whole sky map on small regions, on the basis of some significance tests on the pixel-by-pixel photon counts. *Variable-source-number* models address the problem via more comprehensive approaches, where the number of sources in the whole map are jointly estimated. See [2] for further details.

According to the latter perspective, the aim of this work is to develop a model for detecting high energy emitting sources. While the signal of the sources is known to be blended by a diffuse $\gamma-$ray background which spreads over the entire area observed by the telescope, we assume that background processes have been pre-filtered out via some pre-processing, so that we can explicitly focus on the detection of the emitting sources.

Our goal is pursued via the suitable adaptation of *nonparametric*, or *modal* clustering ideas to the considered framework. With respect to most clustering methods, relying on heuristic ideas of similarity between objects, the nonparametric formulation is built on a probabilistic framework, which guarantees a sounder theoretical ground, and allows, for instance, a natural application of inferential tools. Additionally, the number of clusters is determined itself within the estimation process. Modal clustering relies on the assumption that a probability density underlies the data, and clusters are defined as the domains of attraction of the density modes. Two main issues arise in the operational implementation of these ideas: the density function needs to be appropriately estimated, usually via nonparametric methods, and its modes detected. We address both points by taking advantage of the context at hand.

The rest of the paper is organized as follows. After providing an overview about non-parametric clustering (Section 2.1), we propose and discuss a novel method specifically conceived for high-energy $\gamma-$ray sources detection (Section 2.2), and illustrate its application on a set of data simulated from one of the catalogues released by the Fermi LAT Collaboration (Section 3).

2 Nonparametric clustering for high-energy $\gamma-$ray sources detection

2.1 Overview on nonparametric clustering

Nonparametric, or *modal* clustering hinges on the assumption that the data $(x_1,\dots,x_n)'$ are sampled from a probability density function f. The modes of f represent the archetypes of the clusters, which are in turn described by the surrounding regions. The identification of the modal regions may be performed by associating each cluster to the set of points along the steepest ascent path towards a mode, in turn located by some optimization method. Alternatively, clusters can be identified as the disconnected density level sets of the sample space, without attempting the explicit task of mode detection. This is also the route followed in this work. Specifically, any section of f, at a level λ, singles out the (upper) level set

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}, \quad 0 \leq \lambda \leq \max f$$

which may be connected or disconnected. In the latter case, it consists of a number of connected components, each of them associated with a cluster at the level λ.

While there may not exist a single λ which catches all the modal regions, any connected component of $L(\lambda)$ includes at least one mode of the density and, on the other hand, for each mode there exists some λ for which one of the connected components of the associated $L(\lambda)$ includes this mode at most. Hence, all the modal regions may be detected as the connected components of $L(\lambda)$ by varying λ.

A side consequence of varying λ along its range, is the opportunity of building a hierarchical structure of the high-density sets, known as the *cluster tree*. For each λ, it provides the number of connected components of $L(\lambda)$, and each of its leaves corresponds to a cluster, i.e. the largest connected component of $L(\lambda)$ including one mode only. Figure 1 illustrates a simple example of this idea: clusters associated with the highest modes 2 and 3 are identified by the smallest λ larger than λ_3, while the smallest λ larger than λ_1 identifies the cluster associated to mode 1.

Operationally, a few choices are required to implement the ideas underlying modal clustering. Since f is unknown, a nonparametric estimator $\hat{f}$ is employed which, disregarding the specific choice adopted, is in general governed by some parameter defining the amount of smoothing. It stands to reason the need of selecting this parameter appropriately, as it affects the resulting modal structure of $\hat{f}$. A second choice derives from the lack, in multidimensional sample spaces, of an obvious method to identify the connected components of a level set. The issue is usually addressed by building a suitable graph on the observed data and, conditional to their belonging to a level set, the connected components of the graph are identified. A key matter becomes to suitably define the graph. See [3] and reference therein for a detailed review on modal clustering.

In the following, we take advantage of information available on the photon emissions to address the choices of the smoothing amount and the level set graph.

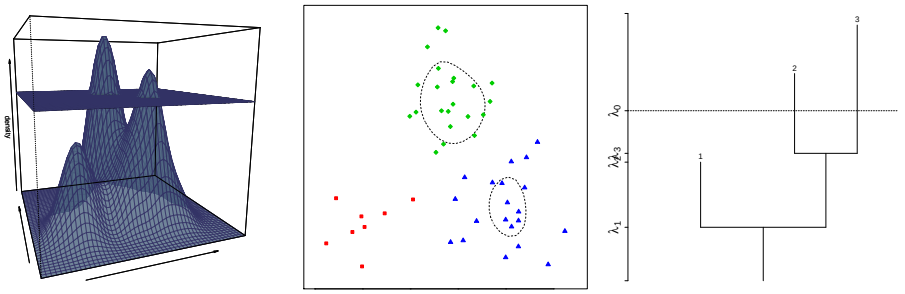


Fig. 1 A section of a density function at a level λ_0 (left), the identified level set (middle panel), formed by two disconnected regions and the associated cluster tree, with leaves corresponding to the modes. The horizontal line is at the level λ_0 (right).

2.2 A novel method for high-energy γ−ray sources detection

Each photon emission i ($i = 1, \dots, n$) is identified by the LAT detector, and the event reconstructed through its trackers and calorimeters. A reconstructed event is finally described by a pair of Galactic coordinates (θ_i, Φ_i), which represent the direction of the emission, along with, among others, the energy of the recorded photon, its incidence angle, and the time of detection. In fact, event reconstruction is subject to several sources of uncertainty due, for instance, to the instrument-pitch and multiple scattering, added in quadrature. Such uncertainty is worth to be accounted for in the analysis as it affects the perceived concentration of the photon emission.

One measure of uncertainty is the 95% *containment angle* (CA95), which broadly speaking depends on the angle of incidence on the LAT surface, the energy, and the event type classification of each individual photon. Information provided by the CA95 is here exploited to perform nonparametric clustering in a twofold way.

The containment angle of each photon $CA95_i$ is roughly defined as the 0.95-quantile of a generalized bivariate Student distribution $t_2(0, \sigma_i^2 \mathbb{I}_2)$, which approximates, at least for high energies, the *point spread function* of the emitting source, measuring the response of the LAT to a point source [1, 4]. An approximated value of σ_i can be then derived as $\sigma_i \simeq CA95_i / t_{2,0.95}$, being $t_{2,0.95}$ the theoretical 0.95-quantile of a $t_2(0, \mathbb{I}_2)$. Then σ_i is used as the bandwidth for an adaptive product kernel density estimator

$$\hat{f}(\theta, \Phi) = \sum_{i=1}^{n} \frac{1}{n\sigma_i^2} K\left(\frac{\theta - \theta_i}{\sigma_i}\right) K\left(\frac{\Phi - \Phi_i}{\sigma_i}\right). \tag{1}$$

Once that the density underlying the data has been estimated via (1), its level sets are determined for varying λ, and its connected components are to be identified. In fact, given $\hat{L}(\lambda)$, finding the connected components of a set is both conceptually and computationally easy in the unidimensional case only, where connected sets are intervals. Conversely, in multidimensional spaces it is immediate to state whether

any given point belongs to $\hat{L}(\lambda)$, while saying how many connected sets comprise $\hat{L}(\lambda)$, and identifying them, is not obvious. We address this issue by shifting the target from the continuous multidimensional space where data are defined, to the finite and discrete set identified by the observations themselves, via the identification of the connected components of a suitable graph built on the data. Specifically, we consider connected two observations i and j when

$$d_a((\theta_i, \Phi_i), (\theta_j, \Phi_j)) < \max(CA95_i, CA95_j),$$

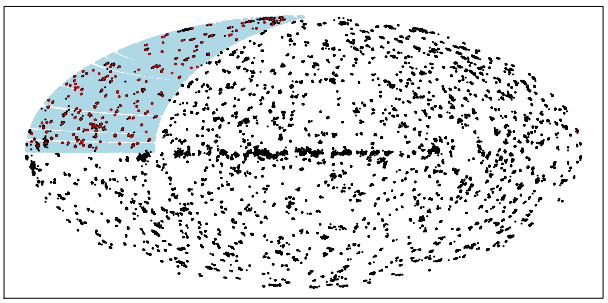
where $d_a(\cdot, \cdot)$ denotes the orthodromic distance function. The underlying rationale is that two photons are likely to be generated by the same source whenever one of them lies within the essential support of the other one. After running nonparametric clustering, detected clusters are associated to emitting sources, whose location is identified by the pertaining density mode.

3 Empirical analysis

This section presents the results of the application of the proposed procedure to a set of data simulated from the 3FHL catalog of the Fermi LAT collaboration and spread on the whole sky map. The sky distribution of the data, illustrated in Figure 2, is quite heteregeneous, with separated sources in the extragalactic sky, overlapping sources in the galactic region and various source sizes. To limit the computational effort, we restrict the analysis to a sub-region of the whole map, as illustrated in Figure 2. Additionally, we remove all the sources with less than four photons emitted. The resulting data set include 6309 photons, emitted by 169 sources, whose size range from 4 to 1015 photons.

As a benchmark, we also apply a single linkage clustering, based on the orthodromic distance between photons. Due to the high concentration of the sources, this methods appears as the most promising, among the standard clustering methods. In this case, there is no obvious method to select the number of detected sources, hence for the sake of comparison, we set this number to the actual number of sources.

Since the data are simulated, we may evaluate the performance of the procedure with respect to the knowledge of the pertaining source of each photon emission and of its location. According to the former aspect, we compute the Adjusted Rand Index (ARI), which takes maximum value equal to one in the case of a perfect classification. As for the second aspect, the estimated source locations, defined by the galactical coordinates of the local maximum of the density within each detected group, are associated to the true source which presents minimum distance. The association is confirmed if the detected source belongs to the range of the positional error available for the true source. As a summarizing measure of the quality of the association, we compute the True Positive Rate (TPR) and the False Positive Rate (FPR). The former index is defined as the proportion of true sources correctly detected, while the latter one corresponds to the proportion of estimated components which are not associated with any source.



Nonparametric method	
ARI	0.985
$\overline{d}(s,\hat{s})$	0.030
TPR	0.675
FPR	0.240
Single linkage	
ARI	0.953
$\overline{d}(s,\hat{s})$	0.409
TPR	0.337
FPR	0.746

Fig. 2 Aitoff projection of the sky map data. The analysis has been restricted to the highlighted area, where location of the true sources is identified by the overimposed red crosses (left). Clustering results; $\overline{d}(s,\hat{s})$ is the average distance between true and detected sources (right).

The procedure identifies 150 out of 169 sources. Results, summarized in Figure 2, show and excellent performance with respect to the allocation of the photons to the pertaining sources with an ARI very close to 1, meaning that the unidentified sources include a very small fraction of photons. The TPR, albeit pretty good, indicates that the precision of source location can be improved, despite the average distance between the true and the detected sources amount to 0.03 degrees only. The procedure outperforms the competitor with respect to all the criteria, despite the competitor has been given an head start by suggesting the true number of sources.

While results have proven more than satisfactory, there is room for improvement. A first aspect which is worth to investigate on derives from the possible use of cartesian coordinates and a kernel function specifically designed for spherical data, as their directional nature cannot be completely caught by the use of galactic coordinates. Additionally, our procedure requires that the background components have been previously filtered out. Future research will focus on how to account for this aspect within the estimation procedure.

Acknowledgements This research was supported by SID 2018 grant "Advanced statistical modelling for indexing celestial object" (BIRD185983) awarded by the Department of Statistical Sciences of the University of Padova.

References

1. Ackermann, M., et al.: Determination of the point-spread function for the Fermi LAT from on-orbit data and limits on pair halos of active galactic nuclei. Astroph. J., 765.1 (2013): 54.
2. Hobson, M.P., Jaffe, A.H., Liddle, A.R., Mukherjee, P., Parkinson, D.: Bayesian Methods in Cosmology. Cambridge University Press (2010).
3. Menardi, G.: A review on modal clustering. Int. Stat. Rev., 84.3 (2016): 413-433.
4. Sottosanti, A., Costantin, D., Bastieri, D. and Brazzale, A.R.: Discovering and locating high-energy extra-galactic sources by Bayesian mixture modelling. New Statistical Developments in Data Science. SIS 2017 (A. Petrucci, F. Racioppi and R. Verde Eds.), Springer Proceedings in Math. & Stat., 288, (2019): 135148.

Statistical Analysis of Macroseismic Data for a better Evaluation of Earthquakes Attenuation Laws

Analisi statistica dei dati macrosismici per una migliore valutazione delle leggi di attenuazione dei terremoti

Marcello Chiodi, Antonino D'Alessandro, Giada Adelfio, Nicoletta D'Angelo

Abstract In this work we propose a statistical approach, based on the joint analysis of macroseismic data of Italian seismic events of the last two centuries, with which we obtain simultaneously maximum likelihood estimates of attenuation laws and coordinates of hypocenters. Our first results encourage us to use in the future more complex models, with a larger number of historical earthquakes.

Abstract *In questo lavoro proponiamo un'approccio statistico, fondato sull'analisi congiunta di dati macrosismici di eventi sismici italiani degli ultimi due secoli, col quale otteniamo simultaneamente stime di massima verosimiglianza delle leggi di attenuazione e delle coordinate ipocentrali. I primi risultati ci incoraggiano ad usare in futuro modelli più complessi per un numero maggiore di terremoti storici.*

Key words: Macroseismic Data, Attenuation Laws, Historical Earthquakes

1 Introduction to macroseismic data and theoretical attenuation laws

The estimation of macroseismic intensity of seismic events is carried out in order to quantify, through observations of the effects on buildings, the environment and

Marcello Chiodi
Università degli Studi di Palermo, e-mail: marcello.chiodi@unipa.it, and Istituto Nazionale di Geofisica e Vulcanologia

Antonino D'Alessandro
Istituto Nazionale di Geofisica e Vulcanologia, e-mail: antonino.dalessandro@ingv.it

Giada Adelfio
Università degli Studi di Palermo, e-mail: giada.adelfio@unipa.it, and Istituto Nazionale di Geofisica e Vulcanologia

Nicoletta D'angelo
Università degli Studi di Palermo, e-mail: nicoletta.dangelo@unipa.it

people, the shaking effects due to a moderate to strong earthquake. The macroseismic intensity is an ordinal variable that describes the seismic damage effects of an earthquake.

Macroseismic intensity is still often the only observed parameter to quantify the level of ground motion severity in many towns where seismometric instruments are not available [1, 2]. Moreover, macroseismic intensities are the only intensity measures available for pre-instrumental historical earthquakes and so are an important source to learn from historical earthquakes. Since 1960s reliable measurements of earthquakes shaking are available thanks to seismic networks. Seismograms analysis allow estimation of the severity of an earthquake by means of magnitude determination, which indirectly measures the amount of energy released by the shock.

The ground shaking observed at a site, when and an earthquake occurs, depends not only on the magnitude of the event, but also on its hypocentral depth and observation distance or epicentral distance. In fact, the wave field generated by an earthquake along propagation is subject to a lot of attenuation phenomena due to geometrical spreading, physical attenuation, reflection, refraction, diffraction, scattering of waves, etc. The phenomenon of seismic attenuation is therefore very complicated to model deterministically. Attenuation laws, generally deduced empirically on the basis of instrumental observations like PGA (Peak Ground Acceleration), show a logarithmic dependence from distance.

In general, it is widely accepted that macroseismic intensity depends significantly on ground shaking even if there is no physical model capable of giving quantitative account of these relationships. Due to their different nature, the two quantities PGA and Intensity show moderate levels of correlation, as widely observed in the current literature [3, 5, 6, 7, 8, 9, 12, 13].

In this paper, after a brief description of a dataset of historical eartquakes, we will propose the estimation of attenuation laws by means of a global estimation of a set of parameters including hypocentral coordinates. After some comments on the results, final remarks and ideas for further studies are outlined.

2 Historical earthquakes and data description

In this article we consider the spatial distribution of the effects caused by an earthquake as expressed by the values of the macroseismic intensity recorded at various locations surrounding the epicentre. The dataset here analysed are macroseismic data extracted from the Parametric Catalog of Italian Earthquakes ($CPTI15_{v2.0}$) and the Italian macroseismic database ($DBMI15_{v2.0}$). The $CPTI15_{v2.0}$ provides homogeneous parametric data, both macroseismic and instrumental, relating to earthquakes with maximum intensity ≥ 5 or magnitude ≥ 4.0 of interest for Italy in the time window 1000-2017. The $DBMI15_{v2.0}$ provides a homogeneous set of macroseismic intensities from different sources related to earthquakes with maximum intensity ≥ 5 and of interest for Italy in the time window 1000-2017. More details about these database can be found in [11].

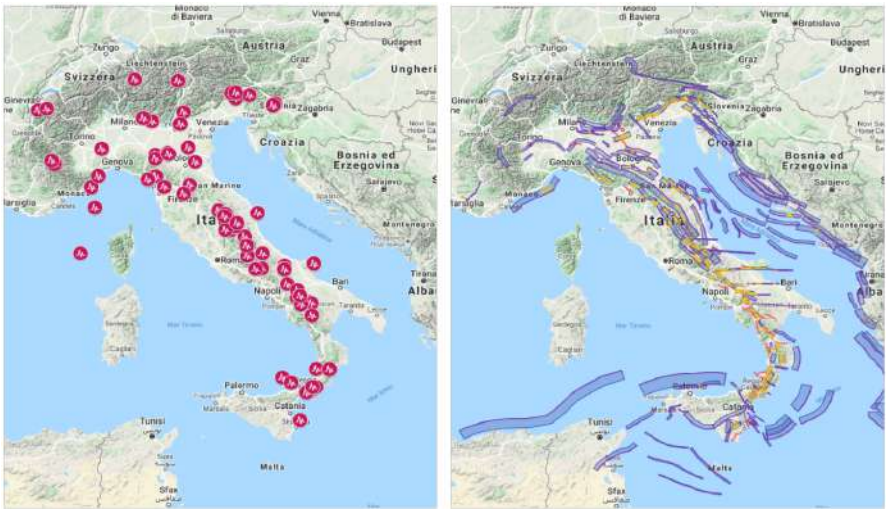


Fig. 1 Epicenters coordinates of 74 historical earthquakes observed from 1783 to 2016 (left); Main seismogenic sources [4]

For this first analysis data consist of 33,588 observations, that is the macroseismic intensities of 74 historical earthquakes (Fig. 1 left) observed from 1783 to 2016. Events have been selected mainly according to the quality of the dataset of historical reports avaliable for each earthquake, so that only events that on $DBMI15_{v2.0}$ have more than 300 macroseismic intensities observed have been used. For each earthquake E_j we have n_j observations ($j = 1,2,...,74$) of macroseismic intensities y_{ij},($i = 1,2,...,n_j$) on a scale ranging from 1 to 11. Observations with missing intensities have been dropped out so that n_j finally ranges from 123 to 1431. Each record contains of course the longitude x_{ij} and the latitude y_{ij} of the place where the macroseismic intensity has been recorded, all around Italy, and approximately centered on the epicenter of the event; estimates $\tilde{x}_{0j}, \tilde{y}_{0j}, \tilde{z}_{0j}$ ($j = 1,2,...,74$) of the hypocentral coordinates are also present in the historical records.

In the following, we illustrate the procedure followed to identify the attenuation law of macroseismic fields relative to the 74 selected earthquakes.

3 Statistical Analysis

To estimate simultaneously parameters of attenuation laws and reasonable location of events, we considered the basic model:

$$I_{ij} = \beta_{0j} + \beta_{1j} \log \sqrt{d^2(P_{ij}, P_{0j}) + z_{0j})^2} + \varepsilon_{ij}, \tag{1}$$

$$j = 1,2,...,k;\ i = 1,2,...,n_j$$

where I_{ij} is the $i-th$ macroseismic intensity, observed in the point P_{ij} of longitude and latitude $P_{ij} : \{x_{ij}, y_{ij}\}$, P_{0j}, is the epicenter, with longitude and latitude $P_{0j} : \{x_{0j}, y_{0j}\}$ and $d^2(P_{ij}, P_{0j})$ is their geodetic distance; z_{0j} is the depth of the $j-$th event and ε_{ij} is a random component (in this first version basic assumptions have been made about their conjoint distribution [1]).

$\beta_{0j}, \beta_{1j}, x_{0j}, y_{0j}, z_{0j}$ $(j = 1, 2, ..., 74)$ must be estimated from observed data.

We used standard optimization methods to obtain simultaneously all Maximum Likelihood estimators. After a preliminary analysis inside each event, We discarded from successive analysis the event number 66, because the estimated values where not reliable: with this elimination the optimization algorythm globally converges and we obtained for each event E_j new hypocentral locations $\hat{x}_{0j}, \hat{y}_{0j}, \hat{z}_{0j}$, together with the parameters which describe the attenuation laws.

4 Comments on first results

We summarized main results in few figures: a very interesting result is the relocation of events according to the Maximum Likelihood estimation of hypocentral parameters Fig. 2.

Particularly the depth $\hat{z}_{0j}$ seems more reliable with respect to original estimates, since many values of $\tilde{z}_{0j}$ were missing or substitued with a value of 10 kilometers. The median distance between estimated and original hypocentral coordinates is of 13.2 km.

Few diagnostic statistical summaries are reported in Fig. 3 residuals have been computed on the final fitted model (model1), where all unknown quantities have been estimated simultaneously for all earthquakes (that is attenuation law parameters β_{0j}, β_{1j} and hypocenter coordinates x_{0j}, y_{0j}, z_{0j}, $j = 1, 2, ..., 74$)

The left plot shows a satisfactory behaviour of residuals distribution with respect to fitted values, given however the elementary assumptions made in this first approach; further on the right we observe that the empirical distribution of residuals shows a marked leptokurtosis [2].

In Fig. 2, on the right, we report the main directions for each event computed on the basis of observed macroseismic intensities and centered on relocated epicenters: we can see that these directions are compatible with th main seismogenetic sources reported in the right panel of Fig. 1.

[1] In a foregoing version more specific assumptions will be made about a possible spatial autocorrelation of the ε_{ij}; some considerations about effects of departure from normality will be made in the final section

[2] Indeed we also estimated a value of the parameter p of a p-normal distribution (also known as Exponential Power Distribution) between 1 and 2, but we will not deal with this problem in the present paper: however a relocation made with maximum likelihood with estimated p led to values very close to normal ones, that is for $p = 2$

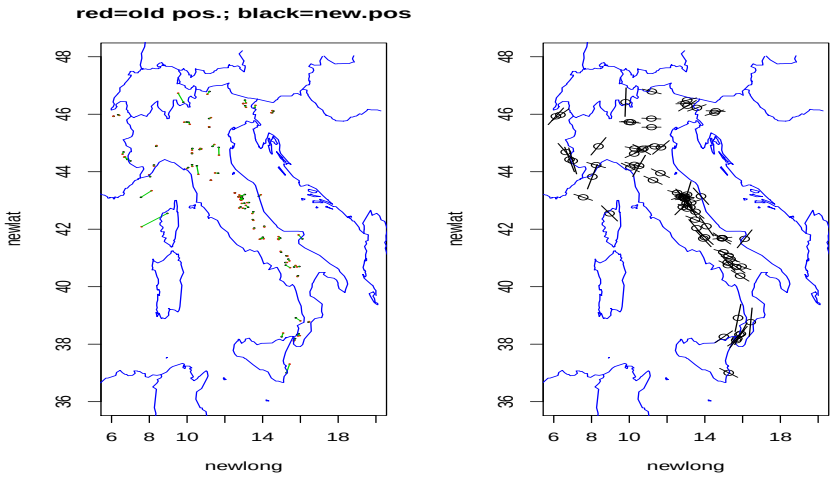


Fig. 2 Relocation of the epicenters on the basis of the model 1. Original location in blank squares, new estimates in black circles. On the right estimation of the main directions for each event on the basis of observed macroseismic intensities

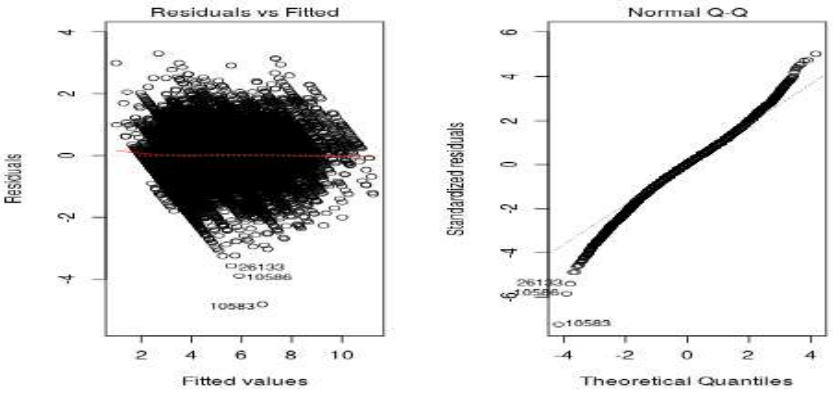


Fig. 3 Final residuals from the model 1. Residuals vs. fitted (left); Q-Q plot of residuals vs a normal distribution

5 Final remarks and future perspectives

The problem of determining quantitative parameters of pre-instrumental earthquakes is an interesting challenge in seismology, because the location of the epicenter must be based on amplitudes since arrival times are not meaningful for macroseismic data, intensity is an ordinal quantity that is only weakly correlated with

the amplitude of the ground motion and the real paths traveled within the Earth's lithosphere by the seismic wave are unknown.

The availability of extensive collections of macroseismic observations, and of modern computing powers, has led to the development of methods that use the spatial distribution of intensities to determine hypocentral parameters and attenuation laws.

In a forecoming study we will analyze information about a larger number of historical earthquakes and we will study the problem of optimal scaling of macroseismic intensity together with a deeper analysis of the spatial components.

References

1. Agostinelli C., Rotondi R.: Analysis of macroseismic fields using statistical data depth functions: considerations leading to attenuation probabilistic modelling, Bull. Earthquake Eng. **14**, 7, 1869–1884 (2015) doi: 10.1007/s10518-015-9778-2.
2. Agostinelli, C, Rotondi, R., Varini, E, . Clustering macroseismic fields by statistical data depth functions: Studies in Classification, Data Analysis, and Knowledge Organization, **2**15879, 145–153, (2018) doi: 10.1007/978-3-319-55708-3-16.
3. Atkinson, G.M.: Linking historical intensity observations with ground-motion relations for eastern North America, Seismol. Res. Lett., **72**, 560–574 (2001)
4. Basili, R., Kastelic, V., Valensise, G., and DISS Working Group 2009, DISS3 tutorial series: Guidelines for compiling records of the Database of Individual Seismogenic Sources, version 3, Rapporti Tecnici INGV, **108**, 20 (2009)
5. Boatwright, J., K. Thywissen and L.C. Seekins . Correlation of Ground Motion and Intensity for the 17 January 1994 Northridge, California, Earthquake, Bull. Seismol. Soc. Amer., **91**, 4, 739–752 (2001)
6. Chiaruttini, C. and L. Siro . The correlation of peak ground horizontal acceleration with magnitude, distance, and seismic intensity for Friuli and Ancona, Italy, and the Alpide Belt, Bull. Seismol. Soc. Amer., **71**, 6, 1993–2009 (1981)
7. Margottini, C., D. Molin and Serva, L.: Intensity versus ground motion: A new approach using Italian data, Engineering Geology, **33**, 1, 45–58 (1992)
8. Masi, A., Chiauzzi, L., Nicodemo, G., Manfredi, V., . Correlations between macroseismic intensity estimations and ground motion measures of seismic events. Bulletin of Earthquake Engineering, 1–34 (2020) doi: 10.1007/s10518-019-00782-2.
9. Panza, G.F., Cazzaro, R. and Vaccari, F.: Correlation between macroseismic intensities and seismic ground motion parameters, Ann. Geophys., **XL**, 5, 1371–1382 (1997)
10. Rotondi, R., Varini, E., Brambilla, C.: Probabilistic modelling of macroseismic attenuation and forecast of damage scenarios, Bulletin of Earthquake Engineering, **14**, 7, 1777–1796 (2014) doi: 10.1007/s10518-015-9781-7.
11. Rovida, A., Locati, M., Camassi, R., Lolli, B., Gasperini, P.: Catalogo Parametrico dei Terremoti Italiani (CPTI15), versione 2.0. Istituto Nazionale di Geofisica e Vulcanologia (INGV). (2019) doi: 10.13127/CPTI/CPTI15.2.
12. Schenk, V., F. Mantlik, M.N. Zhizhin and Timarkin, A.G.: Relation between macroseismic intensity and instrumental parameters of strong motions: a statistical approach, Nat. Hazards, **3**, 111–124 (1990)
13. Wald, D. J.,Quitoriano, V., Heaton, T.H. , Kanamori, H.: Relations between Peak Ground Acceleration, Peak Ground Velocity, and Modified Mercalli Intensity in California, Earthquake Spectra, **15**, 3, 557–564 (1999)

Network Modelling in Biostatistics

Natural direct and indirect relative risk for mediation analysis

Rischio relativo naturale diretto e indiretto per l'analisi di mediazione

Monia Lupparelli and Alessandra Mattei

Abstract We consider a mediation setting involving a mediator which may channeling a part of the treatment effect along the causal pathway between the treatment and the primary outcome. Under a sequential ignorability assumption, we propose a recursive regression framework for binary outcomes so that the total causal relative risk can be decomposed into the natural direct and indirect relative risk by combining model parameters. Inference is performed by maximum likelihood methods.
Abstract *Nel contesto dell'analisi di mediazione si prevede la presenza di un mediatore che potrebbe interferire sull'effetto del trattamento sulla variabile finale di interesse. Sotto l'assunzione di ignorabilitá sequenziale, si propone un modello di regressione ricorsivo per variabile risposta binaria in cui il rischio relativo totale del trattamento rispetto alla variabile di interesse si puó scomporre nel rischio relativo naturale diretto ed indiretto combinando i parametri del modello. Metodi di massima verosimiglianza sono usati per l'inferenza.*

Key words: binary or continuous mediator, potential outcome, causal model

1 Introduction

Casual mediation analysis focuses on understanding the causal pathways by which a treatment affects an outcome, by investigating the mediating role of an intermediate variable, named mediator, in the treatment-outcome relationship. When the effect of the treatment on the response might be channeled by a mediator, the interest is on

Monia Lupparelli
Department of Statistics, Computer Science, Applications
University of Florence, e-mail: monia.lupparelli@unifi.it

Alessandra Mattei
Department of Statistics, Computer Science, Applications
University of Florence, e-mail: alessandra.mattei@unifi.it

disentangling *indirect effects*, that are through the mediator of interest, and *direct effects*, that are through other pathways other than the mediator [4, 3, 2, 8].

We focus on causal relative risks, defining natural direct and indirect effects for binary outcomes on the risk ratio scale [7]. Under a sequential ignorability assumption, which is usually invoked in mediation analysis, we propose to draw inference on the natural direct and indirect relative risks using a recursive regression framework, where a regression model for the binary outcome, based on the log-link function, is coupled with another regression model for binary responses if the mediator is binary and a linear regression model for continuous responses if the mediator is continuous. We show that under the assumption that models are correctly specified, the natural direct and indirect relative risks can be derived by analytically combining model parameters. Inference can be performed by maximum likelihood methods.

The method is applied to the observational Swedish National March Cohort (NMC) study to investigate the causal mechanism between physical activity, body mass index, and cardiovascular disease.

2 Notation and causal estimands

Let Y be a binary outcome, W a binary treatment and M a mediator, a post-treatment intermediate variable lining in the causal pathway between the treatment and the outcome. The outcome and the treatment take value $y, w \in (0,1)$, respectively, and the mediator $m \in \mathscr{M}$, e.g., $\mathscr{M} = \{0,1\}$ whether the mediator is binary and $\mathscr{M} = R$ whether it is continuous. Also, let X be a vector of covariates with generic element $x \in R^k$, where k denotes the size of X.

Let $M(w)$ and $Y(w)$ denote the potential outcomes for the mediator and the primary endpoint, respectively: they are the values of M and Y under assignment to treatment w, $w = 0,1$. Let W, M^{obs} and Y^{obs} denote the observed values of the treatment, the mediator and the outcome: $M^{\text{obs}} = M(W) = WM(1) + (1-W)M(0)$ and $Y^{\text{obs}} = Y(W) = WY(1) + (1-W)Y(0)$. We are interested in disentangling *direct* and *indirect effects* of the treatment, W, on the outcome, Y, in the presence of the mediator, M.

Following the literature on mediation analysis (e.g., [4, 3]), to formalize the concepts of direct and indirect effects, we introduce potential outcomes of the form $Y(w,m)$ and $Y(w,M(w'))$, $w, w' \in \{0,1\}$: $Y(w,m)$ would be the value of the outcome Y if the treatment were set to the level w and the mediator M were set to the value m; and $Y(w,M(w'))$ would be the value of the outcome Y if the treatment were set to the level w and the mediator M were set to the value it would have taken if the treatment had been set to an alternative level, w'.

The causal relative risk conditional on covariates level $X = x$,

$$RR_{Y|x} = \frac{P(Y(1)=1|X=x)}{P(Y(0)=1|X=x)} = \frac{P(Y(1,M(1))=1|X=x)}{P(Y(0,M(0))=1|X=x)} \tag{1}$$

can be decomposed into the product of a 'natural' direct relative risk,

$$RR_{Y|x}^{dir} = \frac{P(Y(1,M(0)) = 1|X = x)}{P(Y(0,M(0)) = 1|X = x)} \tag{2}$$

and a 'natural' indirect relative risk,

$$RR_{Y|x}^{ind} = \frac{P(Y(1,M(1)) = 1|X = x)}{P(Y(1,M(0)) = 1|X = x)} \tag{3}$$

as $RR_{Y|x} = RR_{Y|x}^{dir} \times RR_{Y|x}^{ind}$. Direct and indirect relative risks can be identified under these assumptions [1, 2]:

Assumption 1 *(Ignorability of the treatment).*

$$\{Y(w,m), M(w')\} \perp\!\!\!\perp W \mid X \qquad \textit{for all } w,w' \in \{0,1\}, m \in \mathscr{M}$$

Assumption 2 *(Ignorability of the mediator).*

$$Y(w,m) \perp\!\!\!\perp M(w') \mid W = w', X \qquad \textit{for all } w,w' \in \{0,1\}, m \in \mathscr{M}$$

which [2] jointly refer to as *sequential ignorability*. Ignorability of the treatment states that the treatment assignment mechnism is ignorable conditional on the observed covariates and holds by design in randomized experiments. Ignorability of the mediator states that the mediator is ignorable given the observed treatment and covariates. Under Assumptions 1 and 2, the mediation formula holds [3]:

$$P[Y(w,M(w')) = 1 \mid X = x] = \int_{m \in \mathscr{M}} P(Y^{\text{obs}} = 1 \mid M^{\text{obs}} = m, W = w, X = x) f_{M^{\text{obs}}|w',x}(m \mid w', x) \nu(dm)$$

where $f_{M^{\text{obs}}|w',x}(m \mid w', x)$ is the probability density/mass function for the random variable $M^{\text{obs}} \mid \{W = w', X = x\}$, with $w' \in \{0,1\}$ and $x \in R^k$, and ν is the Lebesgue measure if M is continuous and the counting measure if M is discrete. Therefore, under Sequential Ignorability we can use regression models for estimate natural direct and indirect causal relative risks.

3 The model

3.1 Binary mediator

We consider a recursive regression framework based on the log-link function for the binary random vector $(W^{\text{obs}}, M^{\text{obs}}, Y^{\text{obs}})$:

$$\log P(M^{\text{obs}} = 1 \mid W, X) = \beta_\emptyset + \beta_W W + \beta_X^T X \tag{4}$$

$$\log P(Y^{\text{obs}} = 1 \mid M^{\text{obs}}, W, X) = \alpha_\emptyset + \alpha_W W + \alpha_M M^{\text{obs}} + \alpha_X^T X. \tag{5}$$

For the sake of simplicity, we consider additive models with no interaction terms; the generalization including multiplicative effects is straightforward. Regression coefficients in Equation (4) and (5) correspond to the logarithm of conditional relative risks, for any $x \in R^k$:

$$\beta_W = \log RR^{\text{obs}}_{M|W.x} = \log \frac{P(M^{\text{obs}}=1|W=1,X=x)}{P(M^{\text{obs}}=1|W=0,X=x)},$$

$$\alpha_W = \log RR^{\text{obs}}_{Y|W.m,x} = \log \frac{P(Y^{\text{obs}}=1|W=1,M^{\text{obs}}=m,X=x)}{P(Y^{\text{obs}}=1|W=0,M^{\text{obs}}=m,X=x)}; \qquad m \in \mathscr{M} = \{0,1\}$$

$$\alpha_M = \log RR^{\text{obs}}_{Y|M.w,x} = \log \frac{P(Y^{\text{obs}}=1|W=w,M^{\text{obs}}=1,X=x)}{P(Y^{\text{obs}}=1|W=w,M^{\text{obs}}=0,X=x)}; \qquad w \in \{0,1\}.$$

The $k-$dimensional parameter vector α_X and β_X represent the effect of the individual covariate X on the distribution of Y and of M, respectively.

Then, under Assumptions 1 and 2, and under correct specification of the regression models, the natural direct and indirect relative risks are given by

$$RR^{dir}_{Y|x} = \frac{[1 - e^{(\beta_\emptyset + \beta_X x)}]e^{\alpha_W} + e^{(\beta_\emptyset + \beta_X x)}e^{(\alpha_W + \alpha_M)}}{1 - e^{(\beta_\emptyset + \beta_X x)} + e^{(\beta_\emptyset + \beta_X x)}e^{\alpha_M}} = e^{\alpha_W}$$

and

$$\begin{aligned} RR^{ind}_{Y|x} &= \frac{[1 - e^{(\beta_\emptyset + \beta_W + \beta_X^T x)}]e^{(\alpha_\emptyset + \alpha_W + \alpha_X^T x)} + e^{(\beta_\emptyset + \beta_W + \beta_X^T x)}e^{(\alpha_\emptyset + \alpha_W + \alpha_M + \alpha_X^T x)}}{[1 - e^{(\beta_\emptyset + \beta_X^T x)}]e^{(\alpha_\emptyset + \alpha_W + \alpha_X^T x)} + e^{(\beta_\emptyset + \beta_X^T x)}e^{(\alpha_\emptyset + \alpha_W + \alpha_M + \alpha_X^T x)}} \\ &= \frac{[1 - e^{(\beta_\emptyset + \beta_W + \beta_X^T x)}] + e^{(\beta_\emptyset + \beta_W + \beta_X^T x)}e^{\alpha_M}}{[1 - e^{(\beta_\emptyset + \beta_X^T x)}] + e^{(\beta_\emptyset + \beta_X^T x)}e^{\alpha_M}}. \end{aligned}$$

It is worth noticing that, assuming the convention that $RR^{reg}_{Y|M.w,x} = 1$ for $M^{\text{obs}} = 0$, the natural indirect relative risk can be written as ratio of weighted averages of conditional relative risks of the mediator M^{obs} on the outcome, Y^{obs} with weights given by the conditional probabilities $P(M^{\text{obs}} = m \mid W = w)$, $w = 0, 1$:

$$RR^{ind}_{Y|x} = \frac{\sum_{m=0}^{1} RR^{\text{obs}}_{Y|M.w,x} P(M^{\text{obs}} = m \mid W = 1, X = x)}{\sum_{m=0}^{1} RR^{\text{obs}}_{Y|M.w,x} P(M^{\text{obs}} = m \mid W = 0, X = x)}.$$

3.2 Gaussian mediator

In studies where the mediator is continuous, we assume that $M^{\text{obs}} \mid \{W, X\} \sim N\left(E[M^{\text{obs}} \mid W, X], \sigma^2\right)$ and use the following recursive regression framework:

$$E[M^{\text{obs}} \mid W, X] = \gamma_0 + \gamma_W W + \gamma_X^T X \tag{6}$$

$$\log P(Y^{\text{obs}} = 1 \mid W, M^{\text{obs}}, X) = \alpha_0 + \alpha_W W + \alpha_M M^{\text{obs}} + \alpha_X^T X. \tag{7}$$

Under Assumptions 1 and 2, and under correct specification of the regression models, we can show that the natural direct and indirect relative risks are given by

$$RR_{Y|x}^{dir} = \frac{e^{(\alpha_0+\alpha_W+\alpha_X^T x)} e^{\left[(\gamma_0+\gamma_X^T x)\alpha_M+\sigma^2\frac{\alpha_M^2}{2}\right]}}{e^{(\alpha_0+\alpha_X^T x)} e^{\left[(\gamma_0+\gamma_X^T x)\alpha_M+\sigma^2\frac{\alpha_M^2}{2}\right]}} = e^{\alpha_W}$$

and

$$RR_{Y|x}^{ind} = \frac{e^{(\alpha_0+\alpha_W+\alpha_X^T x)} e^{\left[(\gamma_0+\gamma_W+\gamma_X^T x)\alpha_M+\sigma^2\frac{\alpha_M^2}{2}\right]}}{e^{(\alpha_0+\alpha_W+\alpha_X^T x)} e^{\left[(\gamma_0+\gamma_X^T x)\alpha_M+\sigma^2\frac{\alpha_M^2}{2}\right]}} = e^{\gamma_W \alpha_M}.$$

In case of continuous mediators, both the natural direct and indirect effect are independent of the covariate set. We remark that the indirect effect specified on the odds ratio scale through logistic regression models, is not invariant with respect to the covariates, even for the case of continuous mediators; see [7].

4 Case study

The method is applied to a real data set taken from the observational Swedish National March Cohort (NMC) study, previously analyzed by [6, 5]. The NMC data set includes information on self-reported physical activity (PA) level, body mass index (BMI), and baseline covariates (X), measured for each subject at enrollment. Subjects are followed from year 1997 to 2004 and each cardiovascular disease (CVD) event is recorded. Following [6], we classify each subject as either a "low-level exerciser" ($W = 1$) or a "high-level exerciser" ($W = 0$). Let $Y^{\text{obs}} = 1$ for subjects who report at least one CVD event before end of follow-up, and $Y^{\text{obs}} = 0$ otherwise.

Our focus is on the extent of a causal effect of PA on CVD risk mediated or not mediated through BMI. We answer this research question applying the framework for mediation analysis we propose with BMI both as a continuous mediator on its original scale and as a binary mediator, which classifies subjects as either "Obese" ($M^{\text{obs}} = 1$) if their BMI ≥ 30 or "Not Obese" ($M^{\text{obs}} = 0$) if their BMI < 30. Under Assumptions 1 and 2 the analysis is conducted specifying models in Equations (4) and (5) for binary BMI and in Equations (6) and (7) for continuous BMI conditional on the following covariates: Gender (1 = Male, 0 = Female), age (in years), a binary indicator for smoking (1 = Current/Former smoker, 0 = No smoker), and a binary indicator for alcohol use (1 = Medium/High drinker, 0 = Low/No drinker).

Results are collected in Table 1 both for the case of binary and continuous mediator. If we consider continuous BMI, the estimates of the causal effects are all positive and highly significant. If we consider the model for binary BMI, the value of the estimates of the causal effects are comparable with the previous case, never-

Table 1 NMC study: Estimate and standard error (SE) of the total relative risk of PA on CVD and of the natural direct and indirect relative risks of PA on CVD with mediator BMI

Binary BMI[a]			Continuous BMI		
Estimand	Estimate	SE	Estimand	Estimate	SE
RR^{dir}_{CVD}	1.160	0.116	RR^{dir}_{CVD}	1.125	0.110
$RR^{ind}_{CVD\|x}$	1.024	0.165	RR^{ind}_{CVD}	1.052	0.008
$RR_{CVD\|x}$	1.187	0.225	RR_{CVD}	1.184	0.116

[a] In the case of binary BMI we show the estimate and the standard error of $RR^{ind}_{CVD|x}$ and of $RR_{CVD|x}$ for a woman aged 48 (sample mean of age), who is a no smoker and a low/no drinker.

theless the standard error of the estimate of the indirect causal effect is higher (0.165 instead of 0.008). The greater uncertainty may depend by the dichotomization of the BMI, since the sub-sample with $BMI \geq 30$ shows a greater sparsity.

We consider the model for the continuous mediator easier to interpret since both the natural direct and indirect effect are independent of the covariate set. Nevertheless, the marginal indirect causal effects could be derived even for the model with the binary BMI by summing over the sampling distribution of the covariate set.

References

1. Forastiere, L., Mattei, A., Ding, P.: Principal ignorability in mediation analysis: through and beyond sequential ignorability. *Biometrika*, **105**, 4, 979–986 (2018)
2. Imai, K., Keele, L., Yamamoto, T.: Identification, inference, and sensitivity analysis for causal media- tion effects. *Statistical Science*, **25**, 51–71 (2010).
3. Pearl, J.: Direct and indirect effects. In: Breese, J. S. and Koller, D. (eds.) *17th Conference on Uncertainty in Artificial Intelligence*, pp. 411–420 (2001).
4. Robins, J. M. and Greenland, S.: Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155 (1992).
5. Schwartz, S., Li, F., Mealli, F.: A Bayesian Semiparametric Approach to intermediate variables in causal inference. J. Am. Stat. Assoc., **106**, 13311344 (2011)
6. Sjölander, A., Humphreys, K., Vansteelandt, S., Bellocco, R., Palmgren, J.: Sensitivity Analysis for Principal Stratum Direct Effects, With an Application to a Study of Physical Activity and Coronary Heart Disease. *Biometrics*, **65**, 514520 (2009)
7. Vanderweele TJ, Vansteelandt S.: Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol.*, **172**, 13391348 2010
8. VanderWeele2015, T.J.: *Explanation in Causal Inference. Methods for Mediation and Interaction*. Oxford University Press, New York (2015)

New issues on multivariate and univariate quantile regression

Mixtures of linear quantile regression models for longitudinal data: an R package

Misture di modelli di regressione quantile per dati longitudinali: un pacchetto R

M.F. Marino, M.G. Ranalli, and M. Alfò

Abstract In this paper, we introduce **`lqmix`**, an R package that has been specifically tailored to the estimation of mixtures of linear quantile regression models, based on time-constant and/or time-varying, discrete, random coefficients. An Expectation-Maximization (EM) algorithm is used to obtain maximum-likelihood estimates of model parameters and likelihood-based information criteria are adopted to select the number of mixture components. For the regression coefficients, bootstrap-based standard errors and confidence intervals are also provided.

Abstract *In questo paper, si introduce il pacchetto R* ***`lqmix`****, specificamente pensato per la stima di misture di modelli di regressione quantilica basate sull'utilizzo di coefficienti casuali discreti tempo-costanti e/o tempo-variabili. Un algoritimo EM è utilizzato per la stima dei parametri tramite un approccio di massima verosimiglianza, mentre si considerano criteri standard basati sulla penalizzazione della funzione di verosimiglianza per la scelta del numero ottimale di componenti della mistura finita. In output, si ottengono infine le stime degli errori standard e degli intervalli di confidenza per i parametri del modello impiegando ricampionamento via bootstrap.*

Key words: Random coefficients, NPML, hidden Markov models, EM algorithm.

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: mariafrancesca.marino@unifi.it

Maria Giovanna Ranalli
Dipartimento di Scienze Politiche, Università degli Studi di Perugia, e-mail: giovanna.ranalli@unipg.it

Marco Alfò
Dipartimento di Scienze Statistiche, "Sapienza" Università di Roma, e-mail: marco.alfo@uniroma1.it

1 Introduction

In the statistical literature, quantile regression [5] has become a quite popular and established technique for the analysis of data when the interest is not focused on and goes beyond the (conditional) mean. With longitudinal data, dependence between observations recorded on the same individual need to be taken into account to avoid bias and inefficiency issues in parameter estimates. This can be achieved by including in the quantile regression model individual-specific, time-constant and/or time-varying, random coefficients that, being shared by all measures associated to a given individual, make repeated observations dependent of each other. In this framework, a standard way of proceeding is that of specifying a parametric distribution for the random coefficients [see, e.g. 3, 4]. A further alternative is that of leaving such a distribution unspecified and estimate it directly from the observed data, by exploiting a finite mixture specification of the model. This can be either time-constant or time-varying, as we will detail in the following. Such an approach offers a number of specific advantages: (*i*) it allows us to avoid unverifiable assumptions on the random coefficient distribution; (*ii*) it allows us to account for extreme and/or asymmetric departures from the homogeneous model, as random coefficients are completely free to vary over the corresponding support; (*iii*) the discrete nature of the mixing distribution allows us to avoid integral approximations and considerably reduces the computational effort to derive parameter estimates.

2 Modeling alternatives

Random coefficient models represent a standard approach to model the effect of observed covariates on an outcome repeatedly observed over time. This also holds in the quantile regression framework, where the interest is in modeling the (conditional) quantiles of the outcome distribution as a function of fixed and random coefficients. To ensure flexibility and avoid unverifiable parametric assumptions on the random coefficient distribution, a specification based on the use of (dynamic) finite mixtures represent a viable strategy to adopt. According to the chosen specification, time-constant (time-varying), discrete, random coefficients are added to the model to capture time-constant (time-varying) sources of individual-specific unobserved heterogeneity.

2.1 Linear quantile regression with time-constant random coefficients

To account for sources of individual-specific unobserved heterogeneity, [1] proposed to include in a quantile regression model time-constant random coefficients without a pre-specified parametric distribution. Such a distribution is directly estimated from the data via a NonParametric Maximum Likelihood approach [NPML - 7, 8, 9]. In this framework, the NPML approach leads to the estimation of

a (quantile-specific) discrete mixing distribution defined over a finite number of (quantile-specific) locations. For a given quantile level $\tau \in (0,1)$, the model likelihood resembles that of a finite mixture of linear quantile regressions:

$$L(\cdot \mid \tau) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G[\tau]} \left[\prod_{t=1}^{T_i} f(y_{it} \mid c_{ig} = 1; \tau) \right] \pi_g[\tau] \right\}, \tag{1}$$

where $i = 1, \ldots, n$, and $t = 1, \ldots, T_i$, index individuals and time occasions, respectively. The component label c_{ig} is the generic element of the random vector $\mathbf{c}_i = (c_{i1}, \ldots, c_{iG[\tau]})'$, with $G[\tau]$ being the number of mixture components for the the τ-th quantile level; in particular, c_{ig} is equal to one if unit i belongs to the g-th component of the finite mixture and is zero otherwise. Component probabilities are denoted by $\pi_g[\tau] = \Pr(c_{ig} = 1; \tau)$, $g = 1, \ldots, G[\tau]$, while $f(y_{it} \mid c_{ig} = 1; \tau)$ is the conditional density of the Asymmetric Laplace (AL) distribution, with skewness and scale parameters equal to τ and σ respectively, and location parameter modeled as

$$\mu_{it}[c_{ig}; \tau] = \mathbf{x}'_{it}\beta[\tau] + \mathbf{z}'_{it}\zeta_{c_{ig}}[\tau].$$

Here, $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ denote the vectors of observed covariates associated to fixed ($\beta[\tau]$) and random ($\zeta_{c_{ig}}[\tau]$) coefficients, respectively.

2.2 Linear quantile regression with time-varying random coefficients

To account for the presence of time-varying sources of individual-specific unobserved heterogeneity influencing the outcome of interest, [2] considered in the model specification time-varying random intercepts. These are assumed to evolve over time according to a homogeneous, first order, hidden Markov chain $\{S_{it}[\tau]\}$ defined over the state space $\mathscr{S}[\tau] = \{1, \ldots, m[\tau]\}$, with initial and transition probabilities given by $\delta(s_{i1}; \tau)$ and $q(s_{it} | s_{it-1}; \tau)$, respectively. Clearly, a broader specification based on general, time-varying, random coefficients can also be considered.

As for the model described above, the convenient assumption of AL-distributed responses is introduced and, for a given quantile level $\tau \in (0,1)$, model parameters are estimated by maximizing the following likelihood function:

$$L(\cdot \mid \tau) = \prod_{i=1}^{n} \left\{ \sum_{\mathbf{s}_i} \left[\delta(s_{i1}; \tau) f(y_{i1} \mid s_{i1}; \tau) \right] \left[\prod_{t=2}^{T_i} q(s_{it} | s_{it-1}; \tau) f(y_{it} \mid s_{it}; \tau) \right] \right\}. \tag{2}$$

In the above expression, $\sum_{\mathbf{s}_i}$ is a shorthand for $\sum_{s_{i1}} \cdots \sum_{s_{iT_i}}$, while $f(y_{it} \mid s_{it}; \tau)$ denotes the AL density with scale and skewness equal to τ and σ respectively, while the location parameter is modeled as

$$\mu_{it}[s_{it}; \tau] = \mathbf{x}'_{it}\beta[\tau] + \mathbf{w}'_{it}\alpha_{s_{it}}[\tau].$$

Here, $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$ denote the vectors of covariates associated to fixed ($\beta[\tau]$) and random ($\alpha_{s_{it}}[\tau]$) coefficients, respectively. When looking at equation (2), we may easily recognize an dynamic extension of the likelihood reported in equation (1); that is, it represents the likelihood of a dynamic finite mixture model, where random coefficients associated to a generic individual i vary according to the hidden state he/she visits at each time occasion.

2.3 Linear quantile regression with time-constant and time-varying random coefficients

In some real data applications, both time-constant and time-varying sources of individual-specific unobserved heterogeneity may affect the outcome distribution. This makes the quantile regression models described above no longer appropriate as they account for one source at a time only. To face such situations, [10] proposed to include both types of random coefficients in the model specification. In this framework, the likelihood function is given by

$$L(\cdot \mid \tau) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G[\tau]} \sum_{\mathbf{s}_i} \left[\delta(s_{i1};\tau) f(y_{i1} \mid s_{i1};\tau) \right] \times \right. \\ \left. \times \left[\prod_{t=2}^{T_i} q(s_{it}|s_{it-1};\tau) f(y_{it} \mid c_{ig}=1, s_{it};\tau) \right] \right\} \pi_g[\tau], \tag{3}$$

where $f(y_{it} \mid \mathbf{c}_{ig}=1, s_{it};\tau)$ denotes the AL density with location parameter modeled as

$$\mu_{it}[c_{ig}=1, s_{it};\tau] = \mathbf{x}'_{it}\beta[\tau] + \mathbf{z}'_{it}\zeta_{c_{ig}}[\tau] + \mathbf{w}'_{it}\alpha_{s_{it}}[\tau],$$

while all other quantities are defined as above.

In this framework, unobserved individual-specific features that remain constant over time may be captured thanks to the random coefficients $\zeta_{c_{ig}}[\tau]$. Similarly, sudden temporal shocks in the individual profiles, due to time-varying sources of unobserved heterogeneity may be easily captured by the random coefficients $\alpha_{s_{it}}[\tau]$ in the model. These features render therefore the proposed model more flexible and general than the alternatives described so far, at the cost of a higher computational complexity.

3 The `lqmix` R package

In this paper, we introduce the R [11] package **`lqmix`**, specifically tailored to the estimation of (dynamic) finite mixtures of linear quantile regression models. Separate functions allow to estimate the three model specifications described above, based on time-constant (1), time-varying (2), and time-constant and time-varying (3) random coefficients, respectively. Both random intercepts and random slopes are supported.

An EM algorithm is used to derive estimates in a maximum likelihood framework, even though the parametric specification for the (conditional) distribution for the responses is only introduced as a computational convenient trick to recast estimation in such a framework. Standard penalized likelihood criteria are computed to identify the optimal number of mixture components. Standard errors and confidence intervals for model parameters are obtained according to a non-parametric block-bootstrap. That is, they are obtained by re-sampling individuals and retaining the corresponding sequence of measurements to preserve within individual dependence [6]. Missingness in the responses is also taken into account, according to the implicit assumption of a Missing At Random (MAR) mechanism.

4 Concluding remarks

In this paper, the class of (dynamic) finite mixtures of quantile regression models for the analysis of longitudinal data has been explored. Besides having recently become rather popular in statistics, there is still a lack of support within the most popular statistical software. Here, we have introduced a novel and efficient R package specifically conceived for fitting and making inference on the parameters defining this class of models.

References

[1] M. Alfó, N. Salvati, and M. G. Ranalli. Finite mixtures of quantiles and M-quantile models. *Statistics and Computing*, 27:547–570, 2017.

[2] A. Farcomeni. Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Stat. Comput.*, 22, 2012.

[3] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–54, 2007.

[4] M. Geraci and M. Bottai. Linear quantile mixed models. *Statistics and Computing*, pages 1–19, 2013.

[5] R. Koenker and G. Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[6] S. Lahiri. Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, 27:386–404, 1999.

[7] N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[8] B. G. Lindsay. The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11(1):86–94, 1983.

[9] B. G. Lindsay. The geometry of mixture likelihoods, Part II: the exponential family. *The Annals of Statistics*, 11(3):783–792, 1983.

[10] Maria Francesca Marino, Nikos Tzavidis, and Marco Alfò. Mixed hidden markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statistical methods in medical research*, 27(7):2231–2246, 2018.

[11] R Core Team et al. R: A language and environment for statistical computing. 2013.

Multivariate Mixed Hidden Markov Model for joint estimation of multiple quantiles

Modello Hidden Markov Multivariato ad effetti misti per la stima congiunta di quantili condizionati

Merlo Luca, Petrella Lea and Tzavidis Nikos

Abstract This paper develops a Mixed Hidden Markov Model for joint estimation of multiple quantiles in a multivariate linear regression for longitudinal data. This method accounts for association among multiple responses and study how the relationship between dependent and explanatory variables may vary across different quantile levels of the conditional distribution of the multivariate response variable. Unobserved heterogeneity sources and serial dependence are jointly modeled through the introduction of individual-specific, time-constant random coefficients and time-varying parameters that evolve over time with a Markovian structure, respectively. Estimation is carried out via a suitable EM algorithm without parametric assumptions on the random effects distribution. We assess the empirical behaviour of the proposed methodology through the analysis of the Millennium Cohort Study data.

Abstract Questo lavoro sviluppa un modello di Markov nascosto multivariato ad effetti misti per la stima congiunta di quantili marginali condizionati associati a variabili risposta multivariate, nell'ambito di una regressione lineare per dati longitudinali. La metodologia proposta consente di tenere conto dell'associazione esistente tra le variabili risposta e intende studiare come tale struttura di associazione varia quando si considerano diversi quantili della distribuzione condizionata della variabile risposta. Le fonti di eterogeneità non osservate, costanti e variabili nel tempo, vengono modellate congiuntamente introducendo effetti casuali costanti e coefficienti che variano nel tempo secondo una catena di Markov latente. La stima dei

Merlo Luca
Department of Statistical Sciences, Sapienza University of Rome, e-mail: luca.merlo@uniroma1.it

Petrella Lea
MEMOTEF Department, Sapienza University of Rome, e-mail: lea.petrella@uniroma1.it

Tzavidis Nikos
Department of Social Statistics and Demography, University of Southampton, e-mail: N.TZAVIDIS@soton.ac.uk

parametri è ottenuta tramite l'algoritmo EM senza formulare assunzioni parametriche sulla distribuzione degli effetti casuali. La validità del nostro approccio viene analizzata attraverso un'applicazione empirica con i dati del Millennium Cohort Study.

Key words: Longitudinal data, Mixed Hidden Markov Model, Multivariate Asymmetric Laplace Distribution, Quantile Regression, Random Effects Model

1 Introduction

Ever since quantile regression was first introduced in the seminal work of [6], it has attracted researches' and practitioners' attention. It provides a way to model the conditional quantiles of a response variable with respect to a set of covariates in order to have a more complete picture of the entire conditional distribution compared to the classical mean regression. In a univariate quantile regression analysis, the likelihood based inferential approach to estimate the parameters relies on the introduction of the Asymmetric Laplace (AL) distribution: the maximization of the likelihood associated with the AL density is equivalent (in terms of parameter estimates) to the minimization of the quantile loss function of [6]. When multivariate response variables are concerned, the existing literature on quantile regression is less extensive due to the fact that there is not a unique definition of quantile for a multivariate random variable because there is no "natural" ordering in a p-dimensional space, for $p > 1$. Hence, the concept of multivariate quantile is still a debatable issue (see [7] and the references therein for relevant studies). Recently, [12] generalized the AL distribution inferential approach of the univariate case to a multivariate framework by using the Multivariate Asymmetric Laplace (MAL) distribution defined in [8]. By using the MAL distribution as likelihood based inferential tool, the authors sidestep the problem of defining a multivariate quantile, and meanwhile they implement a joint estimation of the marginal conditional quantiles of a multivariate response variable, taking into account for possible correlation among marginals.

When dealing with longitudinal data, because measurements recorded on the same individuals are likely correlated, the potential association between dependent observations should be taken into account in order to provide correct inferences. In such cases, random effect models have been proposed to accommodate for time-constant, within-subject correlation and between subject heterogeneity (see [10, 5]). However, when the assumption of time-constant random coefficients does not hold, adopting such model specification may lead to biased parameter estimates (see [3]). To account for serial heterogeneity, [4] suggested the use of Hidden Markov Models (HMM). The key assumption is the conditional independence of the response variables given a latent process that follows a Markov chain on a finite number of states. In this context, the application of HMMs is well justified by their versatility and mathematical tractability (see [11]).

The purpose of this article is to extend the work of [12] by introducing a Mixed Hidden Markov Model (MHMM) to the longitudinal data setting to account for the correlation between responses. The MHMM (see [2]) encompasses Generalized Linear Mixed Models and HMMs as it accommodates time-constant and time-varying sources of random variation. Time-constant unobserved heterogeneity is described via individual-specific random coefficients while temporal effects are captured through state-specific effects that evolve over time depending on a hidden Markov chain. In order to prevent inconsistent parameter estimates due to misspecification of the random effects distribution, we adopt the Non-Parametric Maximum Likelihood (NPML) approach of [9] in which it is left unspecified and approximated by a discrete finite mixture distribution. Model parameters are estimated through maximum likelihood by using the Expectation-Maximization (EM) algorithm while standard error estimates rely on bootstrap resampling. From a computational perspective, we provide an efficient version of the EM algorithm with M-step updates in closed form for all model parameters.

2 Methodology

Let $\mathbf{Y}_{it} = (Y_{it}^{(1)}, \ldots, Y_{it}^{(p)})$ be a continuous p-variate response variable vector and $\mathbf{X}_{it} = (X_{it}^{(1)}, \ldots, X_{it}^{(k)})$ be a k-dimensional vector of explanatory variables for every subject $i = 1, \ldots, N$ and time occasion $t = 1, \ldots, T_i$. Given p quantile indexes $\tau = (\tau_1, \ldots, \tau_p)$, with $\tau_j \in (0,1)$, $j = 1, \ldots, p$, let $S_{it}(\tau)$, $i = 1, \ldots, N$, $t = 1, \ldots, T_i$ be a homogeneous, first-order, hidden Markov chain defined over a discrete states space $\mathscr{S} = \{1, \ldots, M\}$ with initial and transition probabilities denoted by $\mathbf{q} = (q_1, \ldots, q_M)$ and $\mathbf{Q} = \{q_{jk}\}$ over $\mathscr{S} \times \mathscr{S}$, respectively. Finally, let $\mathbf{b}_i(\tau)$ be a time-constant, subject-specific, random effects matrix having distribution $f_{\mathbf{b}}(\cdot \mid \tau)$ which, as they are meant to capture different unobserved characteristics, is independent of the hidden Markov chain, $S_{it}(\tau)$, and where $\mathbb{E}(\mathbf{b}_i(\tau)) = 0$ is used for parameter identifiability. We assume that the τ_j-th quantile of each of the j-th components of $\mathbf{Y}_{it}$ can be modeled as a function of some explanatory variables. Let $\beta(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))$ be the $k \times p$ matrix of unknown regression coefficients. Then, the multivariate Mixed Hidden Markov Model (MHMM) is defined as follows:

$$\mathbf{Y}_{it} = \mathbf{X}_{it}\beta(\tau) + \mathbf{Z}_{it}\mathbf{b}_i(\tau) + \mathbf{W}_{it}\alpha_{s_{it}}(\tau) + \varepsilon_{s_{it}}(\tau), \tag{1}$$

where $\mathbf{Z}_{it}$ is a subset of $\mathbf{X}_{it}$, $\mathbf{W}_{it}$ is a further subset of $\mathbf{X}_{it}$ whose effects are assumed to vary over time, $\varepsilon_{s_{it}}(\tau)$ denotes a p-dimensional vector of error terms with univariate component-wise quantiles (at fixed levels $\tau_1, .., \tau_p$, respectively) equal to zero and where the coefficients matrix $\alpha_{s_{it}}(\tau)$ evolves over time according to the hidden Markov chain, $S_{it}(\tau)$, and takes one of the values in the set $\{\alpha_1(\tau), \ldots, \alpha_M(\tau)\}$.

Our objective is to provide joint estimation of the p marginal conditional quantiles of $\mathbf{Y}_{it}$ taking into account for potential correlation among the dependent variables. Conditional on the hidden state occupied at time t, $S_{it}(\tau)$, and on the

individual-specific random coefficients, $\mathbf{b}_i(\tau)$, observations from the same individual are independent and the following equality holds:

$$f_{\mathbf{Y}|S,\mathbf{b}}(\mathbf{y}_{it} \mid \mathbf{y}_{i1:t-1}, s_{i1:t}, \mathbf{b}_i, \tau) = f_{\mathbf{Y}|S,\mathbf{b}}(\mathbf{y}_{it} \mid s_{it}, \mathbf{b}_i, \tau), \tag{2}$$

where $\mathbf{y}_{i1:t-1}$ represents the history of the responses for the i-th subject up to time $t-1$ and $s_{i1:t}$ is the individual sequence of states up to time t.

In order to derive maximum likelihood estimates for the regression model in (1), we consider the Multivariate Asymmetric Laplace (MAL) distribution, $\mathscr{MAL} \sim (\mu_{it}, \mathbf{D}\tilde{\xi}, \mathbf{D}\Sigma\mathbf{D})$ (see [8]), whose conditional density function is given by:

$$f_{\mathbf{Y}|S,\mathbf{b}}(\mathbf{y}_{it} \mid s_{it}, \mathbf{b}_i, \tau) = \frac{2\exp\left\{(\mathbf{y}_{it}-\mu_{it})'\mathbf{D}^{-1}\Sigma^{-1}\tilde{\xi}\right\}}{(2\pi)^{p/2}|\mathbf{D}\Sigma\mathbf{D}|^{1/2}} \left(\frac{\tilde{m}_{it}}{2+\tilde{d}}\right)^{\nu/2} K_{\nu}\left(\sqrt{(2+\tilde{d})\tilde{m}_{it}}\right), \tag{3}$$

where the location parameter μ_{it} is defined by the linear model $\mu_{it} = \mu(s_{it}, \mathbf{b}_i, \tau) = \mathbf{X}_{it}\beta(\tau) + \mathbf{Z}_{it}\mathbf{b}_i(\tau) + \mathbf{W}_{it}\alpha_{s_{it}}(\tau)$, $\mathbf{D}\tilde{\xi} \in \mathbb{R}^p$ is the scale (or skew) parameter with $\mathbf{D} = \text{diag}[d_1, ..., d_p]$, $d_j > 0$ and $\tilde{\xi} = [\tilde{\xi}_1, \tilde{\xi}_2, ..., \tilde{\xi}_p]'$ having generic element $\tilde{\xi}_j = \frac{1-2\tau_j}{\tau_j(1-\tau_j)}$, $j = 1, ..., p$. $\tilde{\Sigma}$ is a $p \times p$ positive definite matrix such that $\Sigma = \Lambda\Psi\Lambda$, with Ψ being a correlation matrix and $\Lambda = \text{diag}[\sigma_1, ..., \sigma_p]$, with $\sigma_j^2 = \frac{2}{\tau_j(1-\tau_j)}$, $j = 1, ..., p$. Moreover, $\tilde{m}_{it} = (\mathbf{y}_{it} - \mu_{it})'(\mathbf{D}\Sigma\mathbf{D})^{-1}(\mathbf{y}_{it} - \mu_{it})$, $\tilde{d} = \tilde{\xi}'\Sigma\tilde{\xi}$, and $K_{\nu}(\cdot)$ denotes the modified Bessel function of the third kind with index parameter $\nu = (2-p)/2$. The constraints imposed on $\tilde{\xi}$ and Λ represent necessary conditions for model identifiability for any fixed quantile level $\tau_1, ..., \tau_p$ and guarantee that $\mu_{it}^{(j)}$ is the τ_j-th conditional quantile function of $Y_{it}^{(j)}$ given $S_{it}(\tau)$ and $\mathbf{b}_i$, for $j = 1, \ldots, p$.

It is worth noting that our methodology reduces to the (multivariate) linear quantile Hidden Markov Model of [4] when $\mathbf{W}_{it} = \mathbf{1}$ and $\mathbf{b}_i(\tau) = \mathbf{0}$ for all $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$; whereas it reduces to the (multivariate) linear quantile Mixed Model of [10] when there is only one state of the hidden Markov chain, i.e. $M = 1$.

In the case of a continuous parametric distribution for the random effects, the likelihood for the model in (1)-(3) involves the integration over the distribution of the random effects, $f_{\mathbf{b}}(\cdot \mid \tau)$. Such integral cannot be solved analytically and maximum likelihood parameter estimates can be obtained through numerical integration techniques. In addition parametric assumptions on the distribution of the random coefficients can be too restrictive and misspecification of the mixing distribution can lead to biased parameter estimates. For these reasons, we may rely on the Nonparametric Maximum Likelihood (NPML) estimation theory of [9]: $f_{\mathbf{b}}(\cdot \mid \tau)$ is left unspecified and we approximate it by using a discrete distribution on $G < N$ locations, $\mathbf{b}_g(\tau)$, with associated probabilities defined by $\pi_g(\tau) = \Pr(\mathbf{b}_i(\tau) = \mathbf{b}_g(\tau))$, $i = 1, ..., N$ and $g = 1, ..., G$. That is, $\mathbf{b}_i(\tau) \sim \sum_{g=1}^{G} \pi_g(\tau)\delta_{\mathbf{b}_g}(\tau)$ where δ_{θ} is a one-point distribution putting a unit mass at θ. In this case, if we suppress the index τ to simplify the notation, the observed data likelihood of the model has the form:

$$L(\Phi_\tau) = \prod_{i=1}^{N} \prod_{g=1}^{G} \sum_{\mathscr{S}^{T_i}} \left\{ \left[\prod_{t=1}^{T_i} f_{\mathbf{Y}|S,\mathbf{b}}(\mathbf{y}_{it} \mid s_{it}, \mathbf{b}_g) \right] q_{s_{i1}} \prod_{t=2}^{T_i} q_{s_{it-1}s_{it}} \right\} \pi_g, \tag{4}$$

where $\Phi_\tau = (\beta, \mathbf{D}, \Psi, \mathbf{b}_1, \ldots, \mathbf{b}_G, \pi_1, \ldots, \pi_G, \alpha_1, \ldots, \alpha_M, \mathbf{q}, \mathbf{Q})$ denotes the vector of model parameters and $f_{\mathbf{Y}|S,\mathbf{b}}(\mathbf{y}_{it} \mid s_{it}, \mathbf{b}_g)$ represents the response distribution of unit i of being in the hidden state s_{it} at time t and of belonging to the g-th component of the finite mixture, which is assumed to have a MAL density in (3) with location parameter given by $\mu_{it} = \mu(s_{it}, \mathbf{b}_g, \tau) = \mathbf{X}_{it}\beta(\tau) + \mathbf{Z}_{it}\mathbf{b}_g(\tau) + \mathbf{W}_{it}\alpha_{s_{it}}(\tau)$.

2.1 Estimation

Given the representation in (4), let us denote by w_{ig} the indicator variable that is equal to 1 if the i-th unit belongs to the g-th component of the finite mixture, and 0 otherwise. Similarly let $u_{it\,j}$ be equal to 1 if unit i is in state j at time t and 0 otherwise; let $v_{it\,jk}$ be equal to 1 if unit i is in state j at time $t-1$ and in state k at time t, and 0 otherwise. Finally, we denote by $z_{it\,jg}$ the indicator of the i-th individual being in state j at time t and coming from the g-th component of the mixture. The log-likelihood for the complete data has the following form:

$$\ell_c(\Phi_\tau) = \sum_{i=1}^{N} \left\{ \sum_{g=1}^{G} w_{ig} \log \pi_g + \sum_{j=1}^{M} u_{i1\,j} \log q_j + \sum_{t=2}^{T_i} \sum_{j=1}^{M} \sum_{k=1}^{M} v_{it\,jk} \log q_{jk} + \sum_{t=1}^{T_i} \sum_{j=1}^{M} \sum_{g=1}^{G} z_{it\,jg} \log f_{\mathbf{Y}|S,\mathbf{b}}(\mathbf{y}_{it} \mid S_{it} = j, \mathbf{b}_g) \right\}. \tag{5}$$

In the E-step of the algorithm, the presence of the unobserved indicator variables $w_{ig}, u_{it\,j}, v_{it\,jk}$ and $z_{it\,jg}$ is handled by taking their conditional expectation given the observed data and the current parameter estimates. Calculation of such quantities may be addressed via an adaptation of the forward and backward variables; see [14]. Subsequently, the M-step solutions are updated by maximizing the conditional expectation of (5) given the observed data and the current parameter estimates with respect to Φ_τ and solving the M-step equations. We derive closed form update expressions of the model parameters, based on the mixture representation of the MAL distribution. Finally, the E- and M-steps are alternated until convergence. To avoid convergence to local maxima, for each value of the pair (G, M), we initialize model parameters using a multi-start strategy.

3 Application

To investigate the behaviour of the proposed methodology, we analyse the data from the Millennium Cohort Study (MCS) which has been studied by [1] and [13] in the context of M-quantile regression with time-constant random-effects. The MCS is a longitudinal survey which aims at better addressing the effects of social disadvan-

tage on children's outcomes in the UK. The two outcomes of interest, emotional problems and behavioural problems, measured by the SDQ internalizing score and by the SDQ externalizing score, respectively, were collected at ages 3, 5 and 7 years. A description of the included demographic and socio-economic covariates can be found in [13]. We used the proposed model with constant random intercepts and time-varying random slopes specified for age. A summary of the results when $\tau = (0.25, 0.25), \tau = (0.50, 0.50)$ and $\tau = (0.75, 0.75)$ is reported in Table 1. The estimated regression coefficients are consistent with those obtained by [13]. In particular, we selected a number of mixture components equal to $G = (3,5,5)$ at quantile levels $0.25, 0.50$ and 0.75 respectively, and we identified a decreasing number of hidden states $M = (5,4,3)$ as the analyzed quantile level increases.

τ-th quantile	(0.25, 0.25) [G = 3, M = 5]		(0.50, 0.50) [G = 5, M = 4]		(0.75, 0.75) [G = 5, M = 3]	
Variable	SDQ_{Int}	SDQ_{Ext}	SDQ_{Int}	SDQ_{Ext}	SDQ_{Int}	SDQ_{Ext}
Intercept	**0.864** (0.126)	**1.846** (0.178)	**1.267** (0.215)	**3.925** (0.251)	**3.790** (0.288)	**4.275** (0.412)
Age year scal	**1.379** (0.010)	**−0.804** (0.016)	**1.234** (0.009)	**−0.634** (0.011)	**−0.583** (0.021)	**−0.244** (0.031)
Age2 year scal	**0.045** (0.005)	**0.193** (0.009)	**0.077** (0.011)	**0.208** (0.013)	**0.104** (0.018)	**0.287** (0.024)
ALE 11	**0.022** (0.008)	**0.036** (0.016)	**0.086** (0.018)	**0.113** (0.019)	**0.116** (0.039)	**0.205** (0.055)
SED 4	**0.070** (0.023)	**0.105** (0.045)	**0.175** (0.030)	**0.221** (0.030)	**0.201** (0.051)	**0.398** (0.076)
Kessm	**0.090** (0.009)	**0.143** (0.012)	**0.167** (0.009)	**0.189** (0.012)	**0.208** (0.018)	**0.299** (0.025)
Degree	**−0.350** (0.109)	**−0.894** (0.149)	**−0.526** (0.114)	**−1.482** (0.171)	**−0.703** (0.160)	**−1.267** (0.232)
GCSE	**−0.217** (0.110)	**−0.582** (0.150)	**−0.352** (0.113)	**−0.430** (0.166)	**−0.413** (0.149)	**−0.427** (0.213)
White	−0.090 (0.061)	−0.143 (0.097)	−0.075 (0.149)	0.059 (0.160)	−0.216 (0.203)	0.320 (0.303)
Male	**−0.062** (0.016)	**0.793** (0.030)	0.027 (0.037)	**0.950** (0.045)	0.082 (0.080)	**0.944** (0.119)
IMDscore	**−0.022** (0.005)	**−0.036** (0.009)	**−0.025** (0.008)	**−0.027** (0.010)	−0.027 (0.020)	−0.045 (0.029)
Eng eth stratum	−0.043 (0.159)	−0.168 (0.174)	0.122 (0.193)	0.144 (0.220)	0.174 (0.241)	−0.025 (0.374)
Eng dis stratum	−0.003 (0.031)	0.003 (0.072)	0.085 (0.047)	**0.106** (0.050)	0.156 (0.111)	0.337 (0.175)

Table 1 Point estimates with non-parametric bootstrap standard errors in parentheses ($B = 1000$ re-samples) for different quantile levels. Parameter estimates are displayed in boldface when significant at the standard 5% level. The number of mixture components G and hidden states M are selected according to the BIC criteria.

References

[1] Alfo', M., Marino, M. F., Ranalli, M. G., Salvati, N. and Tzavidis, N. [2016], 'M-quantile regression for multivariate longitudinal data: analysis of the Millennium Cohort Study data', *ArXiv e-prints* .

[2] Altman, R. M. [2007], 'Mixed Hidden Markov models: an extension of the Hidden Markov model to the longitudinal data setting', *Journal of the American Statistical Association* **102**(477), 201–210.

[3] Bartolucci, F. and Farcomeni, A. [2009], 'A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure', *Journal of the American Statistical Association* **104**(486), 816–831.

[4] Farcomeni, A. [2012], 'Quantile regression for longitudinal data based on latent Markov subject-specific parameters', *Statistics and Computing* **22**(1), 141–152.

[5] Geraci, M. and Bottai, M. [2014], 'Linear quantile mixed models', *Statistics and Computing* **24**(3), 461–479.

[6] Koenker, R. and Bassett, G. [1978], 'Regression Quantiles', *Econometrica: Journal of the Econometric Society* **46**(1), 33–50.

[7] Koenker, R., Chernozhukov, V., He, X. and Peng, L. [2017], *Handbook of Quantile Regression*, CRC press.

[8] Kotz, S., Kozubowski, T. and Podgorski, K. [2012], *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Springer Science & Business Media.

[9] Lindsay, B. G. et al. [1983], 'The geometry of mixture likelihoods: a general theory', *The Annals of Statistics* **11**(1), 86–94.

[10] Liu, Y. and Bottai, M. [2009], 'Mixed-effects models for conditional quantiles with longitudinal data', *The International Journal of Biostatistics* **5**(1).

[11] Maruotti, A. [2011], 'Mixed Hidden Markov models for longitudinal data: an overview', *International Statistical Review* **79**(3), 427–454.

[12] Petrella, L. and Raponi, V. [2019], 'Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress', *Journal of Multivariate Analysis* **173**, 70–84.

[13] Tzavidis, N., Salvati, N., Schmid, T., Flouri, E. and Midouhas, E. [2016], 'Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **179**, 427–452.

[14] Welch, L. R. [2003], 'Hidden Markov Models and the Baum-Welch algorithm', *IEEE Information Theory Society Newsletter* **53**(4), 10–13.

Recent methodological advances in finite mixture modeling with applications (CLADAG-SIS)

Aggregating Gaussian mixture components
Come aggregare le componenti gaussiane di un miscuglio

Roberto Rocci[1]

Abstract The finite mixture of Gaussians is a well-known model frequently used to classify a sample of observations. It considers the sample as drawn from a heterogeneous population where each subpopulation, cluster, is Gaussian and corresponds to one component of the mixture. Whenever such assumption is false, the model may use two or more Gaussians to describe a single cluster. In this case, the researcher has the problem of how to identify the clusters starting from the estimated components. This work proposes to solve this problem by aggregating the components in clusters by optimizing an appropriate criterion based on their posterior probabilities.

Abstract *I modelli miscuglio di gaussiane sono spesso utilizzati nell'analisi dei gruppi. L'idea è quella di considerare la popolazione che ha generato il campione come formata da sottopopolazioni, gruppi, ognuna ben descritta da una componente del miscuglio. Quando questa assunzione risulta falsa, il modello tende ad utilizzare due o più componenti per rappresentare un unico gruppo, creando così il problema di come identificare i gruppi a partire dalle componenti stimate. In questo lavoro proponiamo di risolvere il problema aggregando le componenti in modo da ottimizzare un criterio basato sulle loro probabilità a posteriori.*

Key words: Unsupervised Classification, Finite mixtures of Gaussians, Within and Between deviances.

1 Introduction

In cluster analysis, or unsupervised classification, quite frequently observations are classified by using a finite mixture of Gaussians (see for example Hennig et al., 2015). The idea is to consider the population as heterogeneous, i.e. formed by sub-populations, clusters, which are well represented by the mixture components.

[1]Department of Statistical Science, Sapienza University of Rome; email: roberto.rocci@uniroma1.it

Such approach works well in practice unless one or more subpopulations have a distribution different from, or not well approximated by, a single component of the mixture. In this case, the model may use more than one component, i.e. a sub-mixture of two or more Gaussians, to describe a single cluster destroying the one to one correspondence between clusters and components. This problem is well recognized in practice and several solutions have been proposed.

The first idea is to assume for each component a functional form that is more flexible than the Gaussian (see McNicholas, 2016, for an excellent review about this approach). This solves the problem in many cases in practice. However, it cannot be considered as the definitive solution because even in this case we cannot exclude that more than one component could be necessary to represent a cluster. This derives from the identifiability of the finite mixture model. For example, a cluster that is a finite mixture of two components cannot be represented by only one and vice versa.

A different idea comes up by observing that, if the mixture fits well the data, then the information about the clustering structure is contained in the estimated model and it can be recovered by aggregating, in an opportune way, the components. This allow us to represent a very wide variety of possible distributions for the clusters and to relax the, usually made and sometimes restrictive, assumption of same functional form for the distribution of each cluster. Technically, the model would become a finite mixture of finite mixtures of Gaussians, i.e. a finite mixture where each component is a finite mixture of Gaussians. Unfortunately, such a model is not identified because different aggregations of the components give the same population distribution. The estimation is then possible only by using some constraints on model parameters making the model identified (see for example Di Zio et al. 2007), or by introducing a criterion determining the aggregation. On the latter approach, there are several proposals in the literature where the Gaussian components are hierarchically aggregated into clusters on the basis of a measure of proximity (see Hennig 2010, Comas-Cufi et al. 2017 and references there in for some examples). However, such methods are optimal only locally. They establish what is the best way to merge two components not what is the best way to aggregate, say G, components into, say K, clusters. It is not specified how the "internal cohesion" and "external isolation" (Cormack, 1971) are measured, related and optimized (e.g. total deviance = within deviance + between deviance in K-means). To achieve this goal a partitioning method should be adopted but, as far as we know, only Li (2005) considered a partitioning approach based on the application of the K-means on the mean components. This proposal makes clear the aforementioned aspects but it is based on a measure of dissimilarity between components depending only on their locations. Our purpose is to go beyond the limits of this proposal.

In our paper we are going to propose a new method to aggregate the Gaussian components of a finite mixture model making clear how the identified partition optimize the "internal cohesion" and "external isolation" of the clustering. The plan of the paper is the following. Our proposal, based on the use of the Kullback Leibler divergence to measure the dissimilarity among components, will be presented in section 2. In section 3, some insights on how to extend the technique are presented with particular reference to other dissimilarity measures and its hierarchical version.

2 Partitioning Gaussian components by the Kullback-Liebler divergence: the KL-components method

The finite mixture of Gaussians (McLachlan & Peel, 2000) is based on the assumption that the probability density of a multivariate observation is of the form

$$f(\mathbf{x}_i;\Theta) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i;\vartheta_g)\,. \tag{1}$$

where $\mathbf{x}_i = [x_{i1}, x_{i2},\ldots,x_{iJ}]'$ is a random vector of J variables sampled from a population parametrized by $\Theta = \{\vartheta_1,\vartheta_2,\ldots,\vartheta_G, p_1,p_2,\ldots,p_G\}$, which consists of G groups, or subpopulations, in proportions $\pi_1, \pi_2,\ldots, \pi_G$, where π_g is the prior probability to sample one observation from group g. The density $f_g(\mathbf{x}_i;\vartheta_g)$ of $\mathbf{x}_i$ in the g^{th} group is multivariate normal (Gaussian). Let $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$ be a sample of n independent and identically distributed observations, we can use the above model to classify the observations into G classes. First, we compute the posterior probabilities

$$\mathrm{P}(g \mid \mathbf{x}_i) = \pi_{g|i} = \pi_g f_g(\mathbf{x}_i;\vartheta_g) \Big/ \sum_{h=1}^{G} \pi_h f_h(\mathbf{x}_i;\vartheta_h)\,, \tag{2}$$

then, we use the MAP rule (Maximum A Posterior probability) to assign the observations to the Gaussian components. Usually, the parameter Θ and the number of components G are unknown and estimated from the data.

Our method, named the KL-components technique, originates from the observation that two components, say g and h, are equal, with respect to the data, if and only if their posterior probabilities are proportional. In formulas

$$g = h \Leftrightarrow \boldsymbol{\pi}_g = \mathrm{P}(g \mid \mathbf{x}_i) / \pi_g = \mathrm{P}(h \mid \mathbf{x}_i) / \pi_h = \boldsymbol{\pi}_h\,,\ g,h = 1,\ldots,G \text{ and } i = 1,\ldots,n. \tag{3}$$

It seems quite natural to measure the dissimilarity between components as a function of the diversity between the normalized posteriors, say profiles. In particular, we investigate the use of the Kullback-Leibler (KL) divergence

$$\mathrm{KL}(\boldsymbol{\pi}_g, \boldsymbol{\pi}_h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_{g|i}}{\pi_g} \log\left(\frac{\pi_{g|i}}{\pi_g}\frac{\pi_h}{\pi_{h|i}}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_g(\mathbf{x}_i)}{f(\mathbf{x}_i)} \log\left(\frac{f_g(\mathbf{x}_i)}{f_h(\mathbf{x}_i)}\right), \tag{4}$$

which can be considered an estimate of

$$\mathrm{E}\left(\frac{f_g(\mathbf{x})}{f(\mathbf{x})} \log\left(\frac{f_g(\mathbf{x})}{f_h(\mathbf{x})}\right)\right) = \int f_g(\mathbf{x}) \log\left(\frac{f_g(\mathbf{x})}{f_h(\mathbf{x})}\right) d\mathbf{x}\,, \tag{5}$$

i.e. the KL divergence between the two components. It is well known, and evident from formula (5), that the KL divergence is not symmetric. However, this is not a problem for us because we use it to define a sort of deviance rather than to measure the dissimilarity between two components. In particular, we define the Total deviance as the weighted sum of the KL divergences among the profiles of the mixture components and their barycenter. By noting that the latter quantity is

$$\sum_{g=1}^{G} \pi_g \frac{\pi_{g|i}}{\pi_g} = 1, \quad i = 1, 2, ..., n, \tag{6}$$

the Total deviance results to be

$$\mathrm{D}_T = \sum_{g=1}^{G} \pi_g \mathrm{KL}(\boldsymbol{\pi}_g, \mathbf{1}) = \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n} \pi_{g|i} \log\left(\frac{\pi_{g|i}}{\pi_g}\right). \tag{7}$$

Given a partition of the G components into K clusters according to the binary row stochastic membership matrix $\mathbf{U} = [u_{gk}]$, where u_{gk} is equal to 1 if component g belong to cluster k and 0 otherwise, we note that the posteriors probabilities of cluster k are

$$p_{k|i} = \sum_{g=1}^{G} u_{gk} \pi_{g|i}, \tag{8}$$

the priors are

$$p_k = \sum_{g=1}^{G} u_{gk} \pi_g = \sum_{g=1}^{G} u_{gk} \frac{1}{n} \sum_{i=1}^{n} \pi_{g|i} = \frac{1}{n} \sum_{i=1}^{n} \sum_{g=1}^{G} u_{gk} \pi_{g|i} = \frac{1}{n} \sum_{i=1}^{n} p_{k|i}, \tag{9}$$

and the barycenter is

$$\frac{1}{\sum_{g=1}^{G} u_{gk} \pi_g} \sum_{g=1}^{G} u_{gk} \pi_g \frac{\pi_{g|i}}{\pi_g} = \frac{1}{p_k} \sum_{g=1}^{G} u_{gk} \pi_{g|i} = \frac{p_{k|i}}{p_k}, \quad i = 1, 2, ..., n. \tag{10}$$

According to (7), the Within deviance is defined as

$$\mathrm{D}_W(\mathbf{U}) = \sum_{k,g} u_{gk} \pi_g \mathrm{KL}(\boldsymbol{\pi}_g, \mathbf{p}_k) = \sum_{k,g} u_{gk} \pi_g \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_{g|i}}{\pi_g} \left[\log\left(\frac{\pi_{g|i}}{\pi_g}\right) - \log\left(\frac{p_{k|i}}{p_k}\right) \right]. \tag{11}$$

Formula (11) suggests in a very natural way a partitioning method based on its minimization with respect to **U**. The minimization of (11) guarantees the maximum internal cohesion of the clusters. However, we should ask: what about the external

isolation? To give an answer to this question, we should find a way to measure the external isolation, i.e. the cluster separation. In coherence with (7), we can define the Between deviance as the weighted sum of the KL divergences among the cluster profiles and their barycenter, that is the vector of ones even in this case. In formulas

$$\mathrm{D}_B(\mathbf{U}) = \sum_{k=1}^{K} p_k \mathrm{KL}(\mathbf{p}_k, \mathbf{1}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} p_{k|i} \log\left(\frac{p_{k|i}}{p_k}\right). \tag{12}$$

Once again, we have a formula suggesting, in a very natural way, a partitioning method corresponding to the maximization of (12) with respect to **U**. However, it is possible to show that, as in the case of *K*-means, the minimization of D_W is equivalent to the maximization of D_B because the sum of the two is constant and equal to D_T. The Within deviance (11) can be minimized by using a coordinate descent algorithm that minimize (11) with respect to **U** and the centroids $\mathbf{p}_1$, $\mathbf{p}_2$, …, $\mathbf{p}_K$.

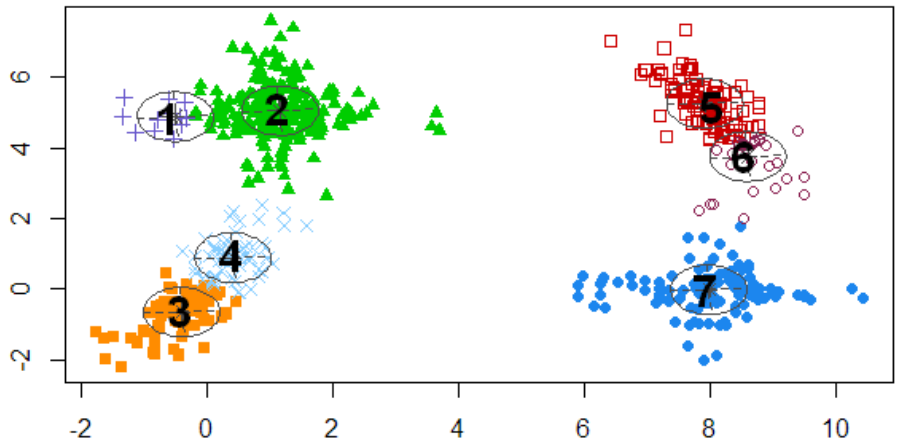


Figure 1: Simulated sample from a 6-component mixture along with the estimated classification in 7 components given by a homoscedastic mixture of Gaussians.

In order to check if the proposed method works properly, we considered the simulated dataset analysed in section 4.1 of Baudry et al. (2010). It is a sample from a finite mixture of 6 bivariate Gaussians. In Figure 1, the data is shown along with the classification in 7 components given by a homoscedastic mixture of Gaussians where the number of components have been selected by using BIC and the parameters estimated by maximum likelihood. Looking at the figure, it is not clear if the number of true clusters is 2 or 4 and then we considered both. The algorithm has been run for $K = 2$ and 4, from several different starting points. The technique aggregated the 7 components as {1,2,3,4} and {5,6,7} for $K = 2$ and as {1,2}, {3,4}, {5,6} and {7} for $K = 4$. From Figure 1, it is clear that in both cases, KL-components has been able to find the correct aggregation of the components.

3 Final comments and extensions

The method here presented has been extended to the use of divergences different from the KL. In particular, we have proven that all the properties of the KL-components technique shown in the previous section do hold if a Bregman divergence (Bregman, 1967) is used. The proofs are not reported here for the sake of space.

In practical applications, especially when the clusters of components are not well separated, the coordinate descent algorithm quite often remains trapped into a local optimum. Frequently, this problem can be simply solved by starting several times the algorithm from different random partitions. However, there is still the need to have a method able to produce good rational, non random, starting points. To this end, a hierarchical clustering procedure has been proposed, not shown here for the sake of space, where at each step two clusters are merged by minimizing the increment of Within deviance. The hierarchical solution is then used to start the partitioning algorithm.

The aim of our method is not to find the true number of clusters, even if it can help us in this task exploring the possible components aggregations. In this respect, further insights can be obtained by looking at the plot of D_W vs K, computing the Calinski-Harabasz index (1974) or any other index based on D_W and/or D_B.

We conclude by noting that the results presented here do not use the assumption that the components of the mixture are normally distributed. It follows that the proposed techniques can be used to cluster components that are not Gaussians.

References

1. Baudry, J.P., Raftery, A., Celeux, A., Lo, K., Gottardo, R.: Combining mixture components for clustering, *Journal of Computational and Graphical Statistics*, 19, 332-353 (2010).
2. Bregman, L. M.: The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* 7: 200–217 (1967)
3. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis.Com-munications in Statistics, 3, no. 1:1–27 (1974).
4. Comas-Cufi M., Martin-Fernandez J.A., Mateu-Figueras G.: Merging the components of a finite mixture using posterior probabilities, *Statistical Modelling*, 19(2), 1–31 (2017)
5. Cormack, R. M.: A review of classification (with discussion). *Journal of the Royal Statistical Society*, A, 134, 321-67 (1971)
6. Di Zio M., Guarnera U., Rocci R.: A mixture of mixture models for a classification problem: the unity measure error. *Computational Statistics and Data Analysis*, 51, 5, 2573-2585 (2007)
7. Hennig, C.: Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification* 4, 3–34 (2010)
8. Hennig, C., Meila, M., Murtagh, F., Rocci R.: *Handbook of Cluster Analysis*, Chapman and Hall/CRC (2015)
9. Li J.: Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 14:547–568 (2004)
10. McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, New York (2000)
11. McNicholas, P.D.: *Mixture Model-Based Classification*, Boca Raton FL: Chapman & Hall/CRC Press (2016)

Local and overall coefficients of determination for mixtures of generalized linear models

Coefficienti di determinazione locali e globali per misture di regressioni lineari

Roberto Di Mari, Salvatore Ingrassia, and Antonio Punzo

Abstract In this paper we define deviance-based local and overall coefficients of determination for mixtures of generalized linear models whose parameters are estimated via the maximum likelihood approach. The aim is to summarize how well the model fits the data both in each group and taken as a whole.

Abstract *In questo lavoro definiamo coefficienti di determinazione locali e globali, basati sulla devianza, per le misture di modelli lineari generalizzati i cui parametri sono stimati attraverso l'approccio della massima verosimiglianza. L'obiettivo è quello di valutare l'adattamento del modello ai dati sia a livello di singolo gruppo che a livello globale.*

Key words: Cluster validation, Mixtures of regressions, Model-based clustering, Maximum likelihood.

1 Introduction

Local and overall coefficients of determination have been first proposed in [3] for mixtures of linear (Gaussian) regressions to take into account different variability within groups. In this paper, we generalize these results to mixtures of generalized linear models. The proposal is based on one of the definitions of coefficient of determination given in [1].

In generalized linear models [5] a monotone and differentiable link function $h(\cdot)$ is introduced to relate the expected value μ of a random response variable Y with respect to a vector $\mathbf{X}$ of J covariates through the relation $h(\mu) = \beta_0 + \beta_1'\mathbf{x}$. Assume now that the regression of Y on $\mathbf{X}$ varies across the k levels (groups or clusters) of

Roberto Di Mari · Salvatore Ingrassia · Antonio Punzo
Department of Economics and Business, University of Catania
Corso Italia 55, 95129 Catania, Italy, e-mail: roberto.dimari@unict.it, s.ingrassia@unict.it, antonio.punzo@unict.it

a categorical latent variable G taking values in $\{1,2,\ldots,k\}$. These data can be conveniently modeled by finite mixtures mixtures of regressions with fixed covariates characterized by the following conditional density function

$$p(y|\mathbf{x};\psi) = \sum_{g=1}^{k} \pi_g q(y|\mathbf{x};\theta_g), \tag{1}$$

where $q(y|\mathbf{x};\theta_g)$ is a component specific density from the exponential family depending on some parameter vector θ_g, with π_g being the mixing weight, where $\pi_g > 0$ and $\sum_{g=1}^{k} \pi_g = 1$, see e.g. [7]. Given a random sample $(\mathbf{x}_1', y_1)', \ldots, (\mathbf{x}_n', y_n)'$ of $(\mathbf{X}', Y)'$, for a fixed number k of groups, the ML estimates of the parameters are typically obtained via the expectation-maximization (EM) algorithm.

2 A deviance-based coefficient of determination R^2

In [1] some R-squared measures, say R^2, are defined for count data regression models based on the scaled deviance $D(\mathbf{y};\widehat{\mu}) = 2\{\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\widehat{\mu};\mathbf{y})\}$, where $\mathscr{L}(\mathbf{y};\mathbf{y})$ and $\mathscr{L}(\widehat{\mu};\mathbf{y})$ are the log-likelihoods computed on the data $\mathbf{y}$ and on the fitted data $\widehat{\mu}$, respectively. Then, considering the decomposition $\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\bar{\mathbf{y}};\mathbf{y}) = \{\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\widehat{\mu};\mathbf{y})\} + \{\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\bar{\mathbf{y}};\mathbf{y})\}$, the coefficient of determination has been defined as

$$R^2 = 1 - \frac{2\{\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\widehat{\mu};\mathbf{y})\}}{2\{\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\bar{\mathbf{y}};\mathbf{y})\}} = \frac{\mathscr{L}(\widehat{\mu};\mathbf{y}) - \mathscr{L}(\bar{\mathbf{y}};\mathbf{y})}{\mathscr{L}(\mathbf{y};\mathbf{y}) - \mathscr{L}(\bar{\mathbf{y}};\mathbf{y})}. \tag{2}$$

In this paper, we extend this measure (2) to mixture of generalized regression models. For simplicity, only the Gaussian and the Poisson cases are here presented.

The Gaussian case. Consider first mixture models with Gaussian components. Thus, conditional on $G = g$ (for $g = 1,\ldots,k$) assume now that $Y|\mathbf{x}$ is Gaussian with parameters $\mu(\mathbf{x};\beta_g)$ and σ_g^2:

$$q(y|\mathbf{x};\beta_g,\sigma_g^2) = \phi(y;\mu(\mathbf{x};\beta_g),\sigma_g^2)$$

where $\beta_g = (\beta_{0g}, \beta_{1g}')'$ is a vector of $J+1$ component specific regression coefficients including the intercept, and $\mu(\mathbf{x};\beta_g) = \beta_{0g} + \beta_{1g}'\mathbf{x}$. Then the maximum likelihood equation is given by

$$\begin{aligned}\mathscr{L}(\psi) &= \sum_{g=1}^{k}\sum_{i=1}^{n} z_{ig} \ln \pi_g + \sum_{g=1}^{k}\sum_{i=1}^{n} z_{ig} \ln\left\{\phi\left[y_i;\mu(\mathbf{x}_i;\beta_g),\sigma_g^2\right]\right\} \\ &= \sum_{g=1}^{k}\sum_{i=1}^{n} z_{ig} \ln \pi_g + \ell(\beta,\sigma^2)\end{aligned} \tag{3}$$

respectively, where $z_{ig} = 1$ if $(\mathbf{x}_i', y_i)'$ comes from component g and $z_{ig} = 0$ otherwise, $\ell(\beta, \sigma^2) = \sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} \ln \left\{ \phi \left[y_i; \mu(\mathbf{x}_i; \beta_g), \sigma_g^2 \right] \right\}$, $\beta = (\beta_1', \ldots, \beta_k')'$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)'$. In [3] the classical total sum of squares decomposition for regression models is generalized when data come from a heterogeneous population and they are modeled through a mixture of regressions with Gaussian components. This relation can be generalized as follows. Let us focus on the model term in the loglikelihood function and set

$$\ell(\mu, \sigma^2; \mathbf{y}) = \sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} \left[-\frac{1}{2} \ln \sigma_g^2 - \frac{(y_i - \mu_{i,g})^2}{2\sigma_g^2} \right], \tag{4}$$

where $\mu_{i,g} = \mu(\mathbf{x}_i; \beta_g) = \beta_{0g} + \beta_{1g}' \mathbf{x}_i$ and the constant term has been omitted in (4). Once the model has been fitted to the given data and parameters have been estimated, let us introduce the following quantities:

$$\begin{aligned}
\ell(\widehat{\mu}, \sigma^2; \mathbf{y}) &= \sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} \left[-\frac{1}{2} \ln \widehat{\sigma}_g^2 - \frac{(y_i - \widehat{\mu}_{i,g})^2}{2\widehat{\sigma}_g^2} \right] \\
\ell(\mathbf{y}, \widehat{\sigma}^2; \mathbf{y}) &= \sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} \left[-\frac{1}{2} \ln \widehat{\sigma}_g^2 - \frac{(y_i - y_i)^2}{2\widehat{\sigma}_g^2} \right] = \sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} \left[-\frac{1}{2} \ln \widehat{\sigma}_g^2 \right] \\
\ell(\bar{\mathbf{y}}, \widehat{\sigma}^2; \mathbf{y}) &= \sum_{g=1}^{k} \sum_{i=1}^{n} z_{ig} \left[-\frac{1}{2} \ln \widehat{\sigma}_g^2 - \frac{(y_i - \bar{y})^2}{2\widehat{\sigma}_g^2} \right]
\end{aligned}$$

and consider the decomposition

$$\ell(\mathbf{y}, \widehat{\sigma}^2; \mathbf{y}) - \ell(\bar{\mathbf{y}}, \widehat{\sigma}^2; \mathbf{y}) = \left\{ \ell(\mathbf{y}, \widehat{\sigma}^2; \mathbf{y}) - \ell(\widehat{\mu}, \sigma^2; \mathbf{y}) \right\} + \left\{ \ell(\widehat{\mu}, \widehat{\sigma}^2; \mathbf{y}) - \ell(\bar{\mathbf{y}}, \widehat{\sigma}^2; \mathbf{y}) \right\}. \tag{5}$$

After some algebras we get

$$\begin{aligned}
\ell(\mathbf{y}, \widehat{\sigma}^2; \mathbf{y}) - \ell(\bar{\mathbf{y}}, \widehat{\sigma}^2; \mathbf{y}) = &\sum_{g=1}^{k} \sum_{i=1}^{n} \widehat{z}_{ig} \frac{(y_i - \widehat{\mu}_{i,g})^2}{2\widehat{\sigma}_g^2} + \sum_{g=1}^{k} \sum_{i=1}^{n} \widehat{z}_{ig} \frac{(\widehat{\mu}_{i,g} - \bar{y}_g)^2}{2\widehat{\sigma}_g^2} + \\
&+ \sum_{g=1}^{k} \hat{n}_g \frac{(\bar{y}_g - \bar{y})^2}{2\widehat{\sigma}_g^2}.
\end{aligned}$$

Let us set

$$\begin{aligned}
\Delta\ell_t(\mathbf{y}, \bar{\mathbf{y}}) &= \sum_{g=1}^{k} \sum_{i=1}^{n} \widehat{z}_{ig} \frac{(y_i - \bar{y})^2}{2\widehat{\sigma}_g^2}, & \Delta\ell_r(\mathbf{y}, \widehat{\mu}) &= \sum_{g=1}^{k} \sum_{i=1}^{n} \widehat{z}_{ig} \frac{(y_i - \widehat{\mu}_{i,g})^2}{2\widehat{\sigma}_g^2}, \\
\Delta\ell_f(\widehat{\mu}, \bar{\mathbf{y}}_G) &= \sum_{g=1}^{k} \sum_{i=1}^{n} \widehat{z}_{ig} \frac{(\widehat{\mu}_{i,g} - \bar{y}_g)^2}{2\widehat{\sigma}_g^2}, & \Delta\ell_b(\bar{\mathbf{y}}_G, \bar{\mathbf{y}}) &= \sum_{g=1}^{k} \hat{n}_g \frac{(\bar{y}_g - \bar{y})^2}{2\widehat{\sigma}_g^2}
\end{aligned}$$

where $\bar{\mathbf{y}}_G = (\bar{y}_1, \ldots, \bar{y}_k)'$. Thus, the decomposition (5) can be written as

$$\Delta\ell_t(\mathbf{y}, \bar{\mathbf{y}}) = \Delta\ell_r(\mathbf{y}, \widehat{\mu}) + \Delta\ell_f(\widehat{\mu}, \bar{\mathbf{y}}_G) + \Delta\ell_b(\bar{\mathbf{y}}_G, \bar{\mathbf{y}}) \tag{6}$$

which generalizes the relation (32) in [3] because here the component variances are now included. According to [1] and [3], we define the local coefficient of determination for the gth group ($g = 1, \ldots, k$) as

$$R^2_g = \frac{\Delta\ell_{f,g}(\widehat{\mu}, \bar{\mathbf{y}}_g)}{\Delta\ell_{r,g}(\mathbf{y}, \widehat{\mu}) + \Delta\ell_{f,g}(\widehat{\mu}, \bar{\mathbf{y}}_g)} \tag{7}$$

and afterwards we define the overall coefficient of determination as

$$R^2 = \frac{\Delta\ell_f(\widehat{\mu}, \bar{\mathbf{y}}_g)}{\Delta\ell_r(\mathbf{y}, \widehat{\mu}) + \Delta\ell_f(\widehat{\mu}, \bar{\mathbf{y}}_g)}. \tag{8}$$

The Poisson case. Conditional on $G = g$ (for $g = 1, \ldots, k$) assume now that $Y|\mathbf{x}$ is Poisson with parameter $\mu_g(\mathbf{x}; \beta_g)$, for some $\beta_g \in \mathbb{R}^{p+1}$; that is, $Y|\mathbf{x}(G = g) \sim \text{Poi}\left[\mu_g(\mathbf{x}; \beta_g)\right]$, and set

$$q(y|\mathbf{x}; \beta_g) = \exp\left[-\mu_g(\mathbf{x}; \beta_g)\right] \frac{\left[\mu(\mathbf{x}; \beta_g)\right]^y}{y!},$$

where $\mu(\mathbf{x}; \beta_g) = \exp(\beta_{0g} + \beta'_{1g}\mathbf{x})$. Using similar arguments like in the previous case, we set

$$\begin{aligned}
\Delta\ell_t(\mathbf{y}, \bar{\mathbf{y}}) &= \sum_{g=1}^{k}\sum_{i=1}^{n} \widehat{z}_{ig}\left\{y_i \ln y_i - y_i - y_i \ln \bar{y} + \bar{y}\right\} \\
\Delta\ell_r(\mathbf{y}, \widehat{\mu}) &= \sum_{g=1}^{k}\sum_{i=1}^{n} \widehat{z}_{ig}\left\{y_i \ln y_i/\widehat{\mu}_{i,g} - (y_i - \widehat{\mu}_{i,g})\right\} \\
\Delta\ell_f(\widehat{\mu}, \bar{\mathbf{y}}_G) &= \sum_{g=1}^{k}\sum_{i=1}^{n} \widehat{z}_{ig}\left\{y_i \ln \widehat{\mu}_{i,g}/\bar{y}_g - (\widehat{\mu}_{i,g} - \bar{y}_g)\right\} \\
\Delta\ell_b(\bar{\mathbf{y}}_G, \bar{\mathbf{y}}) &= \sum_{g=1}^{k}\sum_{i=1}^{n} \widehat{z}_{ig}\left\{y_i \ln \bar{y}_g/\bar{y} - (\bar{y}_g - \bar{y})\right\}
\end{aligned}$$

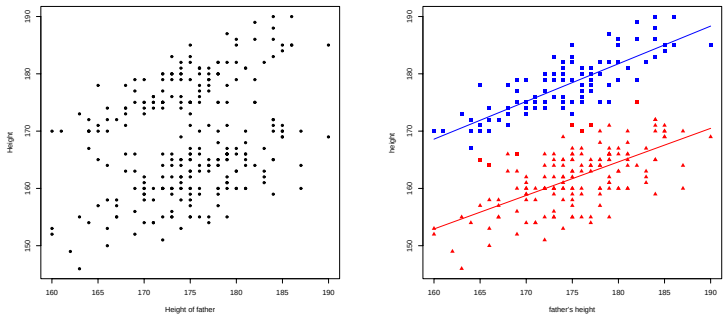
where again $\bar{\mathbf{y}}_G = (\bar{y}_1, \ldots, \bar{y}_k)'$. Therefore, the decomposition (5) can be written as

$$\Delta\ell_t(\mathbf{y}, \bar{\mathbf{y}}) = \Delta\ell_r(\mathbf{y}, \widehat{\mu}) + \Delta\ell_f(\widehat{\mu}, \bar{\mathbf{y}}_G) + \Delta\ell_b(\bar{\mathbf{y}}_G, \bar{\mathbf{y}}).$$

Finally the indices (7) and (8) can be derived accordingly.

3 A real data application

In order to illustrate the proposed fit measures, we analyze a data set from a survey of $n = 270$ students of basic Statistics at the Department of Economics and Business of the University of Catania in the academic year 2011/2012. The data set was first used by [2] and is available from the R package **flexCWM**; see [4]. The aim is to model the height of a student based on the information about the height of his/her father.



(a) Scatter plot **(b)** Scatter plot with fitted components

Fig. 1: Panel 1a is the scatter plot of the variables *Height* and *Height of father* of the *students* data; panel 1b is the scatter plot with fitted mixture component labels and regression lines. Red triangles indicate first group membership, blue squares indicate second group membership.

Figure 1 presents the scatters of the unlabeled data (Panel 1a), and of the labeled data as estimated by the two-component mixture of linear regression models (Panel 1b), including the regression lines. Overall we observe a positive association between *Height of father* and *Height* in both clusters. The cluster labels that are found are close enough to the classification of the gender variable (not reported), with an Adjusted Rand Index of about 0.899.

Cluster	1	2
π_j	0.583	0.417
R^2_j	0.389	0.646

R^2	0.531
$\Delta\ell_t(\mathbf{y},\bar{\mathbf{y}})$	1035.294
$\Delta\ell_r(\mathbf{y},\widehat{\boldsymbol{\mu}})$	133.997
$\Delta\ell_f(\widehat{\boldsymbol{\mu}},\bar{\mathbf{y}}_G)$	151.856
$\Delta\ell_b(\bar{\mathbf{y}}_G,\bar{\mathbf{y}})$	749.441

Table 1: Local and overall fit measures for the finite mixture of two linear regressions fitted to the *students* data.

Table 1 reports the local and overall fit measures for the output model. Interestingly, the regression fit for group 2 (male) has a higher local R^2 than that of group

1 (female) - respectively 0.646 and 0.389. We report an overall R^2 of about 0.5. In addition, from the other overall deviance measures we observe that more than 70% of the total deviance is due to the separation on the y-axis between the groups (as measured by $\Delta\ell_b(\bar{\mathbf{y}}_G, \bar{\mathbf{y}})$).

4 Concluding remarks

The decomposition here proposed concerns an intuitive tool aiming at facilitating the interpretation of the results of statistical modeling based on mixtures of regressions. In this framework, we remark that this approach should not to be used for model selection.

The decomposition involves essentially the response variable. While for simplicity, we have illustrated the approach for mixtures of regressions, the proposal holds in general for other kinds of mixture of regressions. In particular, the results here presented can be easily extended to mixtures of regressions with concomitant values and to Cluster-Weighed Models, see e.g. [6] and [2].

References

1. Cameron, A. C., Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, **14**(2), 209–220.
2. Ingrassia S., Minotti S. C. and Punzo A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, **71**(4): 159–182.
3. Ingrassia, S., Punzo, A. (2019). Cluster validation for mixtures of regressions via the total sum of squares decomposition. *Journal of Classification*, doi: 10.1007/s00357-019-09326-4.
4. Mazza, A., Punzo, A., Ingrassia, S. (2018). **flexCWM**: a flexible framework for cluster-weighted models. *Journal of Statistical Software*, **86**(2), 1–30.
5. McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman & Hall, Boca Raton, Second edition.
6. Dayton, C.M., Macready, G.B.: Concomitant-variable latent-class models. Journal of the American Statistical Association, **83**(401), 173–178 (1988)
7. McLachlan, G.J., Peel, D. (2000). Finite Mixture Models. John Wiley & Sons, New York.

Statistical Analysis of Satellite Data (SDS-SIS)

Functional Data Analysis for Interferometric Synthetic Aperture Radar Data Post-Processing: The case of Santa Barbara mud volcano

Analisi di dati funzionali per il post-processamento di dati interferometrici: Il caso dei vulcanelli di Santa Barbara

Matteo Fontana, Alessandra Menafoglio, Francesca Cigna, Deodato Tapete

Abstract The systematic and widespread availability of cheap computing power and storage space, alongside the drastic improvement in data transmission and active sensor technologies, has triggered an increasing interest by scientific and government institutions in the use of satellite data. One of the most promising applications is to provide emergency management agencies with early warning systems for catastrophic natural events, in order to support decision making. The data used in this work are ground displacement records gathered via interferometric processing of satellite RADAR data. We provide a novel post-processing approach based on a Functional Data Analysis paradigm, and apply it to Santa Barbara mud volcano.

Abstract *La sistematica disponibilità di memoria e potenza di calcolo, insieme al drastico miglioramento nelle tecnologie di tramissione dati e dei sensori attivi, ha stimolato un interesse crescente da parte di istituzioni scientifiche e governative verso l'utilizzo di dati satellitari. Una delle applicazioni più promettenti è fornire sistemi di allerta precoce alle agenzie di gestione delle emergenze, per supportare i processi decisionali. I dati usati in questo studio sono stime di spostamento del terreno ottenute dall'elaborazione interferometrica di immagini RADAR satellitari. Viene proposto un innovativo metodo di post-processamento basato sul paradigma dell'analisi di dati funzionali, e la sua applicazione ai vulcanelli di Santa Barbara.*

Key words: Functional Data Analysis, InSAR, Remote Sensing, Conformal Prediction

---

Matteo Fontana
MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy
e-mail: matteo.fontana@polimi.it

Alessandra Menafoglio
MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy

Francesca Cigna
Italian Space Agency (ASI), Rome, Italy

Deodato Tapete
Italian Space Agency (ASI), Rome, Italy

1 Motivation

In the last years, a significant advance in data transmission technologies and availability of computing power, storage space and sensors, have triggered a boost in terms of possible applications of satellite imaging data. Among the wealth of data that can be acquired by passive or active sensor arrays on-board satellites, Synthetic Aperture Radar (SAR) images and their processing with Interferometric SAR (InSAR) methods are one of the most useful sources of information for natural hazard monitoring [10].

The statistical analysis of the output products of advanced InSAR processing is still at an early stage. We are witnessing several attempts at using standard statistical frameworks such as time-series analysis and geostatistics, as shown in the review by [2]. To the best of our knowledge, no attempts have been made yet to combine the dynamics in time and in space, or to employ more advanced statistical techniques to gather novel insights from these specific data.

Our aim is to provide a proof of concept for the application of advanced statistical techniques in this realm, tackling the issue of geological hazard monitoring and early-warning. We provide deeper explanations about the novelty of our approach with respect to the current state of the art in [6]. We exploit, as a test case, the event of a mud volcano eruption occurred in the village of Santa Barbara, in the eastern sector of the city of Caltanissetta, in Sicily (Italy). On the 11^{th} of August 2008 the area was affected by paroxysmal eruption that caused damage to urban infrastructure as far as 2 km from the main eruptive vent. A more detailed description of the event and its geological features can be found in [5] and [8].

2 The Application: InSAR data of Caltanissetta

We used 32 ENVISAT Advanced SAR scenes acquired along ascending track T172 between 12/10/2002 and 07/06/2008 (i.e. before the mud volcano erupted). These data are in C-band (5.6 cm wavelength, 5.3 GHz frequency) and characterized by a Line-Of-Sight (LOS) with ~23° look angle, VV co-polarization, ~20 m ground resolution and nominal site revisit of 35 days.

InSAR processing was carried out with the Small Baseline Subset (SBAS) technique developed by [1] and parallelized by [3]. The output dataset consists of $n = 1735$ coherent targets, distributed across an area of 150 km^2. For each target, the annual LOS velocity over the monitoring period, LOS displacement time series, temporal coherence, and elevation above the reference ellipsoid were estimated. The position of the targets can be seen in Fig. 1, while a plot of the corresponding temporal dynamics is shown in Fig. 2a. The 2002-2005 ground deformation scenario in Caltanissetta was described by [13]. Previous semi-automated analysis of a different InSAR dataset was carried out in [5], where the computation of Deviation Indices was proposed to identify trend changes in InSAR time series (see also [12]). We present the main findings about the mud volcano area in [6].

3 A Functional Approach for InSAR data post-processing

Each displacement series estimated via InSAR consists of a set of discrete, time-indexed evaluations of a continuous trajectory in time, with some degree of smoothness given by the physics of the deformation phenomenon. We can also assume some measurement error to be present, due to the particular nature of the measuring and processing technique. Moreover, in this specific experimental setting, derivatives of such continuous trajectories carry a lot of information: the first derivative represents the velocity of the measured displacement, while the second derivative is the acceleration profile and, via a multiplicative constant, force values. These specific and nonstandard data features call for the use of statistical methods able to correctly include and model them into the analysis process.

A very good candidate to do so is Functional Data Analysis (FDA) [9], the field of statistics which uses, as the unit of the analysis, one or more continuous functions over a domain, either univariate or multivariate. The first step in every FDA pipeline is the extraction of continuous functions from the discrete, longitudinal data points. In our case this was performed via a smooth B-spline basis. Further details about the procedure, and the choice of the smoothing parameter can be found in [6], while a plot of the smoothed data is shown in Fig. 2b. After the data smoothing, we performed a functional Principal Component Analysis (PCA) to assess and quantify the variability of the displacement curves. A functional clustering based on the K-Mean Alignment (KMA) procedure [11] was also performed, to identify a group structure

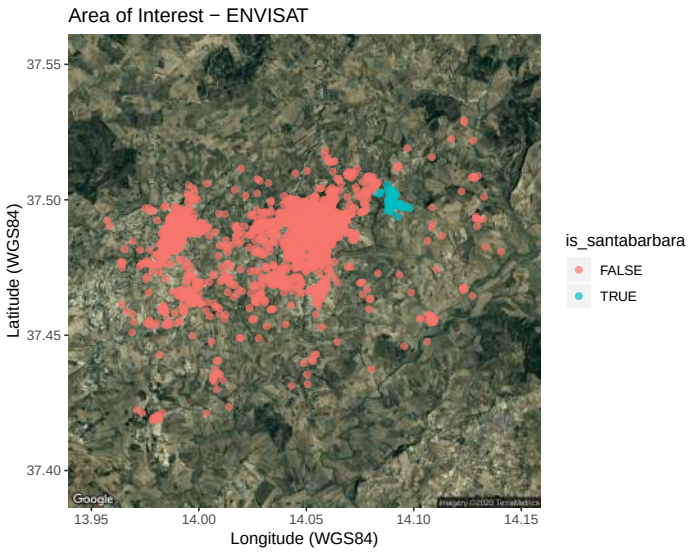
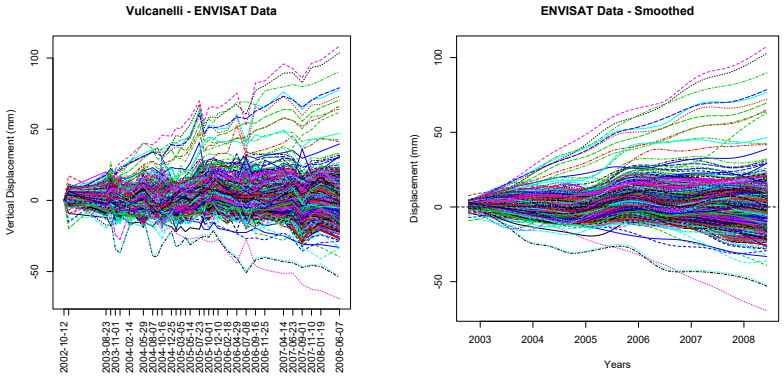


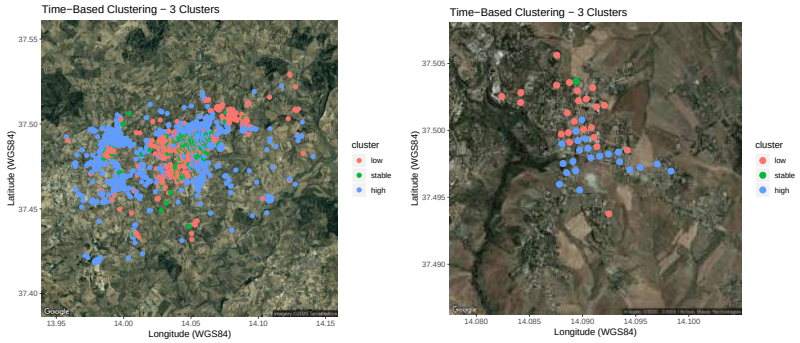
Fig. 1: Map of the geographical position of the target points on a satellite image of the area of Caltanissetta (Italy), with indication of Santa Barbara village.



(a) Unsmoothed Case (b) Smoothed Case

Fig. 2: Plot of the time dynamics of displacement for the target points in the smoothed and unsmoothed case.

in the data. In Fig. 3 we present a spatialization of the clusters obtained via the KMA procedure, with a zoom on the Santa Barbara area. Further details about the results of the exploratory analysis can be found in [6].



(a) Map of the whole area (b) Zoom over Santa Barbara

Fig. 3: A map of the target points over the city of Caltanissetta, where the color depends on the cluster that was assigned to the corresponding displacement curve by the KMA procedure, with a zoom on the Santa Barbara area

After the exploratory analysis, we focused our attention on the issue of set forecasting for these complex data objects, that is defining a prediction region for an infinite-variate data object such as a function. Providing meaningful forecasts with

associated uncertainty for a given data object is, in general, a key problem from a theoretical and applied perspective. If we frame the early warning problem as identifying a time instant at which the time series departs significantly from its expected value, it appears immediately evident how an appropriate set forecast method represents a key stepping stone to realize any kind of outlier identification system.

The simplest and possibly the most natural way to forecast a time series of data objects is to use an autoregressive model. Let δ_t be the displacement surface measured at time t, which is here assumed to be a square-integrable function defined over bidimensional Euclidean space S, $\delta_t \in L^2(S)$. Modeling the surface at $t+1$ with a concurrent functional autoregressive (FAR) model of order 1 $FAR(1)$, we can write

$$\delta_{t+1}(s) = \Phi(s)\delta_t(s) + \varepsilon(s) \tag{1}$$

where $s \in S$ and $\Phi \in L^{\infty}(S)$. The main issue in this setting is the substantial lack of usable and interpretable methods for interval forecasting. A very powerful approach to tackle this kind of problems is represented by Conformal Prediction (CP) [14]. CP was developed in the late 1990s in the Machine Learning community as a method to provide interval forecast for support vector machines. Since then, it has been extended in several ways, and used as a framework to provide distribution-free prediction sets in the scalar case, and very recently also in the functional one [7].

To provide meaningful forecasts for the concurrent FAR(1) model in Equation 1, we extend the framework in [7] in two directions. The first one is moving from the functional univariate case to the bivariate one, required in this application. The second, and most challenging one, is to move from the case of iid observations (described in [7]) to setting of data with (temporal) dependence. This is performed by adapting the ideas described in [4] for the scalar case to the case of functions whose domain is bidimensional. The complete mathematical description of the method, alongside a detailed analysis of the Santa Barbara test case can be found in [6].

4 Conclusions

Motivated by the need to test advanced statistical methodologies to analyse displacement series obtained from satellite InSAR techniques, we illustrate a Functional Data Analysis framework for the analysis of displacement data related with a mud volcano eruption occurred in 2008 close to the city of Caltanissetta, Italy. Aiming to develop an early warning system, we describe an extension (fully shown in [6]) of a newly developed forecasting technique based on Conformal Prediction to the case of dependent functional observations defined over a multivariate domain.

5 Acknowledgements

This work was carried out in the framework of the ASI-POLIMI "Attività di Ricerca e Innovazione" project, grant agreement n.2018-5-HH.0.

References

1. Berardino P., Fornaro G., Lanari R., Sansosti E. 2002. A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. IEEE Trans. Geosci. Remote, 40, 2375-2383, doi:10.1109/TGRS.2002.803792
2. Bernardi M.S., Africa P.C., de Falco C., Formaggia L., Menafoglio A., Vantini S. 2020+. On the use of Interfeometric Synthetic Aperture Radar Data for Monitoring and Forecasting Natural Hazards. Manuscript, Politecnico di Milano.
3. Casu F., Elefante S., Imperatore P., Zinno I., Manunta M., De Luca C., Lanari R. 2014. SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation. IEEE J. Sel. Top. Appl., 7, 3285-3296, doi:10.1109/JSTARS.2014.2322671
4. Chernozhukov V., Wüthrich K., Yinchu Z. 2018. Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. Proc. 31st Conference On Learning Theory, in PMLR 75:732-749.
5. Cigna F., Tapete D., Casagli N. 2012. Semi-automated extraction of Deviation Indexes (DI) from satellite Persistent Scatterers time series: tests on sedimentary volcanism and tectonically-induced motions. Nonlin. Processes Geophys., 19, 643-655, doi:10.5194/npg-19-643-2012
6. Fontana M., Bernardi M., Cigna F., Menafoglio A., Tapete D., Vantini S. 2020+. Interferometric Synthetic Aperture Radar Data Post-Processing: A Functional Data Analysis Perspective. Manuscript, Politecnico di Milano.
7. Fontana M., Vantini S., Tavoni M., Gammerman A. 2020. A Conformal Approach for Distribution-Free Prediction of Functional Data. In: Aneiros G. et al. (eds) Functional and High Dimensional Statistics and Related Fields. Contribution to Statistics. Springer. Forthcoming.
8. Madonia P., Grassa F., Cangemi M., Musumeci C. 2011. Geomorphological and geochemical characterization of the 11 August 2008 mud volcano eruption at S. Barbara village (Sicily, Italy) and its possible relationship with seismic activity. Nat. Hazards Earth Syst. Sci., 11, 1545—1557, doi:10.5194/nhess-11-1545-2011
9. Ramsay J.O., Silverman B.W. 2005. Functional Data Analysis. Second edition. Springer Series in Statistics. New York, NY: Springer.
10. Rosen P.A., Hensley S., Joughin I.R., Li F.K., Rodríguez E., Goldstein R.M. 2000. Synthetic Aperture Radar Interferometry. Proc. IEEE, 88, 3, doi:10.1109/5.838084
11. Sangalli L.M., Secchi P., Vantini S., Vitelli V. 2010. 'K-Mean Alignment for Curve Clustering'. Computational Statistics & Data Analysis 54 (5): 1219–33. https://doi.org/10.1016/j.csda.2009.12.008
12. Tapete D., Casagli N. 2013. Testing computational methods to identify deformation trends in RADARSAT Persistent Scatterers time series for structural assessment of archaeological heritage. In: Murgante B. et al. (eds) Computational Science and Its Applications, ICCSA 2013. Lecture Notes in Computer Science, 7972, 693-707, doi:10.1007/978-3-642-39643-4_50
13. Vallone P., Giammarinaro M.S., Crosetto M., Agudo M., Biescas E. 2008. Ground motion phenomena in Caltanissetta (Italy) investigated by InSAR and geological data integration. Engineering Geology, 98, 3-4, 144-155, doi:10.1016/j.enggeo.2008.02.004
14. Vovk, V., Gammerman A., Shafer G. 2005. Algorithmic Learning in a Random World. New York, NY: Springer.

Recent Contributions to the Understanding of the Uncertainty in Upper-Air Reference Measurements

Recenti Contributi alla Comprensione dell'Incertezza delle Misure Climatiche dell'Alta Atmosfera

Alessandro Fassò

Abstract Measurement uncertainty of climatic variables in the Upper Troposphere/-Lower Stratosphere (UTLS) is a key for the understanding of climate warming. Beyond traditional measurement uncertainty, the focus of the atmospheric science community is moving to an uncertainty which is related to the spatial and temporal scales, and to the comparison of different instruments. This paper reviews some recent contributions to the analysis of data related to radiosonde and satellite observation. In particular, techniques based on functional data analysis approach, spatio temporal models, regularisation, and cross-validation are applied to radiosonde profile data and satellite data.

Abstract *L'incertezza della misura di variabili climatiche nell'alta troposfera/bassa stratosfera (UTLS) è fondamentale nella comprensione del riscaldamento globale. Tale incertezza non riguarda solo la tradizionale incertezza strumentale ma coinvolge anche aspetti legati alla dimensione spaziale e temporale ed al confronto fra strumenti diversi. I metodi statistici considerati riguardano la functional data analysis, i modelli spazio-temporali, anche 4D, e le tecniche di cross-validazione.*

Key words: radiosonde, functional data, Gaussian process, cross-validation

1 Introduction

Climate variables in the Upper Troposphere/Lower Stratosphere (UTLS) may be measured in-situ, for example through a flying radiosonde balloon, or remotely, for example through an optical instrument on-board a satellite. Data products related to the latter are fast increasing in quality and quantity, by providing massive regularly spaced observations. Nevertheless, radiosonde observations are considered crucial, not only because they provide long time series, but also because they are used for

Alessandro Fassò
University of Bergamo, Dalmine (BG), Italy, e-mail: alessandro.fasso@unibg.it

calibration and validation of both satellite observations and climate models such as the ERA-interim re-analysis, [2].

In this frame, an important role is played by GRUAN, the GCOS[1] Reference Upper-Air Network (www.gruan.org). Despite having a limited spatial coverage, GRUAN aims at providing fully traceable reference measurements. To do this, not only has more then 20 stations around the world but also implements data processing covering measurement uncertainty estimation, [1].

For this reason climatological studies are largely based on baseline measurement networks, which have an intermediate metrological quality but a larger spatio-temporal coverage.

In metrology, measurement uncertainty is a characteristics of the statistical dispersion of the values attributed to a measured quantity. In this paper, using a statistical approach, we use the Root Mean Square Error (RMSE) to quantify measurement uncertainty. An empirical assessment of it is often given by means of a statistical model able to describe the natural variability attributable to environmental forcing factors. Hence, the model RMSE is used to assess measurement uncertainty.

The rest of the paper is organised as follows. Section 2 deals with the uncertainty arising in the comparison of partially misaligned profiles using a functional data analysis approach. The first case is related to the comparison of pair of radiosonde profiles from sites in the same climatic area. The second case study considers the comparison of radiosonde and satellite data, which have quite different vertical smoothing properties. Section 3, considers the harmonisation of 40-years global time-series, using a combination of 4D local Gaussian process modelling and fused LASSO. Section 4 considers the uncertainty propagation in interpolation of missing data using a Gaussian process approach coupled with block-bootstrap correction.

2 Co-location uncertainty

In the comparison of two profiles obtained by two different instruments, various mismatch sources should be taken into account: the spatio-temporal mismatch between profiles; the different vertical smoothing and resolution of the two instruments/data sets; the different horizontal smoothing and resolution of the two instruments; and the comparator uncertainty. In particular, considering radiosonde, comparator uncertainty is related to instrument issues, including solar radiation, dry-bias when measuring humidity, ventilation effects, ground calibration effects, and all the other problems detailed in [3].

In Section 2.1, pairs of temperature radiosonde profiles from nearby GRUAN stations are compared, hence the main source of discrepancy is related only to the spatio-temporal mismatch. In Section 2.2, the temperature profiles of baseline quality radiosondes and satellite observations are compared.

[1] GCOS is the Global Climate Observing System of the World Meteorological Organisation (WMO)

2.1 Heteroskedastic Regression for Functional Data

Let us consider the comparison of pairs of temperature profiles, $(y(h), y'(h))$ from nearby stations, where h denotes the altitude, as discussed by [6, 9]. Each profile is modelled as a functional data object. Hence, considering the co-location error $z(h) = y(h) - y'(h)$, the following functional regression is established:

$$z(h) = \beta(h)'x(h) + \varepsilon(h)$$
$$\sigma_\varepsilon^2(h|x) = \gamma(h)'x(h).$$

In this model $\sigma_\varepsilon^2(h|x) = Var(\varepsilon(h)|x(h))$ defines the irreducible co-location uncertainty, which depends on altitude and other variables characterising the local condition of the atmosphere.

After suitable estimation of the unknown parameters, this approach gives the following total profile uncertainty budget:

$$\sigma_z^2(h) = \beta_0^2 + \sigma_{x\sim\beta_0}^2(h) + \sigma_\omega^2(h) + \sigma_{\hat{\beta}}^2(h) + \sigma_{\Delta_\varepsilon}^2$$

whose elements are discussed in the sequel.

1. β_0^2 accounts for co-location bias. This error component may be easily covered with this approach.
2. $\sigma_{x\sim\beta_0}^2(h) = \beta'(h)\Sigma_x(h)\beta(h) - \beta_0^2$ is the (marginal) reducible collocation uncertainty. It is called "reducible" because it depends on environmental forcing factors x. Hence, knowing x allows to correct for this error component.
3. $\sigma_\omega^2(h) = E_x\left(\hat{\sigma}_\omega^2(h|x)\right)$ is the irreducible collocation uncertainty. It is called "irreducible" because, using this model and information in x this term can not be reduced.
4. $\sigma_{\hat{\beta}}^2(h) = E(x(h)'\Sigma_{\hat{\beta}}(h)x(h))$ is the estimation uncertainty or sampling error;
5. $\sigma_{\Delta_\varepsilon}^2 = 2\sigma_\varepsilon^2$ is the measurement error.

2.2 Comparing Radiosonde and IASI profiles

The validation of satellite products is important to ensure their quality for climate and weather applications. To do this, a fundamental step is the comparison with other instruments. In [7], the temperature profiles obtained by the Infrared Atmospheric Sounding Interferometer (IASI) instrument, on board EUMETSAT satellite MetOP-A and –B, are compared to the radiosonde observations within the network of the Universal Rawinsonde Observation Program (RAOB).

In this type of comparison, the mismatch uncertainty is dominated by the difference in vertical smoothing. In fact IASI data products give profiles which are known to have a limited number of degrees of freedom. In some cases the so-called averaging kernels are available, defining the vertical smoothing by a set of weights.

In other cases, for example when handling dated time series, the averaging kernels are not available and [7] estimated such weights using the Maximum Likelihood principle.

To see this, we adopt a simplified version of [7], hence we ignore the so-called comparator uncertainty. The idea is to let each RAOB profile to mimicking the corresponding IASI profile. In other words the true signal of the $k-th$ profile is considered as a smooth function denoted by $s_k(h)$. It is related to the observation $y_{J,k}(h)$ with $J = R, I$ for RAOB and IASI respectively, by the following equations

$$y_{J,k}(h) = s_{J,k}(h) + \varepsilon_{J,k}(h), J = R, I$$

where $\varepsilon_{J,k}(h)$ is Gaussian distributed, $N(0, \sigma^2_{J,k}(h))$ and the two are related by

$$s_{I,k}(h) = \int_{VR} s_{R,k}(h) w(q;h) dq.$$

where VR is the vertical range of the profiles. The best weighting function w resulted to be the generalised extreme value probability density function, selected by cross-validation among a number of candidates.

3 Harmonisation of radiosonde time series

Copernicus is the European union's Earth observation programme (www.coopernicus.eu) and is involved in collecting data from multiple sources. In particular, the Copernicus Climate Change Service (C3S), will provide comprehensive climate information covering a wide range of components of the Earth-system and timescales spanning decades to centuries, [4].

In reconstructing radiosonde global timeseries, harmonisation is essentially made by change detection and adjustment of data for any kind of known and quantifiable inhomogeneities, including bias, change of sensors, calibration drift, local environment changes [8, 11, 10].

In this frame, [5] discusses the harmonisation of 40-year time-series of temperature profiles from the Integrated Global Radiosonde Archive (IGRA), covering about $n = 800$ sites around the world. To do this, a two step procedure is used.

In the first step a 4D local Gaussian Process modelling is developed

$$y(s,t,h) = z(s,t,h) + \varepsilon(s,t,h)$$

where $s = (lat, lon) \in Spherical\ shell,\ h \in [925hPa, 50hPa],\ t \in R^+$. Moreover, for each site $j = 1, ..., n$, the following residuals are computed

$$e(s_j,t,h) = E[y(s_j,t,h)|Y_{\sim j}] - y(s,t,h)$$

where $Y_{\sim j}$ contains all the IGRA information excluding $j-th$ site.

At the second step a fused LASSO change detection on each residual time-series is applied [12]. In particular, ignoring site subscript j, we assume

$$e_t = \beta_t + \zeta_t,\ t = 1,...,T$$

where $\zeta_t \equiv NID(0,\sigma^2)$ and $(\beta_1,...,\beta_T)$ is given by the the (regularised) optimisation of

$$\sum_{t=1}^{T}(e_t-\beta_t)^2 + \lambda_1\sum_{t=1}^{T}|\beta_t| + \lambda_2\sum_{t=2}^{T}|\beta_t-\beta_{t-1}|$$

where the first penalty term controls the number of $\beta_t \neq 0$ and the second one controls smoothness of β_t, hence identifying temporary and permanent changes.

4 Interpolation uncertainty of high-vertical-resolution profiles

Modern radiosondes, such as Vaisala RS41, send data to ground site second by second. For various reasons, missing data are sometimes spread along the atmospheric profile. A common strategy is to compute (linear) interpolation to fill the gaps in order to have a regularly spaced data product. In the perspective of uncertainty full traceability of climate data products, the issue of the resulting interpolation uncertainty is of interest.

To see this, considering a single radiosonde profile, we assume that $y(t)$ is the observation of the true temperature profile, say $s(t)$, at flying time $t = 1,...,T$ and

$$y(t) = s(t) + \varepsilon(t)$$

where $\varepsilon(t) \sim N(0,\sigma_t^2)$ is the measurement error with possibly non constant measurement uncertainty σ_t.

If there is an observation gap in the interval (t^-,t^+), the linear interpolation at time t, for $t^- \leq t \leq t^+$, is given by the straightforward formula

$$\hat{y}(t) = \alpha(t)y^+ + (1-\alpha(t))y^-$$

where, $y^\pm = y(t^\pm)$, and $\alpha(t) = \frac{t-t^-}{t^+-t^-}$. If we assume that $s(t) = \omega(t)$ is a Gaussian process with autocorrelation function $\gamma(t,t')$, the linear interpolation mean squared error is given by

$$MSE_y(t) = E[(\hat{y}(t)-s(t))^2] = \vec{a}'\Sigma\vec{a} + \sigma_t^2$$

where Σ is the 3×3 variance covariance matrix of $(y^+,y^-,y(t))$ and $\vec{a}' = (\alpha(t),1-\alpha(t),-1)$.

Under the above Gaussian process assumption and if σ_t is known, this formula propagates the uncertainty from $y^\pm$ to $y(t)$, hence making the interpolated value fully traceable. Nonetheless, if the Gaussian process is only an approximation, so

that $s(t) = \omega(t) + \delta(t)$ than the above MSE underestimate the true interpolation uncertainty, say $u(t)^2$, by the quantity $E(\delta(t)^2)$. For this reason, using a sample of profiles, an extensive Block-Bootstrap cross-validation exercise may be used to estimate $E(\delta(h)^2)$ as a function of altitude and provide a Bootstrap-based semi-parametric estimate of $u(t)^2$. The same cross-validation design may be used to assess the gain in precision in using a kriging interpolation instead of the simpler linear interpolation.

References

1. Bodeker, G., S. Bojinski, D. Cimini, R. Dirksen, M. Haeffelin, J. Hannigan, D. Hurst, T. Leblanc, F. Madonna, M. Maturilli, A. Mikalsen, R. Philipona, T. Reale, D. Seidel, D. Tan, P. Thorne, H. Vömel, and J. Wang,: Reference upper-air observations for climate: From concept to reality. Bull. Am. Meteorol. Soc., **97**, 123–135 (2016)
2. Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., **137**, 553–597 (2001)
3. Dirksen, R. J., Sommer, M., Immler, F. J., Hurst, D. F., Kivi, R., and Vomel, H. : Reference quality upper-air measurements: GRUAN data processing for the Vaisala RS92 radiosonde. Atmospheric Measurement Techniques, **7**, 4463–4490 (2014)
4. Fassò A., Finazzi F., Madonna F.: Statistical issues in radiosonde observation of atmospheric temperature and humidity profiles. Statistics and Probability Letters. **136**, 97-100.
5. Fassò, A., Huang H.-C., Valli I., Madonna F.: Change detection and harmonisation of atmospheric large spatiotemporal series. Proceeding of the 62nd ISI World Statsitics Congress 2019. Special Topic Session, **2**, 217–221 (2019)
6. Fassò, A., Ignaccolo, R., Madonna, F., Demoz, B. and Franco-Villoria, M.: Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles. Atmos. Meas. Tech., **7**, 1803–1816 (2014) doi:10.5194/amt-7-1803-2014.
7. Finazzi F., Fassò A., Madonna F., Negri I., Sun B., Rosoldi M.: Statistical harmonization and uncertainty assessment in the comparison of satellite and radiosonde climate variables. Environmetrics, **30**(2), 1-17, DOI: 10.1002/env.2528 (2018)
8. Haimberger, L., Tavolato, C., and Sperka, S. 2012. Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations, J. Climate., **25**, 8108–8131.
9. Ignaccolo R., Franco-Villoria M., Fassò A.: Modelling collocation uncertainty of 3D atmospheric profiles. Stoch. Env. Res. and Risk Assess, **29**(2), 417-429 (2015)
10. Sherwood, S.C., Meyer C.L., Allen R.J., Titchner H.A.: Robust tropospheric warming revealed by interactively homogenised radiosonde data. J. Clim. **21**, 5336 – 5352 (2008)
11. Thorne, P.W., Brohan, P., Titchner, H.A., et al.: A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. J. Geophys. Res. **116**, D12116 (2011) doi:10.1029/2010JD015487.
12. Tibshirani R., Saunders M., Rosset A., Heights Y., Zhu J., Arbor A., and Knight K.: Sparsity and smoothness via the fused lasso. J. R. Statist. Soc. B, **67**:1, 91–108 (2005)

Statistical models and methods for Business and Industry

Modelling and monitoring of complex 3D shapes: a novel approach for lattice structures

Modellazione e monitoraggio di forme complesse 3D: un nuovo approccio per strutture lattice

Bianca Maria Colosimo, Marco, Grasso, Federica Garghetti

Abstract The industrial breakthrough of novel manufacturing paradigms, like additive manufacturing, enables the production of innovative shapes, characterized by complex 3D structures going beyond more 2.5D free-form surfaces and other geometries produced with traditional methods. This opens novel challenges for product qualification and statistical process monitoring. This study focuses on a category of complex shapes known as lattice structures (or metamaterials), where a unit cell regularly repeats in space to achieve advanced mechanical and functional performances. We present a novel solution for statistical process monitoring of these structures, aimed at identifying local out-of-control cells within- and between-parts.

Abstract *Lo sviluppo industriale di nuovi paradigmi di produzione, come le tecnologie additive, permette di produrre forme innovative, caratterizzate da strutture complesse 3D che vanno ben oltre le superfici 2.5D e altre forme prodotte con tecnologie tradizionali. Questo apre nuove sfide in termini di qualificazione di prodotto e monitoraggio di processo. Lo studio si concentra su forme complesse note come strutture lattice (o metamateriali), dove una cella si ripete regolarmente nello spazio per ottenere prestazioni avanzate. Presentiamo una nuova soluzione per il monitoraggio statistico di processo con l'obiettivo di identificare celle fuori controllo all'interno di una singola parte o in produzioni in serie.*

Key words: complex shape, lattice structure, metamaterials, statistical process monitoring; additive manufacturing, profile monitoring.

Bianca Maria Colosimo, Dipartimento di Meccanica, Politecnico di Milano; email: biancamaria.colosimo@polimi.it

Marco Grasso, Dipartimento di Meccanica, Politecnico di Milano; email: marcoluigi.grasso@polimi.it

Federica Garghetti, Dipartimento di Meccanica, Politecnico di Milano; email: federica.garghetti@polimi.it

1 Introduction

Thanks to the industrial breakthrough of new manufacturing and inspection technologies, innovative kinds of complex shapes can be made available for several different applications. The level of shape complexity that can be achieved nowadays (e.g., by means of additive manufacturing methods) goes beyond the one of simpler 2.5D free-form surfaces and more traditional products. This yields brand new challenges in the design and use of data modelling and process monitoring methodologies. It also motivates an increasing interest for statistical analysis of complex geometries and spatially dense metrology data.

The problem of statistical process monitoring of free-form surfaces based on 2.5D or 3D point cloud data was investigated by various authors (Del Castillo et al., 2015; Shi et al., 2019). Some authors focused on identifying the nature of part-to-part variation (Shi et al., 2019; Shan and Apley, 2008; Lee and Apley, 2004), whereas other authors focused on the design of statistical quality monitoring methods (Colosimo et al., 2015; Colosimo et al., 2014; Colosimo, 2018; Zang and Qiu, 2018a, 2018b). Some authors focused on the analysis of stochastic textures in the absence of a golden standard shape (Bui and Apley, 2018) and random cellular structures (Menafoglio et al., 2018).

Some novel kinds of complex shapes exhibit much more complicated geometrical features than the ones investigated in previous studies. This motivates the study of novel methods to determine the quality of such complex product (even in the presence of small lots and one-of-a-kind productions) and to monitor the stability of the production process. Lattice structures belong to this category (Wu et al., 2019, Cansizoglu et al., 2008). Lattice structures are complex geometries where a unit cell, usually characterized by a trabecular shape, regularly repeats in space within the part. They gain their functional properties from their structure rather than inheriting them directly from the material they are composed of, and hence they are also called metamaterials (Wu et al., 2019).

This study presents a statistical modelling and monitoring approach for the identification of local geometrical distortions in one (or more) cells within one single part and between parts. The proposed method grounds on the estimation of the "real geometry" of the lattice structure via X-ray computed tomography (CT), and on the comparison with such real geometry with the "nominal geometry", i.e., the originating CAD file. A profile monitoring approach is proposed the model and monitor the evolution of the deviation between the real and nominal geometries of each cell along the building direction (Ramsay, 2004; Woodall, 2007). The method is demonstrated by means of a real case study involving the production of metal lattice structures via Laser Powder bed Fusion (LPBF), an additive manufacturing process where a laser beam is used to locally melt a metal powder bed and produce complex shapes on a layer-by-layer basis.

Section 2 introduces the case study related to the additive production of lattice structures. Section 3 briefly presents an overview of the proposed approach. Section 4 presents the results and Section 5 concludes the paper.

2 The lattice structure case study

Lattice structures combine a lightweight design with high specific stiffness and strength, with appealing performances in industrial sectors like aerospace and racing. Moreover, their isotropic structure and cell geometry-dependent properties make them suitable for several applications, from heat exchange to energy absorption and acoustic insulation. Finally, in the biomedical sector, the trabecular structure is particularly suitable to enhance the osteo-integration and the compatibility with the human tissue. Additive manufacturing methods enable the production of this kind of structures in a wide range of materials and for a wide range of applications. (Colosimo, 2018; Colosimo et al., 2018).

The case study consists of an Al-Si-Mg alloy lattice structure with a designed porosity of 90%, composed by $N = 64$ dodecahedron unit cells of size $l = 10\ mm$ within a specimen of dimension $40 \times 40 \times 40$ mm. Each unit cell consists of 32 prismatic elements with struct diameter of 0.67 mm (Figure 1). The lattice structure was produced via Laser Powder Bed Fusion (LPBF). The as-built structure was inspected by means of an X-ray CT scan system with a resolution or 33 μm.

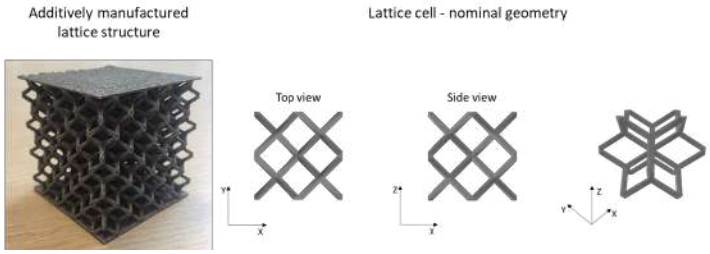


Figure 1: Left panel: picture of the as-built lattice structure; right panel: nominal geometry of the cell

3 Overview of the proposed approach

The underlying idea of the proposed methodology consists of modelling the deviation between the real and nominal geometry for each unit cell of a lattice structure. By modelling such deviation, a statistical monitoring scheme allows one to determine whether one or more unit cells exhibit an out-of-control geometrical distortion with respect to the natural variability of the deviations. Figure 2 shows a schematic overview of the overall method. The real geometry, reconstructed via X-ray CT scan, is sliced along the build direction, Z, into a stack of images: each image represents the real geometry of each cell at a given height along the Z axis. The corresponding image representing the nominal geometry at the same height along Z is obtained by slicing the originating CAD model. By comparing, slice by slice, the real and nominal geometry of each cell, it is possible to estimate a deviation index, which can be represented in terms of a 1D profile along Z. Each unit cell in the structures can therefore be associated to a 1D deviation profile.

The proposed approach consists of using the 1D profile data to synthetize the 3D deviation between the real and nominal geometry of the cell. A profile monitoring approach can then be applied to model these profiles – a B-spline basis was proposed (Ramsay, 2004) – and detect any out-of-control deviation by means of a control charting scheme applied to the B-spline coefficients and the model residuals.

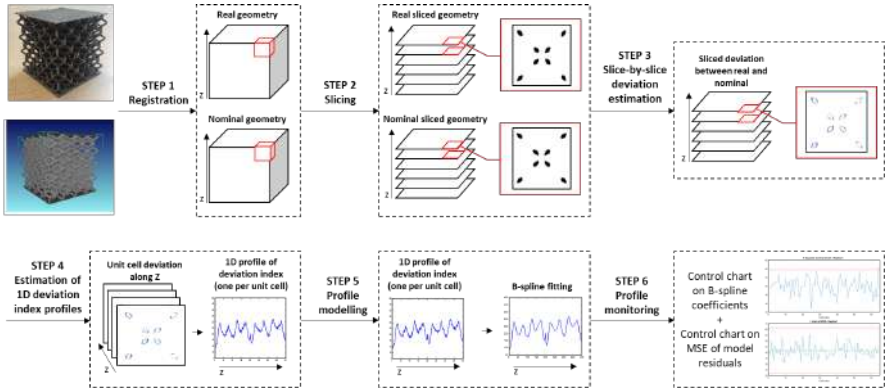


Figure 2: Scheme of the proposed approach

4 Results

Figure 3 shows the deviation profiles for all the unit cells of the lattice structure and the corresponding B-spline model fits. The general pattern of the deviation index profiles can be regarded as a signature of the process, where the main discontinuities correspond to the salient geometrical features of the unit cell (i.e. junctions of the cell struts). The cell-to-cell variability is caused by the small local variations as a result of the additive manufacturing process itself. The general oversizing of the real geometry with respect to the nominal one, can be regarded as a signature of the process too.

The control charts for within-part variability analysis of the printed structure are shown in Figure 4, designed with familywise Type I error α=0.0027.

The T^2 control chart on the B-spline model coefficients signals an out-of-control in correspondence of the 52-th unit cell. In that cell, an actual geometrical distortion was present, possibly caused by an error in the powder recoating while that cell was produced.

The method can be applied to model and analyse the within-part variability, i.e., the cell-to-cell variability in one single part, like in the example here presented, but also to monitor the stability over time of the part-to-part variability in a series production. The capability to identify a geometrical distortion in one (or more) unit cells enables a novel solution for product qualification even in the presence of a one-of-a-kind product, regardless of the complexity and actual geometry of the structure.

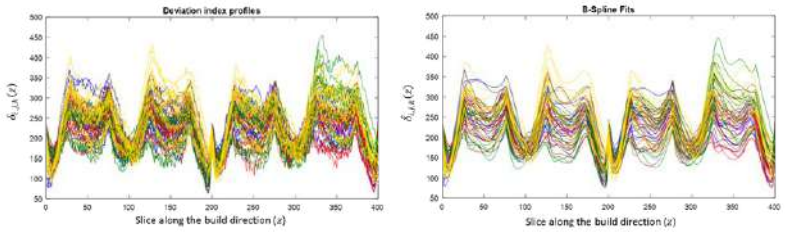


Figure 3: Left panel: original deviation profiles for all the cells; right panel: B-pline fits of the deviation profiles

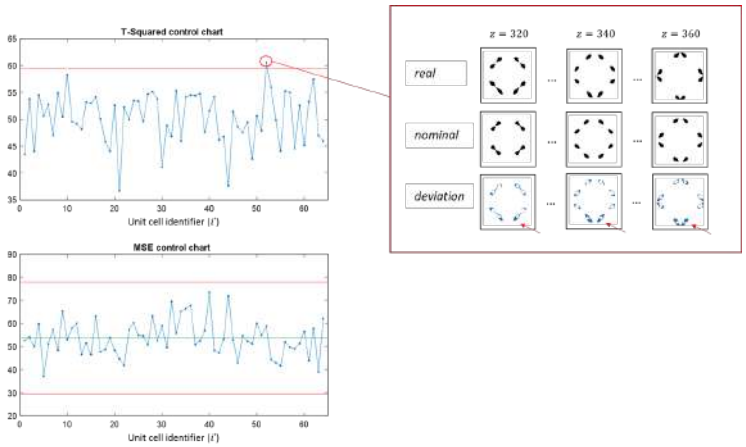


Figure 4: Control charts applied to monitor the within variability of the lattice structure, with an example of the unit cell signalled as out-of-control (real and nominal geometry and deviation).

5 Extendibility to other geometries

One major issue in novel manufacturing paradigm regards the lack of statistical methods to model and monitor the quality of innovative products characterized by complex 3D shapes. This study presented a first approach aimed at tackling this challenge. The underlying idea of studying the deviation between the real and nominal shapes can be extended to different kinds of complex geometries, and it opens up to a wide range of possible applications. This study is inserted into a research framework focused on the quality characterization, modelling and monitoring of complex geometries, not limited to lattice structures. As an example, in another study in collaboration with the Department of Mathematics, Politecnico di Milano (Scimone et al., 2020), the study of methods to synthetize and monitor the deviations between a real and nominal geometry was extended to structures without regularly repeating cells. Generally speaking, the proposed methods can be extended not only to different families of shapes, but also to geometrical reconstructions gathered while the part was produced, rather than at the end of process (Grasso and Colosimo, 2017; Everton et al., 2016; Colosimo and Grasso, 2018). This possibility

is enabled by the layerwise paradigm of additive manufacturing processes and it paves the way to novel in-line and in-situ part qualification capabilities.

Acknowledgements

This research was supported by Accordo Quadro ASI-POLIMI "Attività di Ricerca e Innovazione" n.2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano.

References

1. Bui, A. T., & Apley, D. W. Monitoring for changes in the nature of stochastic textured surfaces. Journal of Quality Technology, 50(4), 363-378. (2018).
2. Cansizoglu, O., Harrysson, O., Cormier, D., West, H., & Mahale, T. Properties of Ti–6Al–4V non-stochastic lattice structures fabricated via electron beam melting. Materials Science and Engineering: A, 492(1-2), 468-474. (2008)
3. Colosimo, B. M., Cicorella, P., Pacella, M., Blaco, M. From profile to surface monitoring: SPC for cylindrical surfaces via Gaussian processes. Journal of Quality Technology, 46(2), 95-113. (2014).
4. Colosimo, B. M. Modeling and Monitoring Methods for Spatial and Image Data. Quality Engineering, 30(1), 94-111. (2018).
5. Colosimo, B. M., Huang, Q., Dasgupta, T., & Tsung, F. Opportunities and challenges of quality engineering for additive manufacturing. Journal of Quality Technology, 50(3), 233-252. (2018).
6. Colosimo B.M., Grasso M. Spatially Weighted PCA for Monitoring Video Image Data with Application to Additive Manufacturing, Journal of Quality Technology, 50(4), 391-417. (2018).
7. Del Castillo, E., Colosimo, B. M., & Tajbakhsh, S. D. Geodesic Gaussian processes for the parametric reconstruction of a free-form surface. Technometrics, 57(1), 87-99. (2015).
8. Everton, S. K., Hirsch, M., Stravroulakis, P., Leach, R. K., Clare, A. T. Review of in-situ process monitoring and in-situ metrology for metal additive manufacturing. Materials & Design, 95, 431-445. (2016).
9. Grasso M., Colosimo B.M. Process Defects and In-situ Monitoring Methods in Metal Powder Bed Fusion: a Review, Measurement Science and Technology, 28(4), 1-25
10. Lee, H. Y., and D. W. Apley. Diagnosing manufacturing variation using second-order and fourth-order statistics. International Journal of Flexible Manufacturing, Systems 16:45–64. (2004).
11. Menafoglio, A., Grasso, M., Secchi, P., & Colosimo, B. M. Profile monitoring of probability density functions via simplicial functional PCA with application to image data. Technometrics, 60(4), 497-510. (2018).
12. Ramsay, J. O. Functional data analysis. Encyclopedia of Statistical Sciences, 4. (2004).
13. Shan, X., and D. W. Apley. Blind identification of manufacturing variation patterns by combining source separation criteria. Technometrics 50:332–343. (2008).
14. Shi, Z., Apley, D. W., Runger, G. C. Identifying and visualizing part-to-part variation with spatially dense optical dimensional metrology data. Journal of Quality Technology, 51(1), 3-20. (2019).
15. Woodall, W. H. Current research on profile monitoring. Production, 17(3), 420-425. (2007).
16. Wu, X., Su, Y., & Shi, J. Perspective of additive manufacturing for metamaterials development. Smart Materials and Structures, 28(9), 093001. (2019).
17. Zang, Y., & Qiu, P. Phase I monitoring of spatial surface data from 3D printing. Technometrics, 60(2), 169-180. (2018a).
18. Zang, Y., & Qiu, P. Phase II monitoring of free-form surfaces: An application to 3D printing. Journal of Quality Technology, 50(4), 379-390. (2018b).
19. Scimone, R., Taormina, T., Colosimo, B.M., Grasso, M., Menafoglio, A., Secchi, P.: Statistical modelling and monitoring of geometrical deviations in complex shapes with application to Additive Manufacturing. Manuscript under submission. (2020)

Open data powered territorial planning - Case study: The Turin historical center

L'Open data per potenziare la pianificazione territoriale – un caso di studio sul Centro storico di Torino

Silvia Casagrande[1], Gianmaria Origgi[1], Alberto Pasanisi[1], Martina Tamburini[1], Pascal Terrien[2], Tania Cerquitelli[3], Alfonso Capozzoli[3]

Abstract This study focuses on the harvesting of energy-related open data to support territorial planning, thus discovering interesting insights useful to support the decision-making process. To this aim, we selected the Energy Performance Certificates because of the heterogeneity of the attributes characterizing buildings, released by the Piedmont region as open data. We derived a data-driven methodology to identify a high-quality set of data to build a baseline energy scenario in a major Italian city. Different urban requalification scenarios were simulated by analyzing financial, environmental, energy and social Key Performance Indicators.

Abstract. Il presente lavoro riguarda il trattamento e l'utilizzo di open data energetici a supporto della pianificazione territoriale, derivandone utili spunti propedeutici al processo decisionale. Sono stati utilizzati gli attestati di certificazione energetica, forniti dalla Regione Piemonte, per via dell'eterogeneità degli attributi degli edifici. Su questa base, è stata sviluppata una metodologia per identificare un set di dati di alta qualità e costruire una scenaristica di riqualificazione energetica per una grande città italiana, analizzando indicatori finanziari, ambientali, energetici e sociali.

Keywords: Open data, energy performance certificates, urban planning.

1 Introduction

A key research challenge in the context of open data (i.e., data freely available to everyone to use and republish as they wish) is to use them in real-life applications

[1] Edison Spa. Foro Buonaparte, 31. 20121 Milano. Contact : alberto.pasanisi@edison.it

[2] EIFER. Emmy-Noether-Str. 11, 76131 Karlsruhe, Germany

[3] Politecnico di Torino. Corso Duca degli Abruzzi, 24. 10129 Torino

(including business cases) with no restrictions from copyright, patents or other control mechanisms. However, the effective exploration of such data in real-life settings is very challenging and requires a lot of expertise in the field of open data and a lot of effort for extracting useful knowledge from poor quality data as frequently open data are.
This study focuses on the harvesting of open data for the characterization of the energy performance of buildings with the final goal of creating value for different stakeholders. In particular, the research is aimed at exploring how open data on energy consumption of buildings can be exploited to design and appraise urban requalification policies. After a detailed census of data related to buildings issued as open, a database of Energy Performance Certificates (EPCs) was selected. It provides a variety of heterogeneous information related to each certified building including the standard-based calculation of energy performance, thermo-physical and geometrical features and geospatial information. The exploration of such energy-related database to support business cases and providing interesting insights to various stakeholders is challenging because of the heterogeneity of the attributes, the need of both energy and data science expertise. The available data set was used to build a baseline energy scenario. Then, several urban requalification scenarios were simulated by analyzing financial, environmental, energy and social KPIs.
The paper is organized as follows. Section 2 describes the work done to transform the raw available open data in valuable input information for building requalification scenarios. Then, in Section 3, the scenarios built and their evaluation (by means of a specific urban planning digital tool) are presented. Finally, Section 4 stresses the main conclusions and perspectives of this study.

2 Treatment of Open Data

In this work a dataset consisting of a number of EPCs was analysed; it was collected in the Piedmont Italian Region during the period from 2009 to 2018. The dataset has been gathered and openly released by CSI Piemonte (the regional Information System Consortium) and regulated by the Piedmont Region authority (Sustainable Energy Development Sector). The dataset includes nearly 270,000 EPCs. In this study, the portion of EPCs related to buildings located in Turin was mainly analysed, because of the high number of data available in this area (47,623).

Data Cleaning. Since EPCs are available as open data, different kind of inconsistencies could be present. In this context, the data pre-processing is a fundamental step to handle good quality data in the next analytics steps. As in [1, 2] two main data cleaning steps were performed: (i) geospatial data cleaning and (ii) outlier detection and removal.

Geospatial data cleaning. Since the final goal of the analytics process consists in displaying the energy performance of the buildings on a map, a good quality of geospatial data is an essential prerequisite. Considering that addresses, house numbers, ZIP codes and coordinates included in EPCs belong to open text fields, the potential number of errors is not negligible. As in [1, 2], the addresses reported in

each EPC were compared with the city's street database and then matched if the Levenshtein similarity is higher than a given threshold (i.e. 0.95 in our case study).

The Levenshtein similarity is based on the computation of the Levenshtein distance, which is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. In case of lower similarity (i.e. the minimum threshold is not satisfied during the comparison), geocoding requests have been sent via the Google GeocodingAPIs[1] to reconstruct EPCs' addresses and coordinates. Thanks to the proposed solution, 99% of the addresses have been resolved correctly.

Outlier detection and removal. Open data are often affected by the presence of outliers, which are observations that deviate markedly from the data. Therefore, a semi-supervised outlier detection was computed based on both univariate and multivariate analysis. Specifically, for the univariate outlier analysis three stages were performed:

1) *Definition of acceptability ranges for each energy-related attribute* defined with a detailed expert analysis. The final acceptability ranges for a relevant set of features are shown in Table 1

Table 1: Example of EPCs' variables with the defined validity range

Feature (Acronim)	***Unit***	***Validity Range***
Normalised Primary heating energy consumption	[KWh/m^2]	[0 - 682]
Aspect ratio (S/V)	[m^{-1}]	[0.1 – 2]
Surface area (SA)	[m^2]	[24.9 - 880]
Floor area (FA)	[m^2]	[21.5 - 296]
Average U-value of the vertical opaque envelope (UO)	[W/m^{2}K]	[0.15 - 3]
Average U-value of the windows (UW)	[W/m^{2}K]	[0.9 - 7]
Heating system global efficiency (ETAH)	--	[0.3 - 1.06]
Construction year (Year)	--	[1700 - 2018]

2) *Univariate outlier detection with generalised Extreme Studentised Deviate (gESD)* [3]. It requires the number of outliers defined by the energy expert. Given the upper bound value for the number of outliers, the gESD test essentially performs r separate tests: a test for one outlier, a test for two outliers, and so on up to *r* outliers. In our use case, *r* was set to 0.5% of EPCs and the critical value to 0.01

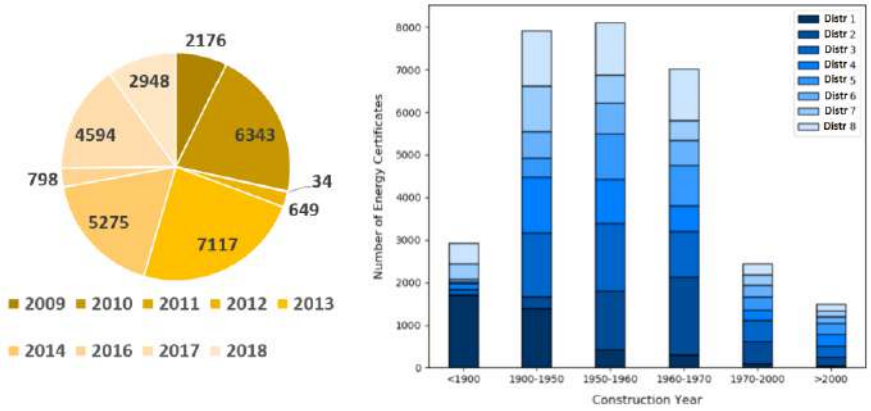
3) *Analysis of the frequency data distribution to identify outliers in the first percentile.* The records belonging to the first percentile were removed, to eliminate inconsistent values with a low deviation from the rest of the observations.

Lastly, the presence of outliers with a multivariate analysis were discovered by applying the Density-Based Spatial Clustering of Application with Noise algorithm (DBSCAN) [4]. In particular, DBSCAN identifies clusters of data by searching higher-density regions separated by lower-density regions. In order to automatically set the DBSCAN input parameters (*MinPoints* and *Epsilon*), the methodology presented in [5] was employed, thus *MinPoints*=5 and *Eps*=0.28 were set. The final

[1] https://developers.google.com/maps/documentation/geocoding/intro

dataset includes roughly 30,000 EPCs, issued in the period 2009-2018 (distributed as shown in Figure 1 (left)), and related to buildings built on different ranges of construction year (see Figure 1 (right)) and distributed across all districts in Turin.

Figure 1: (left): Pie-chart distribution of the number of EPCs for each year under analysis; (right) stacked bar graph of the number of EPCs for each district of the Turin city separately for different ranges of construction year of the dwellings.

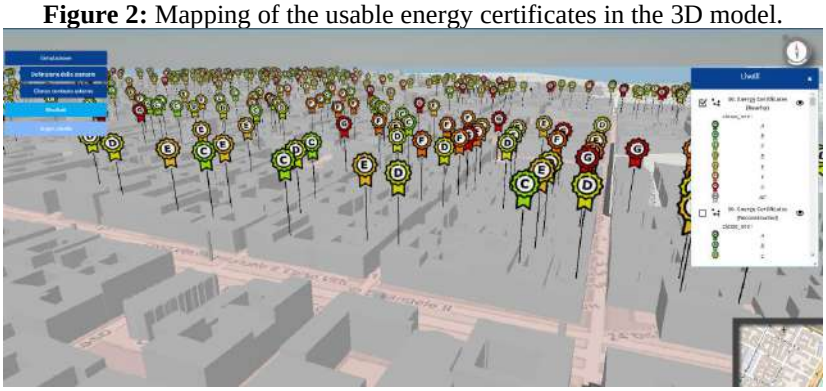


3 Building and evaluating urban renovation scenarios

Starting from the data treatment described in Section 2, the software *City Platform* was customized on a part of the City to build and evaluate renovation scenarios.
City Platform is a digital decision support system tool (DDS) with the main purpose to support urban decision makers in designing, evaluating and comparing different strategies, in order to find the one that best fits a given set of objectives, represented by a number of KPIs (financial, energy-related, environmental, social, quality of life etc.). It was used worldwide [6, 7] with significant applications in Singapore, Berlin, Shanghai, Moscow. The customized version on Turin was concerned with the improvement of the urban buildings stock: reducing the energy consumption, costs and CO_2 emissions, decreasing management costs and creating positive impacts on local economy.
The implementation started with a careful analysis of the city urban fabric, which led to the identification of the historical city centre of Turin as the working area. This area was selected for the presence of several public buildings of interest, the high density of historical buildings with different technological and architectural features and the large potential of renovation.
The following KPIs were considered for this study: (i) private investments (€), (ii) public subsidies (€), (iii) CO_2 emissions (%), (iv) PM_{10} emissions (%) (v) energy consumptions (%), (vi) local jobs created by the investments (n. of employees).
Attributes of interest from EPCs dataset include envelope thermal transmittance, system efficiency, energy class, specific primary energy consumption, architectural

features, the period of construction and the building typology (function, public or private, historical or not, etc.). Additional information necessary to build the model and related to spatial data for describing the urban fabric, was gathered from the topographic database of Piedmont Region.
The study consisted first in creating a 3D model of the selected city district using the regional topographic database. The energy performance and the construction period of each building along with the results of the simulated scenarios were then displayed on this 3D model (see Figure 2).

Figure 2: Mapping of the usable energy certificates in the 3D model.



Different scenarios were designed gathering interventions on energy systems and envelope. For each building, interventions are compliant with architectural restrictions, the existing situation and the eligibility to public subsidy (Ecobonus). A preliminary evaluation of the IRR and ROI also assure that proposed interventions are realistic by an economic viewpoint.
The energy performance of each building after renovation was evaluated by a steady state calculation method according to ISO 13790 standard. For the purpose of the study, 3 scenarios were defined (each one is scheduled on 10 years):
A) Moderate improvement: replacement of all diesel and oil boilers with gas ones the envelope refurbishment and BEMS. It involves 22 buildings.
B) Significant improvement: deep improvement (envelope, boiler, BEMS) of 50% of all the eligible residential stock (approx. 300 buildings). In addition to the buildings refurbished in Scenario A, it includes also non-gas boiler buildings.
C) High improvement: This was the most ambitious scenario. It is like scenario B, but interventions concern 100% of the eligible stock (approx. 600 buildings).
The effectiveness of each combination of measures (scenario) was evaluated by assessing the value of several KPIs (Table 2). The scenario C led to particularly good results considering the number of buildings renovated, the total amount of the investments, the improvement of environmental quality of the city and local job created by the interventions.

Table 2: Main KPI from the analysis of the 3 scenarios of urban renovation

KPI	*Unit*	*Scenario A*	*Scen. B*	*Scen. C*
Private investment after public subsidy	M€	14	54	96

Total investment	M€	32	116	205
Reduction of CO2 emissions	%	18	35	53
Reduction of PM10 emissions	%	23	38	55
Decrease of energy consumptions (heating)	MWh	69,455	137,236	212,307
Jobs created (full time equivalent)	n.	169	611	1081

4 Conclusions and perspectives

This work highlights the huge potential of open data, associated to powerful data cleaning, pre-processing tools and urban simulation capacities. As a main outcome of the study, we mention the impact consequent to the energy renovation policy at urban scale: more than 50% cut on CO_2 emission and 1000 full time equivalent jobs created (based on the simulated scenario). As a major perspective, the analysis reported in this paper can be easily extended to other cities and territories in Italy (in both urban and suburban areas) and in other countries, in particularly in European Union, where ECPs are established since 2002 (Directive 2002/91/EC) as a standard tool to foster energy efficiency. However, the study has also shown that often the quality of available EPCs is not satisfactory. Having a large quantity of good-quality data, coupled with powerful decision support tools (like City Platform) could be a game-changer for local powers, and more generally for urban decision maker. Finally, this study proved the usefulness of tailor-made digital decision support tools in urban planning, to build and evaluate scenarios, but also to define realistic goals to be achieved by urban policies.

Acknowledgements. The authors gratefully thank Alexander Simons, Isaac Boates, Monjur Syed Murshed, Wanji Zhu, Samuel Thiriot (EIFER) and Alexandru Nichersu (KIT), for the great work made on City Platform in this study.

References

1. Cerquitelli, T. Di Corso, E. Proto, S. Capozzoli, A. Bellotti, F. Cassese, M.G. Baralis, E. Mellia, M. Casagrande, S. Tamburini, M. Exploring Energy Performance Certificates through Visualization. In Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference (EDBT/ICDT 2019) Lisbon, 2019.
2. Cerquitelli, T. Di Corso, E. Proto, S. Capozzoli, A. Mazzarelli, D. M. Nasso, A. Baralis, E. Mellia, M. Casagrande, S. Tamburini, M. Visualising high-resolution energy maps through the exploratory analysis of energy performance certificates. In Proceedings of the IEEE SEST 2019, Porto, 2019.
3. Seem, J.E. Using intelligent data analysis to detect abnormal energy consumption in buildings, In Energy and Buildings, vol. 39, no. 1, pp. 52–58, 2007.
4. Ester, M. Kriegel, H.-P. Sander, J. Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996.
5. Ankerst, M. Breunig, M. M. Kriegel, H.-P. Sander, J. Optics: Ordering points to identify the clustering structure, in ACM SIGMOD Record, vol. 28, pp. 49–60, ACM, 1999.
6. Plessis, G. Pons, M. Pasanisi, A. Data for Planning and Monitoring Smart and Sustainable Cities: Downscaling and Validation Perspectives. ENBIS 2017 Conference, Naples, 2017.
7. Pasanisi, A. Plessis, G Koch, A. Data, methods and tools in support of smart and sustainable cities planning: an insight. SMSA2017 Conference, Berlin, 2017.

Process optimization in Industry 4.0: Are all data analytics models useful?

Ottimizzazione dei processi nell'industria 4.0: tutti i modelli di analisi dei dati sono appropriati?

Alberto Ferrer

Abstract: Not all data analytics models are equally useful in industry 4.0. Machine learning models are currently often used in many Big Data projects in Industry 4.0. These models are proven to be very powerful for passive applications (predictive modelling and maintenance, pattern recognition and classification, and process monitoring, fault detection and diagnosis). However, its indiscriminate use, no matter the type of application, can be counterproductive. They should not be used for process optimization unless data come from a design of experiments (what is a severe limitation in industrial practice). On the other hand, predictive methods based on latent variables (such as partial least squares regression) can be used for process optimization regardless of whether the data come from a design of experiments or daily production process (historical/happenstance data).

Abstract: *Non tutti i modelli di analisi dei dati sono ugualmente appropriati nell'industria 4.0. Attualmente, i modelli di Machine Learning sono spesso utilizzati in molti progetti di Big Data nell'Industria 4.0. Questi modelli hanno già dimostrato di poter garantire performance molto soddisfacenti per applicazioni passive come la modellazione e la manutenzione predittiva, il riconoscimento di pattern, la classificazione, il monitoraggio di processo ed il rilevamento e la diagnosi di guasti operazionali. Tuttavia, il loro utilizzo indiscriminato può essere controproducente: non è possibile, ad esempio, ricorrere ad essi al fine di ottimizzare un processo industriale (applicazione attiva) a meno che i dati analizzati non siano stati raccolti a partire da un disegno sperimentale (il che costituisce una grave limitazione dal punto di vista pratico). D'altro canto, i metodi predittivi basati su variabili latenti (come la regressione ai minimi quadrati parziali) possono essere utilizzati per*

[1] Alberto Ferrer, Universitat Politècnica de València (Spain); aferrer@eio.upv.es

l'ottimizzazione del processo indipendentemente dal fatto che i dati provengano da un progetto di esperimenti o da un processo di produzione giornaliero (dati storici).

Key words: process optimization, industry 4.0, latent variable methods, PLS model inversion, quality by design (QbD).

1 Introduction

Industry 4.0 paradigm is being adopted by industry and technology fostered by the Industrial Internet of Things (IIoT). Processes not only produce goods or provide services but also data, and there is a strongly belief that data contain useful information that has to be mined for helping the decision-making process (Ferrer, 2020). This is leading to the so-called Big Data environment, characterized by the four V′s: volume, variety, velocity and veracity. This has not caused only a change in the number of the variables but also a change in the nature of the registered data, coming from customers, quality properties and processes, and even from equipment. There are a wide variety of sensors providing different type of signals: spectra (chemical signals), pressures, temperatures, flows, etc. (physical signals), pH, conductivity, dissolved oxygen, etc. (biochemical signals), electronic eyes (digital images), electronic noses and tongues (potentiometric signals), electronic ears (acoustic signals), and so on. These data mostly collected from routine daily production often exhibit high auto and cross correlation, rank deficiency, low signal-to-noise ratio, multi-stage and multi-way structure, and missing values. In most of the cases they are happenstance data (i.e., data from daily routine production and not generated from any experimental design) and, therefore, correlation does not necessarily mean causation.

Process data in industry, although shares many of the characteristics represented by the four V′s may not really be Big Data in comparison to other sectors such as social networks, sales, marketing and finance. However, the complexity of the questions we are trying to answer with industrial process data is really high, and the information that we wish to extract from them is often subtle. This info needs to be analyzed and presented in a way that is easily interpreted and that is useful to process engineers. Not only do we want to find and interpret patterns in the data and use them for predictive purposes, but we also want to extract meaningful relationships that can be used to improve and optimize a process (García-Muñoz and MacGregor 2016).

Latent variables (LV) models, such as principal component analysis–PCA (Jackson 2003) or partial least squares–PLS (Wold et al 2001), are especially suited for successfully addressing the characteristics 4 V′s of Big Data. They are compressing tools that handle the dimensionality and collinearity issues of the high volume of data. They can cope with the variety of data by using multiblock methods (Westerhuis 1998) for integrating data from different sources (data fusion). LV models can be updated in real time to cope with the speed of data acquisition

(velocity) and are especially suited for outlier detection, missing and noisy data, typical issues for checking the data veracity.

LV methodology exploits the correlation structure of the original variables by revealing the few independent underlying events (latent variables) that are driving the process at any time. This is done by projecting the information in the original variables down onto low-dimensional subspaces defined by a few LV (i.e. scores). The multivariate scores are mathematically orthogonal and optimal summaries of the measured variables. The scores are also less noisy than the measured variables, because they are weighted averages (linear combinations) of the measured variables. Classical statistical assumptions (independency, normality and so on) which in general are not appropriate for the original variables recorded, can be reasonable for the scores, and therefore classical statistical tools are appropriate to analyze them. We could conclude that in the latent space, dealing with big data is easier than in the original variables space (Ferrer 2020).

2 Are all data analytics models useful?

MacGregor (2018) states that to analyze historical data, one needs to make use of models, usually empirical – such as linear regression-LR, machine learning-ML (e.g., deep learning, random forest, etc.) or LV models. All models are wrong but some are useful (Box 1976). But all empirical (i.e. data analytics) models are not equally useful. Whether a data analytics model is useful depends on three issues: i) the objectives of the model (passive vs active); ii) the nature of the data used for the modeling (historical operating data vs data from design of experiments–DOEs); and iii) the regression method used to build the model (ML and classical LR vs LV models such as PLS).

Regarding the objective, there are two major classes of models – those to be used for passive use and those to be used for active use. Models for passive use are intended to be used just to passively observe the process in the future (e.g. predictive modelling and maintenance, pattern recognition and classification, and process monitoring, fault detection and diagnosis). For such passive uses one does not need or even want causal models, rather one wants to just model the normal variations common to the operating process. Historical data are ideal for building such models. On the other hand, models for active use are intended to be used to actively alter the process to gain causal information (i.e., process understanding) from the data (e.g., trouble-shooting, optimization and control). For active use one needs causal models. Causality implies that for any active changes in the adjustable variables in the process, the model will reliably predict the changes in the output of interest (MacGregor 2018).

To guarantee causality when using data-driven approaches, however, independent variation in the input variables is required (Box et al 2005). While a Design of Experiments (DOE) on the plant would provide data satisfying this requirement, in practice such approach may be difficult to carry out, if not unfeasible (i.e. the number of potential factors to consider as inputs can be really

high, and due to the complex correlation structure among them there are a lot of restrictions that prevent moving some factors independently from others). On the other hand, nowadays large amounts of historical data are available in most production processes. The problem is that this data is highly collinear and low rank because the variation in the inputs is commonly not independent (i.e., data are not obtained from a DOE that guarantees this independent variation in the inputs). Therefore, input-output correlation does not mean necessarily causation. In this context, classical predictive models (such as LR and ML), proven to be very powerful in passive applications, cannot be used for process optimization (active use). They cannot be used for extracting interpretable or causal models from historical data for active use. With historical data, there are an infinite number of models that can arise from any of these LR or ML methods, all of which might provide good predictions of the outputs, but none of which is unique or causal. Because the process variables are all highly correlated and the number of independent variations in the process is much smaller than the number of measured variables, one can get many of those models all using different variables and having different weights or coefficients on the variables that give nearly identical predictions. This does not allow for meaningful interpretations, even more so if the results come from averaging or voting on many models, such as in random forest (MacGregor 2018). This is the essence of the Box et al (2005)´s warning: predictive models based on correlated inputs must not be used for process optimization if they are built from observational data (i.e. data not coming from a DOE). Note that this is a mistake that is beginning to occur in industry 4.0 with the indiscriminate use of ML tools.

Methods based on LV, such as PLS regression, allow the analysis of large datasets with highly correlated data. Since they assume that the input (X) space and the output (Y) space are not of full statistical rank, they not only model the relationship between X and Y (as classical LR and ML models do), but also provide models for both the X and Y spaces. This fact gives them very nice properties: uniqueness and causality in the reduced latent space (this is the only space within which the process has varied) no matter if the data come either from a DOE or daily production process (historical/happenstance data) (Liu et al 2011, MacGregor et al 2015). These properties make them suitable for process optimization (active use) no matter where the data come from.

3 Process optimization through latent variables (PLS) models

As already commented, model-based process optimization requires building a causal model that relates changes in the process inputs with those in the process outputs. To this purpose, deterministic (i.e. first principles) models are always desirable. However, the lack of knowledge and the generally ample need of resources required to properly construct such models makes their use unfeasible in a large number of cases, and data-driven models are often resorted to, instead.

Optimizing a process using predictive data-driven methods that directly relate the registered input variables with the output variables (such as LR and ML models) requires causality in this input-output relationship, and this is only guaranteed if data are obtained from a DOE, what is highly difficult (or even unfeasible) to get in industrial practice. Nevertheless, LV models, such as PLS, can model causality in the latent space and, therefore, can be used for process optimization even with happenstance data[2].

By moving the LV one can reliably predict the outputs (Y) But to move the latent variables one cannot just adjust individual X variables, but rather combinations of the X variables that respect the correlation structure of the model[3].

This property makes them suitable to be used in process optimization for finding the so-called design space (DS), i.e. the combinations of input variables that are consistent with the historical correlation structure and region where the process has been operated, and also guarantee the desired outputs (Jaeckle and MacGregor 2000, MacGregor et al 2015).

In order to apply the QbD initiative, two distinct strategies have been proposed in the literature (Palací-López et al 2020): i) defining or estimating the DS as a whole, and ii) solving an optimization problem in an attempt to obtain single sets of process conditions within the DS.

The first of these strategies, when resorting to LV models, relies on the so-called null space (NS), i.e. the subspace in the latent space within which the prediction of the outcome responses does not vary (Jaeckle and MacGregor 2000, García-Muñoz et al 2006), for which the uncertainty in its definition can also be accounted for (Facco et al 2015, Bano et al 2018, Palací-López et al 2019). Regarding the second strategy, since the initial number of variables involved is reduced to a smaller number of uncorrelated LV, the computational cost of any optimization problem in the latent space will decrease with respect to the same problem in the original space (Palací-López et al 2019, Tomba et al 2012).

4 Conclusions

Not all data analytics models are equally useful in industry 4.0. The usefulness of a model depends on the objective of the study (passive or active), the nature of the data (historical operating data vs. data from design of experiments – DOEs) and the technique used (ML and classical LR vs LV models such as PLS). ML models are currently being implemented in many Big Data projects in Industry 4.0. Although

[2] This is not the same as optimizing via LR or ML models. In such case, causality cannot be inferred and there is no guarantee that the solution would respect the correlation structure of the data, leading to unfeasible solutions.

[3] The latent variables cannot be explicitly manipulated by the user, but the original variables can be manipulated in a way that changes on the process conditions are done along the directions of the latent variables, which is equivalent to implicitly "manipulating" the latent variables themselves.

these models are proven to be very powerful for passive applications (e.g. predictions) they cannot be used for extracting interpretable or causal models from historical data for active use (e.g. process optimization). On the contrary, predictive methods based on LV (such as PLS) much less known in Big Data environments, have two important properties: uniqueness and causality in the latent space, which allow them to be used for process optimization regardless of whether the data come from a DOE or daily production process (historical / happenstance data). This is illustrated with several real industrial examples.

References

1. Bano, G., Facco, P., Bezzo, F., Barolo, M.: Probabilistic Design Space Determination in Pharmaceutical Product Development: A Bayesian/Latent Variable Approach. *AIChE J.* 64(7):2438–2449 (2018)
2. Box, G.E.P.: Science and Statistics. *J. Am. Stat. Assoc.* 71(356):791-799 (1976)
3. Box, G.E.P., Hunter, W.G., Hunter, J.S.: *Statistics for Experimenters: Design, Discovery and Innovation*. 2nd ed. Hoboken, NJ: John Wiley and Sons (2005)
4. Facco, P., Dal Pastro, F., Meneghetti, N., Bezzo, F., Barolo, M.: Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development. *Ind. Eng. Chem. Res.* 54(18):5128–5138 (2015)
5. FDA.: *Pharmaceutical CGMPs for the 21s Century—A Risk-Based Approach* (2004). https://www.fda.gov/media/77391/download
6. Ferrer, A.: Latent Structures-Based Multivariate Statistical Process Control: A Paradigm Shift. *Qual. Eng.* 26(1):72-91 (2014)
7. Ferrer, A.: Discussion of "A review of data science in business and industry and a future view" by Grazia Vicario and Shirley Coleman. *Appl. Stochastic Models Bus. Ind.* 1–7 (2020)
8. García-Muñoz, S., Kourti, T., MacGregor, J.F., Apruzzese, F., Champagne, M.: Optimization of Batch Operating Policies. Part I. Handling Multiple Solutions. *Ind. Eng. Chem. Res* 45(23):7856–7866 (2006)
9. García Muñoz, S., MacGregor, J.F.: Big Data. Success Stories in the Process Industries. *Chemical Engineering Progress* 112(3):36-40 (2016)
10. Jaeckle, C.M., MacGregor, J.F.: Industrial applications of product design through the inversion of latent variable models. *Chemom. Intell. Lab. Syst.* 50:199–210 (2000)
11. Jackson, J.E.: *A User's Guide to Principal Components*. New York: Wiley (1991)
12. Liu, Z., Bruwer, M.J., MacGregor, J.F., Rathore, S., Reed, D.E., Champagne, M.J.: Modeling and Optimization of a Tablet Manufacturing Line, *J. Pharm. Innov*. 6: 170–180 (2011)
13. MacGregor J.F. Empirical Models for Analyzing "big" data-what´s the difference. In: *Spring AIChE Conf.*, Orlando, Florida, USA (2018)
14. MacGregor, J.F., Bruwer, M.J., Miletic, I., Cardin, M., Liu, Z.: Latent variable models and big data in the process industries, *IFAC-PapersOnLine*. 28:520–524 (2015)
15. Palací-López, D., Facco, P., Barolo, M., Ferrer, A.: New tools for the design and manufacturing of new products based on Latent Variable Model Inversion. *Chemom. Intell. Lab. Syst.* 194:103848 (2019)
16. Palací-López, D., Facco, P., Barolo, M., Ferrer, A.: Improved formulation of the Latent Variable Model Inversion-based optimization problem for Quality by Design applications. *J. Chemom.* (2020) e3230
17. Tomba, E., Barolo, M., García-Muñoz, S.: General framework for latent variable model inversion for the design and manufacturing of new products, *Ind. Eng. Chem. Res.* 51:12886–12900 (2012).
18. Westerhuis J.A., Kourti T., MacGregor J.F. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics* 12(5):301–321 (1998)
19. Wold S., Sjöström M., Eriksson L. PLS-Regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58:109–130 (2001)

Technology and demographic behaviours (AISP-SIS)

Internet and the Timing of Births

Internet e le Tempistiche di Fecondità

Maria Sironi, Osea Giuntella, and Francesco C. Billari

Abstract Technological innovations directly related to fertility have been linked to the timing of births, i.e. with postponement in the case of contraceptive technology and with "recuperation" in the case of assisted reproductive technology. We argue that the diffusion of the Internet also plays a role as an "enabling" factor in fertility choices, with a particular effect on the timing of fertility. We hypothesize that Internet access contributes to the postponement of fertility during the earlier ages and stages of the life course, and to a recuperation in later ages and stages of the life course. We use data drawn from different sources to test our hypothesis. Consistent with our prior, results provide evidence that high-speed Internet decreases fertility among young women and increases adult fertility among high-skilled women.

Abstract *Le innovazioni tecnologiche correlate alla fecondità sono state direttamente associate all'età a cui si hanno i figli, come per esempio nel caso dei contraccettivi che permettono di posticipare la nascita dei figli o della riproduzione assistita che permette di avere figli in età più avanzata. Questo lavoro ipotizza che anche la diffusione di internet giochi un ruolo importante nelle tempistiche di fecondità, aiutando a posticipare la nascita dei figli in giovane età e a velocizzare il processo in età più matura. Per verificare la nostra ipotesi abbiamo utilizzato diverse fonti di dati e i risultati confermano che internet riduce la fecondità per le donne giovani e la aumenta per le donne adulte e altamente qualificate.*

Key words: Internet, Fertility, Teenage Pregnancy

1 Introduction

Has the effect of technological change on the timing of fertility reached its peak with the introduction of modern contraceptives (and perhaps household appliances)? We here argue that the incredibly fast diffusion of digital technologies, and the Internet in particular, might play an important role in shaping the transition to parenthood. The diffusion of the Internet has been incredibly fast. In the U.S., Internet usage rates grew from 5 percent to 74 percent between 2000 and 2009 (Dettling, 2017). The fast diffusion of the Internet might have had effects on teen birth rates and to the postponement of fertility. These effects are potentially offsetting, and there could be opposite forces at work. Some forces could contribute to an earlier timing of fertility, and potentially to higher teen birth rates. This is the case of partnership formation an easier access to a wider market for partners who can be prospective co-parent, or of the easier access to sexual stimuli. Other forces could contribute to a postponement of fertility, in particular information and social interaction. A third set

[1] Maria Sironi, University College London; email: m.sironi@ucl.ac.uk
Osea Giuntela, University of Pittbsurgh and IZA; email: osea.giuntella@pitt.edu
Francesco C. Billari, Bocconi University; email: francesco.billari@unibocconi.it

of forces could contribute to a "recuperation" effect: if the Internet allows to lower the costs of combining work and parenthood, the effects of the Internet would change and potentially reverse over age. Moreover, social stratification, and the presence of "digital divides", might contribute to differential effects across different social strata, as even "digital natives" are socially stratified in terms of Internet skills and uses (Hargittai, 2010).

In this paper, we use data from multiple sources to analyze the effects of Internet on fertility over the life-cycle. First, using individual-level data from the National Longitudinal Study of Youth 1997 (NLSY97) we study the relationship between Internet access and the propensity to give birth among young men and women (under 31). We test a series of hypotheses on the timing of fertility and on differences between genders, educational and occupational groups. Our findings show a significant postponement effect of the Internet on the timing of births. We complement this analysis using data from the Current Population Survey which have less precise information on fertility, but because of the larger size of the sample allows us to use an instrumental variable approach using regional Internet penetration and to further explore the heterogeneity of the effects by age, education, and marital status. Finally, we provide further evidence on the relationship between Internet and fertility using county-level data. In particular, we exploit data from the Natality Detail Vital Statistics covering the universe of births occurring in counties with at least 100,000 inhabitants and merge these data with information on broadband deployment across US counties drawn from the Federal Communication Commission.

2 Empirical Analysis

2.1 Longitudinal analysis using NSLY97

We analyse fertility of younger adults using the National Longitudinal Survey of Youth 1997 for the United States (NLSY97). The NLSY97 is an ongoing, nationally representative longitudinal study of 8,984 youths who were 12 to 16 years old in 1997. It started in 1997 and people in the sample are interviewed every year. The most recent wave that is available has been collected in 2011-2012, when surveyed individuals were between 27 and 31 years old. Since our aim is to study how the diffusion of the Internet is associated with childbearing and the timing of childbearing, our key independent variables are built upon the question: "Do you currently have access to the internet?"

Because we are interested in fertility timing, and how this decision of having a child is connected to Internet access, we look at two different points in time for which we present our descriptive statistics. We focus first on the period 2003-2006, when the respondents are between 19-23 and 22-26, and then on the period from 2007 to 2011, when they are between 23-27 and 27-31 years old. By dividing the analysis into these two time periods we can test the hypothesis that access to Internet is negatively associated with fertility in young ages, and positively associated to fertility at older ages.

In order to test the hypothesis that effects are more marked for women than for men we run all analyses separately for men and women. Moreover, given that we

expect that the association between Internet access and fertility is stronger for individual with high socioeconomic status and higher educational background, and we want to test whether results are explained by partnership formation, we take into account the highest grade completed by each individual, parents' level of education, and partnership status in our multivariate analysis.

After presenting some descriptive statistics, we use discrete-time event history analysis regression models (Allison, 1982) to study the association between Internet access and timing of births. More specifically, we restrict our analysis to individuals who haven't had a child before 2003, so that everyone in the sample starts with a parity of zero. This leaves us with 4,312 individuals. To take into account the fact that some individuals may be more at risk than others for reasons that are not fully captured by the variables included in the model, we run discrete-time multilevel event history logistic models, in which individual birth episodes are nested within individuals (Barber et al., 2000). Multilevel event history models allow us to introduce random effects, which represent individual-specific unobservables. Recurrent events give a two-level hierarchical structure: episodes (i.e. birth of a child) are clustered into individuals (i.e. mother or father). We follow individuals in the sample over time and we look at when they have their first child and subsequent children.

Our regression models include some time-varying covariates other than the key explanatory variable related to Internet access such as age, partnership status, birth parity, region of residence, if living in an urban or rural area, enrolment in school, and educational attainment. They also include time-constant variables, i.e. race, parents' education, and family income in 1997. All the models are run separately for men and women (see Table 1 as an example for men).

2.1.1 Results using NLSY97

Our first two hypotheses seem to be partially confirmed, with a negative association between Internet access and fertility at younger ages, and a less negative relationship and eventually positive relationship at older ages. Our third hypothesis, that the association of Internet and fertility is stronger for women, doesn't seem to be confirmed. Our fourth hypothesis, that the relationship between the Internet and fertility is stronger among those with higher socio-economic status (highly educated parents), is confirmed for both men and women. And finally, our "partnership' hypothesis is confirmed for women but not for men: the association between Internet access and fertility doesn't change based on the presence of a partner among men, but it does among women. The evidence presented so far suggests a significant negative association between access to high-speed Internet and fertility of young men and women. A main caveat is that these results cannot be interpreted causally as there could be unobserved factors not accounted for by the multi-level event-history models that may be correlated with both access to highspeed Internet and fertility.

To partially address this limitation, we complement the analysis exploiting geographical variation in the penetration of broadband internet across areas. In practice, we estimate the relationship between Internet penetration within a county and childbirth. Due to the small sample size we are unable to conduct a two-stage

relationship instrumenting individual access to Internet with the average access in the county, the instrument is too weak. But we document the reduced-form relationship between average access to Internet in the county and fertility patterns. While the coefficients are not precisely estimated given the small sample size available in the NLSY, the pattern is consistent with the hypothesis that Internet may decrease fertility among the very young women, while it may have an opposite and positive effect among older women.

2.2 Individual Analysis Using CPS Data

To further explore the heterogeneity of effects across sociodemographic groups, we complement the analysis using data from the Current Population Survey. The CPS is a monthly U.S. household survey conducted jointly by the U.S. Census Bureau and the Bureau of Labor Statistic. The CPS contains a battery of labor force and demographic questions and since 1997 periodically collects supplemental information on Internet usage and since 2000 on the type of internet connection (Dial-up, faster-connection type such as DSL (Direct Subscriber Line), cable modem, satellite, mobile broadband service, or fiber-optic service. This information is available in the Computer and Internet Use Supplement for the years 2000 and 2001 and in the October Education Supplement in years 2003, 2007 and 2009. Importantly for the purpose of our project, the CPS also contains information on the number of own children in the household, the age of the youngest own child and then number of own children under age 5 which can be used to analyse fertility decisions. We focus on access to high-speed internet distinguishing fast connections (DSL, fiber optic cable, coaxial cable, wireless technology and satellite) from dial-up Internet.

The large size of the CPS sample allows us to adopt an instrumental variable approach. We instrument Internet access with the average access to high-speed internet in the respondents' state. Our instrument is a strong and relevant predictor of internet access and the first-stage F statistic is well above the conventional weak-instrument thresholds. All estimates include controls for a quadratic in age, race dummies, an indicator for reporting Hispanic ethnicity, dummies for citizenship status, state and year fixed effects. As we expected, results show that the relationship between access to highspeed internet and the number of young children in the household is non-significant among younger and unmarried individuals, and positive and significant among older and married individuals. We find similar results when considering as an alternative outcome a dummy for whether individuals had children younger than 5 years old in the household.

2.3 Internet and fertility rate using county-level data

Finally, we provide further evidence on the heterogeneous relationship between Internet and fertility rates by age and education using data drawn from the Natality Detail Data (1999-2004) and broadband data from the Federal Communication Commission. The Natality Data are drawn from the National Vital Statistics System of the National Center for Health Statistics and provide demographic and health data for births occurring during the calendar year. The microdata are based on information abstracted from birth certificates filed in vital statistics offices of each

State and District of Columbia. Using intercensal population estimates by age and gender from the United States Census Bureau we construct birth rates for different socio-demographic groups and linked these rates with county-level broadband data drawn from the Federal Communications Commission (FCC). As we only have data from the FCC since 1999 and as county information is available only until 2004, we restrict the county-level analysis to the years 1999-2004. It is worth noting that information is publicly available only for counties with more than 100,000 inhabitants and thus the results cannot inform about the effects of Internet on fertility in rural and less populated counties.

The county birth rate is positively associated with broadband penetration. A one standard deviation increase in the percentage of the county with at least one provider is associated with a .26 standard deviation (a 14% increase with respect to the mean) increase in fertility rate among 25-44 old women. The coefficient isn smaller and non-significantly different from zero among low-skilled, unmarried, and teenagers. It is instead largest among married women for whom a 1 standard deviation increase in broadband penetration increases fertility by 0.35 standard deviation (a 19% increase with respect to the mean).

3 Conclusions

Internet has revolutionized the way we organize and use our time. This study analyses the effects of access to high-speed internet on fertility in the US. We presented an analysis of the links between one of the main aspects of the digital revolution that took place at the end of the Twentieth Century, the deployment of the Internet, and fertility decisions. In a sample of young adults from the NLSY97, within the age range 18-31, we found that having access to the Internet tends to be associated with lower fertility. We also found that this negative association tends to fade away with age, indicating a role of the Internet in the postponement of fertility. Moreover, the negative association is weaker for men and women from a higher socioeconomic background. We did find evidence for the role of partnership formation as a mediating factor in the relationship between Internet access and fertility among women, but not among men. These results are mainly in line with the findings of Guldi and Herbst (2017) and with an interpretation that emphasizes the role of information availability and social interaction. This interpretation is in line with the idea that the Internet amplifies the information effects that other, older communication technologies, like the TV, have been shown to have on fertility (Jensen and Oster, 2009; La Ferrara et al., 2012).

Using data from other the CPS and Natality Detail Data we also find evidence of the role of the Internet in the “recuperation” of fertility at later ages. Access to high-speed Internet is positively associated with fertility outcomes of high-skilled women in their thirties, while we find non-significant results among low-skilled, young and unmarried women. These results are consistent with the findings of Dettling (2017) and Billari et al. (2017) and the hypothesis that internet may relax time constraints of high-skilled working women reconciling family, motherhood and work.

References

Allison, Paul D, Discrete-time methods for the analysis of event histories, Sociological methodology, 1982, 13, 61-98.

Barber, Jennifer S, Susan A Murphy, William G Axinn, and Jerry Maples, 6. Discrete- Time Multilevel Hazard Analysis, Sociological methodology, 2000, 30 (1), 201-235.

Billari, Francesco C, Osea Giuntella, and Luca Stella, Does Broadband Internet affect fertility? IZA DP 10935, 2017, (10935).

Dettling, Lisa J, Broadband in the labor market: the impact of residential high-speed internet on married women's labor force participation, ILR Review, 2017, 70 (2), 451-482.

Guldi, Melanie and Chris M Herbst, Offline effects of online connecting: the impact of broadband diffusion on teen fertility decisions, Journal of Population Economics, 2017, 30 (1), 69-91.

Hargittai, Eszter, Digital na (t) ives? Variation in internet skills and uses among members of the net generation, Sociological inquiry, 2010, 80 (1), 92-113.

Table 1: Fertility between 2003 and 2011 (NLSY97, Men)

Y=Having a Child	**(1)**	**(2)**	**(3)**	**(4)**
Internet Access	0.409***	0.424***	0.505***	0.525***
	-0.093	-0.09	-0.116	-0.114
Age	1.043	0.911***	0.998	0.904**
	-0.045	-0.029	-0.054	-0.038
Enrolled In School	0.602***	0.692***	0.622***	0.714***
	-0.068	-0.072	-0.069	-0.074
Years of Education	0.888***	0.923***	0.886***	0.922***
	-0.022	-0.018	-0.022	-0.018
Parity (Ref: 0)				
Parity=1	0.793	1.03	0.88	1.094
	-0.187	-0.162	-0.225	-0.169
Parity=2	0.205***	0.495**	0.243***	0.552**
	-0.082	-0.147	-0.107	-0.162
Parity=3	0.038***	0.189**	0.047***	0.220**
	-0.028	-0.126	-0.038	-0.147
Parents Education (Ref: Less than High School)				
High School Diploma	0.909	0.948	0.616*	0.74
	-0.134	-0.102	-0.174	-0.186
More than High School	0.756*	0.882	0.303***	0.397***
	-0.119	-0.101	-0.088	-0.103
Internet Access*Age	1.154***	1.131***	1.067	1.023
	-0.039	-0.036	-0.045	-0.041
In a Partnership		5.621***		5.565***
		-0.474		-0.47
Internet Access*Parents Education*Age				
No Access - HS Diploma - Age			1.025	0.989
			-0.049	-0.044
No Access - more than HS - Age			1.094*	1.049
			-0.055	-0.05
Access - HS Diploma - Age			1.086**	1.072*
			-0.043	-0.04
Access - more than HS - Age			1.176***	1.164***
			-0.046	-0.044
Constant	0.236***	0.188***	0.394**	0.268***
	-0.09	-0.057	-0.157	-0.088
Rho	0.325	0.056	0.294	0.034
N	2168	2168	2168	2168

* p<0.10, ** p<0.05, *** p<0.01. All models include race/ethnicity, region of residence, urban/rural area and family income.

The Internetization of Marriage: Effects of the Diffusion of High-Speed Internet on Marriage, Divorce, and Assortative Mating

L'Internetizzazione del Matrimonio: Effetti della Diffusione di Internet Veloce sul Matrimonio, Divorzio e sul Processo di Selezione del Partner

Francesco C. Billari, Osea Giuntella and Luca Stella

Abstract The advent of the Internet has radically changed the process of finding a stable partner. We exploit some historical and technological features of the broadband roll-out across Germany to provide instrumental variable (IV) estimates. Our IV results suggest that high-speed Internet increases the hazard of marriage, with the effects being concentrated among the high-educated people. Furthermore, we show that broadband access significantly increases educational homogamy, which we interpret as our potential channel underlying the relationship of interest.

Abstract *L'avvento di Internet ha modificato in modo radicale il processo di selezione di un partner. In questo lavoro sfruttiamo alcune caratteristiche storiche e tecnologiche della rete telefonica in Germania per fornire stime a variabili strumentali. In particolare, troviamo che Internet veloce aumenta la probabilita' di sposarsi, soprattutto tra le persone piu' istruite. Inoltre, troviamo che Internet veloce aumenta l'omogamia per livello di istruzione, il che potrebbe spiegare la nostra relazione di interesse.*

Key words: Internetization, Marriage, Divorce, Assortative Mating

1 Introduction

[1] Francesco C. Billari, Bocconi University; email: francesco.billari@unibocconi.it
Osea Giuntella, University of Pittsburgh and IZA; email: osea.giuntella@pitt.edu
Luca Stella, Catholic University, Dondena and IZA; email: luca.stella@unicatt.it

Since the mid 1990s, the world entered full-speed in what Castells (2001) defined the Internet Galaxy. Daily life has become shaped by the presence of the Internet in all its domains. Partnership dynamics, and marriage, are no exception. What are the consequences of the diffusion of the Internet for partnership formation and dissolution? In this paper, we argue and document with a set of empirical analyses, that the diffusion of the Internet has radically changed the transition to marriage, both in terms of speed and in terms of type of match, with ambiguous effects on marital stability. We have entered the era of the Internetization of Marriage.

In general, the notion of Internetization refers to the pervasive two-way effects of the Internet on social and economic phenomena. This notion, importantly, acknowledges that social and economic phenomena that become internetized become radically different from the pre-internetized era, also in their offline nature. The social theorist Leopoldina Fortunati defines and discusses, for instance, the Internetization of mass media. In the context of mass media, the presence of the Internet on the one side allows, and even forces, traditional media to be active online in order to be visible and seen, and on the other side changes the way classic mass media exist--as they change and their traditional forms (e.g. in print, audio, or video), also become internetized (Fortunati, 2005).

The Internetization of Marriage can therefore be seen as a process with two main channels, and possibly two stages, in analogy to mass media. First, the Internet provides, directly, new ways to find a partner online. Referring to this direct role, Rosenfeld and Thomas (2012) defined the Internet as the rising and new "social intermediary" in the search for mates. For instance, online dating is defined as "the use of websites that provide a database of potential partners - typically in close geographical proximity - that one can browse and contact, generally for a fee" (Sautter et al., 2010, p. 555). Second, the presence of the Internet radically changes partnership dynamics, including partnerships that have been formed offline. If we include this second stage, which has been less studied so far, the effects on marriage of the Internet becomes more ambivalent.

For what concerns the first stage of the Internetization of marriage, the speed at which information can be collected, and the possibility to make targeted search, such as for specific combinations of characteristics of a prospective co-parent, as in the case of online dating (Potarca et al., 2015) might in general imply a quicker transition to a partnership. Online interaction allows to gather more information on prospective partners (within social media or online dating, but also in general including prospective partners met offline), with a decline of the role of distance with respect to the pre-Internet era (Mok and Wellman, 2007). For these reasons, more stable partnership situations can be expected as the outcome of better matches in the era of the Internet (Cacioppo et al., 2013; Rosenfeld, 2017). The width of the partnership market might increase and therefore allow a search from a wider pool, which would take longer and imply a postponement of partnership formation, creating a sort of "information overload" (Rosenfeld, 2017).

For what concerns the second stage of the Internetization of marriage, the empirical literature here is scarce. An exception is a study that uses aggregate-

level data and exploits the timing of broadband diffusion in the US, in which Bellou finds that broadband diffusion is positively correlated to marriage rates (Bellou, 2015). The presence of the Internet is changing relationships in general, in many ways, including communication among partners (Coyne et al., 2011), and brings the possibility of new conflicts emerging within the family (Mesch, 2003), and of larger intergenerational contacts outside the co-residential family (Gubernskaya and Treas, 2016). Using data from the U.S. General Social Survey for the year 2000, Wasserman and Richmond-Abott (2005) showed that married individuals were more likely to access the Internet, while they did not significantly differ in the type of use from individuals who were not married. Using the same data, the quality of a marital relationship, for instance, was found to be a predictor of the use of Internet pornography (Stack et al., 2004).

The Internetization of Marriage has additional implications for social stratification and inequality that might go beyond its effects on assortative mating, which by itself has implications on inequality (Schwartz, 2010). In terms of stratifying factors, there is no particular reason to believe that gender plays a specific role in this case. However, other aspects of social stratification will play a role, in addition to the mere access to the Internet, as prospective parents from upper socioeconomic strata might have a bigger advantage from interactions that start online. A first level of inequality relates to the material "digital divide" in access, and the quality of this access, to the Internet itself (Norris, 2001). A second level of "digital inequality" refers to the knowledge of search strategies, the capability to evaluate the quality of information, and to exploit the potential of the Internet (DiMaggio et al., 2001). Eszter Hargittai has termed this latter inequality in online skills as the "second-level digital divide" (Hargittai, 2002). Potarca (2017), for instance, shows that couple who meet online are less endogamous than couples who meet at school, religious venues, or through family and friends.

Our empirical study is based on the German Socio-Economic Panel (SOEP), a longitudinal panel dataset of the German population containing information on a rich set of individual socio-economic characteristics. Three features of the SOEP are key to our analysis. First, for every respondent, the survey contains a detailed marital history, allowing us to use information on the year of marriage and divorce to construct our main outcomes of interest. Second, the survey not only collects household information on whether Internet access is available, but also data on whether Internet access is based on a broadband (DSL) technology. This information is exploited to build our key explanatory variable. Finally, the SOEP collects not only respondents' education but also educational attainment of their spouses. This enables us to examine assortative mating by education, and thus shed light on the potential mechanism through which high-speed Internet access may influence partnership formation. To address the concern regarding endogeneity of broadband Internet use, we follow the identification strategy adopted by Falck et al. (2014) and later replicated by Billari et al. (2019). Our 2SLS estimates suggest that DSL access increases the hazard of marriage. These results are entirely driven by the high-educated people. We find ambiguous results for divorce. Furthermore, we show that broadband access significantly

increases educational homogamy, which we interpret as our potential channel underlying the relationship of interest.

2 Introduction

As previously stated, the focus of this paper is to investigate how access to broadband Internet affects marriage, divorce and assortative mating. To this end, we estimate the following linear probability model:

$$Y_{ist} = \alpha + \beta DSL_{ist} + \gamma X_{ist} + \mu_t + \eta_s + \varepsilon_{ist} \quad (3)$$

where the index ist denotes an individual i residing in federal state s at the year of interview t. We have a set of outcome variables, Yist, defined as follows: 1) the annual hazard of marriage (i.e., the annual probability of getting married for unmarried individuals); 2) the annual hazard of divorce (i.e., the annual probability of divorce for married individuals); 3) educational homogamy, operationalized by an indicator equal to one if the couple has the same educational level (only for individuals who are married); and 4) the distance (in absolute value) between the educational level of the respondent and his/her partner (only for individuals who are married). Our main explanatory variable of interest is DSList, which represents a dummy variable taking value one if an individual has a high-speed Internet subscription at home, and zero otherwise. Accordingly, the coefficient of interest, β, captures the impact of access to high-speed Internet on the outcome of interest. To address the concern regarding endogeneity of broadband Internet use, we follow the identification strategy adopted by Falck et al. (2014). Their main idea is to exploit historical variation in pre-existing telephone infrastructure which significantly affected the cost of broadband adoption across Germany. In particular, Falck et al. (2014) exploit three unique historical and technological peculiarities of the traditional public telephone network, which influenced the deployment of DSL in German municipalities. Table 1 reports the 2SLS results for marriage (see Panel A) and divorce (see Panel B). The 2SLS coefficient in column 1 of Panel A suggests a positive and significant impact of broadband Internet on the transition into marriage: having a high-speed Internet subscription increases the hazard of marriage by 13.3 percentage points. In columns 2 and 3 of Panel A, we test for the presence of digital inequalities, analyzing whether the effect differs by the educational group of the respondent. The effect of broadband Internet on marriage is driven by the high-educated individuals. Panel B of Table 1 reports the 2SLS parameter estimates of high-speed Internet on the hazard of divorce. Because marriage is a prerequisite of divorce, we conduct this analysis on the sample of married individuals. As shown in columns 1 to 3, we do not find evidence of a significant effect of high-speed Internet on divorce, with the effects

being very small in magnitude. DSL technology reduces the cost of searching and allows for the selection of a similar partner with minimum efforts. Therefore, DSL access provides individuals with more opportunities to match with partners that share the same background. To test this hypothesis, we analyze the effects of DSL access on educational assortative mating. Specifically, we measure assortative mating along two dimensions: homogamy, and the distance (in absolute value) between the educational level of the respondent and her partner. Overall, the results suggest that DSL access is associated with higher homogamy, and thus it reduces the educational distance between the partners (results of this analysis are available from the authors upon request).

Table 1: Effects of Internet on Marriage and Divorce by Education Group, 2SLS

Education group:	(1) All	(2) High-educated	(3) Low-educated
Panel A: Dep. var.: Hazard of marriage			
High-speed Internet subscription	0.133*	0.171**	-0.041
	(0.078)	(0.086)	(0.190)
Observations	7,480	5,669	1,811
Mean of dep. var.	0.059	0.059	0.057
Std. dev. of dep. var.	0.235	0.236	0.232
F-test of excluded instruments	10.250	8.535	1.568
Panel B: Dep. var.: Hazard of divorce			
High-speed Internet subscription	0.015	0.030	-0.026
	(0.034)	(0.035)	(0.070)
Observations	15,909	10,606	5,303
Mean of dep. var.	0.015	0.016	0.014
Std. dev. of dep. var.	0.122	0.123	0.119
F-test of excluded instruments	11.89	10.24	3.735

3 References

1. Bellou, Andriana, "The impact of Internet diffusion on marriage rates: evidence from the broad- band market," Journal of Population Economics, 2015, 28 (2), 265–297.

2. Francesco C. Billari, Osea Giuntella & Luca Stella (2019): Does broadband Internet affect fertility?, Population Studies.

3. Cacioppo, John T., Stephanie Cacioppo, Gian C. Gonzaga, Elizabeth L. Ogburn, and Tyler J. VanderWeele, "Marital satisfaction and break-ups differ across on-line and off-line meeting venues," Proceedings of the National Academy of Sciences, 2013, 110 (25), 10135–10140.

4. Castells, M., The Internet galaxy: Reflections on the Internet, business, and society, Oxford University Press, 2001.

5. Coyne, Sarah M., Laura Stockdale, Dean Busby, Bethany Iverson, and David M. Grant, "I luv u :)!: A Descriptive Study of the Media Use of Individuals in Romantic Relationships," Family Relations, 2011, 60 (2), 150–162.

6. DiMaggio, Paul, Eszter Hargittai, W. Russell Neuman, and John P. Robinson, "Social Implications of the Internet," Annual Review of Sociology, 2001, 27 (1), 307–336.

7. Etemad, Hamid, Ian Wilkinson, and Leo Paul Dana, "Internetization as the necessary condition for internationalization in the newly emerging economy," Journal of International Entrepreneur- ship, Dec 2010, 8 (4), 319–342.

8. Falck, Oliver, Robert Gold, and Stephan Heblich, "E-lections: Voting Behavior and the Inter- net," The American Economic Review, 2014, 104 (7), 2238–2265.

9. Fortunati, L., "Mediatization of the Net and Internetization of the Mass Media," Gazette (Leiden, Netherlands), 2005, 67 (1), 27–44.

10. Gubernskaya, Zoya and Judith Treas, "Call Home? Mobile Phones and Contacts With Mother in 24 Countries," Journal of Marriage and Family, 2016, 78 (5), 1237–1249.

11. Hargittai, Eszter, "Second-Level Digital Divide: Differences in People's Online Skills," First Mon- day, 2002, 7 (4).

12. Mesch, Gustavo S., "The Family and the Internet: The Israeli Case*," Social Science Quarterly, 2003, 84 (4), 1038–1050.

13. Mok, Diana and Barry Wellman, "Did distance matter before the Internet?: Interpersonal contact and support in the 1970s," Social Networks, 2007, 29 (3), 430 – 461. Special Section: Personal Networks.

14. Neves, Barbara Barbosa and Cla´udia Casimiro, Connecting Families?: Information & Communication Technologies, generations, and the life course, Bristol, UK: Policy Press, 2018.

15. Norris, Pippa, Digital divide: Civic engagement, information poverty, and the Internet worldwide, Cam- bridge, UK: Cambridge University Press, 20

16. Potarca, Gina, "Does the internet affect assortative mating? Evidence from the U.S. and Germany," Social Science Research, 2017, 61, 278 – 297.

17. Potarca, Gina, Melinda Mills, and Wiebke Neberich, "Relationship Preferences Among Gay and Lesbian Online Daters: Individual and Contextual Influences," Journal of Marriage and Family, 2015, 77 (2), 523–541.

18. Rosenfeld, M.,"Marriage, Choice, and Couplehood in the Age of the Internet," Sociological Science, 2017,4(20),490–510.

19. Sautter, Jessica M., Rebecca M. Tippett, and S. Philip Morgan, "The Social Demography of Internet Dating in the United States*," Social Science Quarterly, 2010, 91 (2), 554–575.

20. Schwartz, Christine R, "Earnings inequality and the changing association between spouses earnings," American journal of sociology, 2010, 115 (5), 1524–1557.

Solicited Sessions

Advanced Statistical Methods in Health Analytics

Assessing the impact of the intermediate event in a non-markovian illness-death model

Valutazione dell'impatto dell'evento intermedio in un modello multistato "illness-death" non-markoviano

Davide Paolo Bernasconi, Elena Tassistro, Maria Grazia Valsecchi, Laura Antolini

Abstract In the illness-death model the transition from the initial state to the final state (death) involves an intermediate state (illness). When it is of interest to compare the hazards of mortality before and after illness, one could resort to the Mantel-Byar test, an extension of the logrank test that accounts for the time-dependent nature of illness by ascribing all patients to the illness-free group until they (possibly) become ill; at that time, they are moved to the illness group using left truncation. However, in this approach, time is always measured by the clock forward scale and this is not valid under non markovian processes. We propose an alternative approach, suitable for semi-Markov and extended semi- Markov scenarios, based on the adoption of the clock reset scale to measure time after illness. The method is a modification of the logrank test that involves the coefficient of the time to illness from a Cox model fitted only on ill patients to determine the number of patient at risk in time after illness. We compare the two approaches in a simulation protocol.

Abstract *Nel modello "illness-death" la transizione dallo stato iniziale a quello finale (death) include un evento intermedio (illness). Volendo confrontare gli azzardi di morte prima e dopo illness, una possibilità è l'uso del test di Mantel-Byar, un'estensione del logrank test che tiene conto della natura tempo-dipendente di illness. Tuttavia tale approccio è valido solo per processi markoviani poichè il tempo è sempre misurato a partire dallo stato iniziale. Proponiamo quindi un metodo alternativo, sempre basato sul logrank test che, utilizzando una scala temporale che parte dallo stato intermedio, consente di valutare l'impatto dell'evento intermedio in scenari non markoviani. I due metodi sono messi a confronto tramite un protocollo di simulazione.*

Key words: Illness-death model, Mantel-Byar test, Markov property

1 Introduction

Situations where subjects start from an initial condition (state 0) and then can possibly develop a final event (state 2 or *death*), either directly or passing through another intermediate condition (state 1 or *illness*), may be described by the simple *illness-death* multistate model [5], as depicted in Figure 1. This framework may be applied also in situations where the transition to the intermediate state does not consist in the occurrence of a disease but instead represents the administration of a therapeutic intervention. Consider, as an example, patients affected by end-stage cardiomyopathy that enter a waiting list (state 0) to receive heart transplant: they may die (state 2) while still on list or they can indeed be transplanted (state 1) after some time.

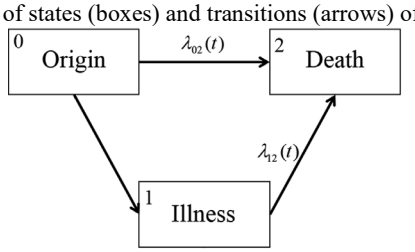
In this context, a key issue is often to assess the impact of the intermediate event on the hazard of mortality which is done, in practice, by comparing two transition hazards: the hazard from state 0 to state 2 $\lambda_{02}(t)$ and the hazard from state 1 to state 2 $\lambda_{12}(t)$. A key aspect to consider to this end is the choice of the scale used to measure time [2]. The most obvious one is the original (or "clock forward") time scale, where time is always measured starting from the entry into the initial state 0. As an alternative, one may resort to the "clock reset" scale, where time is set back to 0 as patients enter the intermediate state 1. In practice, for those who experience the intermediate event, the observation time since origin (denoted as t) can be split in two parts: time elapsed while in state 0 (waiting time to intermediate event, denoted as r) and time elapsed while in state 1 (time since intermediate event, denoted as d).

Depending on the relation between hazards and time scales, three types of data generating process can be defined [6]: Markov, semi-Markov and extended semi-Markov. The process is Markovian if the value of $\lambda_{12}(t)$ for $t>r$ depends only on time t from origin. This means, in other words, that the value of $\lambda_{12}(t)$ depends on time d after intermediate event and on the fixed covariate time to intermediate event r only through their sum $d+r=t$ and not on the values d and r taken separately. The clock forward time scale alone can be conveniently used to measure time even after the intermediate event and the notation $\lambda_{12}(t)$ can be kept, meaning that in this case the hazard depends only on t. The process is semi-Markovian if the value $\lambda_{12}(t)$ for $t>r$ depends only on time d after intermediate event. Thus, the clock reset time scale alone can be conveniently used to measure time after the intermediate event and the new notation $\lambda_{12}(d)$ for $d>0$ is set, meaning that time is measured on the clock reset scale. The process is extended semi-Markovian if the value of $\lambda_{12}(t)$ for $t>r$ depends on time d after intermediate event and on the fixed covariate time to intermediate event r, taken separately. The time d after intermediate event, and thus the clock reset time scale, becomes again the natural way to measure time after the intermediate event and one can set the new notation $\lambda_{12}(r,d)$ meaning that time is measured on the clock reset scale and r has an impact on the hazard $\lambda_{12}(r,d)$ for $d>0$.

In this work we review an extension of the logrank test, namely the Mantel-Byar test, to compare the hazards between two time-dependent groups and show that it is suitable only for markovian processes. Moreover, we propose an alternative method,

based again on the logrank test, that is suitable for non-markovian scenarios. The performance of the two methods is compared in a simulation protocol.

Figure 1: Representation of states (boxes) and transitions (arrows) of the *illness-death model*.



2 Methods

The logrank test is a standard non-parametric tool to compare the hazard of failure between two time-fixed groups, e.g. A and B. Under the null hypothesis H_0: $\lambda_A(t)= \lambda_B(t)$, the distribution of the observed frequency of failures d_{Aj} at each distinct event time $j=1,...,J$ is hypergeometric with $E(d_{Aj})=d_j n_{Aj}/n_j$ and $Var(d_{Aj})= (d_j n_{Aj}/n_j)(1-d_j/n_j)[(n_j - n_{Aj})/(n_j -1)]$. The following test statistics can thus be considered: $Q = U^2/Var(U) = \{\Sigma_j [d_{Aj} – E(d_{Aj})]\}^2/ \Sigma_j[Var(d_{Aj})]$ which follows a Chi-square distribution with 1 degree of freedom.

When the two groups are described by a time-dependent binary indicator, as in the case of an *illness-death model* where patients may enter state 1 after some waiting time, alternative approaches should be considered.

2.1 Mantel-Byar test

An extension of the logrank test to deal with time-dependent groups was proposed back in 1974 by Mantel and Byar [3] in the context of heart-transplant. Time is always measured using the original scale t (time since entry in state 0). At time 0 all patients are in state 0 and thus the risk-set of the intermediate event (i.e. transplant) group is empty.

Suppose a patient is transplanted at time $r=t^*$, since then he/she is moved to the transplant risk-set of the transplanted group. In other words, the observation of that patient is split in two parts: *i)* from time 0 to t^* the patient is ascribed to the waiting-list group and the observation is censored at t^*; *ii)* from t^* onwards the patient is ascribed to the transplanted group and the observation is left-truncated at t^*. If the process is markovian, the hazard of mortality under transplant, after t time units since entry in the waiting list, does not depend on the waiting time to transplant r nor on the time since transplant d. Thus, under markovianity, the Mantel-Byar test is suitable to compare the hazard of mortality of a patient in waiting list with the

hazard of an hypothetical patient transplanted at an arbitrary time r (e.g. at $r=0$). The null hypothesis can be written as H_0: $\lambda_{02}(t) = \lambda_{12;r=0}(t)$.

If the process is not markovian, however, there is an impact of time since transplant d (semi-Markov process) and possibly also of time to transplant r (extended semi-Markov process). In these scenarios, the Mantel-Byar test is not suitable to test the above null hypothesis because the hazard after transplant is calculated by averaging hazards of patients at t time units since entry in list, when patients have spent a different amount of time on list and after transplant [1].

2.2 "Clock reset logrank-type" test

For semi-markovian processes, a simple modification to the Mantel-Byar test can be applied by adopting a double time scale to measure time: the original time scale t before transplant and the clock reset scale d after transplant.

Again, the observation of a patient transplanted at $r=t^*$ is split in two parts: *i)* from time $t=0$ to $t=t^*$ the patient is ascribed to the waiting-list group and the observation is censored at t^*; *ii)* from time $d=0$ onwards the patient is ascribed to the transplanted group. Under the semi-Markov property, the hazard of mortality at time d since transplant does not depend on the waiting time to transplant r. Using this "clock reset logrank-type" test, the hazard after transplant is calculated by averaging hazards of patients at d time units since transplant, disregarding that they spent a different amount of time on list. Thus, under semi-markovianity, the test is suitable to compare the hazard of mortality of a patient on waiting list with the hazard of an hypothetical patient transplanted at an arbitrary time r (e.g. at $r=0$). The null hypothesis can be written as H_0: $\lambda_{02}(t) = \lambda_{12;r=0}(d)$.

When the waiting time on list r has an impact on the hazard after transplant (extended semi-Markov process), this has to be taken into account. Thus, we propose the following two-steps approach. First, fit a Cox model only on post-transplant data to check the impact of time to transplant r on the hazard of mortality after transplant (measured using the clock reset scale) [4]. Then, use the estimated coefficient β_r of r to adjust the size of the risk-set in the transplanted group (n_{12j}) within the "clock reset logrank-type" test. This is done by substituting n_{12j} with $n^*_{12j} = \Sigma_i [exp(\beta_r r_i)]$ (i.e. sum over subjects $i=1,\ldots, n_{12j}$ at risk at time $d=j$) which resembles the expected counts one would observe if patients were all transplanted at $r=0$.

Table 1: Details of the distributions of the transition times used in the simulation protocol. In every scenario: 1000 samples of size 150 were generated; uniform independent censoring was considered; the time to the final state T_{02} was generated from a *Weibull*(0.6;0.7); the time to *illness* R was generated from an *Exponential*(0.5).

Scenario	***D***	***$\beta_r(t)$***
1. H_0: $\lambda_{02}(t) = \lambda_{12;r=0}(d)$	*Weibull*(0.6exp(β_rr/0.7);0.7)	0.5
2. H_0: $\lambda_{02}(t) = \lambda_{12;r=0}(d)$	*Weibull*(0.6exp($\beta_r(t)$r/0.7);0.7)	0.5*log(t/10)
3. H_1: $\beta_{ill} = -log(3)$	*Weibull*(0.6exp(β_{ill}/0.7)exp($\beta_r(t)$r/0.7);0.7)	0.5
4. H_1: $\beta_{ill}(t) \neq 0$	*Weibull*(0.5exp(β_rr/0.5);0.5)	0.5

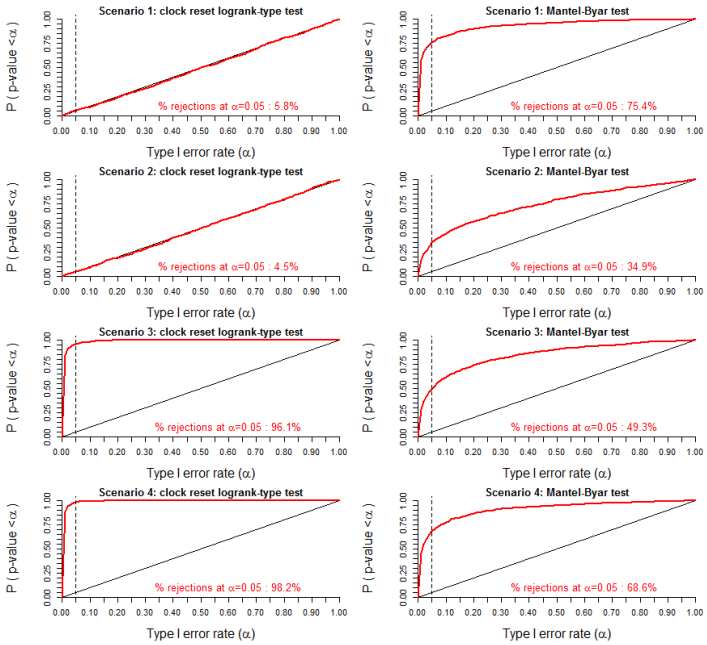
2.3 Simulation protocol

We compared the performance of the Mantel-Byar test vs the "clock reset logrank-type" test in four scenarios under the extended semi-Markov property. In the first two scenarios data are generated under the null hypothesis of equality of the hazards of mortality before and after intermediate event but in scenario 1 the effect of the time to intermediate event *r* is time-fixed while in scenario 2 it is time-dependent. In the last two scenarios data are generated under an alternative hypothesis where the intermediate event has a beneficial effect on mortality but in scenario 3 the effect of *"illness"* is proportional (coefficient β_{ill} is constant) while in scenario 4 it is non-proportional (coefficient β_{ill} is time-dependent). More details are shown in Table 1.

3 Results

The results of the simulation protocol are reported in Figure 2. Under the null hypothesis (scenarios 1 and 2), the percentage of rejections with the "clock reset logrank-type" test are close to the nominal level, while the Mantel-Byar test tends to reject the null too often. Moreover, the "clock reset logrank-type" test has a higher power than the Mantel-Byar test, as shown by the higher percentage of rejections under an alternative hypothesis (scenarios 3 and 4).

Figure 2: Results of the simulation protocol. The R Software (version 3.6.2) was used for the analyses.



4 Concluding remarks

The impact of the occurrence of illness on the hazard of death can be investigated non-parametrically with logrank-type tests.

The null hypothesis H_0: $\lambda_{02}(t) = \lambda_{12;r=0}(t)$ can be tested by the Mantel-Byar test only when the process is Markovian. For non-Markov processes, we proposed a test based on the clock reset time scale (thus suitable for semi-Markov scenarios) that can possibly account for the effect of the time to illness r (thus becoming suitable also for extended semi-Markov scenarios). In our simulation protocol, due space reasons, we only considered an extended semi-Markov data generating process and we only compared two methods: the "classic" Mantel-Byar approach vs the "clock reset logrank-type" test accounting for time to illness. Within this framework, we considered four scenarios to check the performance of the methods under the null (scenarios 1 and 2) or an alternative hypothesis (scenarios 3 and 4) and in the absence (scenarios 1 and 3) or presence (scenarios 2 and 4) of time dependent effects.

As an alternative, one can consider a Cox model where time after transition to state 1 is measured using the clock reset scale and where the time-dependent indicator of transition to state 1 and the fixed time to state 1 are included in the model as covariates [6]. This latter solution is not directly suitable for situations where hazards are not proportional.

References

1. Bernasconi, D.P., Rebora, P., Iacobelli, S., Valsecchi, M.G., Antolini, L.: Survival probabilities with time-dependent treatment indicator: quantities and non-parametric estimators. Statistics in Medicine 35(7):1032-48 (2016)
2. Iacobelli, S., Carstensen, B.: Multiple time scales in multi-state models. Statistics in Medicine 32(30):5315-5327 (2013)
3. Mantel, N., Byar, D.P.: Evaluation of Response-Time Data Involving Transient States: An Illustration Using Heart-Transplant Data. Journal of the American Statistical Association 69:81-86 (1974)
4. Meier-Hirmer, C., Schumacher, M.: Multi-state models for studying an intermediate event using time-dependent covariates: application to breast cancer. BMC Medical Research Methodology 13:80 (2013)
5. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 26: 2389-2430 (2007)
6. Tassistro, E., Bernasconi, D.P., Rebora, P., Valsecchi, M.G., Antolini, L.: Modelling the hazard of transition into the absorbing state in the illness-death model. Biometrical Journal 1:16 (2019)

Big data and AI: challenges and opportunities in healthcare

Big data e Intelligenza Artificiale: sfide e opportunità in sanità

Vieri Emiliani, Gian Luca Cattani, Fabrizio Selmi[1]

Abstract The sustainability of healthcare systems is put at risk by the increase of health and long-term care demand of an aging population. So far, information technology has failed to delivery in healthcare the benefits in terms of efficiency that has been able to provide to other sectors. Conversely, several studies relate the adoption of Electronic Health Records with decreased professional satisfaction and increased risk of professional exhaustion. Big Data and Artificial Intelligence are among the technologies that are driving the so-called digital transformation. The aim of this paper is to provide an overview on how a broader adoption of these new technologies could help addressing these issues, using a first-hand case study to illustrate the results that can be achieved.

Abstract *La crescita della domanda per servizi di cura e salute dovuta all'invecchiamento della popolazione sta mettendo a rischio la sostenibilità dei sistemi sanitari. Ad oggi, l'information technology in sanità ha fallito nel produrre i benefici in termini di efficienza che è stata in grado di realizzare in altri settori. Al contrario, alcuni studi correlano l'introduzione di EHR a un calo della soddisfazione professionale e a un aumento del rischio di burnout. Big Data e Intelligenza Artificiale sono tra le tecnologie che guidano la cosiddetta digital transformation. Lo scopo di questo articolo è fornire un inquadramento di come l'adozione di queste nuove tecnologie può contribuire ad indirizzare questi problemi, usando un caso studio per illustrare i risultati che è possibile raggiungere.*

Key words: healthcare, big data, artificial intelligence

[1] Vieri Emiliani, Maps SpA; vieri.emiliani@mapsgroup.it
Gian Luca Cattani, Maps SpA; gianluca.cattani@mapsgroup.it
Fabrizio Selmi, Maps SpA; fabrizio.selmi@mapsgroup.it

1 Introduction

Advances in medicine and healthcare contribute to increase life expectancy. As a result, the population of the elderly in most countries is increasing, leading to a growing demand of health and long-term care services. The number of Europeans aged 65 or over, reached in 2018 100 million [1]. Eurostat forecasts that people over 65 will grow to 23.9% in 2030, adding 23 million of elderly people that will require complex and expensive care treatments and services. European high standards in providing affordable, accessible and high-quality health systems come at a cost: EU healthcare expenditure relative to GDP in 2016 was already close to 10% [2], while in Italy the total spending for healthcare amounts to 149.5 billion of Euro (8.9% of Italian GDP), 75% of which is covered by public funds [3].

"A radical redesign of health is needed to meet these challenges" [4]. Big Data and Artificial Intelligence (AI) applications are supporting the innovation in many traditional sectors. Our aim is to contribute to the discussion on how these technologies can help to ensure the future sustainability of our health systems by delivering greater efficiency, lower costs and better health outcomes.

The remainder of this paper is structured as follows: in Section 2 we provide an overview of the IT context in healthcare, current issues and how a broader adoption of Big Data and AI could help to address some of them. In Section 3 we present a case study where such technologies have been used to improve the efficiency of allocation of diagnostic procedures in an Italian Local Health Authority. Finally, challenges and barriers that need to be addressed are discussed as our conclusions.

2 Context

From autonomous vehicle control to drone-based aerial intelligence to identify crop pests, insurance claims processing, movie recommendation, fraud detection, Big Data and AI lead this innovation stream called digital transformation. In the last decade, the vast amount of data generated and collected by sensors, mobile phones and computerised systems, combined with the increase of computational power and the commoditisation of data storage, enabled the application of well-known algorithms [5] to a multitude of new use cases that now permeate our lives.

Big Data [6][7] and AI [8][9] research is also very dynamic in Healthcare. AI-powered algorithms outperform physicians in several diagnosis tasks. In 2018 only, more than 1500 papers were published on the topic of AI applied to radiology [10], while in the last year VC firms invested almost $4 billion in Healthcare AI start-ups [11]. Despite this hype, these technologies are not widely applied. 84% of health professionals in Europe either do not use AI tools or are not aware of it, and 59% do not have plans to adopt AI tools over the next 1-3 years [12].

It is acknowledged that adoption of IT tools in healthcare is at least 10 years behind most other sectors [4]. IT systems designed for administrative and management purposes have contributed to a transactional health care model, where poor usability

and rigid workflows limit the efficiency of health care professionals, causing distress and increasing risk of burnout [13],[14].

Healthcare processes are complex, as complex is the data they manage. Industry consensus is that approximately 80% of all healthcare data are unstructured [15]. Computer Vision and Natural Language Processing (NLP) techniques can be applied to convert unstructured data in valuable information for data driven processes, enabling a true digital transformation of healthcare systems [16]. Having tools for the automatic conversion of images and narrative texts in actionable information would relief physicians from the tedious, error prone task of manually encoding diagnoses and procedures and speed up the task of reporting in digital imaging, improving the efficiency of the clinical processes while increasing reliability of the administrative ones (for a review of possible applications please refer to [8]).

In the next section we will illustrate, by means of a first-hand case, how the combination of healthcare data digitisation with the transforming potential of AI have helped a long-term initiative of an Italian Local Health Authority (LHA) in managing quality and appropriateness of referrals for diagnostic tests. In fact, leveraging the availability of referrals in digital format, we fully automated the appropriateness evaluation process, extending the outreach of the initiative from a small fraction of the referrals to their totality, and helped to ensure a more efficient allocation of a scarce and costly resource such as diagnostic imaging.

3 Case study

Inappropriate referral for diagnostic imaging is a worldwide problem for healthcare systems, leading to a waste of resources, increased costs and longer waiting lists, while unnecessarily overexposing patients to radiations. Diagnostic imaging is a co-operative process between the referring clinicians, and the radiologists. The quality of the outcome is strongly influenced by the quality of the diagnostic question and by the appropriateness of the requested diagnostic procedure. Communication challenges, limited knowledge and defensive medicine are all factors that contribute to inappropriate use of diagnostic procedures (DP).

Since 2008, Reggio Emilia LHA (hereafter also AUSL RE) has put in place an initiative targeted to improve the appropriateness of DP referrals. A multi-disciplinary task force developed a set of guidelines for several families of DPs, defining the clinical conditions that are appropriate for referring to a specific DP, and the corresponding level of urgency (priority). The program includes the distribution of educational materials and retraining sessions for the referring physicians.

To provide feedback to doctors, an ex-post review of the referral against the provided guidelines was required. This implies assessing each referral to verify if, given the provided information (gender, age, diagnostic question, diagnostic procedure and priority), it was compliant with the provided criteria. Due to the large number of referrals (more than 90,000 in 2011), this was a daunting task to be performed manually, and therefore only a small sample of the referrals was assessed.

With the digitisation of referrals, an opportunity to automate their assessment using a software tool arose. This led to the development of a software solution, Clinika VAP [17], that analyses the text of the diagnostic question and, given the test and the priority requested, verify the compliance of the referral against the corresponding guidelines. At the core of Clinika VAP there are two main components: first, a patented semantic engine that uses Machine Learning and NLP algorithms to analyse the unstructured clinical information contained in a narrative text. Second, a powerful rule engine that, using a domain specific language, can represent the guidelines criteria and reason about them, to verify if the diagnostic condition is one of those admitted by the referral guideline.

Referral guidelines implementation in Clinika VAP has been progressive, with Neuro MRI being the first to be deployed in 2012, and the others following between 2013 and 2016. Once a guideline is deployed, the LHA can systematically assess all the referral forms, to provide feedback to the GPs, and carry out targeted interventions. In order to measure the effectiveness of the system, more than 400,000 electronic referral forms related to 6 families of DPs have been collected and analysed. For this analysis, a referral was considered as appropriate if and only if both the requested DP and priority were in accordance with the guidelines. The appropriateness trend by referral guideline from 2011 to 2017 is reported in Table 1.

Table 1: Percentage of appropriate referrals by referral guideline, from 2011 to 2017.

		Year *(in grey years preceding guideline implementation)*						
Referral guideline	**Deploy**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Neuro MRI	*2012*	29%	61%	67%	68%	71%	71%	71%
Musculoskeletal MRI	*2013*	28%	64%	70%	71%	75%	73%	74%
Colonoscopy	*2014*	27%	54%	57%	59%	65%	70%	70%
Gastroscopy	*2014*	16%	38%	43%	44%	47%	50%	50%
Neuro CT	*2015*	17%	41%	45%	46%	47%	47%	46%
Musculoskeletal CT	*2015*	13%	32%	35%	42%	44%	52%	58%

Overall, the results show the effectiveness of the adoption of a computerised system for the systematic assessment of referrals. The improvement of the appropriateness rate between 2011 and 2017 ranges from 29% for Neuro CT to 46% for Musculoskeletal RMI, with relative improvements ranging from +145% (Neuro MRI) to +346% (Musculoskeletal CT).

In a recent work [18], the Epidemiology Department of AUSL RE analysed a similar dataset to identify the determinants of inappropriate reporting for a group of imaging and endoscopy DPs, using multivariate log-binomial regression models. Their results confirm that even though "the variability between GPs was the greatest source of inappropriateness variation, promising improvement was observed over time for all procedures under study, consistent with the implementation of several measures of training, shared protocol definition and administrative control".

4 Opportunities and challenges

Big Data and AI could act as a transformational force in healthcare, helping to recovery part of the errors made in the past in the digitisation of healthcare processes. First, by automating the process of extracting meaningful information from vast droves of unstructured data at speed. Second, by systematizing repetitive tasks and by ensuring that all the available knowledge, guidelines and protocols are steadily applied in case management. This would enable a complete redesign of healthcare IT systems, that instead of focussing on the process of data recording, could finally support healthcare professionals in the delivery of care. To deliver these promises, several tough challenges need to be tackled.

Since at the core of this transformation we have data, we need to ensure a transparent and accountable access to it. Given its sensitivity, it is essential to develop policies to balance access to data with individual privacy, to ensure data protection and security, and to restore user control of personal data [19]. New models of data ownership, supported by secure by design data platforms, a clear regulatory framework, establishing strong safeguards and a stable market environment, and increased public awareness on how data are collected and used are among the suggested actions by the EU eHealth Task Force.

Another challenge is posed by the lack of interoperability. Health data are usually siloed in departmental, legacy software such as EHR and EMR software for clinical records, RIS and PACS for radiology, etc. Systems that may use different technologies with limited-to-none interoperability. Despite several initiatives at national [20] and European level [21], lack of technical interoperability and heterogeneity of electronic health record are identified as the major barriers to electronic sharing of health data [22].

The Estonian digital transformation project e-Estonia [23] demonstrates that some of these challenges can be successfully tackled. e-Estonia has digitised over 3000 government services, from tax payments to voting. The Estonian National Health Information System integrates data from all the different healthcare providers in a unified record structure and, since 2015, 99% of health data has been digitised. Patients are in control of their data, managing who can access them, while doctors share a common view on their patients. Whether this approach can be effectively transferred to other countries, or scale at cross-national level remains to be seen.

Acknowledgements

The authors thank Dr. Francesco Venturelli and Dr. Marta Ottone (AUSL Reggio Emilia) for sharing the results of their research and for the valuable explanations.

References

1. Eurostat, “Population structure and ageing - Statistics Explained.” 2020.
2. Eurostat, “Healthcare expenditure statistics - Statistics Explained.” 2020.
3. ISTAT, “Il sistema dei conti della sanità per l’Italia,” *Stat. Rep.*, 2017, doi: 10.1126/science.1102156.
4. European Union, “eHealth Task Force Report: Redesigning health in Europe for 2020,” *Off. Eur. Union*, p. 20, 2012, doi: 10.2759/82687.
5. M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245. pp. 255–260, 2015, doi: 10.1126/science.aaa8415.
6. M. Ambigavathi and D. Sridharan, “Big Data Analytics in Healthcare,” *2018 10th Int. Conf. Adv. Comput. ICoAC 2018*, vol. 2015, pp. 269–276, 2018, doi: 10.1109/ICoAC44903.2018.8939061.
7. N. Mehta and A. Pandit, “Concurrence of big data analytics and healthcare: A systematic review,” *Int. J. Med. Inform.*, vol. 114, no. March, pp. 57–65, 2018, doi: 10.1016/j.ijmedinf.2018.03.013.
8. E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1. Springer US, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
9. A. Esteva *et al.*, “A guide to deep learning in healthcare,” *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019, doi: 10.1038/s41591-018-0316-z.
10. E. West, S. Mutasa, Z. Zhu, and R. Ha, “Global trend in artificial intelligence-based publications in radiology from 2000 to 2018,” *Am. J. Roentgenol.*, vol. 213, no. 6, pp. 1204–1206, 2019, doi: 10.2214/AJR.19.21346.
11. CBInsights, “State of Healthcare: Investment & Sectors to watch.” 2020.
12. HIMSS Analytics, “AI use in European healthcare,” no. May. pp. 1–20, 2018.
13. J. M. Ehrenfeld and J. P. Wanderer, “Technology as friend or foe? Do electronic health records increase burnout?,” *Curr. Opin. Anaesthesiol.*, vol. 31, no. 3, pp. 357–360, Jun. 2018, doi: 10.1097/ACO.0000000000000588.
14. R. S. Patel, R. Bachu, A. Adikey, M. Malik, and M. Shah, “Factors related to physician burnout and its consequences: A review,” *Behavioral Sciences*, vol. 8, no. 11. MDPI Multidisciplinary Digital Publishing Institute, 25-Oct-2018, doi: 10.3390/bs8110098.
15. “Unlocking the value of unstructured patient data.” [Online]. Available: https://www.beckershospitalreview.com/healthcare-information-technology/unlocking-the-value-of-unstructured-patient-data.html. [Accessed: 14-Feb-2020].
16. S. Y. Lin, M. R. Mahoney, and C. A. Sinsky, “Ten Ways Artificial Intelligence Will Transform Primary Care,” *J. Gen. Intern. Med.*, vol. 34, no. 8, pp. 1626–1630, Aug. 2019, doi: 10.1007/s11606-019-05035-1.
17. “Clinika: clinical governance ed efficienza in Sanità.” [Online]. Available: https://clinika.mapsgroup.it/. [Accessed: 26-Feb-2020].
18. F. Venturelli *et al.*, “Using text analysis software to identify determinants of inappropriate clinical question reporting and diagnostic procedure referrals in Reggio Emilia, Italy.,” *Manuscr. Submitt. Publ.*, 2020.
19. P. Kostkova *et al.*, “Who Owns the Data? Open Data for Healthcare,” *Front. Public Heal.*, vol. 4, no. February, 2016, doi: 10.3389/fpubh.2016.00007.
20. M. Ciampi, A. Esposito, R. Guarasci, and G. De Pietro, “Towards interoperability of EHR systems: The case of Italy,” *ICT4AWE 2016 - 2nd Int. Conf. Inf. Commun. Technol. Ageing Well e-Health, Proc.*, no. Ict4awe, pp. 133–138, 2016, doi: 10.5220/0005916401330138.
21. M. Gabriel, “Commission Recommendation on a European Electronic Health Record exchange format,” vol. 4, no. March 2011, p. 8, 2019.
22. European Commission, “Consultation: Transformation Health and Care in the Digital Single Market,” 2018, doi: 10.1111/wej.12101.
23. “e-Estonia - We have built a digital society and we can show you how.” [Online]. Available: https://e-estonia.com. [Accessed: 18-Feb-2020].

Statistical methodology for volume-outcome studies

Metodologia statistica per studi di volume

Marta Fiocco and Floor van Oudenhoven

Abstract The recurrent marked point process is used to study a longitudinal volume-outcome association of clustered data. Methodological issues in the selection of aggregate and non-aggregate and yearly measures for hospital volume are considered. An additional aspect associated with hospital volume data, concerns the presence of informative cluster size, where outcome depends on cluster size conditional on covariates. The concept of informative cluster size within a volume-outcome study presents a unique situation since hospital volume is both the covariate of primary interest under study and it is closely linked to cluster size. A new method suitable for volume-outcome studies in which informative cluster size is presented. A simulation study to assess the performance of the method is performed.

Abstract *II recurrent marked point process viene utilizzato per studiare un'associazione longitudinale con il volume nei dati di cluster. Misure aggregate, non aggregate e annuali riguardanti il volume degli ospedali vengono discusse. I risultati sui dati sul volume dipendono dalle dimensioni del cluster in base alle covariate. Il concetto di dimensione informativa dei cluster all'interno di uno studio sul volume presenta una situazione unica poiché il volume dell'ospedale non solo è la covariata di interesse primario oggetto di studio, ma è anche strettamente associato alla dimensione dei cluster. Un nuovo metodo adatto per studi di volume in cui viene considerata la dimensione informativa del cluster viene presentato. Uno studio di simulazione è stato condotto.*

Marta Fiocco
Mathematical Institute Leiden University, Leiden The Netherlands
Dept. of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
Trial and Data Center, Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
e-mail: m.fiocco@math.leidenuniv.nl

Floor van Oudenhoven
Nutricia Research, Utrecht, Erasmus Medical Centre, Rotterdam, The Netherlands
e-mail: floorvanoudenhoven@hotmail.com

Key words: Volume-outcome , Recurrent marked point process, Informative cluster size

1 Introduction

A growing body of literature studies the association between measures of hospital volume and patient outcomes to evaluate whether hospitals with large case volumes are associated with better outcomes. Applying the appropriate statistical methodology to volume-outcome studies erases several challenges such as the selection of a longitudinal estimation method and the specification of an appropriate measure for hospital volume. In daily practice hospital volume is analysed as a categorical variable, neglecting its time-dependent nature. The recurrent marked point process to approach a longitudinal volume- outcome analysis of clustered data is used in this work. Statistical issues in the selection of both non-aggregate and yearly aggregate measures for hospital volume are considered. An additional aspect sometimes associated with clustered data concerns the presence of informative cluster size, where outcome depends on cluster size conditional on covariates. The concept of informative cluster size within a volume-outcome study presents a unique situation since hospital volume is both the covariate of primary interest under study and it is closely linked to cluster size. Within cluster resampling (WCR) is an appropriate method to analyse informative cluster size data. The novelty of this work is to apply WCR in the framework of a recurrent marked point process to study a longitudinal volume-outcome association. A simulation study has been performed to asses the performance of the proposed method and to evaluate whether the use of aggregate measures for hospital volume leads to bias in the estimation of the volume- outcome association.

2 Data description

Data from the Netherlands Cancer institute for patients after oesophageal cancer surgery between 1989 and 2010, covering all hospitals in the country, is used. To improve survival, it is suggested that surgery should be performed in specialized centres with adequate annual volume.

All 10,0025 patients in the dataset underwent oesophageal cancer surgery performed at 148 different hospitals. The outcome of interest is death from any cause within 6 months since surgery. A binary outcome variable is modelled, indicating whether or not the patient is still alive, by using logistic models. Figure 1 shows observed patients' outcomes after 6 months follow-up since surgery per different categories of cumulative hospital size. This figure suggests a possible association between cumulative hospital size and post-treatment outcome. Although many

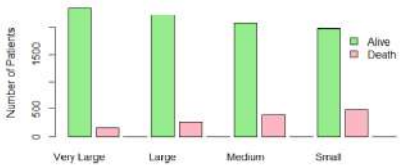


Fig. 1: Observed patient outcome (dead or alive) after 6 months since surgery for each category of cumulative hospital size.

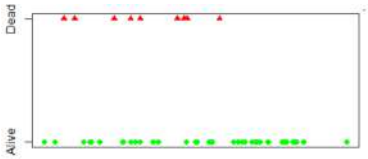


Fig. 2: Marked point process for a specific hospitals with mark space M=Alive, Dead.

volume-outcome studies analyse hospital volume as a categorical variable, this strategy requires great caution. Here hospital volume is analyzed as a continuos variable.

3 Application of a recurrent marked point process

The recurrent marked point process is used as an approach to model volume-outcome associations [1]. This process may be a suitable approach since patients outcome are only observed for patients who underwent surgery. This means that marks are only observed at point locations. Since an outcome only exists when an event of surgery takes place, measurement times differ considerably between hospitals. A recurrent marked point process can cope with different time points, whereas observations in traditional longitudinal data analysis are usually at fixed time points. Use of a recurrent marked point process enables us to interpret the case under study as a point process in one dimension (i.e. time). In this context, surgeries can be seen as ”points” and the corresponding point locations capture information about time the surgery is performed. Figure 2 represents marked point processes for a specific hospitals in the population under study.

Regression methods such as Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMM) can be used to provide estimates under the recurrent marked point process setting by taking into account the dependence within hospitals. However, an assumption of independence between previous outcome and future number of events is required for covariance weighted methods to provide unbiased estimates.

3.1 Fitting a recurrent marked point process model

Different measures for hospital volume are considered. Non-aggregate specification for hospital volume, defined as the cumulative number of surgeries performed at

hospital i through time t, are available at each surgery time. Yearly aggregate measures, defined in three different ways (yearly total volume; cumulative yearly total volume and running average volume), are only available at the end of each year. Figure 3 shows different volume measures over time for a specific hospital in the population under study. Volume specification is the primary challenge of a volume-outcome study [1].

3.2 Statistical model

Let $Y_i(t)$, $X_i(t)$ and $N_i(t)$ denote patient outcome (i.e. mark), covariates and hospital volume respectively, for hospital i $(i = 1,\ldots,n)$ at time point t $(t = 1,\ldots,T)$. According to a marked point process approach a surgery event must occur for an outcome to exist. A surgery which took place at time point t is denoted as $\Delta N_i(t) = N_i(t) - N_i(t-1) = 1$. The complete history is as follows

$$\chi_i(t) = \{X_i(s)|s \le t\};\ \mathscr{N}_i(t) = \{N_i(s)|s \le t\ \};\ \mathscr{Y}_i(t) = \{Y_i(s)|s \le t\}. \qquad (1)$$

A marginal regression model which describes the association between hospital volume and the average outcome among patients satisfying the criteria $\Delta N_i(t) = 1$ is fitted. The partly conditional model which quantify the marginal association between the complete history of the event-time process and the mark process after adjusting for a full history of the covariate process is estimated. The logit link function is used This may be represented as

$$\mu_i(t) = E[Y_i(t)|\Delta N_i(t) = 1, \chi_i(t), \mathscr{N}_i(t)] = g^{-1}(\beta_0 + \beta_1 X_i(t) + \beta_2 N_i(t)) \qquad (2)$$

The parameters β_1 and β_2 quantify the association between the covariate and event-time processes and the average outcome among patients for whom an event of surgery takes place. The model is fitted by Independent Estimating Equation (IEE) and GEE by assuming an exchangeable correlation structure. Both IEE and GEE are fitted by using the `R` package `gee`. A GLMM with hospital specific random intercepts (GLMM-RI), and a GLMM with hospital specific random intercepts and slopes for hospital volume (GLMM-RS) are fitted by using the package `lme4`.

3.3 Results

The table below shows estimated associations between different measures for hospital volume and the odds-ratio of dying within 6 months since surgery for ten-patient increase. Results based on IEE and GEE are population-averaged parameters, quantifying the average volume-outcome associations among the whole population of patients. Results from GLMM-RI (random intercept) and GLMM-RS (random slope) are hospital-specific, assessing the average volume-outcome associations among a

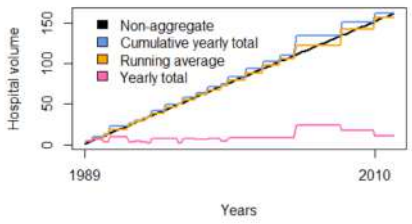


Fig. 3: Volume measures over time for a specific hospital in the population under study.

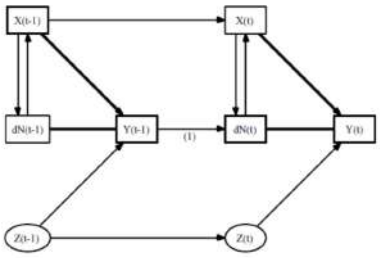


Fig. 4: Underlying framework for the simulation study

population of hospitals. Discrepancies in results might be the consequence of bias, induced in the estimation process because of violation of the assumption about independence between previous patient-outcome and future number of events. In this context the cluster (hospital) size is informative. For the non aggregate cumulative total volume the odds ratio for IEE, GEE. GLMM-RI and GLMM-RS are 0.984(0.979-0.989), 0.997(0.994-1.000), 0.993(0.987-0.999) and 0.958(0.945-0.971) respectively.

4 New approach: within cluster resampling in combination with recurrent marked point process

Within cluster resampling (WCR) [2] is a proper method to analyse informative cluster size data due to one-per-cluster sampling scheme. The novelty of this work is to use WCR in the framework of a recurrent marked point process to study a longitudinal volume-outcome association. The proposed method consists of the following steps

Step 1: Repeatedly (Q times) sample one observation from each cluster I, forming Q resampled datasets.

Step 2: Fit the marginal model on each of the Q resampled datasets, quantifying the volume-outcome associations between hospital volume $N_i(t)$ and patient outcome Yi(t). The estimated parameter $\hat{\beta}_q$ represents the average volume-outcome association within the Q resampled dataset. The variance estimators can take negative values when number of clusters (i.e. hospitals) is too small for the asymptotic variance approximation to hold. This is a drawback of using WCR.

Step 3: The WCR parameter is obtained by taking the average $\widehat{\beta}_{WCR} = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\beta}_q$. WCR is a method suitable to analyse informative cluster size data due to one-per-cluster sampling scheme.

5 Simulation Study

For each hospital i, at each time point t, it is simulated whether or not an event occurs. In case a event takes place a patient-level covariate $X_i(t)$, denoting patient's risk and outcome $Y_i(t)$ are generated. Figure 4 shows conditional relations between patient's risk, the event-time process, patient's outcome and unmeasured error $Z_i(t)$ [3]. Unmeasured error includes serial correlation $W_i(t)$, hospital specific random effects $\gamma_i(t)$ and measurement error $\varepsilon_i(t)$. For each iteration, 1000 hospitals, together with their corresponding patients' risk and outcomes at time t $(1,\ldots,100)$ are randomly selected and volume-outcome analysis is performed. Special focus is on the parameter β_2 since this parameter in the mark process quantifies the effect of hospital volume on patient outcome. WCR, based on 5000 resamplings, is used to obtain mean point estimates for the regression parameter β_2, along with the corresponding mean squared error (MSE), mean standard error, empirical standard error and estimated coverage of 95% confidence intervals. Simulations show that when informative cluster size is present, the proposed method estimates the parameter for volume with small bias. Bias might be introduced when an aggregate measure for present hospital volume is used. Code for simulations study is written in the `R`-software environment and is available upon request.

6 Conclusion

Statistical issues in the specification of both aggregate and non-aggregate measures were considered. Different estimation methods are used to provide volume-outcome estimates under the recurrent marked point process. Results showed an inverse relation between hospital volume and patient mortality, many of them being significant.

Specification of an appropriate measure for hospital volume which takes into account the time-dependent character and bias that may occur in the estimation process when assumptions are violated are also discussed in this work. A new method suitable for volume-outcome studies when informative cluster size is present is proposed. In this research within cluster resampling is used within the framework of the recurrent marked point process, providing cluster-based parameters.

References

1. French, B., Farjah, F., Flum, D. R., Heagerty, P. J.: A general framework for estimating volume-outcome associations from longitudinal data. Statistics in medicine, **31(4)**, 366–382 (2012).
2. Hoffman, E. B., Sen, P. K., Weinberg, C. R.: Within-cluster resampling. Biometrika, **88(4)**, 1121–1134 (2001).
3. French, B., Heagerty, P. J.: Marginal mark regression analysis of recurrent marked point process data. Biometrics, **65(2)**, 415–422 (2009).

Advances in textual data mining

Distance measures for exploring pairs of novels in a large corpus of Italian literature

Misure di distanza per l'analisi di coppie di romanzi di un grande corpus di letteratura italiana

Matilde Trevisani and Arjuna Tuzzi[1]

Abstract In text clustering most distance-based methods summarize the occurrences of a set of linguistic features to obtain a distance. It should decrease when texts are written by the same author, however, there are further properties that might influence the result: gender of the authors, their age, their geographical origin, publication date of the novels, their size, etc. In this study, regression analyses compare the performance of three distances and highlight, among available covariates, the preeminent effect of the author's hand but also interesting patterns in the effect of novels' size.

Abstract *Nel text clustering la maggior parte dei metodi distance-based sfruttano le occorrenze di un insieme di caratteristiche linguistiche per ottenere una distanza, che dovrebbe diminuire quando i testi sono scritti dallo stesso autore, ma ci sono ulteriori proprietà che potrebbero influenzarne il risultato: il genere degli autori, l'età, le origini, la data di pubblicazione dei romanzi, le dimensioni del testo, etc. Questo studio attraverso analisi di regressione confronta tre distanze ed evidenzia come, tra le variabili a disposizione, il fattore autore sia predominante, ma anche pattern interessanti nell'effetto della dimensione dei romanzi.*

Key words: text distances, large corpora, subset selection, average marginal effects, relative importance of regressors

1 Introduction and data

In text mining the classification task plays a relevant role in the exploration of text corpora. The most common distance-based methods include strategies to summarize

[1] Matilde Trevisani, Università di Trieste; email: matilde.trevisani@deams.units.it
Arjuna Tuzzi, Università di Padova; email: arjuna.tuzzi@unipd.it

the occurrences of a set of linguistic features in order to obtain a distance for each pair of texts [8,9]. A distance measure should significantly decrease when a pair of texts have been written by the same author, however, there are potentially further properties that might influence the result: gender of the two authors, their age, their geographical origin, publication date of the novels, their size, etc. Within the specific frame of bag-of-words approaches, this study aims at comparing the performance of three measures to identify what properties of the pair of texts they are able to detect.

In addition to the distance, further choices concerning how many (e.g. all words in the vocabulary or a limited number of most frequent words) and which lexical features (e.g. words, 2-grams, content words) should be considered. Each distance should be tested in different settings thus leading to a large number of comparisons. In this study the distances between pairs of novels have been calculated by three text distances and in settings that involve a large number of words. By means of regression models we started a first exploration of data to understand which properties of the pair of texts emerge as significant in the examples considered.

The corpus of this study represents a reduced version of a large corpus of Italian contemporary literature that has been exploited in previous studies [10]. It includes 143 novels by 39 authors born from 1920 to 1982 in several cities and regions.

A distance matrix is obtained by choosing from three distances: Labbé [5], Delta [1], and Cosine [3]. Delta and Cosine are calculated on most frequent words with at least 143 occurrences (at least one per novel on average) and Labbé has been tested in four variants: the basic version with frequency threshold set to 1 (Labbé-1) and without threshold (Labbé-0), the iterative version [2] based on 200 replications of measures computed on equal-sized text chunks of 5,000 tokens, considering either the whole vocabulary (Labbé) or grammatical words only (Labbé-gramm). Combining novel data (on author: name, gender, year/ city/ region of birth, age at text publication; on text: year of publication, number of word-tokens (N), size (*large* if N>100,000, *small* otherwise)) and distance matrix, a derived matrix is obtained which, for each (unique) pair (a,b) of texts (10,153), records: *y*: distance; *dauthor*: *same* if texts are by the same author, *differ* otherwise; *diffyear*: difference between publication years; *dyear*: *similar* if texts were published in the same period (i.e., *diffyear*<10 years), *diss* otherwise; *dgender*: *same* if gender of authors is the same, *differ* otherwise; *diffbirth*: difference of authors' age; *dbirth*: *similar* if authors are of the same generation (i.e., *diffage*<10 years), *diss* otherwise; *diffage*: age difference between authors at publication date (*diffyear*±*diffage*); *dage*: *similar* if age difference between authors at pubblication date<10 years, *diss* otherwise; *dregion*: *same* if authors are from the same region, *differ* otherwise; *dcity*: *same* if authors are from the same city, *differ* otherwise; *diffN*: difference (in absolute value) of number of word-tokens ($|N_a - N_b|$); *diffN_r*: relative difference in number of word-tokens ($|N_a - N_b| / (N_a + N_b)$); *dsize*: *similar* if texts have similar size category (which can be *large* or *short*, see above), *diss* otherwise; *ratioN*: $\min(N_a, N_b)/\max(N_a, N_b)$.

2 Text distances at comparison

A linear regression model has been tuned for each distance method by a step-wise procedure. (A) Because of high multicollinearity, a preliminary reduction of design matrix *X* has been performed by carrying out an exhaustive subset selection with each regressor forced-in at time across subsets of any size: besides of the goodness-of-fit (R^2) of regression model, sign consistency and significance (*p*-value) of coefficient estimate as well as variance inflation factor (VIF) for regressor have been jointly analyzed to decide which regressors should be removed in sequence. (B) Variable transformation has been opportunely carried out to linearize the relationship between *y* and *X* (Labbé-gramm and Cosine distances as well as all continuous covariates have been square-root transformed due to moderate to serious positive skewness). (C) An exhaustive subset selection is carried out on the reduced *X* to fit the best model without as well as (D) with interaction effects with *dauthor*.

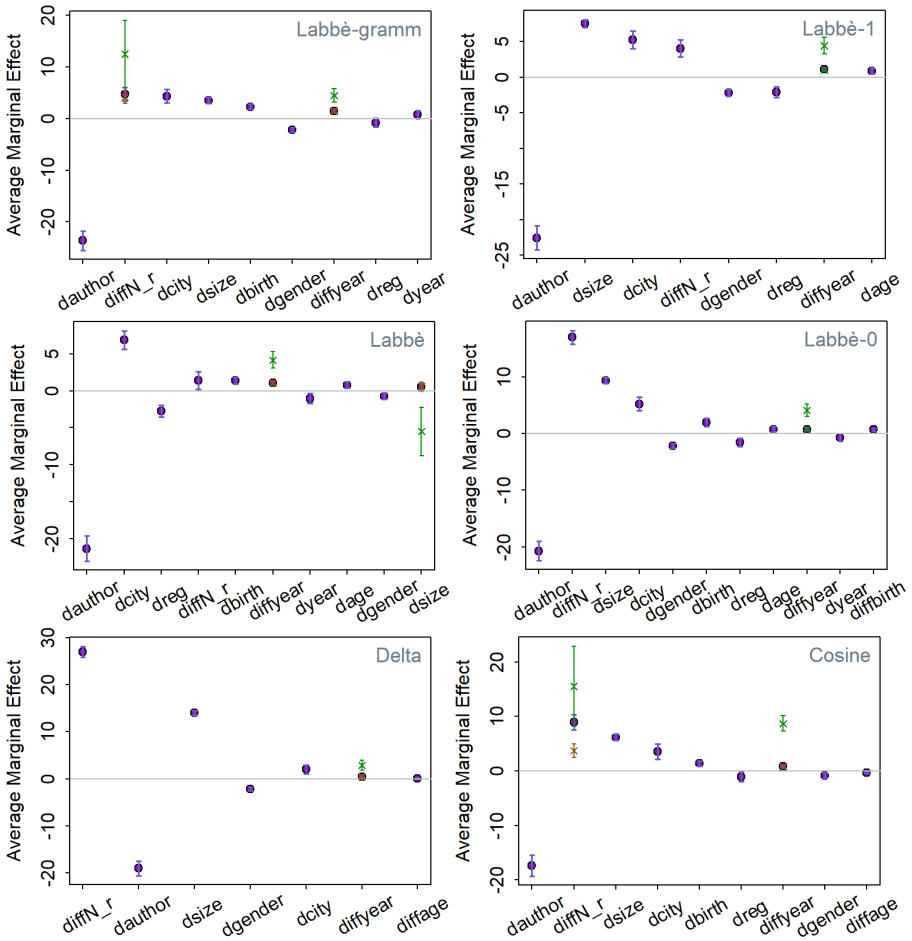
Two measures have been used to compare the effectiveness of distance methods to grasp the author's hand: (1) Average Marginal Effects (AMEs) - average derivative/change (by changing infinitesimally a continuous variable or by 1 an indicator variable) at the values of the covariates as they were observed [6]; (2) relative importance of regressors [4] as measured by LMG method - R^2 partitioned by averaging over orders [7]. AME is useful when model contains non only main effects but more complicate relationships like polynomial (e.g., a quadratic form for *diffN_r* is needed with Delta, Cosine and Labbé-0 distances) or interaction effects (see (D) above). Notice that to compare distances having different ranges, marginal effects have been calculated as % of distance range. LMG overcomes the issue of calculating partial R^2 given the order of regressors' inclusion. To complete the comparison, First - each regressor contribution when included first, which is just the squared covariance between *y* and the regressor - has been added although it is a rawer measure of comparison. In Table 1, rankings of text distances have been highlighted by colouring metric value with different intensity. Adjusted R^2 (R^2_a) summarizes the overall importance of regressors to explain *y* variation.

Table 1: AME %, LMG, First for dauthor, R^2_a for the selected model

Measure/Distance	Labbé-gr	Labbé-1	Labbé	Labbé-0	Delta	Cosine
AME %	23.6	22.5	21.3	20.7	19.0	17.3
LMG	6.1	6.2	5.7	5.6	4.7	3.3
First	6.6	7.0	5.9	6.8	5.7	3.5
R^2_a	11.4	17.1	10.3	20.8	30.6	8.9

In Figure 1, regressors have been ordered according to AME (note that *p*-values are not proper: a statistically significant result may not be practically significant, however, predictors are significant almost everywhere - see 95% CI). Interaction effects are indicated wherever present in the final model selected in (D). In Figure 2, regressors are ordered according to their relative importance in the final model selected in (C).

Figure 1: AMEs with 95% CI (% of distance range). Distances are ordered according to AME-based ranking (from top-left to bottom-right). Interaction effects are added wherever present (green if texts are written by the same author, orange if written by different authors).

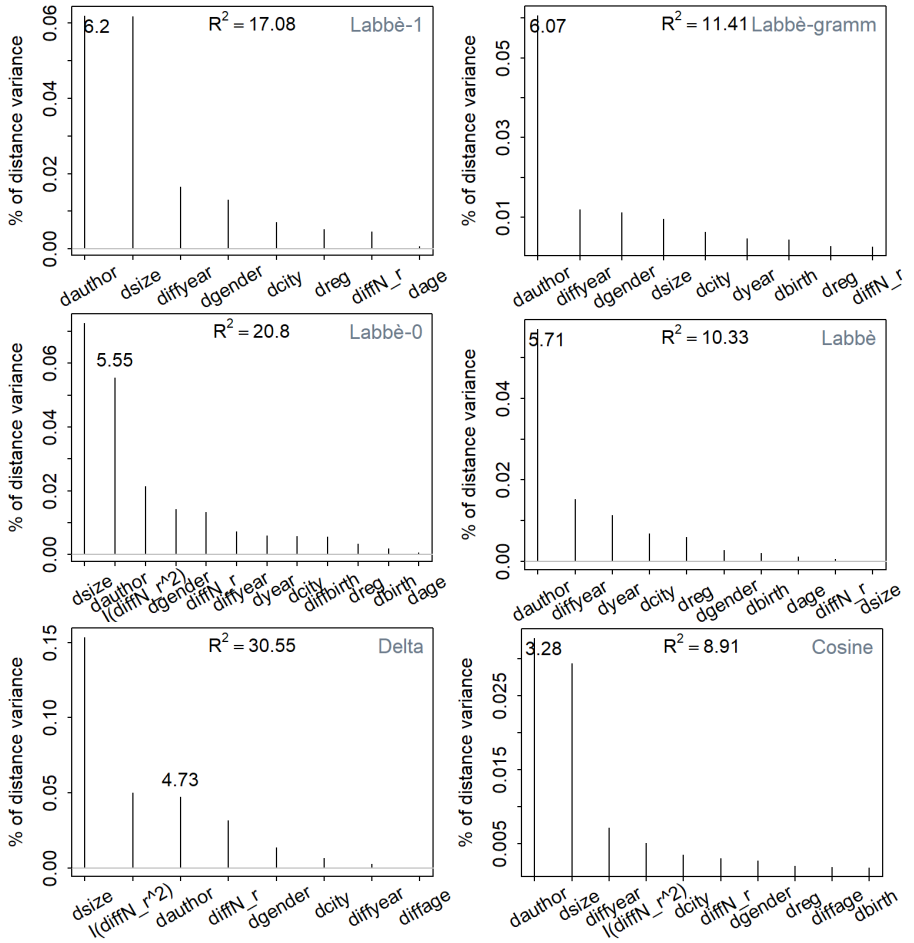


In general, the author factor is highly significant for all distance methods (Figure 1), with a degrading effect from Labbé-gramm (distance drops by -24%, Table 1) to Cosine (-17%). Moreover, relative importance of regressors highlights how the author factor is the most important one in almost all cases (Figure 2), confirming once again a ranking of methods: from about 6% for Labbé-gramm and Labbé-1 down to 3% for Cosine (Table 1). Notice also that the relative importance of author factor weights differently on the overall distance variation explained by the model (R^2_a): more than half for Labbé-gramm and Labbé down to less than one sixth for Delta (passing through one third for Labbé-1 and Cosine and a quarter for Labbé-0).

Another regressor that (except for Labbé and Labbé-gramm) plays a relevant role in explaining distance is text size (in particular for Delta where it is the most

important regressor). The difference in publication year as well as authors' gender and place of birth show a moderate influence almost everywhere. For novels by the same author, difference in year publication – always – and text size – in half of the cases – affect distance (see interaction effects superimposed in Figure 1, AMEs for same author are higher).

Figure 2: Relative importance of effects (LMG method). Distances are ordered according to LMG-based ranking (from top-left to bottom-right).



3 Conclusions

In almost all distance methods the author factor is the most important among the regressors. However, Labbé methods show to be the most effective in detecting the

author's hand (with Labbé-gramm and Labbé-1 at the top), Delta being at an intermediate position, and leaving Cosine as the last choice. Cosine also appears as the distance least explained by the overall factors hypothesised as important (R^2_a≈9%). On the opposite side there is Delta (R^2_a≈31%), Labbé-0 and Labbé-1 are intermediate (≈20%), while Labbé-gramm and Labbé (≈11% and ≈10%) appear to be almost totally explained by the sole author factor (Figure 2). Text size emerges as a second preeminent factor, difference in year publication as well as author's gender and place of birth play a not negligible role. Distance between texts written by the same author is significantly affected by their temporal distance (difference in year publication) and, for some methods, by text size.

This is only a first step in a territory that has been already explored with many methods but needs further systematic investigation. Moving from these first results the future direction of this explorative study consists in increasing the number of distances under investigation and testing their performances in different settings and with different text genres.

References

1. Burrows, J.F.: Delta: A measure of stylistic difference and a guide to likely authorship. Lit. Linguist. Comput. **17**(3), 267-287 (2002)
2. Cortelazzo, M.A., Nadalutti, P., Tuzzi, A.: Improving Labbé's Intertextual Distance: Testing a Revised version on a Large Corpus of Italian Literature. J. Quant. Linguist. **20**(2), 125-152 (2013) doi: 10.1080/09296174.2013.773138
3. Feldman R., Sanger J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)
4. Grömping, U.: Relative Importance for Linear Regression in R: The Package relaimpo. J. Stat. Softw. **17**, 2-27 (2006)
5. Labbé, C. and Labbé, D.: Inter-textual distance and authorship attribution Corneille and Molière. J. Quant. Linguist. **8**, 213-231 (2001)
6. Leeper, T.J.: margins: Marginal Effects for Model Objects. R package version 0.3.25. (2018)
7. Lindeman, R.H., Merenda, P.F, Gold, R.Z.: Introduction to Bivariate and Multivariate Analysis. Scott Foresman, Glenview Ill (1980)
8. Rudman, J.: The state of authorship attribution studies: Some problems and solutions. Comput. Humanities **31**, 351-365 (1997) doi: 10.1023/A:1001018624850
9. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. J. Am. Soc. Inf. Sci. Tec. **60**, 538-556 (2009) doi: 10.1002/asi.21001
10. Tuzzi, A., Cortelazzo, M.A.: What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer. Digit. Scholarsh. Hum. **33**(3), 685-702 (2018) doi: 10.1093/llc/fqx066

Supervised vs Unsupervised Latent Dirichlet Allocation: topic detection in lyrics

Allocazione di Dirichlet latente supervisionata vs. non supervisionata: identificazione di topic da testi musicali

Mariangela Sciandra, Alessandro Albano, Irene Carola Spera

Abstract Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a fixed number of topics starting from words in each document modeled according to a Dirichlet distribution. In this work we are going to apply LDA to a set of songs from four famous Italian songwriters and split them into topics. This work studies the use of themes in lyrics using statistical analysis to detect topics. Aim of the work is to underline the main limits of the standard unsupervised LDA and to propose a supervised extension based on the Correspondence Analysis (CA) association theory.

Abstract *Il topic modeling è un metodo analitico per estrarre gruppi lessicali chiamati "topic" da un insieme di documenti sulla base di calcoli statistico-probabilistici. La Latent Dirichlet Allocation (LDA) è un esempio di topic model che viene usata per analizzare dei testi che si riferiscono ad argomenti specifici. Questo metodo individua un numero fissato di "topic" a partire dalle parole in ogni testo, assumendo una distribuzione Dirichlet. In questo lavoro applichiamo la procedura LDA ad un dataset costituito dalle canzoni di quattro famosi cantautori italiani. Lo scopo del lavoro è sottolineare i limiti principali della procedura LDA standard non supervisionata e proporre un'estensione supervisionata basata sulla Correspondence Analysis (CA).*

Key words: Topic modeling, LDA, Correspondence Analysis, Music mining

---

Mariangela Sciandra
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Viale delle Scienze Edificio 13, 90129 PALERMO, e-mail: mariangela.sciandra@unipa.it

Alessandro Albano
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Viale delle Scienze Edificio 13, 90129 PALERMO, e-mail: alessandro.albano@unipa.it

1 Introduction

Nowadays our collective knowledge is mainly based on extracting information from different written resources in the form of news, blogs, Web pages, scientific articles, books, images, but also from sound, video, and social networks. This large amount of document corpora that can be easily stored is leading to a growing need to analyze large collections of electronic text and makes more difficult the search of what we are looking for. New computational tools are then necessary in order to organize, search, and understand these vast amounts of information. This has led to considerable interest in applying statistical methods able in deriving high-quality information from text. Typical tasks of this process, also known as text mining or text data mining, include text categorization, text clustering, production of taxonomies and sentiment analysis [7]. A common step of all these procedures consists, essentially, in turning text into data for analysis, via application of Natural Language Processing (NLP) [5], different types of algorithms and analytical methods [4]. The final goal this process is the interpretation of the gathered information. The process of learning, recognizing, and extracting meaning hierarchically from words to sentences to paragraphs to documents is known as "*topic modeling*" [1]. Topic models are algorithms for discovering the latent semantic structure ("topics") that pervade a large and otherwise unstructured collection of documents. Once topics have been detected the collection can be organized according to the discovered themes. There are multiple methods of going about doing this in the NLP framework but in this work we will focus on the Latent Dirichlet Allocation (LDA) [3]. LDA is a topic model which assumes that the words of each document arise from a mixture of topics. The topics are shared by all documents in the collection; the topic proportions are document-specific and randomly drawn from a Dirichlet distribution.
Topic modeling algorithms can be adapted to many kinds of data. In this work we apply the standard unsupervised LDA procedure to find patterns in lyrics from different songwriters and we address a limitation of the procedure in making predictions on singers starting from the topics detected on a set of songs. To avoid this problem we propose a supervised LDA obtained by combining standard LDA with a correspondence analysis applied on the LDA results (LDA-CA).
The article is structured as follows: after this introduction, section 2 explains the method employed, followed by section 3, where the main results are described and discussed trough an example on real Genius data.

2 Topic models: unsupervised and supervised approach

Topic modeling is part of a class of text analysis methods that analyze "bags" or groups of words together—instead of counting them individually–in order to capture how the meaning of words is dependent upon the broader context in which they are used in natural language. The starting assumption is that documents have latent semantic structure ("topics") and the challenge is to infer them from word–document

co-occurrences. This depends heavily on the quality of text preprocessing and the strategy of finding the optimal number of topics that is out of the interest of this work. Formally, a topic is a probability distribution over terms in a vocabulary. To introduce the notation let $Pr(z)$ be the probability distribution of topics in a particular document, $Pr(w|z)$ be the probability distribution given a topic z;$Pr(z_i = j)$ will be the probability that the j-th topic will be extracted for the i-th word and $Pr(w_i|z_i = j)$ the probability of the word under topic j. The following distribution of words in a document is derived:

$$Pr(w_i) = \sum_{j=1} Pr(w|z = j)Pr(z = j) \tag{1}$$

Equ. (1) shows as the topic model uses a Bayesian approach: the probability of finding a word in a text comes from the product between the probability of finding a certain topic, $P(z)$, and the probability of finding the same word conditioning on the chosen topic, $P(w|z)$.

2.1 Latent Dirichlet Allocation

The LDA model is a topic modelling technique that was first described by Blei, Ng and Jordan in 2003 [2]. In the LDA model topics are considered to be probability distributions over the finite vocabulary. We denote by V the vocabulary of the corpus and $W_1, W_2, \ldots, W_D$ the documents in the corpus (each assumed to contain N_d words). We denote by $W_{d,n}$ (for $1 \leq d \leq D$; $1 \leq n \leq N_d$) the n-th entry in W_d, i.e. the n-th word in the d-th document of the corpus. Furthermore, we denote the (n-1)-simplex by Δ^{n-1} and $\theta_1, \ldots, \theta_D \in \Delta^{K-1}$ are the K topic proportions (which are distributions over the K topics): an n-vector θ lies in the $(n-1)$- simplex if $\theta_i \geq 0$, $\sum_{i=1}^{n} \theta_i = 1$; let $\beta_1, \ldots, \beta_k \in \Delta^{V-1}$ be the topics (which are distributions over words), and $Z_{d,n} \in 1, \ldots, K$ for $1 \leq d \leq D$, $1 \leq n \leq N$, the topic of word n in document d. Standard LDA assumes that producing a document is a random process described by the following generative model:

1. Choose topics β_1, β_2, …, $\beta_K \sim Dir(\eta)$, where $\eta \in \mathbb{R}^+$ a parameter .
2. For $d = 1, \ldots, D$, choose the topic distribution of document $d \sim Dir(\alpha)$, as:
   (a) Choose the topic of the n-th word, $Z_{d,n} \sim$ Multinomial(θ_d).
   (b) Choose the n-th word, $W_{d,n} \sim$ Multinomial($\beta_{Z_{d,n}}$).

2.2 Latent Dirichlet Allocation-CA

Standard LDA can be considered an 'unsupervised' machine learning process because it does not require a predefined list of tags or training data that has been previously classified by researchers. However, its main limit is that it cannot guarantee

final accurate results, which is why many users opt to invest in *topic classification models*. As opposed to text modeling, topic classification needs to know the topics of a set of texts before analyzing them and so they can be considered "supervised" techniques. In this work we propose a supervised LDA procedure that uses the topics identified in a first standard LDA and then it applies a Correspondence Analysis (CA) [6] that takes advantage of the *a priori* information about clusters. The proposed procedure is named $LDA-CA$ as it consists of this two main steps:

1. Find topics $\beta_1, \beta_2, \ldots, \beta_K$ by applying a standard unsupervised LDA procedure;
2. Use a classical Correspondence Analysis to explore relationships among topics identified by the LDA and the known cluster information (where M is the number of real clusters). In particular:
   a. Read in the M (rows) by K (columns) data matrix, L. Note that the elements of K must be non-negative and that none of the row or column totals is zero.
   b. Compute the proportion matrix, P, by dividing the elements of K by the total of all numbers in K. Mathematically, we write $P = \{p_{ij}\} = \{k_{ij}/k_{..}\}$.
   c. Compute the totals of the rows of P and the columns of P, putting the results in the vectors r and c that will be respectively the row and column profiles on which distances will be calculated once coordinates on the reduced space will be evaluated.

Although topic classification is more complex, this topic analysis technique delivers more accurate results than unsupervised techniques, which means more valuable cluster predictions could be derived.

3 LDA-CA in practice: topics in lyrics

The dataset is made up of 100 lyric songs extracted from genius Web API of 4 famous Italian songwriters (25 songs per singer): Antonacci, Battiato, Ligabue e Nomadi. The corpus was preprocessed and after a tokenization it was cleaned by removing stop words. Table 1 shows the final number of words used by each author:

Antonacci	Battiato	Ligabue	Nomadi	Total
2245	1396	1730	1696	7067

Table 1: Number of words per songwriter.

Then next step is to put together all the 100 songs in a unique corpus such that the individual songs are unlabeled. LDA was applied to discover how songs cluster into distinct topics, each of them presumably representing a specific songwriter. In this example, each document d is represented by a single song. The aim is to know how topics are associated with each document, starting from the per-document-per-topic estimated probabilities $\hat{\theta}_i$.

Topic from the LDA are plotted in Fig. 1 (word-clouds showing the 25 most important words).

(a) Topic 1 (b) Topic 2 (c) Topic 3 (d) Topic 4

Fig. 1: Wordclouds of Topics

Except for topic 1, the word "Amore" is the most common frequent word.
For each song, we get an estimated proportion of words that are generated from the identified topics. Fig. 2 shows the distributions of words in each of the four topics conditioning on the four singers. It could give an idea about how well our unsupervised learning works: the ideal situation should be the one in which songs from a songwriter belong to the same topic.

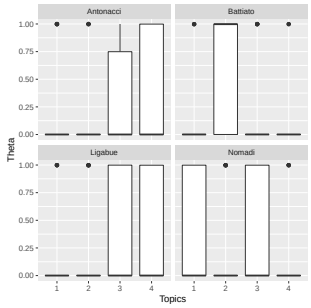


Fig. 2: The θ probabilities for each song within each songwriter.

Unfortunately we notice that almost all of the songs by Antonacci, Ligabue and Nomadi do not not belong to unique topic; on the contrary, Battiato is mostly identified with topic 2. Crossing the topic information from LDA and the known singer clustering we obtain contingency Table 2 that we use as starting point of the CA.

	Topic			
Songwriter	1	2	3	4
Antonacci	414	349	615	867
Battiato	146	792	335	123
Ligabue	297	383	546	504
Nomadi	526	396	470	304

Table 2: Number of words per songwriter and topic.

A global view of CA results useful for interpretation is shown in Fig. 3:

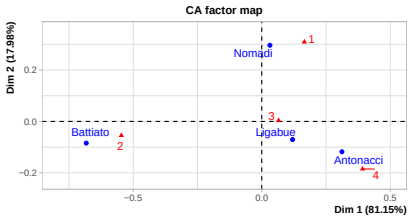


Fig. 3: CA biplot from data in Tab. 2.

Songwriter	Dim 1	Dim 2	Dim 3
Antonacci	25.2	16.0	27.0
Battiato	72.2	5.6	2.5
Ligabue	2.5	3.6	69.5
Nomadi	0.1	74.8	1.0

Table 3: Singer contributions to the CA dimensions.

Fig 3 confirms the goodness of CA results showing the association between topic clustering and songwriters: the percentage of variance explained by Dimension 1 is very high (81.15%). In particular, the result is due to Battiato's songs, for the first dimension, and to Nomadi for the second one.

Conclusions

In conclusion, our method can be used for predicting song classification once topics have been generated. Future work will be addressed in quantifying total classification error and comparing the proposed procedure with a classification tree approach.

References

1. BLEI, D. M. Probabilistic topic models. *Commun. ACM 55*, 4 (Apr. 2012), 77–84.
2. BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, null (Mar. 2003), 993–1022.
3. BLEI, D. M., NG, A. Y., JORDAN, M. I., AND LAFFERTY, J. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 2003.
4. IGNATOW, G., AND MIHALCEA, R. *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. SAGE Publications, 2017.
5. JURAFSKY, D., AND MARTIN, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artifi. Pearson Prentice Hall, 2009.
6. NENADIC, O., AND GREENACRE, M. Correspondence analysis in r, with two- and three-dimensional graphics: The ca package.
7. WEISS, S., INDURKHYA, N., ZHANG, T., AND DAMERAU, F. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer New York, 2010.

Advances in the interaction between artificial intelligence and official statistics

Automated Land Cover Maps from Satellite Imagery by Deep Learning

Mappe di Copertura del Suolo Automatiche da Immagini Satellitari mediante Deep Learning

Fabrizio De Fausti, Francesco Pugliese and Diego Zardetto

Abstract The Italian National Institute of Statistics (Istat) is currently investigating whether Deep Learning methods could be used to derive automated Land Cover estimates of satisfactory quality from Sentinel-2 satellite images. A prototype software system is being developed within the scope of this research. This paper focuses on "automated land cover maps", a very relevant output artefact of the system.
Abstract *Istat sta studiando se sia possibile derivare stime automatizzate di Copertura del Suolo di qualità accettabile da immagini satellitari Sentinel-2 processate con metodi di Deep Learning. Nell'ambito del progetto è in via di sviluppo un sistema software. Questo articolo si concentra su un importante output del sistema: le mappe di Copertura del Suolo automatiche.*

Key words: Land Cover, Satellite Imagery, Deep Learning, Convolutional Neural Networks

1 Introduction

Timely and frequently updated Land Cover (LC) information is of paramount importance to modern National Statistical Institutes (NSI).

As far as Europe is concerned, two flagship LC projects exist: CORINE [2, 3], currently run by the Copernicus Program, and LUCAS [1, 4], managed by Eurostat. Despite these projects address the study of land cover very differently – CORINE in a cartography (i.e. full-coverage) perspective, LUCAS in a statistical estimation (i.e. sample survey) perspective – they suffer common shortcomings. Both are very costly, have very complex production pipelines, rely heavily on clerical work, and produce their outputs with a rather low time frequency. Most of the shortcomings affecting CORINE and LUCAS depend on the huge amount of *human workload* they require. It is, therefore, very tempting to try to overcome these shortcomings through *process automation*. Given an input satellite image depicting a portion of territory, a fully automatic system should ideally be able to (i) *classify* the territory according to some standard LC taxonomy, and to (ii) *quantify* the area (or the proportion) of territory covered by each LC class, without any human intervention.

[1] Fabrizio De Fausti, Istat; defausti@istat.it
Francesco Pugliese, Istat; frpuglie@istat.it
Diego Zardetto, Istat; zardetto@istat.it

The Italian National Institute of Statistics (Istat) is currently investigating whether Deep Learning [5] methods could be used to derive automated Land Cover estimates of satisfactory quality from Sentinel-2 satellite images. A prototype software system is being developed within the scope of this research. This paper focuses on "automated land cover maps", a very relevant – though quite specific – output artefact of the system.

2 Methodology

Istat research goal is to design and develop an automatic LC estimation system. Such a system should be able to take as input a satellite image depicting a portion of territory, and to return as output a table of LC statistics.

Although LC estimation is a quantification problem rather than a classification one, we decided to implement our system according to a 'classify-and-count' design. The main driver of this design choice was to incorporate into our system a Convolutional Neural Network (CNN), so as to take advantage of its tremendous performance in image classification tasks. CNNs [7, 8] are cutting edge Deep Learning architectures that have recently reached superhuman accuracy in many Computer Vision tasks and whose topology was originally inspired by the organization of the visual cortex of mammals.

Without going into technical details, our classify-and-count design can be summarized as follows:

(0) Train a CNN to predict the LC class of a satellite image 'tile' (i.e. a small, fixed-size sub-image).

(1) Divide the satellite images covering a 'target area' (i.e. the territory for which LC statistics have to be computed) into tiles.

(2) Use the trained CNN to predict the LC class of all the generated tiles.

(3) Obtain LC statistics for the target area by simply computing the relative frequencies of predicted LC classes.

It ought to be clear that phases (1), (2), (3) have to be repeated each time LC statistics are requested for a new target area, whereas the CNN's training phase is carried out only once (whence the (0) index in the list).

3 Training Data

We decided to adopt the EuroSAT dataset [8] as training set for our CNN. EuroSAT contains 27,000 manually labelled image patches of size 64 x 64 pixels. These patches have been cropped from carefully selected Sentinel-2 satellite images covering 34 European countries. EuroSAT images are multispectral (all 13 Sentinel-2 bands are provided) but we have so far restricted our interest to Red, Green and Blue bands only (i.e. to RGB color images). Since the resolution of Sentinel-2 images in the R, G and B bands is 10 meters per pixel, each 64 x 64 EuroSAT patch represents a ground area of 640^2 square meters, i.e. about 41 hectares.

The LC classification according to which EuroSAT patches have been manually labelled entails 10 classes: 1) 'Annual Crop', 2) 'Forest', 3) 'Herbaceous Vegetation', 4) 'Highway', 5) 'Industrial', 6) 'Pasture', 7) 'Permanent Crop', 8) 'Residential', 9) 'River', 10) 'Sea & Lake'. EuroSAT authors have defined this LC taxonomy following the principle that the

patterns of each class should be visible at the resolution of 10 meters per pixel. The dataset is roughly balanced with respect to the 10 classes, as class cardinalities range from 2,000 to 3,000 patches.

4 CNN Model Training and Accuracy

To implement the classification engine of the system, we are currently using a cutting-edge, highly sophisticated CNN model named Inception-V3 [9], which we customized and trained on the EuroSAT dataset. As far as the training stage is concerned, we randomly split the EuroSAT data into training set and test set according to a 75/25 proportion. The generated training set and the test set contain 20,250 and 6,750 image patches, respectively.

Fig. 1 below reports the Confusion Matrix obtained contrasting the LC classes predicted by our best model and the true LC labels of the 6,750 image patches belonging to the test set. We achieved an accuracy of 98.43%.

PRED. / TRUE	Ann. Crop	Forest	Herb. Veg.	Highway	Industrial	Pasture	Perm. Crop	Residential	River	Sea Lake
Ann. Crop	740	0	1	1	0	4	4	0	0	0
Forest	0	747	2	0	0	0	0	0	0	1
Herb. Veg.	0	3	738	0	0	2	7	0	0	0
Highway	1	1	0	613	1	1	0	2	6	0
Industrial	0	0	0	1	621	0	1	2	0	0
Pasture	6	0	11	0	0	479	3	0	0	1
Perm. Crop	13	0	12	0	0	2	598	0	0	0
Residential	0	0	1	0	1	0	0	748	0	0
River	1	1	0	8	0	1	0	1	613	0
Sea Lake	1	0	0	0	0	1	0	0	1	747

Figure 1: The Confusion Matrix obtained contrasting the LC classes predicted by our Inception-V3 model and the true LC labels of the 6,750 image patches belonging to the test set. The trace of the matrix gives the overall number of exact predictions (i.e. 6,644), which implies an accuracy of 6,644/6,750, i.e. 98.43%.

5 Land Cover Estimation Algorithm

Once the CNN has been trained on the EuroSAT dataset, our automatic LC estimation system can be fed with a satellite image and return LC statistics for the corresponding territory. To do so, a classify-and-count algorithm is used, whose main logical steps can be summarized as follows:

S1) The input Sentinel-2 image is split into a set of (possibly overlapping) tiles of size 64 x 64 pixels. These tiles are generated by cropping the input image along a regular spatial grid, through a 'sliding window' algorithm.

S2) The trained CNN classifies one tile at a time and logically links the predicted LC class to the corresponding area of the original image. The output of the whole process is a 'classification matrix': each element of this matrix corresponds to a tile of the original image and stores its predicted LC class.

S3) The area share of each LC class for the whole territory depicted in the input satellite image is estimated by the relative frequency of the corresponding label within the classification matrix.

S4) A land cover map of the territory depicted in the input satellite image is obtained by rendering the classification matrix as a raster image.

The working mechanism of the sliding window algorithm mentioned in **S1)** is schematically illustrated in Fig. 3. Basically, a window of 64 x 64 pixels slides horizontally and vertically over the input image with a stride (i.e. step length) of *s* pixels, starting from its upper-left corner. For each step of the window, one tile is generated by cropping the area of the input image that is framed by the window. This way, the algorithm actually produces a *systematic spatial sample* of tiles drawn from the input image. Note that, since each generated tile corresponds to a specific area of the input image, the output sample has an intrinsic geometrical structure. More specifically, the generated tiles are naturally arranged according to a regular spatial grid (see Fig. 2).

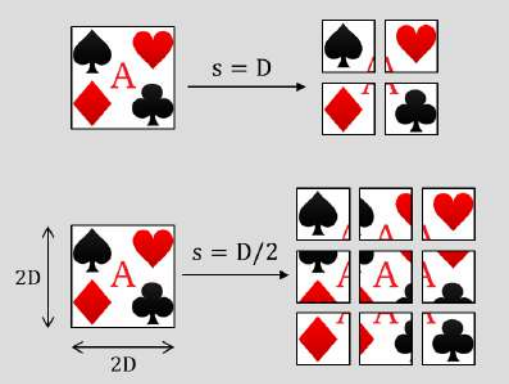


Figure 2: Illustration of the sliding window algorithm. A convenience image of size 2D x 2D is split into tiles of size D x D. In the upper panel, the window slides horizontally and vertically with a stride of length D, giving rise to 4 non-overlapping tiles arranged according to a 2 x 2 grid. In the lower panel, the stride is reduced to D/2: this generates 9 partially overlapping tiles arranged along a 3 x 3 grid. Note that reducing the stride from D to D/2 allowed to resolve more image details: for instance the red 'A', which in the upper panel was not framed in any tile of the grid, now pops up in the central tile of the lower panel grid.

In **S2)** the trained CNN is used to predict the LC class of all the tiles generated in **S1)**. Note, incidentally, that different tiles can be processed independently, allowing our system to take advantage of high-performance parallel computing architectures (GPUs).

In **S3)** the system calculates output LC statistics from the classification matrix. In accordance with our classify-and-count approach, this is accomplished by simply computing class frequencies. If we indicate with W and H the input image width and height in pixels, with c a generic LC class and with f_c the proportion of class c within the classification matrix, then the corresponding area and area share are estimated by:

$$\begin{cases} Area_c = (f_c \cdot W \cdot H) \cdot 100m^2 \\ AreaShare_c = f_c \end{cases} \quad (1)$$

Note that in the upper equation of (1) we took into account that the resolution of the satellite images processed by our system is 10 meters per pixel.

While the LC statistics calculated in **S3)** have to be regarded as the main output of our system, a further interesting artefact can be distilled, as a by-product, from the classification matrix. Indeed, as mentioned in **S4)**, a land cover map can be produced by simply rendering the classification matrix as a raster image. It is worth stressing that this is only possible because of

the geometric structure of the systematic spatial sample of tiles generated by the sliding window algorithm. Clearly, the smaller the stride, the larger will be the dimension of the classification matrix and, therefore, the resolution of the obtained land cover map.

6 Automated Land Cover Maps

We tested and validated our system on several Sentinel-2 images representing quite different Italian territories. Test images have been cropped from Sentinel-2 products downloaded from Copernicus Open Access Hub, namely TCI (True Color Image) objects encoded in JPEG2000 format. Due to space limitations, we focus here on just one test image, which we will refer to as the *"Lecce image"*. We briefly analyze here the automated LC maps that our system generated from the Lecce image as by-products of LC estimation. Recall that our system produces LC maps by simply rendering as a raster image the classification matrix computed for LC estimation. The success of this approach entirely rests on the inherent spatial structure of the sample of tiles determined by the sliding window algorithm (Section 5). Since both the LC estimates and the LC maps produced by our system improve as the stride of the sliding window decreases, we provide here results obtained by setting the stride to its minimum value of 1 pixel. This setting generated (39 x 47) = 1,833 tiles for the Lecce image.

Fig. 3 shows the Lecce image (left panel) and its automated LC map (right panel). Overall, the map exhibits a high degree of spatial consistency, in that the main structures (urban centers, industrial areas, highways, crops and vegetation) have been correctly detected and nicely reconstructed.

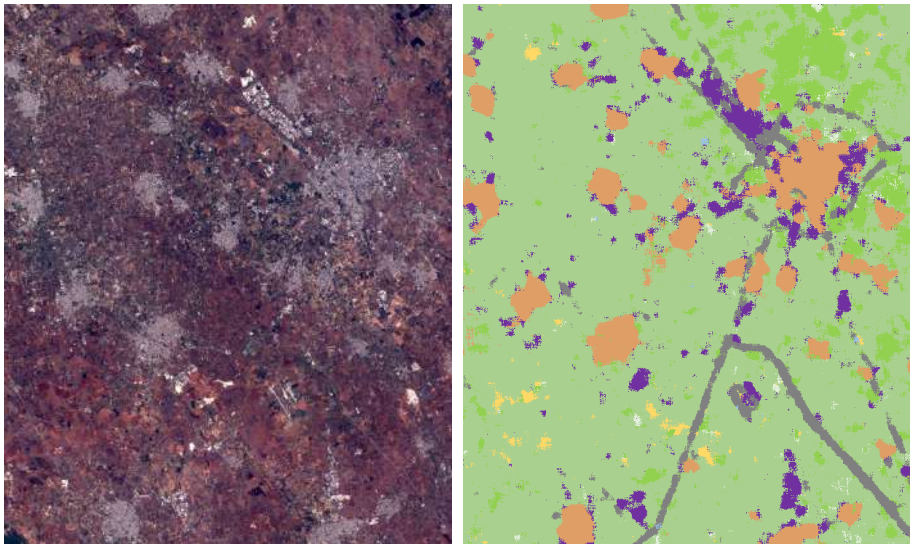


Figure 3: Automated LC map (right panel) of the territory depicted in the Lecce image (left panel).

For instance, focusing on residential areas (brown pixels on the map) and comparing visually the map with the original image, one can observe that the sizes, the shapes and the relative positions of cities are all described fairly well by the map.

Fig. 4 offers an easier way to appreciate the spatial consistency of our LC map with respect to urban areas. To build the left panel of this figure, we first extracted the edge of the

'Residential' areas from the LC map using the Canny Edge Detector algorithm. Then, we overlaid the obtained edges on the Lecce image using a GIS. Evidently, the green edge-line outlines with remarkable accuracy all the cities that are visible in the Lecce image.

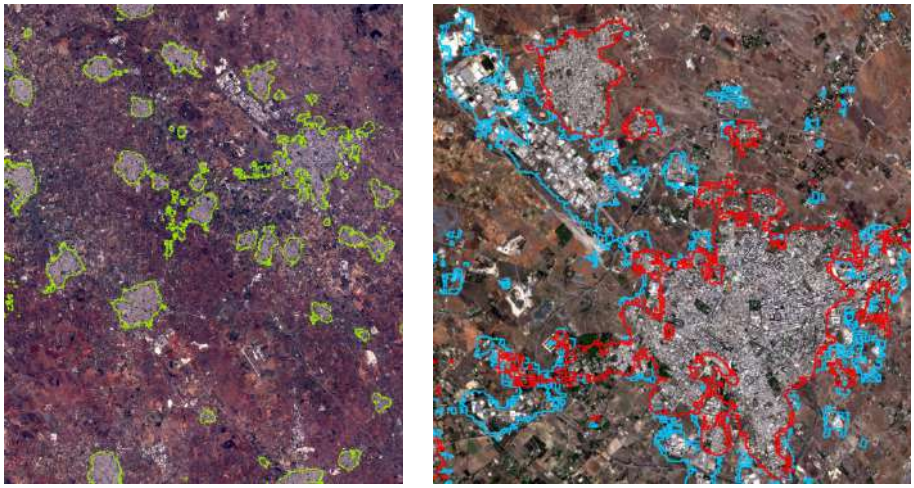


Figure 4: Left panel: the Lecce image overlaid with the edge of the 'Residential' class (green line) extracted from the automated LC map in Fig. 3. Right panel: a detailed view of the city of Lecce overlaid with the edges of the 'Residential' (red line) and 'Industrial' (blue line) classes extracted from the automated LC map in Fig. 3.

The right panel of the figure shows a detailed view of Lecce (taken, of course, from the input satellite image) overlaid this time with two edge-lines: the red one for the 'Residential' class and the blue one for the 'Industrial' class. The segmentation ability of our system emerges very neatly.

References

1. Bettio M, Delincé J, Bruyas P, Croi W, Eiden G. Area frame surveys: aim, principals and operational surveys. Building Agri-environmental indicators, focussing on the European Area frame Survey LUCAS. 2002:12-27.
2. Bossard M, Feranec J, Otahel J. CORINE land cover technical guide: Addendum 2000.
3. Büttner G. CORINE land cover and land cover change products. In Land use and land cover mapping in Europe 2014 (pp. 55-74). Springer, Dordrecht.
4. EUROSTAT. The Lucas survey - European statisticians monitor territory. Office for Official Publications of the European Communities. 2003 Aug 17.
5. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016 Nov 10.
6. Helber P, Bischke B, Dengel A, Borth D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2019 Jun 14.
7. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989 Dec; 1(4):541-51.
8. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. 1995 Apr; 3361(10):1995.
9. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).

CROWD4SDG: Crowdsourcing for sustainable developments goals

CROWD4SDG: Crowdsourcing per gli obiettivi di sviluppo sostenibile

Barbara Pernici

Abstract While a set of measurable indicators has been defined for Sustainable Development Goals (SDG), in some cases the collection of data is problematic. The CROWD4SDG project proposes to support this data collection using Citizen Science. The project is based on three pillars: research for Citizen Science based on new IT tools, research by Citizen Science to facilitate the generation of bottom-up projects, and research on Citizen Science for the development of high-quality Citizen Science projects identifying best practices. The CROWD4SDG project will focus on a set of SDGs, centering on Climate action goals, in particular targeting climate hazards, in combination with SDGs on gender equality, sustainable cities, and rights. The project will exploit also the research results obtained in the E2mC project, which developed an approach to information extraction from social media based on geographic information management, AI, and crowdsourcing in the context of emergency situations after natural disasters.

Abstract *Mentre è stata definita una serie di indicatori misurabili per gli Obiettivi di sviluppo sostenibile (SDG), la raccolta dei dati per valutare questi indicatori è in alcuni casi problematica. Il progetto CROWD4SDG propone di supportare questa raccolta di dati utilizzando la Citizen Science o Scienza dei cittadini. Il progetto CROWD4SDG si basa su tre pilastri: ricerca per la Citizen Science basata sullo sviluppo di nuovi strumenti IT, ricerca con la Citizen Science per facilitare la generazione di progetti bottom-up e ricerca su Citizen Science per lo sviluppo di progetti di Citizen Science di alta qualità con l'identificazione di best-practice nel settore. Il progetto si concentrerà su una serie di obiettivi di sviluppo sostenibile, incentrati sugli obiettivi per il cambiamento climatico, in combinazione con obiettivi di sviluppo sostenibile sulla uguaglianza di genere, sulle città sostenibili e sui diritti. Il progetto si baserà anche sui risultati delle ricerche del precedente progetto E2mC che ha portato allo sviluppo di un approccio per l'estrazione di informazioni dai social media basata sulla gestione delle informazioni geografiche, l'intelligenza*

Barbara Pernici
Politecnico di Milano, DEIB, piazza Leonardo da Vinci 32, Milano, Italy, e-mail: barbara.pernici@polimi.it

artificiale e il crowdsourcing nel contesto di situazioni di emergenza dopo catastrofi naturali.

Key words: Crowdsourcing, social media, SDG, climate change, natural disasters

1 Introduction

The 17 Sustainable Development Goals (SDG) defined by the United Nations[1] were adopted by "all United Nations Member States in 2015, as a call for action by all countries to promote prosperity while protecting the environment". A set of Measurable indicators[2] has been defined in association to the SDG goals and their targets, however in some cases the collection of relevant data is problematic and further sources of data are needed. The goal of the CROWD4SDG is to propose to develop new IT tools and methodological approaches using a Citizen Science approach.

Citizen Science has been advocated as a resources for collecting information to assess SDG [4]. In fact, it can improve coverage and frequency of data collection, allow managing spatial variations across a country, and support the veracity of information. Several citizen-generated projects are emerging, and as reported by SciStarter.org, several activities in Europe address also climate issues.

In particular, CROWD4SDG will focus on SDG 13 climate action, and in particular target 13.1 Strengthen resilience and adaptive capacity to climate-related hazards and natural disasters in all countries, with the following indicators: "13.1.3 Proportion of local governments that adopt and implement local disaster risk reduction strategies in line with national disaster risk reduction strategies; 13.1.1 Number of deaths, missing persons and persons affected by disaster per 100,000 people; 13.1.2 Number of countries with national and local disaster risk reduction strategies". SDG 13 will be analyzed in different yearly rounds, focusing each year on a specific combination with another goal: SDG 11 Sustainable cities and communities; SDG 5 Gender equality; SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels, focusing on rights.

In fact, the impact of climate change in not only limited to environmental issues but it has a deep impact on society, as discussed in detail in [2], that discusses how climate change and weather-related disasters impact on gender-based violence and discrimination and inequality.

In CROWD4SDG the goal is both to collect information to assess SDG indicators and also to collect from Citizens proposals for actionable actions to achieve SDGs.

In Section 2, we illustrate the methodological approach of CROWD4SDG; in Section 3, we illustrate the results achieved in a previous project, E2mC, combining

[1] https://sustainabledevelopment.un.org

[2] https://unstats.un.org/sdgs/indicators/indicators-list/

different approaches to collect first hand information in the first hours after natural disasters, and finally we conclude with open issues and future developments.

2 The Crowd4SDG approach

The project is based on three pillars:

- *Research for Citizen Science based on new IT tools*. Starting from previous research of the participants in the project, tools for organizing online communities, discussion platforms, and mining information from social media will be enhanced with AI-based modules. In particular, we will focus on facilitating the community participant interactions in large communities and in learning from crowdsourced information to improve data mining.
- *Research by Citizen Science* to facilitate the generation of bottom-up projects will start from the experience of the OpenSeventeen challenges (O17)[3], to generate challenges and select interesting projects related to SDGs and guiding the selection of tools to support crowdsourced activities.
- *Research on Citizen Science* for the development of high-quality Citizen Science projects identifying best practices and developing new methodological approaches.

3 Mining social media with the support of AI and crowdsourcing

In the project, in particular in the initial phases, the focus is on climate changes and natural disasters, in particular in urban settings. In this context, the project will be based on the research results obtained in the E2mC[4] project, using information extraction from social media based on geographic information management, AI, and crowdsourcing in the context of emergency situations after natural disasters. In E2mC, a set of tools have been developed to select and analyze information from social media or contributed through crowdsourcing tools and display them on a WebGIS interface [6], [7].

In particular, social media crawlers have been developed to select information, and in particular images relevant to a natural disaster, from social media [6], using data mining with topic extracting for Twitter posts and mining through triangulation techniques and clustering to extract keywords from one social media such as Flicker and apply them in another source, such as for instance YouTube [1]. Automatic geolocation of images is performed analyzing the text of posts with Natural Language Processing techniques and locations are disambiguated using OpenStreetMap [5]

[3] http://openseventeen.org/

[4] https://www.e2mc-project.eu/

and the CIME disambiguation algorithm developed in the project [7]. Kernel density analysis can be then performed to define hot spots for an event [9]. In addition to data mining techniques, evaluation of the relevance selected images and the precision of the geolocation can be assessed using crowdsourcing, in conjuction with crowd information evaluation techniques [8].

Starting from the E2mC results, the CROWD4EMS project will emphasize AI techniques, to improve crowd results evaluation [3] and for learning new keywords from crowdsourcing results and for improving the precision of the geolocation of posted images, using also image analysis techniques to extract relevant features to identify useful images.

4 Open challenges and future work

Several research challenges are open in providing AI enhanced tools to Citizen Science projects. In addition to tools availability and stability, it is important to provide a basis for the agile deployment of an environment in case of new emergency events, tailoring the tools to the specific needs of each situation in a very short time.

We will further enhance existing tools with AI and ML mechanisms, in particular to benefit from the information that can be gathered from the crowdsourcing activities. However, we will also need to study some of the issues emerging from the use of such tools in particular in the case of vulnerable populations, following and advancing the guidelines provided in [10], which discusses the use of AI in those contexts.

Acknowledgements This work was funded by the European Commission H2020 projects E2mC "Evolution of Emergency Copernicus services", project No. 730082, and Crowd4SDG "Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience", project no. 872944. The CROWD4SDG project is coordinated by University of Geneva, partners in the project are the Spanish National Research Council (CSIC), Politecnico di Milano, the Center For Research and Interdisciplinarity (CRI) of Universite Paris Descartes, the United Nations Institute for Training and Research (UNITAR), and the European Organization For Nuclear Research (CERN). The author thanks all the project participants for their contributions and discussions.

This work expresses the opinions of the author and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

References

1. Autelitano, A., Pernici, B., Scalia, G.: Spatio-temporal mining of keywords for social media cross-social crawling of emergency events. GeoInformatica **23**(3), 425–447 (2019). DOI 10.1007/s10707-019-00354-1. URL https://doi.org/10.1007/s10707-019-00354-1
2. Camey, I.C., Sabater, L., Owren, C., Boyer, A.: Gender-based violence and environment linkages. The violence of inequality. IUCN, Gland, Switzerland (2020). URL

https://portals.iucn.org/library/sites/library/files/documents/2020-002-En.pdf
3. Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., Allahbakhsh, M.: Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Comput. Surv. **51**(1), 7:1–7:40 (2018). DOI 10.1145/3148148. URL https://doi.org/10.1145/3148148
4. Fritz, S., See, L., Carlson, T., Haklay, M.M., Oliver, J.L., Fraisl, D., Mondardini, R., Brocklehurst, M., Shanley, L.A., Schade, S., et al.: Citizen science and the united nations sustainable development goals. Nature Sustainability **2**(10), 922–930 (2019)
5. Haklay, M.M., Weber, P.: OpenStreetMap: User-generated street maps. IEEE Pervasive Computing **7**(4), 12–18 (2008)
6. Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J.L., Achte, T.V., Zeug, G., Mondardini, M.R.R., Grandoni, D., Kirsch, B., Kalas, M., Lorini, V., Rüping, S.: E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. Sensors **17**(12), 2766 (2017)
7. Pernici, B., Francalanci, C., Scalia, G., Corsi, M., Grandoni, D., Biscardi, M.A.: Geolocating social media posts for emergency mapping. In: demo paper, Proc. SWDM. Los Angeles, CA (2018). URL https://arxiv.org/abs/1801.06861
8. Ravi Shankar, A., Fernandez-Marquez, J.L., Pernici, B., Scalia, G., Mondardini, M.R., Di Marzo Serugendo, G.: Crowd4ems: A crowdsourcing platform for gathering and geolocating social media content in disaster response. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **42**, 331–340 (2019)
9. Spasenovic, K., Carrion, D., Migliaccio, F., Pernici, B.: Fast insight about the severity of hurricane impact with spatial analysis of Twitter posts. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences (2019)
10. Wright, J., Verity, A.: Artificial Intelligence Principles for Vulnerable Populations in Humanitarian Contexts. Digital Humanitarian Network (2020). URL https://www.digitalhumanitarians.com/artificial-intelligence-principles-for-vulnerable-populations-in-humanitarian-contexts/

Permanent Population Census: evaluation of the effects of regional strategies on the process efficiency. The direct experience of Tuscany

Il Censimento Permanente della Popolazione: valutazione delle strategie regionali sull'efficienza del processo

L. Porciani, L. Francovich, L. Faustini, A. Valentini

Abstract The current challenges of Official Statistic are related to various global changes interesting also social behaviors; and conducting a Population Census is a public and social action has several interconnections with the features of the societies in which it is planned. This work has the main aim to identify some best practices in the management of a complex and completely new statistical process such as the Permanent Population Census (PPC) at regional level, in Tuscany. The analysis were conducted using paradata extracted by the Survey Management System, a web system implemented by the Italian National Statistical Institute (Istat) to manage the survey process.

Abstract *Le attuali sfide della Statistica Ufficiale sono legate a cambiamenti globali che impattono anche sui comportamenti sociali: condurre un censimento della Popolazione è un'azione pubblica che implica la considerazione delle caratteristiche delle società in cui esso viene realizzato. Questo lavoro ha l'obiettivo principale di individuare alcune buone pratiche nella gestione di un processo statistico complesso e completamente nuovo, come il Censimento Permanente della Popolazione, a partire dall'esperienza della Regione Toscana. L'analisi è stata condotta utilizzando i paradata estratti dal Sistema di Gestione delle Indagine, un sistema web implementato da Istat per la gestione del processo di indagine.*

Key words: Permanent Population Census, regional strategies, improvement of the efficiency of statistical process

[1] Linda Porciani, Istat, RTE (Florence), porciani@istat.it
Lisa Francovich, Istat, RTE (Florence), francovi@istat.it
Luca Faustini, Istat, RTE (Florence), faustini@istat.it
Alessandro Valentini, Istat, RTE (Florence), alvalent@istat.it

1 Introduction

Istat introduced Permanent Population Census (PPC) in Italy in 2018, after two pilots survey in 2015 and 2017. PPC represents one of the core data collection process of the new course proposed in the ESS visions strategy 2020[1], based primarily on the Integration of data sources for Official Statistic in order to reduce response rate, response burden and costs and to increase timeliness. In addition, the project of a complex process, such as a PPC, has to take into consideration the impact of social changes on statistical process due to the globalization era[2]. The general decrease in trusting institutions could be one of the reasons for the continuous reduction in response rate; the huge increase of data producers of the criticalities of the response burden; and the widespread of the sentiment of uncertainty of the growing request of a primary attention to the right of privacy and data confidentiality[3]. These elements affect the statistical process, pushing the National Statistical Institute towards the improvement of the evaluation process and, consequently, to share good practices to figure out the major criticalities.
In this perspective, this work has the main aim to identify some strengths (and weakness) of the management of a complex and completely new statistical process such as the PPC, starting from the regional experience of the Istat office of Tuscany.

2 Data and methods

PPC in Italy has been projected on 4 years survey process (2018-21) and on two simultaneous sampling- surveys – namely Areal Survey (AS) and List Survey (LS) – performed every year (roughly) from October to December[4]. Both have a strategy sampling based on a first step of Municipalities and on a different second step: AS is based on an area frame sampling strategy (the initial population is composed by the addresses), whereas LS has a traditional list frame sampling strategy (the initial population is composed by the households extracted from Population Register). In terms of number, the sample counts at national level almost 2800 Municipalities (on more than 8000 ones), 40% of them involved every year, and 1,4 millions of households (2/3 in the LS and 1/3 estimated in AS). In Tuscany region, 120 Municipalities are sampled (69 every year), for almost 110000 households (74% for LS and the rest for AS).

[1] Read more on https://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020
[2] Oleg CARA, 2014. "Challenges Of Official Statistics In The Globalization Era," ECONOMY AND SOCIOLOGY: Theoretical and Scientifical Journal, Socionet; Complexul Editorial "INCE", issue 1, pages 121-127
[3] MacFeely, Steve, 2016. The Continuing Evolution of Official Statistics: Some Challenges and Opportunities. Journal of official statistics. 32. 10.1515/jos-2016-0041
[4] Read more on:
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2019/mtg1/6_Solari_ENG.pdf

Their routinely implementation requires to respect a very detailed and mandatory scheduling that greatly affect the organizational effort of the Regional Census Office for Tuscany (RCOT) both in terms of training of census network operators and in terms of timely interventions to sustain response.

The present analyse was conducted on three aspect of RCOT work:

a) the main strategies adopted at regional level to improve the data quality and the efficiency of the process;
b) the profile of the field operators to evaluate the effects of some socio demographic features on the quality of the field work;
c) the tempus and quantum of response rate through a statistical model based.

The analysis used paradata[1] derived from the Survey Management System (SMS): a web system created to manage the PPC by the side of the census operators. The SMS collect information on process related variables such as: empty houses, non respondents, not eligible families for health problems or relocation, on-going (not finished) interviews, refusals, number of appointments for having an interview and so on.

3 Preliminary results

3.1 Istat regional strategies

RCOT main strategies could be structured into three type of actions aimed to improve the efficiency of the statistical process and to guarantee the data quality: training process, interinstitutional relationships and steady support to the network census.

a) The guidelines of training process has organized at national level: it changed from 2018 to 2019 following to the results of the 2018 evaluation survey. In the current year, the training process passed from 2 days in presence and a general e-learning session, to 1 day in presence and a customised e-learning session. In this framework, in Tuscany the training session were organized depending on the role of the actors in the Census (responsible and coordinator and interviewers) and on the geographically location of the operators. Seven persons managed 37 classes, composed by a mean number of 25 operators (min. 12 , max. 55). The training process took places from June to September 2019 in all the 10 chief town, from 0 to

[1] Kreuter F. (eds), 2013, Improving Surveys with Paradata. Analytic Uses in Process Information, University of Maryland, Wiley Series in Survey Methodology, J.W. and sons, Canada.

140 km away from the Istat Regional Office. Almost 1000 operators where trained in total.

b) The building of a strong interinstitutional network through coordinated actions to reach a unique objective: a good response rate combined with data quality. Several meetings in persons between RCOT and Provincial Offices of Ministry of Interiors were organized and a weekly follow up of the survey process was shared, in order to identify rapidly the criticalities and to figure out them with mixed modes of interventions. The combined actions affected the increase of response rate.

c) RCOT assaulted time resources in the reinforcement of a steady support to the different nodes of the census network. It sent specific email Notes in order to recall process deadlines, the delay of some of them, or to suggest some specific actions to respect the process guidelines (17 Notes). In addition, RCOT managed a direct assistance by email (about 500 assisted email).

3.2 The operators profile

Regarding 2019 wave, on the field were active 319 men (37,3%) and 538 women (62,7%). The mean age was 45,8 years for men ad almost the same (45,2) for women (min. 26 – max. 60 years old). Two thirds of them (68%) were interviewers, the rest were coordinators, chief of the census office or back office operators. Almost half of them have a high school certificate, and 43,1% a university degree; women are more educated: 47.0% have an university degree (men 36,3%).
During the work on the field for the AS, operators had a mean of 33,4 expected families to be contacted: at the end of the survey period each operator completed 26 questionnaires on average[1].

Table 1 - *Work field - AS- Tuscany 2018 and 2019. Mean value*

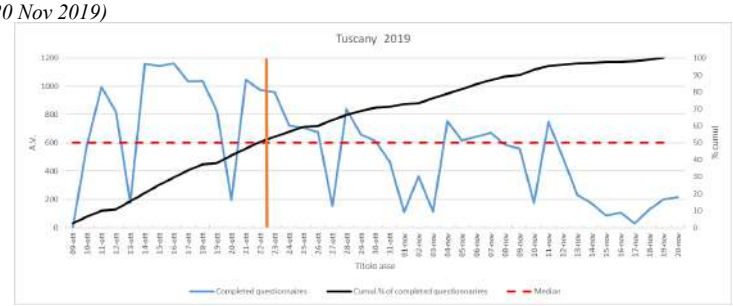
	No. addresses	***Estimated families***	***Contacted Families***	***Questionnaires***
2019	***23,2***	***33,5***	***39,1***	***25,9***
2018	***19,7***	***24,3***	***23,9***	***21,9***

3.3 The statistical model based evaluation of the tempus and quantum of response rate

[1] The same analyse will be conducted for LS.

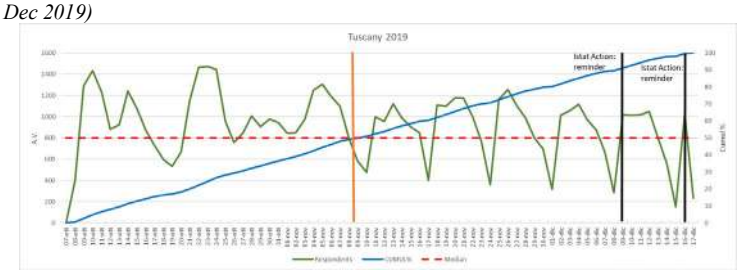
The timing of the two surveys is crucial for the success of the PPC. To evaluate the goodness of the process, it is essential to consider separately AS and LS, because in AS process, the most of the work is up to the operators (09 Oct. – 20 Nov), whereas the LS has a mixed mode data collection strategy: CAWI (7 Oct - 13 Dec), CAPI and CATI (8 Nov - 20 Dec). For the AS, an analysis of the timing shows that the median value of the distribution of the completed questionnaires in Tuscany drop on the 14th day of fieldwork (on a span of 42 days available)[1] represented by an orange bar on Figure 1[2]. After 25 days, 75% of the questionnaires were compiled. The RCOT entered actively in the process through a series of punctual emails and calls to specific municipalities, and with a final reminder on the 19 of November, keeping them to the end of the survey with around 100% of successful contacts (for AS 84,5% of the estimated families – 89,8% in 2018 - and 98,9% in the LS– 96,1% in 2018).

Figure 1 – *Distribution of the questionnaires (absolute values and % cumulated distribution) of the AS (8 Oct - 20 Nov 2019)*



As regards the LS, in the Figures 2 and 3[3], it is evident the spontaneous and non-spontaneous response rate. The median drop on the 34th day on a span of 72 days.

Figure 2 – *Distribution of the respondents (absolute values and % cumulated distribution) of the LS (7 Oct – 17 Dec 2019)*



Two active actions were taken by RCOT at the end of the period to stimulate the operators in order to close the work in time (black vertical line).

A second exploratory analysis has been conducted on the available variables of the SMS 2018 using a logistic regression model approach with the aim to identify elements affecting the quality of the process, more than some proxies of non-

[1] Provincial analyses will be conducted.
[2] In 2018 was the 26th day, mainly due to an interruption for technical problem of SSM.
[3] In Additional Figures file uploaded in the SIS website.

sampling errors. The response variable, Y, is calculated on the median number of days elapsed to get a final solution for a specific survey unit by municipality: 1 in case the median number of days for a municipality is lower than the regional one and zero in the opposite case. The regional median value is 26 days. The explanatory variables selected are the following, the amount of survey units per municipality (x1); number of census operators by municipality (x2); the average number of records for operator per municipality (x3); being in Union of Municipalities (x4); percentage of not completed surveyed units (x5); percentage of completed surveyed units (x6). The selected variables could be ascribed to three main thematic areas: the structure of the Municipality Bureau of Census (x1,x2,x4), the census operator effort (x3) and the general territorial context (x5,x6). Preliminary results show that the local context appears to play a key role in promoting the data collection process (x6), followed by the efforts of Municipal Census Bureau (x1), underlying the centrality of local actors in the complex process of PPC[1].

4 Citations and References

1. Eurostat, ESS 2020 vision strategy, https://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020
2. Eurostat, MANUALS AND GUIDELINESEU legislation on the 2021 population and housing censuses, EXPLANATORY NOTES, 2019 edition, https://ec.europa.eu/eurostat/documents/3859598/9670557/KS-GQ-18-010-EN-N.pdf/c3df7fcb-f134-4398-94c8-4be0b7ec0494
3. Kreuter F. (eds), 2013, Improving Surveys with Paradata. Analytic Uses in Process Information, University of Maryland, Wiley Series in Survey Methodology, J.W. and sons, Canada
4. MacFeely S., 2016. The Continuing Evolution of Official Statistics: Some Challenges and Opportunities. Journal of official statistics. 32. 10.1515/jos-2016-0041
5. Oleg C., 2014. "Challenges Of Official Statistics In The Globalization Era," ECONOMY AND SOCIOLOGY: Theoretical and Scientifical Journal, Socionet; Complexul Editorial "INCE", issue 1, pages 121-127
6. Porciani L., Faustini L., Valentini A., Martelli B.M., Italian e-census: a regional analysis of web response, Istat Working Paper, no. 10/2015
7. UNECE,https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2019/mtg1/6_Solari_ENG.pdf

[1] In Additional Figures file uploaded in the SIS website.

Capture-recapture methods

Bayesian Model Averaging for Latent Class Models in Capture–Recapture

Model Averaging Bayesiano per Modelli a Classi Latenti in Cattura-Ricattura

Davide Di Cecco

Abstract Model selection appears to be crucial in capture-recapture problems as it is common that different models with an equally good level of adaptation to the observed data lead to rather different estimates of the undercounts. We consider log–linear Latent Class Models as our capture-recapture model and propose Bayesian model averaging to overcome the difficulties of model selection within this class. We show that, by focusing on graphical decomposable models, we can design a simple Gibbs–based MCMC to sample over the space of eligible models.
Abstract *In problemi di cattura–ricattura, la selezione del modello risulta essere un aspetto delicato e difficoltoso. Non è inusuale, infatti, trovare modelli con valori di bontà di adattamento molto simili tra di loro che conducono a delle stime del numero di unità non catturate molto diverse. In questo lavoro trattiamo una famiglia estesa di modelli a classi latenti che rilassa l'ipotesi di indipendenza condizionata modellando le interazioni tra le variabili attraverso un modello log–lineare con una variabile latente. Per superare la difficoltà di scelta del modello all'interno di questa classe ampliata, proponiamo un model averaging Bayesiano. Mostriamo come, se ci limitiamo a considerare i modelli log-lineari decomponibili, è possibile costruire un semplice Gibbs sampler per ottenere la distribuzione a posteriori della numerosità della popolazione d'interesse.*

Key words: Bayesian Model Averaging, Latent Class Models, Capture-Recapture

1 Introduction

As pointed out by many authors, see, e.g., [4], the problem of estimating the size of a population in a capture-recapture model is essentially a problem of forecasting. As a consequence, it is not unusual that different models with a comparable level of

Davide Di Cecco
Sapienza University, viale del Castro Laurenziano 9, e-mail: davide.dicecco@uniroma1.it

goodness of fit lead to rather different estimates of the total population count. Given the lack of specific criteria for model choice, and the impossibility to validate the estimates, it is not unusual in capture-recapture practice to simply rule out a model resulting in unrealistic estimates. We think that a reliable procedure to deal with model selection in a more automatic way would certainly be of interest.

We treat the case of Multiple Record System, that is, the data consists of a set of capturing lists, usually originating from different sources, reporting partial listing of the same target population. In this setting it is common to assume different capture probabilities for the various sources. As a consequence, log-linear models are the tool of choice in capture–recapture modeling, and Latent Class Models (LCM) represent the natural extension when one wants to include unobserved heterogeneity. The use of LCM in capture–recapture dates back at least to [1] with many developments thereafter. The simplest formulation of LCM envisages the conditional independence assumption (CIA) which appears to be too restrictive in many situations. There are many proposals in literature to relax the CIA resulting in more flexible models. We focus on log-linear LCM where the additional dependencies are directly modeled by interaction parameters. Previous works on this class include [3], [13], [12]. We propose a Bayesian approach to the class as previously introduced in [6] and [7]. To overcome the difficulty of model selection within this class, we propose Bayesian model averaging to analyze the posterior distribution of the population count over a set of eligible models. Usually a full Bayesian approach to model averaging requires the use of a Reversible Jump algorithm ([8]) which is in general hard to implement. See [11] for an example of use of the algorithm within the class of log-linear models (without a latent variable). We show that, if we restrict ourselves to the subclass of decomposable models, it is possible to implement a simple Gibbs–based MCMC. Some preliminary results on simulated data (not shown in this work for space limitation), seem to indicate that the restriction to that subclass does not affect the efficacy of the procedure.

2 The model

Consider k capturing variables $\mathbf{Y} = (Y_1, \ldots, Y_k)$, where $Y_i = 1$ if a certain unit is listed in the i-th source and 0 otherwise, and let X be the latent variable taking values in $\{1, \ldots, m\}$ identifying the latent classes of our population. The LCM under the CIA can be equivalently expressed as the mixture model

$$P(\mathbf{Y} = \mathbf{y}) = p_{\mathbf{y}} = \sum_{x=1}^{m} p_x \prod_{i=1}^{k} p_{y_i|x}, \tag{1}$$

where $p_{y_i|x}$ indicates the conditional probability $P(Y_i = y_i|X = x)$, or as the log-linear model

$$[XY_1][XY_2]\cdots[XY_k], \tag{2}$$

where we use the classic notation reporting only the higher order interactions (generators) of the model. The dependence graph of this model is a star–shaped graph, where the node representing X is connected to all other nodes, like the one in Figure 1 on the left. Any additional interaction term with respect to (2), (i.e. any additional arc in the graph), constitutes a relaxation of the CIA.

Denote the number of observed units as n_{obs}, and the number of units presenting the capture profile $\mathbf{y} \in \{0,1\}^k$ as $n_{\mathbf{y}}$. Let $n_{x,\mathbf{y}}$ be the number of units having that profile which belong to latent class x, so that $\sum_x n_{x,\mathbf{y}} = n_{\mathbf{y}}$, and $n_{\mathbf{x},\mathbf{y}}$ be the vector $(n_{x,\mathbf{y}})_{x=1,\ldots,m}$. Let $n_{\mathbf{0}}$ be the number of uncaptured units to be estimated and N the total population count such that $\sum_{\mathbf{y}} n_{\mathbf{y}} + n_{\mathbf{0}} = N$. Let M be the (random variable associated to) the model to be chosen in a pre determined set $\mathscr{M}$, and Θ_M the set of parameters associated to model M.

2.1 Prior distributions

Obviously, the choice of a prior on the set of models $\mathscr{M}$, as on any other parameter, is subjectively arbitrary. However, we usually just want to exclude some cases, and set a uniform prior on all remaining cases. As we have said, we focus on decomposable models and, as an elementary additional criterion, we rule out all unidentifiable models. In [14] we can find a necessary and sufficient condition for the identifiability of any graphical model (hence of any decomposable model too).

In practice, to utilize the proposed methodology, we have to list out all possible models for a given number of variables. That is, all identifiable models represented by decomposable graphs containing the star–shaped graph relative to the CIA. Consider the case $k = 4$: in Figure 1 we have all decomposable graphs grouped by isomorphism. This leaves us with 19 eligible models. When $k = 5$ the number of identifiable decomposable models goes up to 355.

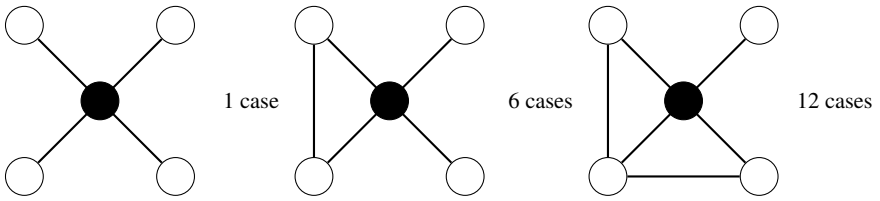


Fig. 1 Identifiable decomposable graph models with 4 manifests (empty nodes) and a latent (black node) grouped by classes of isomorphism

As for the prior distributions over the parameters Θ_M of each model M, we utilize the Hyper Dirichlet distribution described in [5]. Such a choice allows us to exploit

the result of [9] giving an analytical formula for the posterior probabilities of the models given the data.

3 Model averaging

We analyze the posterior distribution of N given the observed data $\{n_{\mathbf{y}}\}$ with $\mathbf{y} \neq \mathbf{0}$,

$$\pi(N \mid \{n_{\mathbf{y}}\}) = \sum_{M \in \mathscr{M}} \pi(N \mid M, \{n_{\mathbf{y}}\})\, \pi(M \mid \{n_{\mathbf{y}}\}).$$

The posterior probability of model M is given by:

$$\pi(M \mid \{n_{\mathbf{y}}\}) \propto \pi(M)\pi(\{n_{\mathbf{y}}\} \mid M) = \pi(M) \int \pi(\{n_{\mathbf{y}}\} \mid M, \Theta_M)\pi(\Theta_M \mid M) d\Theta_M.$$

In practice, given the computational complexity of calculating the last integral quantity, one can settle for the simplest (first order) approximation of the marginal likelihood of a model based on the Bayesian Information Criteria (BIC), that is, $\exp(-BIC/2)$, which, given equal prior probabilities for the models, leads to the following approximation of the weights $\pi(M \mid \{n_{\mathbf{y}}\})$:

$$\frac{\exp(-BIC_M/2)}{\sum_{M \in \mathscr{M}} \exp(-BIC_M/2)}, \tag{3}$$

where BIC_M is the BIC of model M. Then, one can use those weights in computing the averaged mean

$$E[N \mid \{n_{\mathbf{y}}\}] = \sum_{M \in \mathscr{M}} \widehat{N}_M\, \pi(M \mid \{n_{\mathbf{y}}\}), \tag{4}$$

where $\widehat{N}_M$ is the posterior mean of N under model M.

By using formula (4), one should keep in mind that the approximation quality can be poor in some cases, and, in any case, we limit ourselves to a point estimate of N. A full Bayesian approach to the problem, on the other hand, would result in an estimate of the whole posterior distribution of N marginalized over $\mathscr{M}$.

4 The Gibbs sampler

In this section we outline a Gibbs–based MCMC algorithm to sample from the joint distribution of $(N, N_{\mathbf{x},\mathbf{y}}, M, \Theta_M)$, conditioned on the observed data $\{n_{\mathbf{y}}\}$. Note that we cannot obtain the full conditionals for all terms: as pointed out in [2] and in [10], given $n_{\mathbf{x},\mathbf{0}}$, the value of N is deterministically defined. As a workaround, they propose to consider the conditional distribution of the couple $N, N_{\mathbf{x},\mathbf{0}}$ conditionally

on the rest. Similarly, we cannot obtain the conditional distribution of M given the parameters Θ_M. For these reasons, the algorithm loops over the following steps:

1) sample $n^{(t)}_{\mathbf{x},\mathbf{y}}$ from $\pi(N_{\mathbf{x},\mathbf{y}} \,|\, N, M, \Theta_M, \{n_{\mathbf{y}}\})$, for all $\mathbf{y} \neq \mathbf{0}$,

$$N_{\mathbf{x},\mathbf{y}} \sim Mult\left(n_{\mathbf{y}}, p_{\mathbf{x}|\mathbf{y}}\right),$$

where the $p_{\mathbf{x}|\mathbf{y}}$ are calculated according to the current value of M and Θ_M;

2) sample a couple $\left(N^{(t)}, n^{(t)}_{\mathbf{x},\mathbf{0}}\right)$ from

$$\pi(N, N_{\mathbf{x},\mathbf{0}} \,|\, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(N \,|\, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\})\, \pi(N_{\mathbf{x},\mathbf{0}} \,|\, N, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}).$$

Note that

$$\pi(N \,|\, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(N \,|\, M, \Theta_M, n_{obs}) \propto \pi(N) \binom{N}{n_{obs}} p_{\mathbf{0}}^{N-n_{obs}} (1-p_{\mathbf{0}})^{n_{obs}},$$

then, if we choose the improper prior $\pi(N) \propto 1/N$, the conditional distribution of N results in a Negative Binomial distribution, and we simply have to

- sample $N^{(t)}$ from $NegBin\,(n_{obs}, 1-p_{\mathbf{0}})$;
- sample $n^{(t)}_{\mathbf{x},\mathbf{0}}$ from $Mult\big((N^{(t)} - n_{obs}), p_{\mathbf{x}|\mathbf{0}}\big)$.

where $p_{\mathbf{x}|\mathbf{0}}$ and $p_{\mathbf{0}}$ are calculated according to the current value of M and Θ_M;

3) sample from $\pi(M, \Theta_M \,|\, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(\Theta_M \,|\, M, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})\, \pi(M \,|\, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$. That is,

- calculate the posterior probability of each eligible decomposable model and sample $M^{(t)}$ from $\pi(M \,|\, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$. The posterior probability of each model M is defined as a product of Gamma functions (see [9]);
- then sample all parameters $\Theta^{(t)}_M$ from their posterior conditional distribution $\pi(\Theta_M \,|\, M, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$, which is a product of Dirichlet distributions.

5 Conclusions

The proposed algorithm can be just used for model selection by simply inspecting the generated values of M, as the relative frequency of each model constitutes an estimate of its posterior probability, and select the best model accordingly. However, model averaging seems to be the best choice in capture-recapture problems. Compared to the usual approximation techniques, our estimates should be more accurate, and allow to inspect the whole posterior distribution of N at the cost of some additional computational effort which appears nonetheless reasonable. The restriction to decomposable models may seem a severe limiting factor, as they constitutes a minority fraction of the possible models, and many frequently used models, such as the one with all second order and no higher order interactions, are left out. How-

ever, some preliminary results based on simulations (not shown in this work for space limitation), appears to be encouraging. In fact, the proposed approach seems to work well even when used on data generated from non decomposable models.

References

[1] A. Agresti. Simple capture–recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, 50(2):494–500, 1994.

[2] S. Basu and N. Ebrahimi. Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279, 2001.

[3] A. Biggeri, E. Stanghellini, F. Merletti, and M. Marchi. Latent class models for varying catchability and correlation among sources in Capture-Recapture estimation of the size of a human population. *Statistica Applicata*, 11(3):1–14, 1999.

[4] B.A. Coull and A. Agresti. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55(1):294–301, 1999.

[5] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21(3):1272–1317, 1993.

[6] D. Di Cecco. Estimating population size in multiple record systems with uncertainty of state identification. In *Analysis of Integrated Data*, pages 169–196. Chapman and Hall/CRC, 2019.

[7] D. Di Cecco, M. Di Zio, and B. Liseo. Bayesian latent class models for capture–recapture in the presence of missing data. *Biometrical Journal*, 2020.

[8] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[9] D. Madigan and J. C. York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19–31, 1997.

[10] D. Manrique-Vallier. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254, 2016.

[11] A. M. Overstall and R. King. `conting`: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58(7):1–27, 2014.

[12] E. Stanghellini and M. G. Ranalli. Population size estimation using a categorical latent variable. In *Capture-Recapture Methods for the Social and Medical Sciences*, pages 277–290. Chapman and Hall/CRC, 2017.

[13] E. Stanghellini and P. G. M. van der Heijden. A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics*, 60(2):510–516, 2004.

[14] E. Stanghellini and B. Vantaggi. Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli*, 19(5A):1920–1937, 2013.

Combining "signs of life" and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts

Un modello a classi latenti per combinare segnali di presenza e dati di indagine nel contesto di stima cattura-ricattura dei conteggi di popolazione in presenza di sovra-copertura

Marco Fortini, Antonella Bernardini, Marco Caputi, Nicoletta Cibella[1]

Abstract The new permanent census strategy integrates population registers with sample surveys to obtain the counts of the usual resident population at a reference date, in such a way clearing Census counts of coverage errors affecting the register. A Lincoln-Petersen estimator is used after adjusting for people erroneously included in the register (over-coverage). We show a latent class model integrating the 2018 survey data affected by both missing values and response errors with administrative signals of job and school education (signs of life) in order to estimate over-coverage rates. We apply the model to a random sample of municipalities of Piedmont. We will see how and to what extent the signs of life are useful in determining usual residence of individuals and differences between Italians and foreigners in terms of over-coverage errors.

Abstract *La nuova strategia di censimento permanente integra il registro della popolazione con indagini campionarie per ottenere i conteggi della popolazione abitualmente dimorante alla data di riferimento, in modo da eliminare dal conteggio del censimento gli errori di copertura del registro. Viene utilizzato uno stimatore Lincoln-Petersen dopo l'aggiustamento per le persone erroneamente incluse nel registro (sovra copertura). In questo lavoro mostriamo un modello a classi latenti che integra i dati del censimento del 2018, affetti da valori mancanti ed errori di risposta, con segnali amministrativi di lavoro e di istruzione scolastica (segnali di presenza) al fine di stimare i tassi di sovra copertura. Il modello è applicato a un campione casuale di comuni piemontesi. Vedremo come, e in che misura, i segnali di presenza sono utili per determinare la dimora abituale delle persone e le differenze tra italiani e stranieri in termini di errori di sovra copertura.*

Key words: Census, administrative data, Lincoln-Petersen estimator, dual frame, multiple-records system

[1] Marco Fortini, Istat, fortini@istat.it; Antonella Bernardini, Istat, anbernar@istat.it; Marco Caputi, Istat, caputi@istat.it; Nicoletta Cibella, Istat, cibella@istat.it

1. Introduction

The new permanent census of the Italian population integrates administrative data and surveys. It will be carried out every year on a sample of Municipalities and households and its aim is updating data contained in the integrated system of registers, taking into account for possible coverage errors affecting them. Two types of errors are considered, called over- and under-coverage, which can cause bias in population counts when they are not perfectly balanced each other. Although the coverage error affects population counts across all territorial levels, the adjustment at the municipal level is considered here, which is the most detailed territorial domain reported in the census figures. In this way, the estimate of coverage errors only relates to this level of detail, while the corrected population at higher levels is obtained as a sum from the corrected municipal-level counts.

An estimate of corrected (Nirel and Glickman, 2009) population count $\widehat{N}_{ch}$ for stratum h and municipality c is obtained starting from the corresponding register count R_{ch} and the error estimation components as

$$\widehat{N}_{ch} = R_{ch} \frac{\hat{p}_{D|Rch}}{\hat{p}_{R|Dch}}$$

where $\hat{p}_{D|Rch}$and $\hat{p}_{R|Dch}$ are, respectively, the proportions of actual residents among people rostered by register and of those included into the register among people surveyed on the field. A sample of individuals drawn from the register and subsequently checked on the field can estimate the first error component while the second one is estimated by people found in the register among those enumerated in a random sample of municipal areas (first stage sample units), under the conditions of validity of the Lincoln-Petersen estimator.

Here we focus on determining $\hat{p}_{D|Rch}$ when some of the units assigned to the enumerator cannot be determined with certainty. In practice, while the interviewer is quite certain of the respondent's usual residence, the status of "not more dwelling" for deceased or moved people is less reliable, being it communicated by a family member or neighbour and not directly verifiable at the survey date. Even more, the status remains unknown when the interviewer cannot obtain information on the actual residence of the subjects due to a lack of contact.

Since a simple exclusion of undetermined cases from the analysis can cause severe bias on estimates, external information coming from administrative data is collected in order to build a secondary variable representing the so called 'signs of life', which can offer clues on the actual presence of the units on the field.

These two variables can be used as indicators of individual's dwellings, both affected by error, and included in a latent class model (Biemer, 2011). A dichotomous latent variable represents in this case the true dwelling status of the individuals. Further descriptive variables, such as the citizenship (Italian or foreign), age class and membership of a single or multi-component family have also been considered in the analysis with the dual aim to better investigate the phenomenon and make the model identifiable. In the following, we describe the data used, then we introduce the latent class model and finally show some results concerning a sample of municipalities in the Piedmont region.

2. Data settings

The new Census is designed as a two phases Master Sample (MS), based on two balanced

and coordinated sampling surveys, called L and A survey respectively[2]. It is planned for supporting the Istat Population Register (PR) in order to increase the amount of provided statistical information and to improve the level of coverage and quality.
With the aim of identifying the "usual resident population", ISTAT linked the PR to subject-specific administrative sources (Labour and Education registers, Tax Returns register, Earnings, Retired, and Non-Pension Benefits registers, Permits to Stay archive). As a results of this trial, an Integrated Archive of Usual Resident Population (AIDA) was obtained. This Archive allows for the assessment of coverage measures of the permanent census and could be used to correct the population counts estimates by using individuals' "signs of life" on the Italian territory (ONS, 2019). Whereas the absence of signals for people in the population register could denote over-coverage cases, it is possible to assume that, in case of strong association between administrative data and Census survey, "signs of life" can be used to predict or to correct the results of the Census survey for the estimation phase. This approach can reduce the possible under coverage of the survey without inflating over-coverage of the PR in Census results.
In this paper the focus is on the L survey of the MS, the list sample, which is designed on a list of households and has the purpose of producing the thematic integration of surveys and registers so as to estimate the hypercube [3] which cannot be obtained using the replaceable information coming from registers. The L component is based on a yearly sample, with the size of about 950,000 households, drawn from 2,850 municipalities out of 7,950 of which around 1,150 are self-representative; in four years' time all the Italian municipalities will be selected. The analysis is conducted on a sample of 205 municipalities in the region Piedmont, further sampled so to select 141,241 individuals within them.
For each individual, the residence status is based on the field contact results and the corresponding signs of life from administrative sources are compared to detect possible errors in the survey results.
Table 1[4] shows the unweighted frequency distribution of the individuals, according to their municipality and the dwelling status identified by the surveyor at the time of the survey.
The variables considered for data analysis have been classified as follows:
C – Residence status from field contact (1, Usual resident; 2, moved or deceased; 3, Undetermined)
Signs of life (SOL from now on), which are specifically relative to labour and education, are classified on the basis of the number of distinct months with at least one signal during 2017, the year preceding the survey, as: Strong (8-12 months), Medium (4-7 months) and Weak (1-3 months).
The variable S summarises the kind of administrative signals associated to an individual as follows:
1 - No signals at all or person living abroad
2 - Strong SOL and all of the administrative signals are geo-referenced in the Labour Area Market (LAM) of residence
3 - Strong SOL, but not all of the administrative signals are in the LAM of residence or some information is missing

[2] L survey adds variables not already available from administrative sources while A survey aims to correct population counts for under-coverage affecting the PR.
[3] Census hypercubes: census tables for the cross classification of the set of variables.
[4] All tables and figures of the present contribution are available online as supplementary material at the following url: https://doi.org/10.6084/m9.figshare.12471707.

4 - Medium SOL or university students and all of the administrative signals are geo-referenced in the LAM of residence
5 - Medium SOL or university students but some of the administrative signals are not in the LAM of residence or some information is missing
6 - Dependent family members resident in the same municipality of their own tax declarant
7 - Retired persons having the administrative source reporting the same municipality of official residence
8 - Weak SOL
9 - No SOL but other administrative signals
The variable is built sequentially so, except for the residual class 1, once an individual is classified with a lower number he cannot be further classified with a higher number.
Other variables considered in the analysis are: E for age classes (0-19; 20-34; 35-64; 65+), F for family size (1 if the individual is a single-member family; 0 otherwise), N for citizenship (1 if the individual is foreigner; 0 otherwise), P for the size of the municipality (<2,000 inhabitants; 2,001-5,000 inhabitants; 5,001-20,000 inhabitants; 20,000-50,000 inhabitants; >50,000 inhabitants).
The latent variable X represents the true residence status and has been categorised as 1 for "not dwelling" and 2 for "dwelling in" the same municipality of their official residence.

3. The statistical model

We applied a log linear model with latent variables on the 141.241 sample units where manifest variables are those listed above while the latent variable is with the true residence status X. For practical reasons the model has not been weighted with inclusion probabilities for sample units. The observed data distribution is factorised by the following path analysis model:

$$\Pr(C,S,F,E,N,P) = \sum_X \Pr(C|X,P)\Pr(S|X,E)\Pr(X|F,E,N,P)\Pr(F,E,N,P)$$

where $\{PFE, PFN, PEN, FEN\}$ are interaction terms considered for the hierarchical log-linear component. We also imposed a structural zero in table {FE} to consider absence of cases in the "0-19 years old" and "single component household". Concerning the logit model describing structural effect of manifest variables on the latent variable we imposed {XPE, XPN, XF} interactions, while marginal effects {SX, SE} and {CX, CP} were considered for logits implied by the two measurement models $\Pr(C|X,P)$ and $\Pr(S|X,E)$. Finally, a structural zero has been considered for people found by interviewer as resident at the address (C=1) being not dwelling at the same address of its official residence (X=1), since such an error by the interviewer is supposed markedly rare.
Data analysis was accomplished with the LEM software (Vermunt, 1997). The model definition followed a forward search strategy where interactions in logit and log-linear models have been tested by a compromise between BIC and index of dissimilarity, the latter given by the sum of absolute differences between the observed and the expected frequency cells, divided by twice the number of statistical units. The final model uses 145 parameters leaving 2014 DoF and getting a dissimilarity index of 0.06975.

[5] A value under 0.1 is generally considered as a good fitting.

4. Main results

While the use of the model for the purpose of the permanent census is mainly for forecasting purposes, the approach here will be more explanatory. The data showed here refer only to the L survey sample and therefore it is noted that the (unweighted) average values reported cannot be directly extended to the entire Piedmont population. From the structural component of the model we obtain the conditional probabilities that people are not truly dwelling (X=1) although they result resident in the municipality's population register.

Model fitting appears satisfactory as shown by the plot of expected against observed frequencies in figure 1.

Figure 2 shows the log-odds of being a dweller given age category and class of population size for people that hold the residence in the municipality. It is apparent that risk of over-coverage is higher for large cities and in the "20-34" age class. Older people show a low and almost equal risk across municipalities.

Figure 3 describes the log-odds of being a dweller given citizenship and municipality population class of people's residence. It comes out that being a foreigner person is by far the most important risk factor for over-coverage.

Figure 4 highlights the higher risk of over-coverage for people that live in single-sized households than those living in larger ones. As expected, the risk also depends on population size.

The measurement model also allows to evaluate the effectiveness of field contact and signs of life in identifying the over-coverage. As far as field contact is concerned, a constraint imposed on the model required that all individuals identified by the interviewer as dwellers should be considered as such with certainty. This information helps to determine the meaning of the latent classes and by assumption assigns to dwellers the 90% of the whole sample.

Figure 5 shows the log odds (LO) of field contact outcomes for dwellers vs non dwellers, given class of municipality population size. An arbitrary value of 3 in the graph was assigned to LLR1 to people found as resident at home during the field contact, which would have been instead infinite given the structural zero considered in the model.

As expected, both the other contact outcomes are clues of not being resident. However, the information provided by a proxy about the move or death of not surveyed individuals seems no more informative than the non-contact differently from what it is expected. In this sense, it could be interesting to consider cases where proxy respondent is a family member from those where information is given by a municipality officer or an acquaintance of the subject (e.g. a next-door neighbour or the doorman).

Concerning LO in figure 6 for signs of life and four age classes, they seem definitely accurate in identify dwellers when there is a strong evidence of work or school attendance in the same LMA (2) or retired persons resulting in the same municipality of the official residence (7). Strong signals outside LMA of the official residence municipality (3) continues to weakly support for the actual presence of people at their actual residence, meaning perhaps that many people keep their life centre around their official residence even when their job is far than this place. This effect could concern people returning to their families at the end of the working week and should be better investigated.

By contrast, evidence supporting for not dwelling at the official residence is given for people without signals or with signal of living abroad (1), people with weak signals in LMA of their official residence municipality (8) and people with residual and not geo-referenced signs of

life (9). In addition, class 4 (medium-level signals in LMA), 5 (medium-level signals outside LMA) and 6 (dependent relatives) are not very informative in this case and need further refinement on the basis of the collected evidence. Finally we performed a cross validation by randomly excluding 40 municipalities with less than 20,000 inhabitants and then estimating the model on the remaining 165 municipalities, with 107,402 statistical units. In doing so, we then estimated the model fitting at municipality level for those retained in analysis, while we evaluated the predictive power of the model on the 40 excluded municipalities. The observed people classified by the result of field contact in each municipality are plotted in figure 7 against their correspondent expected numbers, representing the excluded municipalities by red dots. The graphs show that both the fitting of the model and its predictive power seem acceptable for the higher frequencies, while they are less reliable for small frequencies. Since, as hoped, the lower frequencies concern the case of missing information, for these we notice a greater variability.

Finally in figure 8 we compare the coverage rates estimated by the model for each municipality with corresponding rates computed under the missing at random (MAR) assumption after the exclusion of not contacts (C=3). Again, red dots are municipalities excluded from the model estimation phase and used only during the prediction stage.

While we cannot see remarkable differences for "predicted" municipalities when compared to those involved into model estimation, a general improvement of coverage rate is obtained by latent class model as revealed by position of almost all the dots at the right of the 'equivalence' line with the MAR estimates.

5. Final remarks

Latent structure models are useful tools when a phenomenon is measured by multiple indicators affected by errors or that relates to different aspects of the same fact. Here this approach has been applied to the check of usual residence on a sample of residents in the Piedmont municipalities, in order to overcome the non-contact issue by interviewers using information coming from administrative sources. The model provides coherent and reasonable results and could be used both to interpret the components related to over-coverage and within a population estimation strategy based on capture-recapture models. However, the analyses carried out here need to be repeated in different census social and territorial contexts before they can be reliably used for direct estimation in official statistics. This will help to understand the strength and limitations of administrative signals on the territory and how they can be used to support statistical surveys. Furthermore, it is essential to continue to refine the model given its sensitivity, well known in the literature, in order to clarify how it can be better integrated into a production and quality control perspective.

References

1. Biemer, Paul P. Latent class analysis of survey error. Vol. 571. John Wiley & Sons, 2011. http://dx.doi.org/10.1002/9780470891155
2. Nirel, R., and H. Glickman. "Chapter 21-Sample Surveys and Censuses." Handbook of Statistic, Elsevier (2009). https://doi.org/10.1016/S0169-7161(08)00021-7
3. ONS (2019). Annual assessment of ONS's progress towards an Administrative Data Census post-2021, https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs.
4. Vermunt, Jeroen K. "LEM: A general program for the analysis of categorical data." Department of Methodology and Statistics, Tilburg University (1997).

Population size estimation with interval censored counts and external information

Stima della numerositá di una popolazione in presenza di conteggi con censura intervallare e informazioni esogene

Alessio Farcomeni

Abstract We discuss Bayesian log-linear models for incomplete contingency tables with both missing and interval censored cells, with the aim of obtaining reliable population size estimates. External information on the censoring probability may substantially reduce uncertainty. We are motivated by an original example on estimation of prevalence of multiple sclerosis in the metropolitan area of Rome; where five out of six lists have interval censored counts. External information comes from mortality rates.

Abstract *Introduciamo modelli log-lineari per tabelle di contingenza incomplete in cui siano presenti sia celle mancanti che con censura intervallare. Lo scopo é ottenere una stima della somma di tutte le celle. Mostriamo come usare informazioni esogene sul meccanismo di censura, riducendo sostanzialmente l'incertezza. La motivazione nasce dal stima della prevalenza di sclerosi multipla nell'area metropolitana di Roma. Cinque liste su sei sono censorate. Le informazioni esogene riguardano i tassi di mortalitá.*

Key words: Incomplete tables; capture-recapture; external information; left-censored counts; right-censored counts.

1 Introduction

Population size estimation often involves specifying a log-linear model for the incomplete contingency table of cross-counts, where the cell(s) corresponding to exclusion from all lists is (are) missing. The parameter of interest is the population size N, which corresponds to the sum of observed and unobserved counts (Fienberg, 1972; Cormack, 1989; Fienberg et al, 1999; Farcomeni, 2016). In this work we fo-

Alessio Farcomeni
Department of Economics and Finance, University of Rome "Tor Vergata", Via Columbia, 2, Roma, Italy, e-mail: alessio.farcomeni@uniroma2.it

cus on the case in which observed counts are censored, that is, they only provide an upper bound for the true underlying counts. This happens for instance when some individuals not belonging to the population of interest might have been included. This situation is rather common in epidemiology and social sciences research, but it is oftentimes overlooked. Formally, we extend works by Overstall et al (2014); Overstall and King (2014); Alunni Fegatelli et al (2017), which deal with right-censored counts in one list, to the case of more than one list, interval-censored counts, and possible presence of external information. Surprisingly enough, use of external information can be advantageous even when it is somehow misspecified (Dotto and Farcomeni, 2018). Our motivating example comes from an original study on the prevalence of MS in the metropolitan area of Rome (Farcomeni et al, 2018). Six centers provided us with lists of their patients, and several patients appear in more than one list. All centers also provided us with information regarding gender, age at prevalence date, date of last visit. The prevalence date was set at the last day of December, 2015. One center (MS unit of Fondazione Santa Lucia) has also carefully verified that patients were alive at the prevalence date, but the other ones have not. For patients not included in the Santa Lucia list we can only note that subjects whose last visit occurred after the prevalence date were clearly alive at the prevalence date. For the other patients, we can not confirm this. We thus have five out of six lists with some censored counts. True counts therefore have both upper (the total number of patients in the list) and lower (the total number of patients that could be confirmed to be alive at the prevalence date) bounds. Source code to implement the proposed approaches, and the data set, are available as Supporting Information in the accompanying paper (Farcomeni, 2020).

2 Method

Suppose there are n cells in an incomplete contingency table, of which n_O are observed precisely, n_U are completely unobserved, and n_C are censored. Denote with Y_{U_j}, $j = 1,\ldots,n_U$, the counts, to be estimated, in the unobserved cells and Y_{C_j} the true but unobserved cell counts in the censored cells. The observed counts in the censored cells are denoted Z_{C_j}, $j = 1,\ldots,n_C$. Finally, Y_{O_j}, $j = 1,\ldots,n_O$ denote the precisely observed counts. For each censored cell we also have an upper and a lower bound, so that $L_{C_j} \leq Y_{C_j} \leq U_{C_j}$ for $j = 1,\ldots,n_C$. In the most simple scenario, $L_{C_j} = 0$ and $U_{C_j} = Z_{C_j}$, the observed left-censored count. We denote with $Y = (Y_O, Y_C, Y_U) = (Y_1,\ldots,Y_n)$ the collection of true (observed, censored, and unobserved) cell counts. Common assumptions on Y include that they are Poisson or multinomial distributed. In order to make inference we follow a Bayesian framework. Note that the observed data is given by uncensored observed counts Y_O, and by lower and upper bounds L_C and U_C. After some algebra and using obvious conditional independence assumptions (see Farcomeni (2020) for more details) it can be seen that

$$\pi(Y_U, Y_C, \beta | Y_O, L_C, U_C) \propto \pi(Y|\beta)\pi(\beta)\pi(U_C, L_C | Y_C, \beta); \quad (1)$$

which can be simply marginalized to obtain $\pi(Y_U, Y_C | Y_O, L_C, U_C)$. A simple assumption for the last factor in (1) is that U_{C_i} and L_{C_i} are conditionally independent, and that U_{C_i} is uniformly distributed in the interval $[Y_{C_i}, \infty)$, while L_{C_i} is uniformly distributed in the interval $[0, Y_{C_i}]$. As prior for the β parameters we use $\beta_{-0} \sim N(0, n(X'_{-1}X_{-1})^{-1})$, which is a restricted unit information prior (Ntzoufras et al, 2003) (UIP), where β_{-0} indicates all parameters except the intercept and X_{-1} the design matrix without the first column of ones. The model will not in general be known *a priori*. Model selection can be performed using Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993, 1997; Ntzoufras et al, 2000). Let γ denote a binary vector. We assume

$$\beta_{-0} \sim N(0, \Gamma n(X'_{-1}X_{-1})^{-1}\Gamma), \quad (2)$$

with $\Gamma = \text{diag}\left(\sqrt{\gamma_j \tau_{1j}^2 + (1-\gamma_j)\tau_{0j}^2}\right)$. In this work we fix $\tau_{1j} = 1$ for all j so that when an effect is included in the model, it follows the UIP marginally; and $\tau_{0j} = 1e-3$. For γ we assume prior independence and $\Pr(\gamma_j = 1) = 0.5$. The posterior distribution for the population size is a direct by-product of the MCMC sampling scheme we describe below. Indeed, $\Pr(N|Y_O, L_C, U_C) = \Pr(\sum_j Y_{C_j} + \sum_j Y_{U_j} = N - \sum_j Y_{O_j} | Y_O, L_C, U_C)$. Since with SSVS more than one design matrix is used throghout the algorithm, the final resulting population size estimate can be obtained either via model averaging (Hoeting et al, 1999) (that is, averaging over all sampled models), or after model selection. In this case we suggest including all effects with a posterior probability of inclusion larger than 50%, the so called median model, which has been shown to be optimal from a predictive perspective (Barbieri and Berger, 2004) and model consistent under mild assumptions (Farcomeni, 2010).

In order to generate an MCMC sample from the posterior distribution (1) we employ a data augmentation procedure, where the missing data Y_U and Y_C is sampled at each iteration conditionally on the current value of the parameters. A Metropolis-within-Gibbs allows us to sample from the full conditional for β. When the model is not fixed the inclusion indicators are then sampled from their full conditional distribution, which is a Bernoulli. Finally, in order to update the unobserved cell entries we use their full conditional distribution, which simply corresponds to a Poisson. Similarly, the full conditional distribution for censored cells corresponds to a Poisson distribution restricted to the support given by lower and upper bounds.

We conclude this section by discussing external information on the probability of censoring, which in some cases is available. We suggest that external information is seen as an aggregated data measurement D (as in Bayesian meta-analysis, for instance). A consequence is that external information induces conditional independence between Y_{C_j} and β. Hence, $\pi(Y_{C_j} = x | D, L, U, \beta) = \pi(Y_{C_j} = x | D, L, U)$, with support $x = L_{C_j}, L_{C_j} + 1, \ldots, U_{C_j}$. These probabilities can be computed as sums of probability of censoring for the $U_{C_j} - L_{C_j}$ uncertain individuals, which under independence correspond to a Poisson-binomial (a binomial under further assumptions of homogeneity).

3 Simulations

We generate data fixing $S = 6$ lists and $N = 5000$, with 5 censored lists. The full simulation study can be found in Farcomeni (2020). Upper bounds are obtained by adding a binomial random variable with maximum value five times the true unobserved count, and success probability p_s. Lower bounds by subtraction of an independent identically distributed binomial. When we use wrong external information we further misspecify the censoring probability p_x by sampling it from a uniform in the interval $(p_x * 0.8, \min(p_x * 1.2, 0.995))$. We report the square root of the median squared error (RMSE) when estimating the true N. Results are in Table 1.

Table 1 RMSE for $\hat{N}$ for different approaches and success probabilities for censoring counts (p_s). The first 5 lists are censored. INC-C (Overstall et al, 2014) only uses upper bound information, IGN-C treats censored counts as truly observed, INC-B-C uses our lower and upper bound plus SSVS approach, INC-B-EXT-C additionally uses external information, INC-B-EXT-C (Wrong) uses wrong external information. Results are averaged over $B = 1000$ replicates.

p_s	INC-C	IGN-C	INC-B-C	INC-B-EXT-C	INC-B-EXT-C (Wrong)
0.10	1746.61	2039.75	1658.60	43.50	401.22
0.25	4676.05	5103.40	4579.26	110.50	377.43
0.33	6234.28	6736.40	6136.06	144.90	393.43

The simulation results clearly indicate that INC-B-C outperforms INC-C and IGN-C in all settings. Use of external information can additionally dramatically improve the performance of INC-B-C, even when this is somehow misspecified; as noted also in a different context in Dotto and Farcomeni (2018).

4 Data analysis

According to official (even if outdated) prevalence estimates the total number of patients with MS in the metropolitan area of Rome is slightly above 4,000. Results obtained with our data (i) ignoring the issue of censored counts, and (ii) using log-linear models with unobserved heterogeneity in a frequentist framework, lead to $\hat{N} = 4,610$ (Farcomeni et al, 2018). The resulting estimated prevalence would be 143 per 100,000. As external information we use official age-gender-specific mortality rates for the general Italian population as published by ISTAT, the national statistical agency, adjusted using standardized mortality ratios specific to multiple sclerosis (Scalfari et al, 2013). In our analysis we will compare, in addition to INC-C and IGN-C, with classical Chao estimators and Generalized Chao (GC) estimators with covariates (Böhning et al, 2013; Farcomeni, 2018); both ignoring the censoring problem (IGN- methods) and restricting only to subjects guaranteed to be alive (LB-methods). Results are reported in Table 2.

Our approach with use of lower bounds leads to very reasonable and stable estimates, regardless of the use of covariates and external information. Our fi-

Table 2 Population size estimates $\hat{N}$ with 95% Credibility Intervals (CI) for different approaches applied to the multiple sclerosis data set. IGN- methods treat censored cells as truly observed, INC-C methods use upper bound information, LB-C fix censored counts at their lower bounds, INC-B-C use both lower and upper bound information as proposed and -EXT- use also external information. Finally -Covs methods additionally stratify on gender and age class.

Method	$\hat{N}$	CI-low	CI-up
IGN-C	4612	3767	5665
INC-C	224	180	420
LB-C	1012	1007	1043
IGN-Chao	4813	4544	5082
LB-Chao	1320	1188	1453
IGN-C-Covs	4655	3777	5736
INC-C-Covs	NA	NA	NA
LB-C-Covs	312	308	328
IGN-Chao-Covs	5003	4715	5290
LB-Chao-Covs	1380	1237	1523
INC-B-C	4212	4036	4347
INC-B-C-Covs	4188	4042	4375
INC-B-EXT-C	4227	4115	4338
INC-B-EXT-C-Covs	4225	4111	4339

nal estimates are indeed more credible than the previously published figure (Farcomeni et al, 2018) and have reasonably narrow 95%CI. Our final estimate for the prevalence of MS in the metropolitan area of Rome is then 131.2 per 100,000 $(95\%CI : 127.6 - 134.8)$. We note that previously published $\hat{N} = 4,610$ lies outside all credibility intervals obtained with our method; hence this estimate might be deemed to have included dead patients in the prevalence estimate.

References

Alunni Fegatelli D, Farcomeni A, Tardella L (2017) Bayesian population size estimation with censored counts. In: Bohning D, van der Heijden PGM, Bunge J (eds) Capture-recapture methods for the social and medical sciences, CRC Press, Boca Raton, FL, pp 371–385

Barbieri MM, Berger JO (2004) Optimal predictive model selection. Annals of Statistics 32:870–897

Böhning D, Vidal-Diez A, Lerdsuwansri R, Viwatwongkasem C, Arnold M (2013) A generalization of Chao's estimator for covariate information. Biometrics 69:1033–1042

Cormack RM (1989) Log-linear models for capture-recapture. Biometrics 45:395–413

Dotto F, Farcomeni A (2018) A generalized Chao estimator with measurement error and external information. Environmental and Ecological Statistics 25:53–69

Farcomeni A (2010) Bayesian constrained variable selection. Statistica Sinica 20:1043–1062

Farcomeni A (2016) A general class of recapture models based on the conditional capture probabilities. Biometrics 72:116–124

Farcomeni A (2018) Fully general Chao and Zelterman estimators with application to a whale shark population. Journal of the Royal Statistical Society (Series C) 67:217–229

Farcomeni A (2020) Population size estimation with interval censored counts and external information: prevalence of multiple sclerosis in Rome. Biometrical Journal doi:10.1002/bimj.201900268 (in press)

Farcomeni A, Cortese A, Sgarlata E, Alunni Fegatelli D, Marfia GA, Buttari F, Mirabella M, De Fino C, Prosperini L, Pozzilli C, Grasso MG, Iasevoli L, Di Battista G, Millefiorini E (2018) The prevalence of multiple sclerosis in the metropolitan area of Rome: a capture-recapture analysis. Neuroepidemiology 50:105–110

Fienberg SE (1972) The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika 59:591–603

Fienberg SE, Johnson MS, Junker BW (1999) Classical multilevel and Bayesian approaches to population size estimation using multiple lists. Journal of Royal Statistical Society, Series A 162:383–405

George E, McCulloch R (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association 88:881–889

George E, McCulloch R (1997) Approaches for bayesian variable selection. Statistica Sinica 7:339–373

Hoeting J, Madigan D, Raftery A, Volisnky C (1999) Bayesian model averaging: a tutorial. Statistical Science 14:382–417

Ntzoufras I, Forster JJ, Dellaportas P (2000) Stochastic search variable selection for log-linear models. Journal of Statistical Computation and Simulation 68:23–37

Ntzoufras I, Dellaportas P, Forster JJ (2003) Bayesian variable and link determination for generalised linear models. Journal of Statistical Planning and Inference 111:165–180

Overstall A, King R (2014) conting: An R package for Bayesian analysis of complete and incomplete contingency tables. Journal of Statistical Software 58:1–27

Overstall AM, King R, Bird SM, Huchinson SJ, Hay G (2014) Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Statistics in Medicine 33:1564–1579

Scalfari A, Knapperts V, Cutter G, Goodin DS, Ashton R, Ebers GC (2013) Mortality in patients with multiple sclerosis. Neurology 81:184–192

Changes in environment extremes and their impacts

FPCA Clustering of rainfall events

Classificazione di dati pluviometrici basata su FPCA

Gianluca Sottile, Antonio Francipane, Leonardo Noto, Giada Adelfio

Abstract The increasing occurrence of flood events in some areas of the Southern Mediterranean area (e.g., Sicily), over the last few years, has contributed to raise the importance of characterizing such events and identifying their causes. Since most of these events can be related to high-intensity rainfalls due to convective events, it is very important to understand which factors could be recognized as drivers of such extreme rainfalls. Nevertheless, the way to distinguish between convective and stratiform rainfall is still today an open issue and not easy to solve. With this regard, starting from precipitation time series recorded at different rain gauge stations of Sicily, which is the greatest island of the Mediterranean Sea, this work proposes an algorithm capable to classify precipitation distinguishing between their convective and stratiform components.

Abstract *Negli ultimi anni si sono osservati sempre più eventi pluviometrici di natura estrema anche nell'area del Sud Mediterraneo. Sembra pertnato necessario un maggiore approfondimento delle loro caratteristiche e delle principali cause. Ad oggi la distinzione tra eventi convettivi e stratiforme è ancora un problema aperto. In questo lavoro si presenta un algoritmo basato sulla natura funzionale dei dati al fine di classificare e caratterizzare gli eventi osservati in una regione dell'area analizzata.*

Key words: FPCAC, clustering of curves, rainfall events

G. Sottile and G. Adelfio
Dipartimento di Scienze Economiche Aziendali e Statistiche, Università di Palermo, e-mail: gianluca.sottile@unipa.it

A. Francipane and L. Noto
Dipartimento di Ingegneria, Università di Palermo e-mail: vleonardo.noto@unipa.it

1 Introduction

Since the impacts of climate change on the environment have been constantly rising over the last decades, scientists have paid much attention to understand the effects of this phenomenon (Kunkel et al. 1999). Climate change leads to different kinds of extremes, such as heavy rainfall events, characterized by short duration and high intensity, and drought, which can cause the problem of water scarcity over a certain area. These types of extreme events cause several damages for the affected areas since they can result in loss of human lives and economic damages. Heavy rainfall events, for example, because of their characteristics, especially when they hit small catchments with low times of concentration, may result in flash floods (Aronica et al. 2012; Forestieri et al. 2016) and cause economic damages and, more relevantly, human lives losses. Because of their high intensity and short duration, these kind of events are often associated to convective rainfalls. In the last years, the increase of occurrence of these kinds of phenomena in many areas of Europe, including Sicily, which is the largest island of the Mediterranean Sea, has contributed to raising the importance of understanding which factors could be recognized as drivers of these events. Arnone et al. (2013) studied the changes in rainfall statistics over Sicily, finding out an increasing trend in the occurrence of shortest-duration rainfall events. Nonetheless the distinction between convective and stratiform precipitation is a very important task today, it is not easy to deal with and still today a big challenge to face (Kyselý et al. 2016). In fact, such a distinction is connected with the characteristics of rainfall events, such as the rain rate and duration, which are usually different for a convective and a stratiform precipitation (Liu et al. 2013). Still today, despite some studies tried to set up different criterions to make a net distinction between the convective and the stratiform rainfall events (Fiori et al. 2014; Rulfová and Kyselý 2013), others affirm that in some cases there could be a coexistence of these two kinds of events (Houze 1997; Liu et al. 2013). As an example, Houze (1997) affirmed that, in some cases, convective phenomena are embedded in stratiform regions. In this case, the spatial patterns are similar to those of a stratiform precipitation, even though they also present a convective component. It is clear, then, that a clear-cut separation of the events in only two classes is improper and difficult to carry out and then, with this regard, Rulfová and Kyselý (2013) introduced a class of mixed/unresolved events to classify all of those events.

In this paper a k-means clustering algorithm denoted FPCAC and based on principal component analysis rotation of curves data is proposed to overcome the limitation of the cross-correlation, as well as an alternative to methods based on the interpolation of data by splines or linear fitting (Sangalli et al. (2010), Adelfio et al. (2011), Garcìa-Escudero and Gordaliza (2005)). This technique is applied to rainfall data collected from 2003 to 2018 in Sicily, from six rain gauge stations more or less uniformly distributed over the region.

The rest of the paper is organized as it follows. In section 2 a brief description of the characteristics of pluviometric events in Sicily is reported. In section 3 functional data analysis basic notation, focusing on PCA for functional data and the proposed clustering approach are reviewed. The application of the proposed clus-

tering approach to highlight common characteristics of rainfall data is proposed in section 4. Some conclusions are provided in section 5.

2 Pluviometric events in Sicily

Sicily (Italy) is the largest island of the Mediterranean Sea and covers an area of about 25,000 km^2. The elevation varies a lot across the island, ranging from 0 m a.s.l. along the coast to more than 3,000 m a.s.l. at the volcano Etna. Sicily has always experimented a high spatial and temporal variability of precipitation. With regard to the spatial variability, Di Piazza et al. (2011) obtained a spatial distribution of the mean annual precipitation (MAP) over Sicily with higher MAP recorded in the northeast of the region, where it reaches about 1,900 *mm*, and lower MAP in the southeastern part of the island (about 360 *mm*). The overall mean of the MAP is about 700 *mm*. With reference to the temporal variability, instead, rainfall is mostly concentrated in winter, whilst the summer season (i.e., June, July, and August) is usually rainless. With regard to air temperature, Sicily is a region with a temperate-mesothermal (Mediterranean) climate, with a dry summer, having an average temperature in the hottest month greater than 22°C, with a precipitation regime more intense in the coldest season. The highest values of mean annual temperature (MAT) are around 18.5 ± 19.5°C along the coast, while the lowest values ($10.5/13.5$°C) characterize higher elevations, with a minimum above the Etna volcano. The warmest areas are the flat lands in the North West nearby the city of Trapani and in the South East close to the city of Catania, both with a relevant agricultural tradition. In the last years many areas of the island have been experiencing some very intense rainfall events, usually concentrated in between the end of the summer and the beginning of the autumn, that cause urban floods and flash floods with consequent economic damages and, sometimes, human lives losses.

3 Clustering of rainfall data by FDA

When data are observed as functions of time we refer to as functional data, referring to n pairs (t_i, y_i) where y_i is the value of an observable variable x at time t_i, and focusing on a set of functions defined on $[0,T]$, such that:

$$\{y_i = x_i(t); i = 1,2,\cdots,I; 0 \leq t \leq T\} \tag{1}$$

Therefore, assuming that a functional for replication i can be represented by a set of discrete measured values $y_{i1}, y_{i2}, \cdots, y_{in}$ the first task is to convert these values to a function x_i with values $x_i(t)$ computable for any t, called functional objects. In the functional context the counterparts of variable values are functional values $x_l(t), l = 1, \cdots, p$ and the discrete index j in the multivariate context is now replaced

by the continuous index s, such that:

$$f_l = \int_{\Omega_s} \beta(s) x_l(s) ds \tag{2}$$

with $\beta(s)$ weight functions and Ω_s a subset of R. In the literature, the term harmonic is used to refer to principal component of variation in curves analysis (see Ramsey and Silverman (2006) for more details).

The functional PCA-based clustering approach, denoted as the FPCAC algorithm and proposed by Adelfio et al. (2011), introduces a variation of the trimmed k-means Robust Curve Clustering (RCC) algorithm (Garcìa-Escudero and Gordaliza, 2005), that is a kind of robust version of k-means methodology through a trimming procedure.

In particular, FPCAC looks for clusters of functions according to the direction of largest variance, finding a linear approximation of each curve by a finite p dimensional vector of coefficients defined by the FPCA scores, assigning event to the cluster on the basis of a distance measure, considering the matrix of FPCA scores instead of the coefficients of a linear fitting to B-spline bases.

4 FPCAC for rainfal data

A dataset from the regional agency SIAS (Servizio Informativo Agrometeorologico Siciliano - Agro-meteorological Information Service of Sicily) has been used because of its high temporal resolution, quality, and availability of up-to-date data. The database includes more than 100 rain gauge stations distributed over the entire island that collect the data with a temporal resolution of 10 minutes. Specifically, data from six rain gauge stations (i.e., Enna, Montalbano Elicona, Palermo, Ragusa, Sciacca e Trapani Fontanasalsa), more or less uniformly distributed over the region, have been collected for the period 2003 – 2018 (Figure 1). For each rain gauge, starting from the original dataset, the rainfall events have been identified and extracted within the observation period providing, for each event, the duration (hours) and depth (*mm*). Two subsequent rainfall events have been defined as different if they are separated by at least equal to three hours and have a total depth of at least equal to 1 mm.

For each of the six rain gauge stations across the Sicily, we applied the FPCA clustering algorithm using 5 harmonics to discriminate the identified rainfall events between convective and stratiform precipitations. In figure 2, just as an example, we provide results of the clustering method for Palermo.

In table 1, we reported the numbers of rainfall events classified as stratiform and convective precipitations; minimum, average (standard deviation) and maximum rainfall intensity (mm/hr).

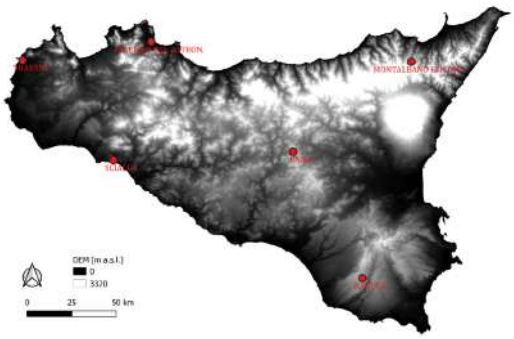


Fig. 1 Location of the considered SIAS rain gauge stations.

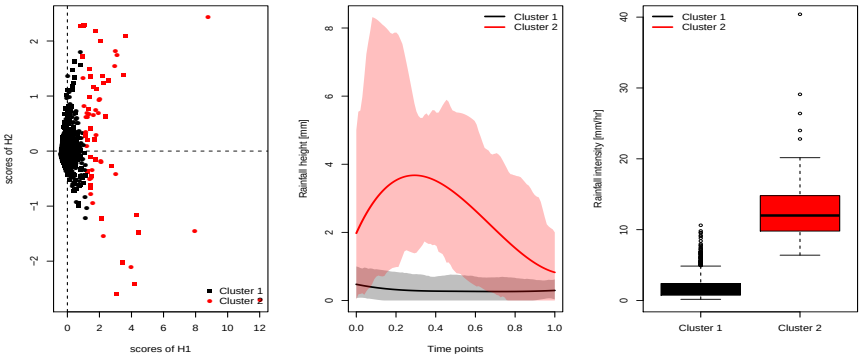


Fig. 2 FPCA clustering method applied to Palermo rain gauge station. Black and red color refer to stratiform and convective precipitation. The left panel shows scores of the first harmonic vs the second one; middle panel shows mean rainfall height in mm of the two clusters and the interquartile range for each normalized time point; the right panel compares boxplots of rainfall intensity (mm/hr)

Table 1 Descriptive statistics of the stratiform and convective precipitations identified applying the FPCAC method to the six rain gauge stations.

Stations	Stratiform				Convective			
	N.	Min	Avg (SD)	Max	N.	Min	Avg (SD)	Max
Enna	1155	0.13	2.17 (2.64)	19.31	20	15.84	26.70 (11.59)	63.60
Montalbano Elicona	1343	0.13	1.47 (1.20)	10.00	43	5.00	11.23 (6.22)	32.64
Palermo	1402	0.17	1.81 (1.55)	10.62	65	6.40	13.62 (7.02)	45.60
Ragusa	1121	0.11	2.02 (2.03)	16.20	29	9.60	22.45 (11.52)	64.80
Sciacca	1216	0.16	1.99 (2.08)	15.90	28	8.70	18.50 (7.57)	46.60
Trapani Fontanasalsa	1319	0.16	1.79 (1.62)	13.52	53	6.40	14.38 (7.39)	43.20

5 Discussion and conclusion

The proposed algorithm seems to clearly distinguish the two precipitation components. However, deeper analysis could be carried out in order to individuate the main causes and features of these two components, based on dependence models accounting for spatial information.

The results in table 1 suggest that the stratiform precipitations have a similar behaviour among the six rain gauge stations, whilst the convective ones are more heterogeneous, as a function of the geographical site.

Finally, one of the advantages of the procedure is related to an immediate use of PCA for functional data avoiding some objective choices related to the splines fitting.

References

1. Adelfio, G., Chiodi, M., D'Alessandro, A. and Luzio, D.: FPCA algorithm for waveform clustering. Journal of Communication and Computer, **8**(6), 494–502 (2011).
2. Arnone E., Pumo D., Viola F., Noto L. V., La Loggia G.: Rainfall statistics changes in Sicily. Hydrol. Earth Syst. Sci. **17**, 2449–2458 (2013).
3. Aronica, G.T., Brigandí, G., Morey, N.: Flash floods and debris flow in the city area of Messina, north-east part of Sicily, Italy in October 2009: the case of the Giampilieri catchment. Nat. Hazards Earth Syst. Sci. **12**, 1295–-1309 (2012).
4. Di Piazza A., Lo Conti F.L., Noto L.V., Viola F., La Loggia G.: Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily. Italy International Journal of Applied Earth Observation and Geoinformation. **13**, 396–408 (2011).
5. Fiori E. et al.: Analysis and hindcast simulations of an extreme rainfall event in the Mediterranean area: The Genoa 2011 case. Atmospheric Research. **138**, 13–29 (2014).
6. Forestieri A., Caracciolo D., Arnone E., Noto L.V.: Derivation of Rainfall Thresholds for Flash Flood Warning in a Sicilian Basin Using a Hydrological Model. Procedia Engineering. **154**, 818–825 (2016).
7. Garcìa-Escudero, L. A. and Gordaliza, A.: A proposal for robust curve clustering, Journal of classification, **22**, 185–201 (2005).
8. Houze: Stratiform Precipitation in Regions of Convection: A Meteorological Paradox? Bulletin of the American Meteorological Society **78**, 2179–2196 (1997).
9. Kunkel K.E., Pielkke R.A. Jr., Changnon S.A.: Temporal Fluctuations in Weather and Climate Extremes That Cause Economic and Human Health Impacts: A Review. Bulletin of the American Meteorological Society. **80**, 1077–1098 (1999).
10. Kyselý J., Rulfová Z., Farda A., Hanel M.: Convective and stratiform precipitation characteristics in an ensemble of regional climate model simulations. Climate Dynamics. **46**, 227–243 (2016).
11. Liu P., Li C., Wang Y., Fu Y.: Climatic characteristics of convective and stratiform precipitation over the Tropical and Subtropical areas as derived from TRMM PR. Science China Earth Sciences. **56**, 375–385 (2013).
12. Ramsey, J. O. and Silverman, B. W.: Functional Data Analysis. Springer, New York (2006).
13. Rulfová Z., Kyselý J.: Disaggregating convective and stratiform precipitation from station weather data. Atmospheric Research. **134**, 100–115 (2013).
14. Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V.: K-means alignment for curve clustering. Computational Statistics and Data Analysis. **54** (5), 1219–1233 (2010).

Trends in rainfall extremes in the Venice lagoon catchment

Tendenze nelle precipitazioni estreme nel bacino scolante della laguna di Venezia

Ilaria Prosdocimi and Carlo Gaetan

Abstract A high number of unprecedented extreme weather events have occurred in recent years in the Venice lagoon and the surrounding region. These events have triggered the questions of whether such large events are linked to climate change. In this study we analyse daily rainfall records from long-term gauging stations in the Venice lagoon catchment. The records are analysed using extreme value model based on the peaks-over-threshold approach: several models are employed to investigate the presence of changes in the extreme rainfall in the catchment. The impact of the parametrisation of the suggested models on the representation of change in high quantiles of the rainfall distribution is investigated.

Abstract La laguna di Venezia e la regione circostante sono state colpite negli ultimi anni da molti eventi climatici estremi. La possibilità che la frequenza e intensità di questi eventi sia legata al cambiamento climatico è quindi di interesse. In questo studio vengono analizzate serie temporali di accumulazioni di piogge misurate da pluviometri in funzione da lungo tempo nel bacino scolante della laguna di Venezia. Vengono in particolare utilizzati modelli di analisi dei valori estremi basati sui picchi sopra una soglia che permettano di caratterizzare il cambiamento nella distribuzione delle piogge estreme. Quindi, viene studiato l'impatto della parametrizzazione utilizzata sulla caratterizzazione del cambiamenti dei quantili.

Key words: Peak-over-threshold, quantiles, rainfall, Venice lagoon

Ilaria Prosdocimi
Ca' Foscari University of Venice, DAIS e-mail: ilaria.prosdocimi@unive.it

Carlo Gaetan
Ca' Foscari University of Venice, DAIS e-mail: carlo.gaetan@unive.it

1 Extreme value analysis for risk assessment

To manage risk related to natural hazards one needs to quantify the relationship between magnitude and frequency of intense events. This relationship is quantified by means of the quantile function: $q(1-p) = x_p$, so that $P(X \leq x_p) = 1-p$, where X denotes the random variable of interest, e.g. daily rainfall accumulation. Typically the distribution of X is unknown and needs to be determined statistically. Further, when assessing risks connected to natural hazards such as rainfall, the interest lies in the frequency of extreme events: the statistical characterisation of X is therefore carried out using only the part of the record which can be considered in some way extreme and for which extreme value theory results hold. In this work we focus on the peaks over threshold approach: only rainfall events which exceed a certain high threshold are considered to be extreme and used to estimate the distribution of extreme daily rainfall. Denoting with X the daily rainfall accumulation and with u a fixed high threshold it can be shown that $X-u$, conditional on X being larger than u, is approximated by a Generalised Pareto (GP) distribution [2]. Denoting with $Y = X-u$ the threshold exceedances we have that: $(Y|X > u) \sim GP(\sigma, \xi)$, with CDF: $F(y) = 1-(1+\xi y/\sigma)$, where $\sigma > 0$ and $\xi \in \mathbb{R}$ are the scale and shape parameter of the distribution. Notice that when $\xi = 0$ the GP distribution reduces to an exponential distribution.

The quantile function for the Generalised Pareto (GP) distribution has the form:

$$q(1-p;u,\sigma,\xi) = \begin{cases} u + \frac{\sigma}{\xi}\left(p^{-\xi}-1\right) & for \xi \neq 0 \\ u + \sigma \log p & for \xi = 0 \end{cases} \quad (1)$$

Once suitable estimates are found for σ and ξ the quantile function is used to estimate the relationship between rainfall totals and the probability of such totals to be exceeded. In engineering, the quantiles of interest are typically denoted by their annual exceedance probability (AEP) $p = P(X > x_p)$ and the closely related return period T, which is the expected time between two events larger than x_p. If the distribution of X is unchanged in time and rainfall events are independent of each other the return period T is the reciprocal of the AEP: $T = 1/p$.

When assessing risk using the peaks over threshold approach one typically assumes that the variable of interest X_i is recorded for a long period of time, with recordings at time t_i independent of other recordings at time t_j ($i \neq j$). For a large threshold u, $Y_i = X_i - u$ is approximated by a GP distribution, and it is typically assumed that Y_i and Y_j are iid. Nevertheless, the increasing evidence of global changes in the climate, puts the assumption that climate variables are identically distributed in time into question. Therefore models which allow for the distribution to change as a function of time or some other forcing variable (e.g. CO_2 emissions, temperature, etc.) have been proposed. For the peaks over threshold approach these mostly consist of two types of models: models which allow the GP parameters or the threshold u to change [2, 1, 3].

In this work four time varying models are employed to investigate whether the distribution of rainfall extremes in the Venice lagoon catchment has changed:

1. a model in which the threshold is constant and the scale parameter for the GP distribution changes linearly in time: $Y_i = (X_i - u) \sim GP(\sigma_0 + \sigma_1 t_i, \xi)$;
2. a model in which the threshold changes linearly in time and the scale parameter for the GP distribution is constant: $Y_i = (X_i - (\alpha + \beta t_i)) \sim GP(\sigma, \xi)$;
3. a model in which the threshold and scale parameter change linearly in time independently of each other: $Y_i = (X_i - (\alpha + \beta t_i)) \sim GP(\sigma_0 + \sigma_1 t_i, \xi)$;
4. a model in which the threshold changes linearly in time and with the scale parameter for the GP distribution changes proportionally to the threshold: $Y_i = (X_i - (\alpha + \beta t_i)) \sim GP(\tau(\alpha + \beta t_i), \xi)$.

All models keep a constant shape parameter in time: the shape parameter is a often found to have high variability and it is rarely allowed to change [2]. Notice also that when estimating the models in practice one needs to ensure that the threshold and the scale functions are positive. While model 1. to 3. have already been employed in the literature the last model has not been used for the detection of trends, but has the advantage of being a parsimonious model to describe changes in both the threshold and the scale parameter of the GP distribution. Further the model has the advantage of implying changes in the quantile function which are constant for all return periods and therefore easy to interpret.

When assessing changes in environmental extremes, the interest often lies in how the distribution quantiles, rather than the parameters, are changing. The different time varying models imply different functional forms of change over time for the quantile function. These can be characterised by taking the ratio of the quantile function estimated at different points in time, say $(t_0 + \Delta t)$ and t_0:

$$M(p, t_0, \Delta t) = \frac{q(p; u(t_0 + \Delta t), \sigma(t_0 + \Delta t), \xi)}{q(p; u(t_0), \sigma(t_0), \xi)}. \tag{2}$$

For example, the ratio of the quantile function on 01/01/2015 over the value on the 01/01/2005 could be of interest to assess what changes have occurred in ten years. When the scale is allowed to change proportionally to the threshold (model 4.), the relative change is constant across all return periods, i.e. all exceedance probabilities p, since the ratio in (2) can be determined to be $M(p, t_0, \Delta t) = 1 + \beta \Delta t / (\alpha + \beta t_0)$.

2 Data

The data analysed in this study have been provided by the Veneto regional environment protection agency (ARPAV). In particular, daily rainfall accumulation records from rain gauges located in the Venice lagoon catchment were made available. From all available records only stations which were operational till the end of 2018 and which have at least 20 years of valid recordings have been kept. A year was deemed to have valid recordings if at least 274 days in the year had a record. In total 20

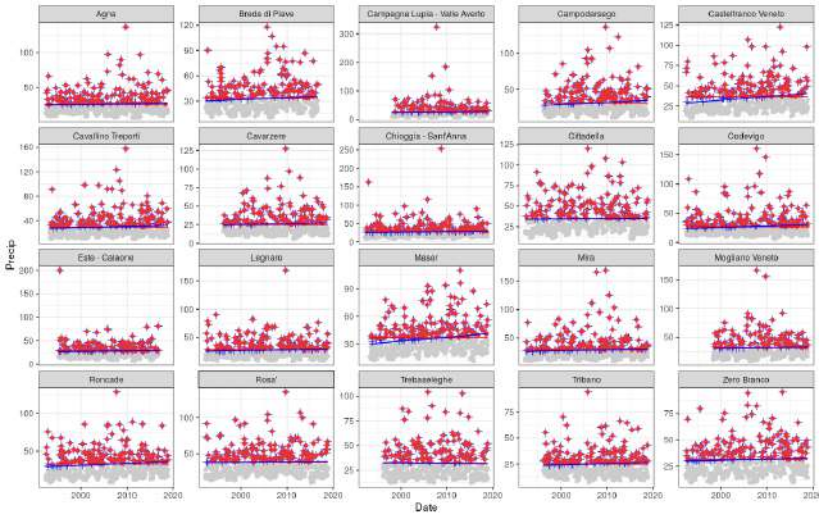


Fig. 1 High daily rainfall data with thresholds and peaks over threshold. Red lines and dots indicate the case of constant threshold. Blue lines and crosses indicate the case of varying threshold.

stations are included in the study and the models are fitted to each station separately, using a constant and a time varying thresholds. At each station, the constant threshold u is selected such that an average of 5 independent peaks in a year is extracted from the daily records. Peaks are considered to be independent when they are separated by at least 5 days. Additionally, a time varying threshold $u_i = u(t_i)$ was determined by means of quantile regression [4] allowing rainfall accumulation quantiles to change as a linear function of time. The varying threshold was also chosen so that an average of 5 independent peaks per year are extracted.

Figure 1 shows the top 20 independent events per year at each station included in the study, together with the constant and varying threshold and the selected peaks. Notice that for each station when using a constant or a varying threshold different peaks are included in the analysis. From the Figure it is clear that some stations have been operating for a longer period of time with several stations starting in 1992.

3 Results

Figure 1 shows that most varying threshold are increasing in time. Only the station in Trebaseleghe shows a slightly negative change in the higher quantile. Nevertheless, the quantile regression slope was found to be significantly different from zero at the 10% level only for one station of the 20 under study (Maser). A map showing the test statistic for trend in the quantile regression (i.e. the estimated slope divided by

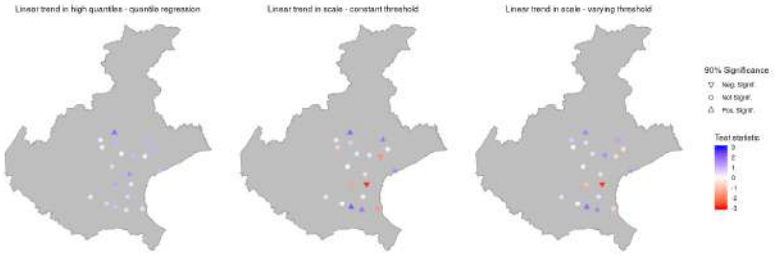


Fig. 2 Left panel: map of test statistic for trend in high quantile - quantile regression. Central panel: map of test statistic for trend in the scale parameter - model with constant threshold. Right panel: map of test statistic for trend in the scale parameter - model with a varying threshold.

its estimated standard deviation) is shown in Figure 2. The test statistic for trend for the scale parameter in model 1. is also shown in the central panel. The model (as all other GP models) was fitted via the maximum likelihood approach. Some increasing and decreasing trends are visible, with two significant decreasing trend and four significantly increasing trends (at the 10% significance level). A similarly scattered result is also found when both the threshold and the scale parameter are allowed to change independently, although only three stations exhibit slopes which are significantly different from zero. The evidence for changing behaviour of rainfall extremes when analysing stations independently is weak.

Figure 3 shows the relative change between the estimated quantiles for the year 2015 and 2005 for the four models at each rainfall station. The changes derived in models in which the scale is allowed to change (in the left panels) tend to be more variable across the different return periods compared to the other models. The relative change for events with higher return period appear to be larger: when rainfall accumulations have been predicted to increase (decrease) in time large events are estimated to increase (decrease) relatively more than frequent events. When only the threshold is allowed to change, little variability can be seen in the implied relative changes in quantiles across return periods.When the threshold, rather than the scale is allowed to change, the long return period events appear to be increasing relatively less than the frequent, low magnitude events. When the scale is allowed to change proportionally to the threshold, the relative change is constant across all return periods, with a fairly high variability across the different stations.

4 Conclusions and perspectives

A peaks-over-threshold analysis of potential changes in rainfall daily accumulation in the Venice lagoon catchment finds little evidence of change in the distribution of extreme rainfall. Several models were employed: significant changes were identified

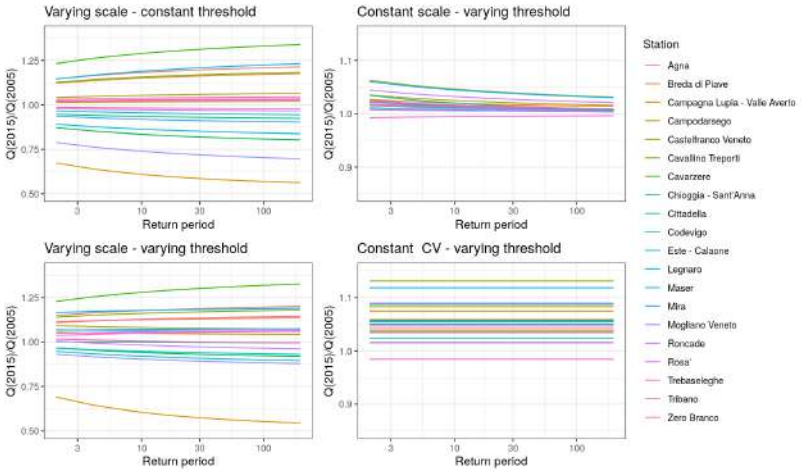


Fig. 3 Relative change across return period (in log scale) for the four different time varying models. Changes are over a ten year period between 2005 and 2015.

only in a handful of stations, with no clear pattern of changes in the area. Further work will focus on a spatial analysis, rather than the separate study of each station.

It is proposed that when comparing the suitability of regression models to describe changes in extremes, measures based on changes implied in the quantile function might be useful. A model is proposed in which the scale parameter changes proportionally to the threshold. The model implies constant relative changes across return periods: this feature improves the interpretability of the derived changes in quantiles. The applicability of the model, and the suitability of its assumptions will be further explored in other applications.

Acknowledgements

The research presented was funded by CORILA - research program VENEZIA 2021. The authors thank ARPAV for providing the rainfall data.

References

1. Coelho, C., Ferro, C., Stephenson, D., Steinskog, D.: Methods for exploring spatial and temporal variability of extreme events in climate data. Journal of Climate (2008) 10.1175/2007JCLI1781.1
2. Coles, S.G.: An Introduction to Statistical Modeling of Extreme Values. Springer, London (2001)
3. Eastoe, E.F., Tawn, J.A.: Modelling non-stationary extremes with application to surface level ozone. Journal of the Royal Statistical Society: Series C (2009) 10.1111/j.1467-9876.2008.00638.x
4. Koenker, R.: Quantile Regression. Cambridge University Press, London (2005)

Copulas: models and inference

Analysis of district heating demand through different copula-based approaches

Analisi della domanda del teleriscaldamento attraverso diversi approcci basati sulla copula

F. Marta L. Di Lascio and Andrea Menapace

Abstract The main purpose of this paper is to investigate the multivariate dependence relationship between meteorological variables and thermal energy demand through copula modelling. In order to keep into account both the multivariate dependence structure and the diverse bivariate relationships observed in the data, mainly two approaches can be followed. One is to adopt a mixture of copulas that allows us to combine different copula families and to generate complex dependence structures not captured by the existing models. The other way is to use vine copulas that make it possible to analyse multivariate probability distributions through bivariate and conditional bivariate copulas organised in a suitable tree. A comparison of the two approaches is discussed.

Abstract *Lo scopo principale del presente contributo è quello di proporre uno studio sulla dipendenza multivariata tra la domanda di calore e le variabili meteo attraverso modelli copula. Al fine di tenere in considerazione sia la struttura di dipendenza multivariata che le relazioni bivariate osservate, è possibile seguire diversi approcci. Un primo approccio si basa sulle misture di copulae che permettono di combinare diversi modelli copula per generare strutture di dipendenza complessa non appartenenti a modelli noti. Un secondo approccio riguarda l'uso delle copulae vine che permettono di analizzare distribuzioni di probabilità multivariate utilizzando copulae bivariate e copulae bivariate condizionate organizzate in strutture ad albero. Un confronto tra questi due approcci è discusso.*

Key words: Copula function, Meteorological variables, Mixture copula, R-vine copula, Thermal energy demand

F. Marta L. Di Lascio
Faculty of Economics, Free University of Bozen-Bolzano, e-mail: marta.dilascio@unibz.it

Andrea Menapace
Faculty of Science and Technology, Free University of Bozen-Bolzano, e-mail: andrea.menapace@unibz.it

1 Introduction

The European Green Deal recently presented by the European Commission aims at transforming Europe in the "first climate-neutral continent by 2050". To achieve this purpose, investments in green technologies and sustainable solutions, like the district heating (DH hereafter) that should be designed to efficiently distribute thermal energy through smart grids for any heat request in the urban area [5], are crucial. The efficient planning and managing a DH system requires an in-depth knowledge of the relationship between thermal energy demand (ED hereafter) and meteorological variables. [2] and [3] are the only studies that address this issue using copulas [8, 9] to keep into account the complex multivariate dependence structure underlying the data generator process. Copula models have been successfully used for analysing several engineering phenomena, e.g. for environmental risk assessment [7], food processes controlling [1], and energy production and demand analyses [10].

In this paper, we present a copula-based analysis between the ED of a DH system and the weather conditions, specifically outdoor temperature (OT hereafter) and solar radiation (SR hereafter). We start from the analyses presented in [3] and investigate the performance of R-vine copula models in capturing a three-variate dependence with different kinds of pair dependencies. We, hence, analyse the dataset used in [3] and compare the goodness of fit of our approach with that used in [3], which is based on mixture copulas.

The rest of the paper presents the problem statement in Section 2, the dependence modelling through R-vine copula in Section 3, and final remarks in Section 4.

2 Problem statement

The main purpose of [3] is to develop a statistical model able to keep into account the relationships between thermal energy demand and meteorological variables with the final goal of supporting district heating operators in a proper planning of the heat production and distribution.

The dataset used in [3] concerns the city of Bozen-Bolzano (Italy) and it includes the daily thermal ED of consumers connected to the DH of Bozen-Bolzano and the corresponding daily average of OT and SR as detected by the weather station of S. Maurizio. The period of observation goes from 2014 to 2017 and the analysis is performed only on the data of the heating season (from October 15th to April 15th according to the Italian law). For further information on the collected dataset see Section 3 in [3].

The methodology proposed in [3] consists in three steps and it can be summarised as follows:

1. the analysis of the serial correlation of ED, OT and SR time series taken separately that is performed through seasonal autoregressive integrated moving average (SARIMA hereafter) models;

2. the investigation of the dependence relationship between the three time series of uncorrelated SARIMA residuals using copula models;
3. the analytical derivation of the copula-based conditional probability function of ED given information on OT and SR.

Focusing on Step 2., it can be solved in different ways. [3] follow an approach based on a finite mixture of heterogeneous parametric copulas and specify a mixture of a three-dimensional unstructured Student-t copula $(C^t(\cdot))$ and a three-dimensional flipped Clayton copula $\left(C_{001}^{\text{Clay}}(\cdot)\right)$ for the variable vector $\mathbf{U} = (U_1, U_2, U_3)$ containing the SARIMA residual time series of OT, SR, and ED, respectively:

$$\begin{aligned} C_{\text{mix}}^{t\text{-Clay}}(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\pi}) &= \pi_1 C^t(U_1, U_2, U_3) + \pi_2 C_{001}^{\text{Clay}}(U_1, U_2, U_3) \\ &= \pi_1 (t_{3,\nu}(t_\nu^{-1}(U_1), t_\nu^{-1}(U_2), t_\nu^{-1}(U_3); \boldsymbol{\theta}_t)) + \\ &\quad \pi_2 \left(\left(U_1^{-\theta_{\text{Clay}}} + U_2^{-\theta_{\text{Clay}}} - 1 \right)^{-\frac{1}{\theta_{\text{Clay}}}} - \left(U_1^{-\theta_{\text{Clay}}} + \right. \right. \\ &\quad \left. \left. + U_2^{-\theta_{\text{Clay}}} + (1 - U_3)^{-\theta_{\text{Clay}}} - 2 \right)^{-\frac{1}{\theta_{\text{Clay}}}} \right) \end{aligned} \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2)$, with $0 \leq \pi_k \leq 1, \forall k$ and $\sum_{k=1}^{2} \pi_k = 1$, is the vector of weights, $t_{p,\nu}(\cdot)$ is the standard p-variate Student-t distribution with ν degrees of freedom (recall that ν controls the heaviness of the tails) and unstructured correlation matrix $\boldsymbol{\theta}_t$, $t_\nu^{-1}(\cdot)$ denotes the inverse univariate Student-t distribution function, θ_{Clay} is the Clayton dependence parameter, and $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_t, \theta_{\text{Clay}}\right)$ is the unknown dependence parameter vector. The selected model is consistent with the empirical bivariate and multivariate relationships as shown in Fig. 1. The estimated weights in Eq. (1) result $\hat{\pi}_1 = 0.847$ and $\hat{\pi}_2 = 0.153$, and the whole model, which has been selected among a set of models (see, for details, Section 5.2 in [3]) on the basis of the Akaike Information Criterion (AIC hereafter), has AIC equals to -260.98.

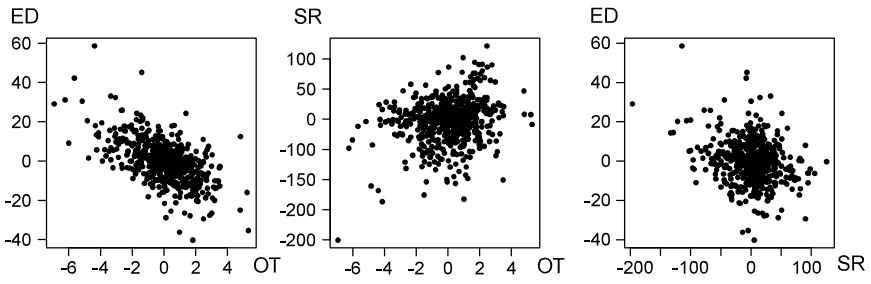


Fig. 1 Scatter plots of the residual time series of outdoor temperature in $^\circ C$ (OT), solar radiation in W/m^2 (SR), and thermal energy demand in MWh (ED).

Since the weight $\hat{\pi}_2$ of the flipped Clayton component in Eq.(1) is small and the selected model is not very parsimonious (7 parameters), we address the three-variate analysis through alternative copula-based approaches. We perform a comparison between different approaches and we stress their advantages and disadvantages.

3 R-vine copula modelling

The simplest way to model the relationships between the residual time series of ED, OT, and SR is to use a single parameter copula model, that is very parsimonious since it expresses the three-variate relationship in only one dependence parameter. This model, however, postulates that all the bivariate marginal models are the same ones, going against the observed bivariate relationships (see Tab. 1) that cannot be reproduced by only one dependence parameter.

Table 1 Estimated Kendall's correlation coefficient $\hat{\rho}_\tau$ of each pair of the residual time series of thermal energy demand (ED), outdoor temperature (OT), and solar radiation (SR), and the p-value of the corresponding test.

	ED and OT	ED and SR	SR and OT
$\hat{\rho}_\tau$	-0.405	-0.135	0.106
p-value	< 0.0001	< 0.0001	0.0002

In light of this consideration, an alternative approach is based on vine copulas [4]. Vines provide the possibility to model multivariate dependencies through a tree of bivariate copulas making it possible to both keep into account all the observed pairwise relationships and define a parsimonious model. We specify a three-dimensional R-vine copula consisting of two trees: the first tree represents the dependence between the residuals of ED and OT, and between the residuals of SR and OT, while the second tree represents the dependence between the residuals of ED and SR conditionally on the residuals of OT. See Fig. 2 for a representation of the described R-vine tree. Based on the AIC, the three bivariate copulas composing the vine model and best fitting the pairwise relationships in our data are: the Student-t copula with $\nu = 9$ and dependence parameter equals to -0.595 for the residual ED and OT, the survival Clayton copula with dependence parameter equals to 0.25 for the residual SR and OT, and the Gumbel copula rotated by 270 degrees with dependence parameter equals to 1.15 for the residual ED and SR, given OT. The selected R-vine copula has the AIC equals to -251.85. Tab. 2 shows the estimation results and Fig. 3 the contour plots of the bivariate densities composing the selected R-vine copula that are in agreement with the patterns observed in the data (see Fig. 1). The analysis of the vine copulas has been carried out using the R package `vinecopula` [6].

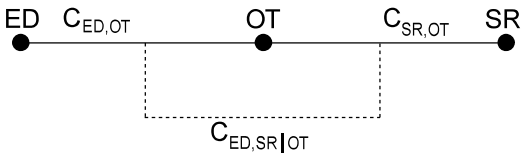


Fig. 2 R-vine copula tree regarding the bivariate copulas ($C_{(\cdot,\cdot)}$) and the conditional copula ($C_{(\cdot,\cdot|\cdot)}$) of the residual time series of ED, OT, and SR according to the model in Tab. 2.

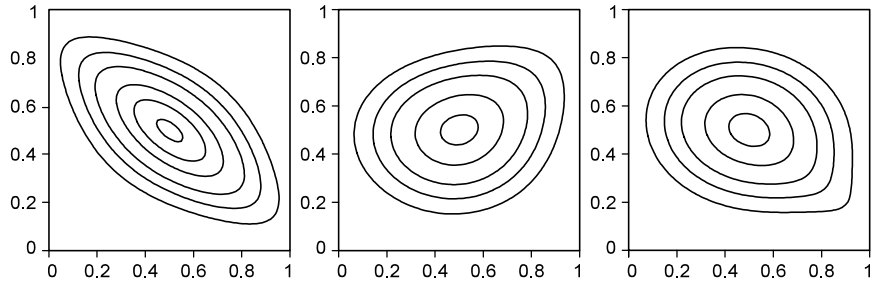


Fig. 3 Bivariate copula density contour plot of the residual time series of ED (x-axis) and OT (y-axis) (left), SR (x-axis) and OT (y-axis) (middle), ED (x-axis) and SR (y-axis) given OT (right) composing the R-vine model in Tab. 2.

Table 2 Three-variate R-vine copula model: estimation results of the bivariate copulas of the residual time series of ED, OT, and SR. Loglik is the value of the maximized log-likelihood of the copula fit.

Tree	Edge	Variables	Copula Model	Rotation	Dependence Parameter	$\hat{\rho}_\tau$	Loglik
1	1	ED,OT	Student-*t*	0	-0.595	-0.406	118.76
1	2	SR,OT	Clayton	180	0.250	0.111	11.77
2	1	ED,SR|OT	Gumbel	270	1.151	-0.130	15.96

4 Conclusion

The mixture copula model is to be preferred to the R-vine copula model on the basis of the AIC that is equal to -260.98 and -251.85, respectively. In addition, the interpretation of the mixture of copulas makes it possible to easily assess the impact of both the meteorological variables on the energy demand, that is the most important goal to support district heating operators in terms of both the production schedule and the storage management of the thermal energy. Hence, even though the R-vine copula model is more parsimonious than the mixture of copulas, we deem that the latter is more advantageous.

Acknowledgements The first author acknowledges the support of the Free University of Bozen-Bolzano via the project "The use of Copula for the Analysis of Complex and Extreme Energy and

Climate data" (CACEEC). Both the authors acknowledge the support of the Free University of Bozen-Bolzano via the project "Techno-economic methodologies to investigate sustainable energy scenarios at urban level" (TESES-Urb).

References

1. Arvanitoyannis, I., Chalhoub, C., Gotsiou, P., Lydakis-Simantiris, N., Kefalas, P.: Novel quality control methods in conjunction with chemometrics (multivariate analysis) for detecting honey authenticity. Crit. Rev. in Food Science and Nutr. **45**(3), 193–203 (2005)
2. Di Lascio, F.M.L., Menapace, A., Righetti, M.: Joint and conditional dependence modelling of peak district heating demand and outdoor temperature: a copula-based approach. Stat. Methods & Appl. pp. 1–23 (2019). DOI 10.1007/s10260-019-00488-4
3. Di Lascio, F.M.L., Menapace, A., Righetti, M.: Analysing the relationship between district heating demand and weather conditions through conditional mixture copula. BEMPS - Bozen Economics & Management Paper Series (2020). URL https://econpapers.repec.org/paper/bznwpaper/bemps68.htm
4. Joe, H.: Multivariate Models and Dependence Concepts. Chapman and Hall, London (1997)
5. Lund, H., Werner, S., Wiltshire, R., Svendsen, S., Thorsen, J.E., Hvelplund, F., Mathiesen, B.V.: 4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems. Energy **68**, 1–11 (2014). DOI 10.1016/j.energy.2014.02.089
6. Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E., B., G., Erhardt, T.: Vinecopula: Statistical inference of vine copulas. r package version 2.2.0.
7. Salvadori, G., Durante, F., De Michele, C., Bernardi, M., Petrella, L.: A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. Water Resour. Res. **52**(5), 3701–3721 (2016)
8. Sklar, A.: Fonctions de répartition à *n* dimensions et leures marges. Publ. de l'Inst. de Stat. de L'Univ. de Paris **8**, 229–231 (1959)
9. Trivedi, P.K., Zimmer, D.M.: Copula Modeling: An Introduction for Practitioners, vol. 1. Found. and Trends in Econom. (2005)
10. Wang, Y., Infield, D.G., Stephen, B., Galloway, S.J.: Copula-based model for wind turbine power curve outlier rejection. Wind Energy **17**(11), 1677–1688 (2014)

CoVaR and backtesting: a comparison between a copula approach and parametric models

Analisi del backtesting del CoVaR: un confronto tra funzioni copula e modelli parametrici

Michele Leonardo Bianchi and Giovanni De Luca and Giorgia Rivieccio

Abstract In this paper, we estimate the backtestable version of the Conditional Value-at-Risk (i.e. CoVaR$^{\leq}$) by fitting different multivariate parametric models that capture four stylized facts about multivariate financial time series of equity returns: heavy tails, negative skew, asymmetric dependence, and volatility clustering. While the volatility clustering effect is got by AR-GJRGARCH dynamics, the other stylized facts are captured through both non-Gaussian multivariate models and copula functions. The CoVaR$^{\leq}$ is computed for four Italian assets on the basis of the multivariate normal, the multivariate normal tempered stable model, the multivariate generalized hyperbolic model and the AIC-based best copula function among a set of Elliptical and Archimedean copulas.

Abstract *Il lavoro contiene una stima della versione del Value-at-Risk condizionato (ovvero CoVaR$^{\leq}$) sottoponibile al backtest. Si considerano diversi modelli parametrici multivariati che catturano quattro fatti stilizzati su serie temporali multivariate di rendimenti finanziari: code pesanti, asimmetria negativa, dipendenza asimmetrica e volatility clustering. Mentre l'effetto del volatility clustering è ottenuto da un modello AR-GJRGARCH, gli altri fatti stilizzati vengono catturati attraverso entrambi i modelli multivariati non gaussiani e funzioni copula. Per quattro asset italiani, il CoVaR$^{\leq}$ viene calcolato considerando il modello normale multivariato, il modello multivariato normale temperato stabile, il modello multivariato generalizzato iperbolico e la migliore funzione copula selezionata secondo il criterio di Akaike (AIC) tra un insieme di copule ellittiche e archimedee.*

Michele Leonardo Bianchi
Regulation and Macroprudential Analysis Directorate, Bank of Italy, e-mail: micheleleonardo.bianchi@bancaditalia.it
This publication should not be reported as representing the views of the Bank of Italy. The views expressed are those of the author and do not necessarily reflect those of the Bank of Italy.

Giovanni De Luca
University of Naples Parthenope e-mail: giovanni.deluca@uniparthenope.it

Giorgia Rivieccio
University of Naples Parthenope e-mail: giorgia.rivieccio@uniparthenope.it

Key words: Systemic risk, Value-at-Risk, Conditional Value-at-Risk, copula functions, backtesting.

1 Introduction

Assuming that the Conditional Value-at-Risk (CoVaR$^{\leq}$) is a proper systemic risk measure, we explore to which extent the model assumptions on the univariate financial institution log-returns and on the dependence structure affect the estimation of this risk measure. Additional, we conduct a backtesting analysis to obtain a more robust model comparison.

In [1], the distress of a financial institution is defined as the event $(y^j = VaR^j_\alpha)$, where y^j is the random variable representing the log-returns of the financial institution j and VaR^j_α the corresponding Value-at-Risk (VaR) at tail level α. Here we consider the CoVaR$^{\leq}$ measure, that is the Conditional Value-at-Risk where the conditioning event is the distress of a financial institution represented through $(y^j \leq VaR^j_\alpha)$. This allows us to have a robust systemic risk measure which can be backtested without a great effort (see [14] and [2]).

In this work we assume that the univariate timeseries have AR-GJRGARCH dynamics and then we compare different dependence structures. We consider the multivariate normal (MN) distribution (as a benchmark) and some non-normal multivariate distributions: the multivariate normal tempered stable (MNTS) and the multivariate generalized hyperbolic model (MGH), and the best copula function in terms of AIC among normal, t, BB1 and BB7 copulas, as described in [12] and [13]. Both non-normal multivariate distributions and copula functions are widely known in the financial literature. [9] and [10] analyzed both MNTS e MGH models applied to risk assessment and portfolio optimizations (see also [8] and [7]). [16] developed a model based on the MNTS distribution to estimate the CoVaR.

2 Methodology

For each institution j, the random variable y^j_t represents the log-returns of the market value of equity. Superscript *sys* denotes the entire financial system, i.e. a stock market index or the capitalization-weighted portfolio of all financial institutions in the selected sample.

Given the VaR of financial institution j, with tail level α, for a given tail level β, the $\text{CoVaR}^{\leq j}_{\beta,\alpha,t}$ of the financial system conditional on financial institution j being in distress (i.e. market returns of bank j are less or equal to its VaR_α) is equal to

$$P\left(y^{sys}_t \leq CoVaR^{\leq j}_{\beta,\alpha,t} \,|\, y^j_t \leq VaR^j_{\alpha,t}\right) = \beta \tag{1}$$

The financial institution j contribution to systemic risk is then defined by

$$\Delta CoVaR^{\leq j}_{\beta,\alpha,t} = CoVaR^{\leq j}_{\beta,\alpha,t} - CoVaR^{\leq j}_{\beta,0.5,t}.$$

where a level α equal to 50% denotes a normal, or median, state. $\Delta\text{CoVaR}^{\leq j}_{\beta,t}$ captures the negative externality that financial institution j imposes on the financial system.

The systemic risk measure estimation is divided in three steps. In the first step we estimate a univariate AR-GJRGARCH model to the time-series of log-returns and compute the Value-at-Risk (VaR) at the given tail level α. In the second step we calibrate the bivariate dependence structure by applying different multivariate approaches. In the third step we estimate the systemic risk measure.

Let S_t be the stock price process of a given financial institution and

$$y_t = \log \frac{S_t}{S_{t-1}}$$

be its log-return process for which we assume an AR(m)-GJRGARCH(p,q) model. The most frequent case, the AR(1)-GJRGARCH(1,1) model, is given by

$$\begin{aligned} y_t &= a y_{t-1} + \sigma_t \varepsilon_t + c \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \left(|\sigma_{t-1}\varepsilon_{t-1}| - \gamma(\sigma_{t-1}\varepsilon_{t-1})\right)^2 + \beta_1 \sigma_{t-1}^2 \end{aligned}$$

where the innovation ε_t are independent and identically distributed random variables with zero mean and unit variance. As observed in [15], the follow equality holds

$$-VaR_\alpha(y_{t+1}) = a y_t + \sigma_t(-VaR_\alpha(\varepsilon_{t+1})) + c.$$

It should be noted that a numerical inversion is usually needed to compute these quantiles.

While closed formula for the CoVaR are available under normal distributional assumptions (see [3]), for non-normal models a numerical integration procedure is needed (see [14] and [4]). More in details, equation (1) can be written as

$$P\left(y_t^{sys} \leq CoVaR^{\leq j}_{\beta,\alpha,t}, y_t^i \leq VaR^j_{\alpha,t}\right) = \alpha\beta.$$

In the MNTS and MGH cases, given the density f_t^j of the bivariate random variable defined by y_t^{sys} and y_t^j, the equality

$$\int_{-\infty}^{CoVaR^{\leq j}_{\beta,\alpha,t}} \int_{-\infty}^{VaR^j_{\alpha,t}} f_t^j(y^{sys}, y)dxdy = \alpha\beta$$

can be considered to obtain an estimate of $CoVaR^{\leq}$.

In the copula function cases, as shown in [3]), if one considers the dependence structure between X_t^{sys} and y_t^j, then equation (1) can be rewritten as

$$C\left(F_{y_t^{sys}}(CoVaR^{\leq j}_{\beta,\alpha,t}), F_{y_t^j}(VaR^j_{\alpha,t})\right) = \alpha\beta \tag{2}$$

where C is a given copula function. By following the approach described in [3], it is possible to obtain an estimate of the CoVaR by means of the inversion of the copula function. Equation (2) has been proved under weaker conditions on the copula in [5].

The definition (1) allows one to perform a two-steps backtesting. In both steps it is possible to follow the approach proposed in [11]. First we conduct a preliminary VaR backtest by considering the entire observation window and defining a first hit sequence (1 if the loss of the financial institution on that day was larger than its predicted VaR level, and zero otherwise). Then we define a subset of observations on the basis on the distress of the financial institution j (i.e. $X_t^i \leq VaR_{\alpha,t}^j$). Thus, by looking at this subset, we can backtest the CoVaR. More in details, we compare the CoVaR forecasts with the ex-post losses of the financial system and define a second hit sequence which is 1 if the loss of the financial system on that day is larger than its predicted CoVaR level, and zero otherwise. The accuracy of forecasted VaR and CoVaR is then analyzed performing the unconditional coverage (UC) and the conditional coverage (CC) tests.

3 Data and results

We have analyzed daily dividend-adjusted closing prices from January 2000 through January 2020 for four Italian assets (two banks, Intesa SanPaolo and Unicredit, and two insurance companies, Generali and UnipolSai). Figure 1 shows the behaviour of the log-returns. For the systemic risk we have selected the FTSEMIB index.

One-step-ahead forecasts of VaR and CoVaR are obtained after estimating the time series in the previous five years. The first forecasts are referred to January, 3, 2005, while the rolling samples include 1267 observations.

The results are presented in terms of p-values of the UC and CC tests with $\alpha = \beta = 0.05$.

The backtesting of the VaR is generally satisfactory, except when the MN is applied to UnipolSai.

The backtesting of the CoVaR shows very high p-values. In particular, for Intesa SanPaolo and Unicredit the copula approach shows better results, while for Generali and UnipolSai the p-values provided by the MNTS and MGH cases are slighlty superior. Moreover, it is worthwhile to stress that all the methods appear to be robust in presence of some abnormal values of UnipolSai returns.

References

1. Adrian, T., Brunnermeier, M.K.: CoVaR. Am. Econ. Rev. **106**, 1705–1741 (2016)
2. Banulescu, D., Hurlin, C., Leymarie, J., Scaillet, O.: Backtesting marginal expected shortfall and related systemic risk measures. Swiss Finance Institute Research Paper. **19-48** (2019)

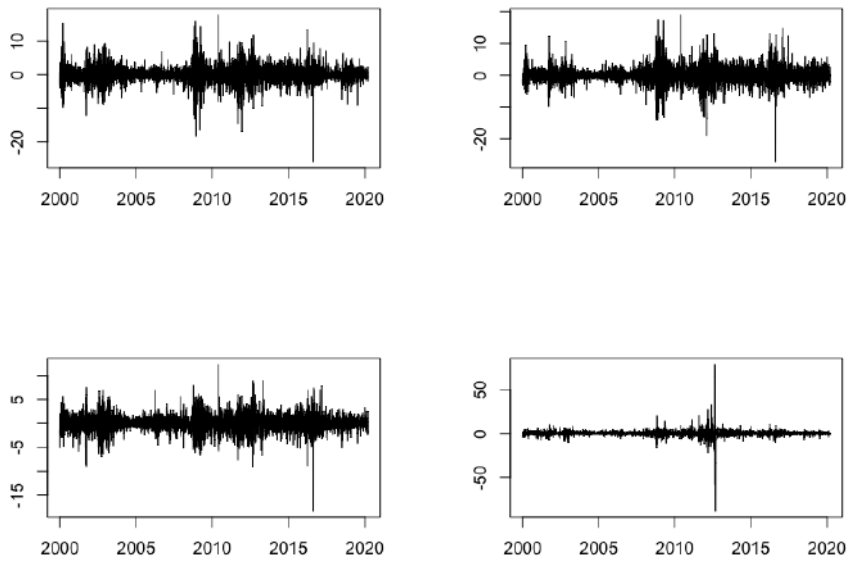


Fig. 1 Time-series log-returns of Intesa SanPaolo (top-left), Unicredit (top-right), Generali (bottom-left) and UnipolSai (bottom-right).

3. Bernard, C., Brechmann, E.C., Czado, C.: Statistical assessments of systemic risk measures. In Fouque, J.P., Langsam, J. (Eds.), The handbook of systemic risk, Cambridge University Press (2012)
4. Bernard, C., Czado, C.: Conditional quantiles and tail dependence. J. Multivariate Anal. **138**, 104–126 (2015)
5. Bernardi, M., Durante, F., Jaworski, P.: CoVaR of families of copulas, Statist. Probab. Lett., **120**, 8–17 (2017)
6. Bianchi, M.L., Sorrentino, A.M.: Measuring CoVaR: An empirical comparison, Comput. Econ. **55**, 511–528 (2020)
7. Bianchi, M.L., Hitaj, A., Tassinari, G.L.: Multivariate non-Gaussian models for financial applications. Preprint (2020)
8. Bianchi, M.L., Stoyanov, S.V., Tassinari, G.L., Fabozzi, F.J., Focardi, S.M.: Handbook of Heavy-Tailed Distributions in Asset Management and Risk Management. World Scientific (2019)
9. Bianchi, M.L., Tassinari, G.L., Fabozzi, F.J.: Riding with the four horsemen and the multivariate normal tempered stable model. International Journal of Theoretical and Applied Finance. **19**, 1650027 (2016)
10. Bianchi, M.L., Tassinari, G.L.: Forward-looking portfolio selection with multivariate non-Gaussian models and the Esscher transform. Quantit. Financ., to appear.
11. Christoffersen, P.: Backtesting, Encyclopedia of Quantitative Finance. Wiley (2010)
12. De Luca, G., Rivieccio, G.: Conditional Value-at-Risk: a comparison between quantile regression and copula functions, Book of short papers SIS2018 (2018)
13. De Luca, G., Rivieccio, G., Corsaro, S.: Value-at-Risk dynamics: a copula-VaR approach. Eur. J. Financ. **26**, 223–237 (2020)

Table 1 p-values for the unconditional coverage test (UC) and the Conditional Coverage (CC) test.

			Intesa SP	Unicredit	Generali	UnipolSai
MN	VaR	UC	0.586	0.195	0.749	0.078
	($\alpha = 0.05$)	CC	0.823	0.318	0.949	0.015
	CoVaR	UC	0.000	0.000	0.000	0.000
	($\alpha = 0.05$, $\beta = 0.05$)	CC	0.000	0.000	0.000	0.000
MGH	VaR	UC	0.397	0.475	0.586	0.586
	($\alpha = 0.05$)	CC	0.689	0.677	0.823	0.118
	CoVaR	UC	0.728	0.298	0.678	0.678
	($\alpha = 0.05$, $\beta = 0.05$)	CC	0.647	0.451	0.636	0.636
MNTS	VaR	UC	0.318	0.475	0.535	0.397
	($\alpha = 0.05$)	CC	0.605	0.677	0.795	0.129
	CoVaR	UC	0.753	0.298	0.690	0.728
	($\alpha = 0.05$, $\beta = 0.05$)	CC	0.652	0.451	0.639	0.647
Copula	VaR	UC	0.743	0.133	0.800	0.580
	($\alpha = 0.05$),	CC	0.948	0.278	0.965	0.117
	CoVaR	UC	0.829	0.605	0.631	0.678
	($\alpha = 0.05$, $\beta = 0.05$)	CC	0.814	0.586	0.555	0.580

14. Girardi, G., Ergün, A.T.: Systemic risk measurement: Multivariate GARCH estimation of CoVaR. J. Bank. Financ. **37**, 3169–3180 (2013)
15. Kim, Y.S., Rachev, S.T., Bianchi, M.L., Mitov, I., Fabozzi, F.J.: Time series analysis for financial market meltdowns, J. Bank. Financ. **35**, 1879–1891 (2011)
16. Kurosaki, T., Kim, Y.S.: Systematic risk measurement in the global banking stock market with time series analysis and CoVaR. Investment Management and Financial Innovations. **10**, 184–196 (2013)
17. Paolella, M. S., Polak, P., Walker, P.S.: Regime switching dynamic correlations for asymmetric and fat-tailed conditional returns. J. Econometrics **213**, 493–515 (2019)

Estimating Asymmetric Dependence via Empirical Checkerboard Copulas

Stima della Dipendenza Asimmetrica attraverso le copule checkerboard empiriche

Wolfgang Trutschnig and Florian Griessenberger

Abstract This contribution sketches how empirical checkerboard copulas can be used to construct an asymmetric, scale-free dependence estimator quantifying the information gain of a random variable Y by observing a random variable X (and vice versa) given samples $(x_1,y_1),\ldots,(x_n,y_n)$ from (X,Y). The most important statistical properties of empirical checkerboard estimators (ECBE) are discussed, a quick simulation study using the R-package qad (short for 'quantification of asymmetric dependence') which implements the developed estimator illustrates small sample properties of ECBEs.

Abstract *Questo contributo descrive come la copula checkerboard empirica può essere usata per costruire uno stimatore di dipendenza asimmetrico e invariante per cambiamenti di scala quantificando l'apporto d'informazione di una variabile casuale Y dall'osservazione di una variabile casuale X (e viceversa) dato un campione $(x_1,y_1),\ldots,(x_n,y_n)$ generato da (X,Y). Verranno esaminate le più importanti proprietà statistiche degli stimatori checkerboard empirici (ECBE), verrà presentato un veloce studio di simulazione utilizzando il pacchetto R qad (abbreviazione di 'quantification of asymmetric dependence') che implementa lo stimatore presentato illustrandone le proprietà per piccoli campioni di ECBE.*

Key words: dependence measure, copula, asymmetry, consistency

Wolfgang Trutschnig
University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria, e-mail: wolfgang@trutschnig.net, wolfgang.trutschnig@sbg.ac.at

Florian Griessenberger
University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria, e-mail: florian.griessenberger@sbg.ac.at

1 Introduction

Standard dependence measures like Pearson correlation, Spearman correlation, and Schweizer and Wolff's σ (see [10]) are symmetric, i.e. they assign each pair of random variables (X,Y) the same dependence as they assign the pair (Y,X). Independence of two random variables is a symmetric concept (modelling the situation that knowing X does not change our knowledge about Y and vice versa) - dependence, however, is not. From the authors' point of view, notions 'measuring' dependence should not necessarily be symmetric since in many situations the dependence structure is highly asymmetric - think, for instance, of a sample $(x_1,y_1),\ldots,(x_n,y_n)$ in the shape of the letter V, in which case it is without doubt easier to predict the y-value given the x-value than vice versa.

The copula-based, hence scale-invariant dependence measure ζ_1 introduced in [12] (also see [11]) was developed in order to overcome this problem. ζ_1 detects asymmetries in the dependence structure and clearly separates independence and so-called complete dependence describing the exact opposite - the situation where $Y = f \circ X$ for some measurable $f : \mathbb{R} \to \mathbb{R}$ (i.e. the situation of maximal information gain about/predictability of Y when knowing X).

Considering that ζ_1 is based on conditional distributions (Markov kernels) and that (in the continuous setting) estimating conditional distribution is a difficult endeavor, it is a-priori unclear if good estimators can be derived at all. It is therefore to a certain extend surprising that so-called empirical checkerboard estimators (ECBEs) can be shown to be strongly consistent in the general setting, i.e., without any smoothness assumptions.

The rest of this contribution paper is organized as follows: Section 2 gathers some preliminaries and notations, Section 3 recalls the definition of empirical copulas and checkerboard aggregations as well as the main result saying that ECBEs are strongly consistent if the resolution (used for the aggregation) is chosen suitably. The aforementioned small simulation study concludes the paper.

2 Notation and Preliminaries

Throughout the whole contribution $\mathscr{C}$ will denote the family of all two-dimensional copulas (for background on copulas we refer to [1, 8]). For every copula $A \in \mathscr{C}$ the corresponding doubly stochastic measure will be denoted by μ_A. As usual, $d_\infty(A,B)$ will denote the uniform metric on $\mathscr{C}$, i.e. $d_\infty(A,B) := \max_{(x,y)\in[0,1]^2} |A(x,y) - B(x,y)|$, A^t will denote the transpose of $A \in \mathscr{C}$. For every metric space (Ω,d) the Borel σ-field will be denoted by $\mathscr{B}(\Omega)$, λ will denote the Lebesgue measure on $\mathscr{B}([0,1])$. A mapping $K : \mathbb{R} \times \mathscr{B}(\mathbb{R}) \to [0,1]$ is called a Markov kernel from $\mathbb{R}$ to $\mathscr{B}(\mathbb{R})$ if $x \mapsto K(x,B)$ is measurable for every fixed $B \in \mathscr{B}(\mathbb{R})$ and $B \mapsto K(x,B)$ is a probability measure for every fixed $x \in \mathbb{R}$. Suppose that X,Y are random variables on $(\Omega,\mathscr{A},\mathbb{P})$. Then a Markov kernel $K : \mathbb{R} \times \mathscr{B}(\mathbb{R}) \to [0,1]$ is called regular conditional distribution of Y given X if for every $B \in \mathscr{B}(\mathbb{R})$

$$K(X(\omega),B) = \mathbb{E}(\mathbb{1}_B \circ Y|X)(\omega)$$

holds $\mathbb{P}$-a.s. It is well known that a regular conditional distribution of Y given X exists and is unique $\mathbb{P}^X$-almost sure. For every $A \in \mathscr{C}$ the corresponding regular conditional distribution (i.e. the regular conditional distribution of Y given X in the case that $(X,Y) \sim A$) will be denoted by $K_A(\cdot,\cdot)$ and considered as function from $[0,1] \times \mathscr{B}([0,1])$ to $[0,1]$. Note that for every $A \in \mathscr{C}$ and Borel sets $E,F \in \mathscr{B}([0,1])$ we have

$$\int_E K_A(x,F)d\lambda(x) = \mu_A(E \times F).$$

For more details and properties of conditional expectations and regular conditional distributions see [5, 6].

In the current paper we will mainly work with the metrics D_1 and D_∞ introduced in [12]. These metrics are defined by

$$D_1(A,B) := \int_{[0,1]} \underbrace{\int_{[0,1]} |K_A(x,[0,y]) - K_B(x,[0,y])|\, d\lambda(x)}_{=:\Phi_{A,B}(y)}\, d\lambda(y) \tag{1}$$

$$D_\infty(A,B) := \sup_{y\in[0,1]} \Phi_{A,B}(y) \tag{2}$$

respectively. It can be shown that both metrics generate the same topology (without being equivalent), that the metric space $(\mathscr{C},D_1)$ is complete and separable. Furthermore, firstly, $D_1(A,\Pi)$ attains only values in $[0,\frac{1}{3}]$ and that, secondly, $D_1(A,\Pi)$ is maximal if and only if A is completely dependent, i.e. if a λ-preserving transformation $h : [0,1] \to [0,1]$ exists such that $K_A(x,\{h(x)\}) = 1$ for λ-a.e. $x \in [0,1]$. In the sequel we will let $\mathscr{C}_d$ denote the family of all completely dependent copulas, and write A_h and $K_h(\cdot,\cdot)$ for the completely dependent copula and the Markov kernel of the completely dependent copula induced by the λ-preserving transformation h respectively. For equivalent definitions and properties of completely dependent copulas we refer to [12] and the references therein.

Letting (X,Y) denote a pair of continuous random variables with joint distribution function H and copula A the dependence measure ζ_1 is defined by (see [12])

$$\zeta_1(X,Y) = \zeta_1(A) := 3D_1(A,\Pi). \tag{3}$$

As a direct consequence of the properties of D_1 it follows that for all continuous random variables X,Y we have $\zeta_1(X,Y) \in [0,1]$, that $\zeta_1(X,Y) = 0$ if and only if X and Y are independent, and that $\zeta_1(X,Y) = 1$ if and only if the copula A of (X,Y) is completely dependent.

3 Empirical checkerboard estimators

Let (X,Y) be a random vector with joint distribution H, margin distributions F,G and copula $A \in \mathscr{C}$. Given a sample $(x_1,y_1),\ldots,(x_n,y_n)$ of (X,Y) we want to estimate

$\zeta_1(X,Y) = \zeta_1(A)$. The natural idea of simply calculating the empirical copula $\hat{E}_n$ (see, e.g. [2, 3]) and considering $\zeta_1(\hat{E}_n)$ does not yield a reasonable estimator. In fact, it is straightforward to show that for every copula A we have

$$\lim_{n\to\infty} \zeta_1(\hat{E}_n) = 1$$

with probability one (the reason being that empirical copulas are 'close' to complete dependence since the conditional distributions have an interval of length $\frac{1}{n}$ as support).

A natural and simple way to overcome this problem is to smooth or aggregate the empirical copula. Aggregation leads to so-called checkerboard copulas with which we are going to work in the sequel. To simplify notation, for every $N \in \mathbb{N}$ and $i,j \in \{1,\ldots,N\}$ define the square R_{ij}^N by $R_{ij}^N = \left[\frac{i-1}{N}, \frac{i}{N}\right] \times \left[\frac{j-1}{N}, \frac{j}{N}\right]$.

Definition 1. Suppose that $N \in \mathbb{N}$. A copula $A \in \mathscr{C}$ is called N-checkerboard copula if A is absolutely continuous and (a version of) its density k_A is constant on the interior of each square R_{ij}^N. We call N the resolution of A.

Letting $\mathfrak{CB}_N(A)$ denote the so-called N checkerboard approximation of A, i.e.,

$$\mathfrak{CB}_N(A)(x,y) := \int_0^x \int_0^y N^2 \sum_{i,j=1}^{N} \mu_A(R_{ij}^N) \mathbb{1}_{R_{ij}^N}(s,t)\, d\lambda(t) d\lambda(s)$$

then according to [7] (also see [12]) for every copula $A \in \mathscr{C}$ it can be shown that

$$\lim_{N\to\infty} D_1(\mathfrak{CB}_N(A), A) = 0 = \lim_{N\to\infty} D_1(\mathfrak{CB}_N(A)^t, A^t).$$

Having this we can use the fact that, according to [3], for continuous random variables with underlying copula A

$$d_\infty(\hat{E}_n, A) = O\left(\sqrt{\frac{\log(\log(n))}{n}}\right) \tag{4}$$

holds with probability one, and choose the resolution N adequately as a function of the sample size n to prove the following consistency result (for a proof based on several lemmata linking d_∞, D_1 and D_∞ we refer to [4]).

Theorem 1. *Suppose that $(X_1,Y_1),(X_2,Y_2),\ldots$ is a random sample from (X,Y) and assume that (X,Y) has continuous joint distribution function H and copula A. Furthermore suppose that $s \in (0,\frac{1}{2})$ and set $N(n) := \lfloor n^s \rfloor$ for every $n \in \mathbb{N}$. Then:*

$$\lim_{n\to\infty} \zeta_1(\mathfrak{CB}_{N(n)}(\hat{E}_n)) = \zeta_1(A) \ [\mathbb{P}]. \tag{5}$$

In what follows we will refer to $\zeta_1(\mathfrak{CB}_{N(n)}(\hat{E}_n))$ as empirical checkerboard estimator (ECBE) for $\zeta_1(A)$. Summing up, our chosen procedure for estimating

$\zeta_1(X,Y) = \zeta_1(A)$ based on a sample $(x_1,y_1),\ldots,(x_n,y_n)$ of (X,Y) with underlying copula A is as follows:

1. Calculate the empirical copula $\hat{E}_n$.
2. Select the resolution $N = N(n)$ and calculate the empirical N-checkerboard $\mathfrak{CB}_N(\hat{E}_n)$.
3. Estimate $\zeta_1(A)$ by $\zeta_1(\mathfrak{CB}_N(\hat{E}_n))$.

Remark 1. Notice that Theorem 1 holds without any smoothness assumptions about the copula A, not even continuous first order partial derivatives (essential in the context of the empirical copula process, see [9]) are required. Numerous simulations insinuate that $\mathfrak{CB}_{N(n)}(\hat{E}_n)$ might also be a strongly consistent estimator for more flexible choices of $N(n)$, particularly for the case $N(n) := \lfloor n^s \rfloor$ and some $s \geq \frac{1}{2}$ - a clarification of this question is future work, in the R-package 'qad' $s = \frac{1}{2}$ was considered.

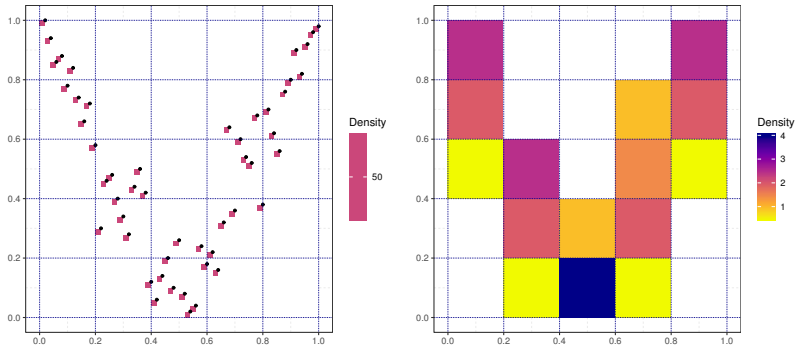


Fig. 1 Left panel: Pseudoobservations of the original sample of size $n = 50$ (black points) and density of the empirical copula $\hat{E}_n$ (magenta squares), Right panel: Density of the empirical checkerboard copula $\mathfrak{CB}_5(\hat{E}_n)$.

4 Simulations

In order to illustrate the performance of the estimator $\zeta_1(\mathfrak{CB}_{N(n)}(\hat{E}_n))$ we consider the Marshall-Olkin copula $M_{0.3,1}$ and the completely dependent copula A_h, where h is given by $h(x) = 5x\,(mod 1)$. Figure 2 and Figure 3 (left panel) depict a sample of size $n = 5000$ of these two copulas. According to Theorem 1 in both cases the empirical checkerboard estimator converges with probability one to $\zeta_1(M_{0.3,1})$ and $\zeta_1(A_h)$, respectively, the theorem does, however, not provide information on the speed of convergence. Generating samples of sizes $n \in \{10,50,100,500,1000,5000,10000\}$ a total of 1000 times yields the results depicted in Figure 2 and Figure 3, respectively.

Acknowledgements The second author gratefully acknowledges the support of the Austrian FWF START project Y1102 'Successional Generation of Functional Multidiversity'.

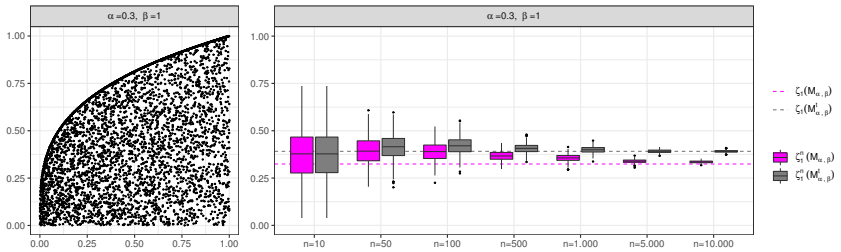


Fig. 2 Boxplots summarizing the 1000 obtained estimates for $\zeta_1(M_{0.3,1})$ (magenta) and $\zeta_1(M^t_{0.3,1})$ (gray). The dashed lines depict $\zeta_1(M_{0.3,1}) \approx 0.324$ and $\zeta_1(M^t_{0.3,1}) \approx 0.391$.

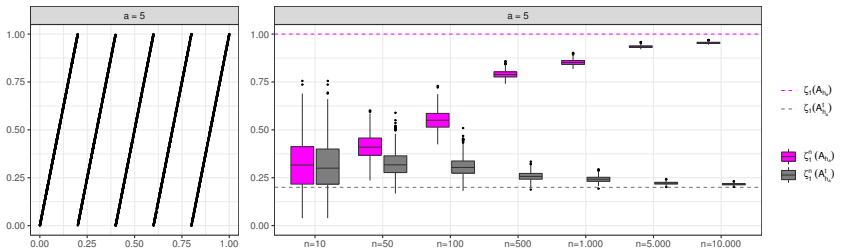


Fig. 3 Boxplots summarizing the 1000 obtained estimates for $\zeta_1(A_h)$ (magenta) and $\zeta_1(A^t_h)$ (gray). The dashed lines depict the true dependence measure $\zeta_1(A_h) = 1$ and $\zeta_1(A^t_h) = \frac{1}{5}$.

References

1. Durante, F., Sempi, C.: Principles of copula theory, CRC Press, New York (2016)
2. Genest, C., Neshlehova, J.G., Remillard, B.: On the empirical multilinear copula process for count data, Bernoulli **20** (3), 1344-1371 (2014)
3. Janssen, P., Swanepoel J., Ververbeke, N.: Large sample behaviour of the Bernstein estimator, *J. Stat. Plan. Infer.* **142**, 1189–1197 (2012)
4. Junker, R.R., Griessenberger, F., Trutschnig, W.: A scale-invariant measure for quantifying direction and asymmetry in dependence, preprint on arXiv, available under `https://arxiv.org/abs/1902.00203`
5. Kallenberg, O.: Foundations of modern probability, Springer-Verlag New York Heidelberg Berlin (1997)
6. Klenke, A.: Wahrscheinlichkeitstheorie, Springer-Verlag Berlin Heidelberg (2008)
7. Li, X., Mikusinski, P., Taylor, M.D.: Strong approximation of copulas, *J. Math. Anal. Appl.* **255**, 608-623 (1998)
8. Nelsen, R.B.: An introduction to copulas, Springer, New York (2006)
9. Segers, J.: Asymptotics of empirical copula processes under non-restrictive smoothness assump-tions, Bernoulli **18** (3), 764-782 (2012)
10. Schweizer, B., Wolff, E.F.: On nonparametric measures of dependence for random variables, Ann. Stat. **9** (4), 879-885 (1981)
11. Siburg, K.F., Stoimenov, P.F.: A measure of mutual complete dependence, Metrika **71** (2), 239-251 (2010)
12. Trutschnig, W.: On a strong metric on the space of copulas and its induced dependence measure, J. Math. Anal. Appl. **384**, 690-705 (2011)

Strong Convergence of Multivariate Maxima
Convergenza Forte Di Massimi Multivariati

Michael Falk and Simone A. Padoan and Stefano Rizzelli

Abstract It is well known and readily seen that the maximum of n independent and uniformly on $[0,1]$ distributed random variables, suitably standardised, converges in total variation distance, as n increases, to the standard negative exponential distribution. We extend this result to higher dimensions by considering copulas. We show that the strong convergence result holds for copulas that are in a differential neighbourhood of a multivariate generalized Pareto copula. Sklar's theorem then implies convergence in variational distance of the maximum of n independent and identically distributed random vectors with arbitrary common distribution function.
Abstract *È noto come il massimo di n variabili casuali indipendenti e uniformemente distribuite su* $[0,1]$*, opportunamente standardizzate, converga all'aumentare di n alla distribuzione esponenziale negativa standard, secondo la distanza totale di variazione. Estendiamo questo risultato in dimensioni maggiori di uno utilizzando la teoria delle copule. Mostriamo che il risultato di convergenza forte vale per le copule multivariate che sono differenziabili e si trovano in un intorno della famiglia Pareto multivariata generalizzata. Quindi, utilizzando il teorema di Sklar, mostriamo la convergenza rispetto alla distanza totale di variazione del vettori di massimi (per componente) di n vettori casuali indipendenti e distribuiti secondo una comune distribuzione arbitraria.*

Key words: Strong convergence; total variation; copula; generalized Pareto copula.

Michael Falk
Institute of Mathematics, University of Würzburg, Würzburg, Germany
e-mail: michael.falk@uni-wuerzburg.de

Simone A. Padoan
Department of Decision Sciences, Bocconi University, Milan, Italy
e-mail: simone.padoan@unibocconi.it

Stefano Rizzelli
Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: stefano.rizzelli@epfl.ch

1 Introduction

Let U be a random variable (rv), which follows the uniform distribution on $[0,1]$, say $V(u)$, $u \in [0,1]$. Let $U^{(1)}, U^{(2)}, \ldots$ be independent and identically distributed (iid) copies of U. Then, clearly, we have for $x \leq 0$ and large $n \in \mathbb{N}$ (natural set),

$$P\left(n\left(\max_{1\leq i\leq n} U^{(i)} - 1\right) \leq x\right) \to_{n\to\infty} G(x), \tag{1}$$

where $G(x)$ is the standard negative exponential distribution function (df). Thus, we have established convergence in distribution of the suitably normalised sample maixmum, i.e.

$$n\left(M^{(n)} - 1\right) \to_D \eta, \tag{2}$$

where $M^{(n)} := \max_{1\leq i\leq n} U^{(i)}$, $n \in \mathbb{N}$, the arrow "$\to_D$" denotes convergence in distribution, and the rv η has df G. Note that, with $v(x) := V'(x) = 1$, if $x \in [0,1]$, and zero elsewhere, we have

$$\begin{aligned} v_n(x) &:= \frac{\partial}{\partial x}\left(V^n\left(1+\frac{x}{n}\right)\right) = V^{n-1}\left(1+\frac{x}{n}\right) v\left(1+\frac{x}{n}\right) \\ &\to_{n\to\infty} g(x) := G'(x) = \begin{cases} \exp(x), & x \leq 0 \\ 0, & x > 0 \end{cases}, \end{aligned}$$

i.e., we have pointwise convergence of the sequence of densities of normalised maximum $n\left(M^{(n)} - 1\right)$, $n \in \mathbb{N}$, to that of η. Scheffé's lemma, see e.g. [4, Lemma 3.3.3], now implies convergence in total variation:

$$\sup_{A\in\mathbb{B}} \left| P\left(n\left(M^{(n)} - 1\right) \in A\right) - P(\eta \in A)\right| \to_{n\to\infty} 0, \tag{3}$$

where $\mathbb{B}$ denotes the Borel-σ-field in $\mathbb{R}$.

Let now X be a rv with *arbitrary* df F and $F^{-1}(q) := \{t \in \mathbb{R} : F(t) \geq q\}$ with $q \in (0,1)$ be the usual *quantile function* or *generalized inverse* of F. Then, we can assume the representation $X = F^{-1}(U)$. Let $X^{(1)}, X^{(2)}, \ldots$ be independent copies of X. Again we can consider the representation $X^{(i)} = F^{-1}(U^{(i)})$, $i = 1, 2, \ldots$. The fact that each quantile function is a nondecreasing function yields

$$\begin{aligned} \max_{1\leq i\leq n} X^{(i)} &= \max_{1\leq i\leq n} F^{-1}\left(U^{(i)}\right) = F^{-1}\left(\max_{1\leq i\leq n} U^{(i)}\right) \\ &= F^{-1}\left(1 + \frac{1}{n}\left(n\left(\max_{1\leq i\leq n} U^{(i)} - 1\right)\right)\right). \end{aligned}$$

The strong convergence in equation (3) now implies the following convergence in total variation:

$$\sup_{A\in\mathbb{B}}\left|P\left(\max_{1\le i\le n}X^{(i)}\in A\right)-P\left(F^{-1}\left(1+\frac{1}{n}\eta\right)\in A\right)\right|\to_{n\to\infty}0. \tag{4}$$

Finally, assume that F is a continuous df with density $f=F'$. We denote the right endpoint of F by $x_0:=\sup\{x\in\mathbb{R}:F(x)<1\}$. Assume also that $F\in\mathscr{D}(G^*_\gamma)$, i.e. F belongs to the domain of attraction of a generalised extreme-value df G^*_γ [3, p. 21]. This means, for $n\in\mathbb{N}$, there are norming constants $a_n>0$ and $b_n\in\mathbb{R}$ such that

$$F^n(a_nx+b_n)\to_{n\to\infty}\exp\left(-(1+\gamma x)_+^{-1/\gamma}\right)=:G^*_\gamma(x), \tag{5}$$

for all $x\in\mathbb{R}$, where $(x)_+=\max(0,x)$ and $\gamma\in\mathbb{R}$ is the so-called *tail index*. Such a coefficient describes the heaviness of the upper tail of the probability density function corresponding to G^*_γ see [3] for details. Furthermore, in this general case, we also have the pointwise convergence at the density level, i.e.

$$f^{(n)}(x):=\frac{\partial}{\partial x}F^n(a_nx+b_n)\to_{n\to\infty}\frac{\partial}{\partial x}G^*_\gamma(x)=:g^*_\gamma(x) \tag{6}$$

for all $x\in\mathbb{R}$, see e.g. Proposition 2.5 in [5]. In particular, if (6) holds true, Scheffé's lemma entails that

$$\sup_{A\in\mathbb{B}}\left|P\left(a_n^{-1}\left(\max_{1\le i\le n}X^{(i)}-b_n\right)\in A\right)-P(Y\in A)\right|\to_{n\to\infty}0, \tag{7}$$

where Y is a rv with distribution G^*_γ and $X^{(i)}$, $i=1,\ldots,n$ are independent copies of a rv X with distribution F, with $F\in\mathscr{D}(G^*_\gamma)$ see [2] for all the details.

In the recent paper [2], we have extended the results in (3), (4) and (7) to higher dimensions. In Theorem 2, we demonstrate that the strong convergence result holds for copulas that are in a differential neighbourhood of a multivariate generalized Pareto copula. As a result of this, we also establish strong convergence of the copula of the maximum of n iid random vectors with arbitrary common df to the limiting extreme-value copula (Corollary 1). These results are discussed in Section 2. The proofs of Theorems 1, 2 and Corollary 1 are available in [2]. Sklar's theorem is then used in the paper [2] to derive convergence in variational distance of the maximum of n iid random vectors with arbitrary common df and of its normalised versions. We do not report this last result here for brevity. All the results developed in [2] address some still open problems in the literature on multivariate extremes.

2 Strong Results for Copulas

Suppose that the random vector (rv) $\boldsymbol{U}=(U_1,\ldots,U_d)$ follows a *copula*, say C, on $\mathbb{R}^d$, i.e., each component U_j has a uniform df V_j. Let $\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\ldots$ be independent copies of $\boldsymbol{U}$ and put for $n\in\mathbb{N}$

$$\boldsymbol{M}^{(n)} := \left(M_1^{(n)},\dots,M_d^{(n)}\right) := \left(\max_{1\le i\le n} U_1^{(i)},\dots,\max_{1\le i\le n} U_d^{(i)}\right). \tag{8}$$

In the sequel the operations involving vectors are meant componentwise, furthermore, we set $\mathbf{0}=(0,\dots,0)$, $\mathbf{1}=(1\dots,1)$ and $\boldsymbol{\infty}=(\infty,\dots,\infty)$. Finally, hereafter, we denote the copula of the random vector in (8) by $C^n(\boldsymbol{u})$, $\boldsymbol{u}\in[0,1]^d$. Suppose there exists a nondegenerate df G on $\mathbb{R}^d$ such that for $\boldsymbol{x}=(x_1,\dots,x_d)\le\mathbf{0}\in\mathbb{R}^d$

$$\begin{aligned} P\left(n\left(\boldsymbol{M}^{(n)}-\mathbf{1}\right)\le\boldsymbol{x}\right) &= P\left(n\left(M_1^{(n)}-1\right)\le x_1,\dots,n\left(M_d^{(n)}-1\right)\le x_d\right)\\ &\to_{n\to\infty} G(\boldsymbol{x}). \end{aligned} \tag{9}$$

Then, G is necessarily a *multivariate max-stable* or *multivariate extreme-value df*, with *extreme-value copula* C_G and standard negative exponential margins G_j, $j=1,\dots,d$ [2, see the reference therein]. In the sequel we refer to the df G in (9) as *standard* multivariate max-stable df. Precisely, the form of G is

$$G(\boldsymbol{x}) = C_G(G_1(x_1),\dots,G_d(x_d)),$$

where the copula C_G can be expressed in terms of $\|\cdot\|_D$, a *D-norm* on $\mathbb{R}^d$, via

$$C_G(\boldsymbol{u}) = \exp\left(-\|\log u_1,\dots,\log u_d\|_D\right), \quad \boldsymbol{u}\in[0,1]^d, \tag{10}$$

while the margins G_j, $j=1,\dots,d$, are exponentially distributed, see e.g. [1]. Therefore, the distribution in (9) has the representation

$$G(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x}\|_D\right), \qquad \boldsymbol{x}\le\mathbf{0}\in\mathbb{R}^d. \tag{11}$$

The convergence result in (9) implies that $C^{(n)}(\boldsymbol{u}) := C^n(\boldsymbol{u}^{1/n}) \to_{n\to\infty} C_G(\boldsymbol{u})$, for all $\boldsymbol{u}\in[0,1]^d$ x. For brevity, with a little abuse of notation we also denote this latter fact by $C\in\mathscr{D}(C_G)$. The convergence result in (9) is also equivalent to the expansion

$$C(\boldsymbol{u}) = 1-\|\mathbf{1}-\boldsymbol{u}\|_D + o(\|\mathbf{1}-\boldsymbol{u}\|) \tag{12}$$

as $\boldsymbol{u}\to\mathbf{1}\in\mathbb{R}^d$, uniformly for $\boldsymbol{u}\in[0,1]^d$, see Proposition 3.1.5 in [1]. In a first step we drop the term $o(\|\mathbf{1}-\boldsymbol{u}\|)$ in expansion (12) and require that there exists $\boldsymbol{u}_0\in(0,1)^d$, such that

$$C(\boldsymbol{u}) = 1-\|\mathbf{1}-\boldsymbol{u}\|_D, \qquad \boldsymbol{u}\in[\boldsymbol{u}_0,\mathbf{1}]\subset\mathbb{R}^d. \tag{13}$$

A copula, which satisfies the above expansion is a *generalized Pareto copula* (GPC). The df of $n\left(\boldsymbol{M}^{(n)}-\mathbf{1}\right)$ is, for $\boldsymbol{x}<\mathbf{0}\in\mathbb{R}^d$ and n large so that $\mathbf{1}+\boldsymbol{x}/n\ge\boldsymbol{u}_0$,

$$P\left(n\left(\boldsymbol{M}^{(n)}-\mathbf{1}\right)\le\boldsymbol{x}\right) = \left(1-\frac{1}{n}\|\boldsymbol{x}\|_D\right)^n =: F^{(n)}(\boldsymbol{x}).$$

Suppose that the norm $\|\cdot\|_D$ has partial derivatives of order d. Then the df $F^{(n)}(\boldsymbol{x})$ has for $\mathbf{1}+\boldsymbol{x}/n \geq \boldsymbol{u}_0$ the density

$$f^{(n)}(\boldsymbol{x}) := \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F^{(n)}(\boldsymbol{x}) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} \left(1 - \frac{1}{n}\|\boldsymbol{x}\|_D\right)^n. \tag{14}$$

As for the standard multivariate max-stable df G in (11), its density exists and is given by

$$g(\boldsymbol{x}) := \frac{\partial^d}{\partial x_1 \ldots \partial x_d} G(\boldsymbol{x}) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} \exp\left(-\|\boldsymbol{x}\|_D\right), \qquad \boldsymbol{x} \leq \mathbf{0} \in \mathbb{R}^d. \tag{15}$$

We are now ready to state our first multivariate extension of the convergence in total variation in equation (3). For brevity, we occasionally denote with the same letter a Borel measure and its distribution function.

Theorem 1. *Suppose the rv $\boldsymbol{U}$ follows a GPC C with corresponding D-norm $\|\cdot\|_D$, which has partial derivatives of order $d \geq 2$. Then*

$$\sup_{A \in \mathbb{B}^d} \left| P\left(n\left(\boldsymbol{M}^{(n)} - \mathbf{1}\right) \in A\right) - G(A) \right| \to_{n\to\infty} 0,$$

where $\mathbb{B}^d$ denotes the Borel-σ-field in $\mathbb{R}^d$.

Next we extend Theorem 1 to a copula C, which is in a *differentiable neighborhood* of a GPC, defined next. Suppose that C satisfies expansion (12), where the D-norm $\|\cdot\|_D$ on $\mathbb{R}^d$ has partial derivatives of order d. Assume also that C is such that for each nonempty block of indices $B = (i_1, \ldots, i_k)$ of $\{1, \ldots, d\}$,

$$\frac{\partial^k}{\partial x_{i_1}, \ldots, \partial x_{i_k}} n\left(C\left(\mathbf{1} + \frac{\boldsymbol{x}}{n}\right) - 1\right) \to_{n\to\infty} \frac{\partial^k}{\partial x_{i_1}, \ldots, \partial x_{i_k}} \phi(\boldsymbol{x}), \tag{16}$$

holds true for all $\boldsymbol{x} < \mathbf{0} \in \mathbb{R}^d$, where $\phi(\boldsymbol{x}) = -\|\boldsymbol{x}\|_D$.

Theorem 2. *Suppose the copula C satisfies conditions* (12) *and* (16)*. Then we obtain*

$$\sup_{A \in \mathbb{B}^d} \left| P\left(n\left(\boldsymbol{M}^{(n)} - \mathbf{1}\right) \in A\right) - G(A) \right| \to_{n\to\infty} 0,$$

where G is the standard max-stable distribution with corresponding D-norm $\|\cdot\|_D$, i.e., it has df $G(\boldsymbol{x}) = \exp(-\|\boldsymbol{x}\|_D)$, $\boldsymbol{x} \leq \mathbf{0} \in \mathbb{R}^d$.

Example 1. Consider, the *Gumbel-Hougaard family* $\{C_p : p \geq 1\}$ of Archimedean copulas, with generator function $\varphi_p(u) := (-\log(u))^p$, $p \geq 1$. This is an extreme-value family of copulas. In particular, we have

$$C_p(\boldsymbol{u}) = \exp\left(-\left(\sum_{j=1}^d (-\log(u_j))^p\right)^{1/p}\right) = 1 - \|\mathbf{1} - \boldsymbol{u}\|_p + o(\|1 - \boldsymbol{u}\|),$$

as $\boldsymbol{u} \in (0,1]^d$ converges to $\mathbf{1} \in \mathbb{R}^d$, i.e., condition (12) is satisfied, where the D-norm is the logistic norm $\|\cdot\|_p$ and the limiting distribution is $G(\boldsymbol{x}) = \exp(-\|\boldsymbol{x}\|_p)$. The copula C_p also satisfies conditions (16). For the proof see [2].

Example 2. Consider the copula

$$C(\boldsymbol{u}) = 1 - d + \sum_{j=1}^{d} u_j + \sum_{2 \le i \le d} \left((-1)^i \sum_{\substack{B \subseteq \{1,\ldots,d\} \\ |B|=i}} \left(\sum_{j \in B} \frac{1}{1-u_j} - d + 1 \right)^{-1} \right). \quad (17)$$

This provides the d-dimensional version (with $d \ge 2$) of the 2-dimensional copula associated to the df discussed in [5, Example 5.14]. It can be checked that $C \in \mathscr{D}(C_G)$, where C_G is, for all $\boldsymbol{u} \in [0,1]^d$, the extreme-value copula

$$C_G(\boldsymbol{u}) = \exp\left(\sum_{j=1}^{d} \log u_j + \sum_{2 \le i \le d} \left((-1)^{i+1} \sum_{\substack{B \subseteq \{1,\ldots,d\} \\ |B|=i}} \left(\sum_{j \in B} \frac{1}{\log u_j} - d + 1 \right)^{-1} \right) \right).$$

Then, by Proposition 3.1.5 and Corollary 3.1.12 in [1], the copula in (17) satisfies condition (12). See [2] for the details.

Let C be a copula and C^n be the copula of the corresponding componentiwise maxima, see (8). We recall that $C^{(n)}(\boldsymbol{u}) := C^n(\boldsymbol{u}^{1/n})$, $\boldsymbol{u} \in [0,1]^d$. Assume that $C \in \mathscr{D}(C_G)$, where C_G is an extreme-value copula. A readily demonstrable result implied by Theorem 2 is the convergence of $C^{(n)}$ to C_G in variational distance.

Corollary 1. *Assume C satisfies conditions* (12) *and* (16)*, with continuous partial derivatives of order up to d on $(0,1)^d$, then*

$$\sup_{A \in \mathbb{B}^d \cap [0,1]^d} |C^{(n)}(A) - C_G(A)| \to_{n \to \infty} 0.$$

References

1. Falk, M.: Multivariate Extreme Value Theory and D-Norms. Springer, New York, (2019)
2. Falk, M., Padoan S.A., Rizzelli, S.: Strong Convergence of Multivariate Maxima. J. Appl. Probab. **53**, 314–331, (2020)
3. Falk, M., Padoan, S. A. and Wisheckel, F.: Generalized Pareto copulas: a key to multivariate extremes. J. Multivar. Anal. **174**, 104538, (2019)
4. Reiss, R.-D.: Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics, New York, (1989)
5. Resnick, S. I.: Extreme Values, Regular Variation and Point Processes. Springer, New York, (2008)

Data Science:
when different expertise meet

Bayesian stochastic modelling of the temporal evolution of seismicity

Modellazione stocastica Bayesiana dell'evoluzione temporale della sismicità

Elisa Varini and Renata Rotondi

Abstract Seismic hazard assessment is based on the research achievements developed in a lively multidisciplinary context involving seismologists, geologists, and geophysicists as well as computer scientists, mathematicians, statisticians, and other scientists. The statistical approach to seismological issues mainly contributes to deal with uncertainty, including of course data uncertainty. Another important source of uncertainty comes from the lack of a physical model capable of satisfactorily describing the earthquakes process. In this paper we present our most recent research which focuses on stochastic modelling of temporal seismic sequences and earthquakes forecasting at long-middle time scale. In particular we favour the Bayesian approach for data analysis as it is especially suitable for the treatment of uncertainties.

Abstract *La valutazione del pericolosità sismica è frutto di risultati di ricerca sviluppati in un vivace contesto multidisciplinare che coinvolge sismologi, geologi e geofisici, nonché informatici, matematici, statistici e altri studiosi. Il principale contributo dell'approccio statistico alla risoluzione di problematiche sismologiche consiste nella gestione dell'incertezza. Oltre all'incertezza dei dati, un'importante fonte di incertezza deriva dalla mancanza di un modello fisico in grado di descrivere in modo soddisfacente il processo dei terremoti. In questo lavoro presentiamo la nostra ricerca più recente sulla modellazione stocastica di sequenze sismiche temporali e la previsione dei terremoti a medio-lungo termine. In particolare, adottiamo l'approccio bayesiano per l'analisi dei dati in quanto è particolarmente adatto per il trattamento delle incertezze.*

Elisa Varini
CNR, Istituto di Matematica Applicata e Tecnologie Informatiche *Enrico Magenes*, via Bassini 15, 20133 Milano, e-mail: elisa@mi.imati.cnr.it

Renata Rotondi
CNR, Istituto di Matematica Applicata e Tecnologie Informatiche *Enrico Magenes*, via Bassini 15, 20133 Milano e-mail: reni@mi.imati.cnr.it

Key words: Point processes, Failure models, Bayesian inference, Seismic hazard assessment

1 Introduction

A long debate on the intrinsic unpredictability of earthquakes is ongoing, but a scientifically convincing outcome has not yet been achieved based on the limited current knowledge of seismic phenomena. To date the standard approach for assessing seismic hazard and forecasting earthquakes has been necessarily probabilistic in order to properly deal with uncertainty; from this perspective, the stochastic modelling of seismic data should be appropriate to incorporate as much geological and geophysical information as possible.

Historical and instrumental seismic catalogs are definitely important sources of information on earthquakes. What emerges from seismic catalogs all over the world is that the earthquakes process broadly evolves at different spatio-temporal scales, i.e. long-term scale for background seismicity and short-term scale for earthquake clusters (e.g. sequences of aftershocks due to strong earthquakes, seismic swarms). The most part of the stochastic models for earthquake occurrences in the literature targets one of these spatio-temporal scales, in particular self-correcting point processes are used for long-term modelling and self-exciting point processes for short-term modelling. However, a better description of seismicity is certainly expected from stochastic models that overcome the duality introduced by considering separately seismic phenomena at different scales and, for example, are able to model the clustering effects observed at both long and short space-time scales.

In Section 2 we first focus on long-term modelling of earthquake sequences emphasizing the contribution provided by the available geophysical and tectonic knowledge in the stochastic modelling, then in Section 3 we propose a new stochastic model that can operate at different time scales, specifically at long-middle time scale. Applications of both models to the Italian historical seismicity are briefly illustrated.

2 Long-term stochastic modelling of the Italian seismicity

We consider the stress release model (SRM), a well-known self-correcting point process based on the Reid's elastic rebound theory and suitable for long-term seismic hazard assessment ([3] and references therein). According to Reid's theory, some tectonic stress X slowly accumulates in a seismic region until it exceeds the strenght of the medium; then it drops sharply causing an earthquake. In the SRM, stress is assumed to increase linearly with time at a constant loading rate ρ:

$$X(t) = X_0 + \rho t - S(t) \tag{1}$$

where X_0 denotes an initial unknown level of stress and $S(t) = \sum_{t_i<t} X_i$ is the stress release due to all previously occurred earthquakes. SRM is completely defined by its conditional intensity function $\lambda(t \mid \mathscr{H}_t) = exp\{\nu + \beta X(t)\}$, where $\mathscr{H}_t$ is the seismic history up to time t.

The stress release X due to an earthquake is not directly measurable so far; however it can be approximated by several empirical relations involving earthquake magnitude M, which is a measure of the strength of the earthquake based on its effects, e.g. some characteristics of the seismic wave and damages.

In [2] we explored four possible proxy measures of stress release X among those that are evaluable for historical events. The first proxy measure is the Benioff strain $X_B \propto 10^{0.75M}$, which was proposed in the original version of the SRM [4]; it is an approximation of the square root of the energy release. The remaining three proxy measures consider different features of the energy release, such as heat loss, coseismic slip, and damaging potential of the seismic wave. They are: the seismic moment $M_0 \propto 10^{1.5M_w}$, the seismic energy $E \propto X_M^{1.5}/\sqrt{A}$, and the scaled energy $E_S \propto E/M_0$, where A is the area of the fault surface that ruptures during an earthquake. Rupture area A is approximated by using the regressions of Wells and Coppersmith [5].

We proposed four versions of the SRM, where X is respectively defined by the Benioff strain, the seismic moment, the seismic energy, and the scaled energy; each of them was then fitted to the historical earthquakes occurred in the eight Italian macro-regions (MRs) depicted in Fig. 1a (large polygons). Each MR has a tectonically homogeneous behaviour according to the Database of Individual Seismogenic Sources (DISS, version 3.0.2), a large repository of geological, tectonic and active

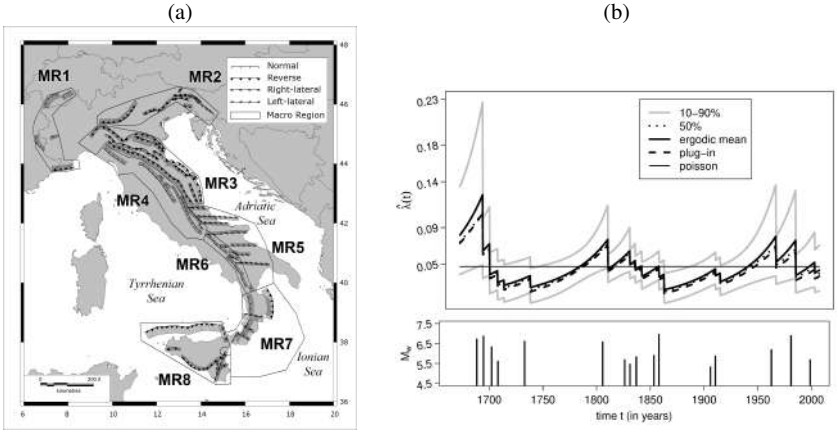


Fig. 1 (a) Map of the Composite Seismogenic Sources from DISS database, version 3.0.2, classified according to the faulting mechanism. Shaded area: vertical projection of the fault plane to the ground surface. The outlined polygons are the identified macro-regions (MRs). (b) Estimated conditional intensity function of the SRM based on scaled energy as for MR4. Ergodic mean (solid line), plug-in (dashed line), and median (dotted line) are practically indistinguishable from each other; 10% and 90% quantiles (gray line). The Poisson rate is shown for comparison (horizontal solid line). The bottom panel shows the magnitude versus the occurrence time of the data.

Italian fault data. In particular, the Composite Seismogenic Sources (CSS) from DISS database were very helpful in delineating the eight MRs, because each CSS identifies an active fault system capable of releasing destructive earthquakes for Italy and their predominant faulting mechanism (small dark polygons in Fig. 1a). Each MR was then associated with a dataset of historical strong earthquakes ($M \geq 5.3$) drawn from the Parametric Catalog of the Italian Earthquakes (CPTI04).

The four versions of the SRM were fitted to each dataset by following the Bayesian paradigm so as to obtain the corresponding posterior distribution of the model parameters [3, 2]; Fig. 1(b) shows the estimated conditional intensity function of the SRM based on the scaled energy as for MR4.

Moreover, we found that the distribution of the waiting time of the next event from any time t follows a Gompertz distribution, in which the shape parameter is history dependent or, in other words, dependent on the current level of stress in the region [3]; this enabled us to perform both retrospective and prospective forecasting tests. According to both Bayes factor and forecasting performance, the best model turns out to be the SRM based on scaled energy.

3 A new multi-scale compound model

We proposed a new stochastic model for earthquake occurrences, hereinafter named compound model, which combines knowledge gained at the different space-time levels [1]. The model assumes the existence of two categories of events, the mainshocks and the subordinate events. The mainshocks are identified among the strongest events occurred in a region and are assumed to follow the SRM based on scaled energy. The subordinate events are especially expected to cluster in the space-time neighborhood of the mainshocks, so as to represent possible premonitory events (foreshocks) and aftershocks; this suggests a bathtub-shaped hazard function for their occurrence times between two consecutive strong earthquakes. Since the hazard functions of some generalized Weibull distributions admit a great variety of

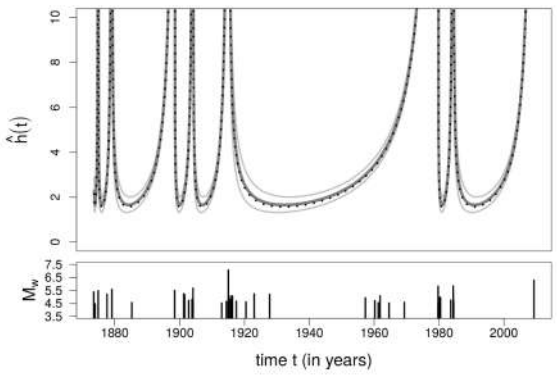


Fig. 2 Compound model: estimated hazard function of the modified Weibull distribution, as ergodic mean (red line), plug-in (dotted line) estimate, and median (black line). These are practically indistinguishable from each other. Also shown: 10% and 90% quantiles (gray line). Magnitude versus occurrence time of the data.

shapes including the bathtub one, we considered two options from this large family: the modified Weibull model and the additive Weibull model.

The compound model was applied to CSS 25 of DISS database, which is part of MR4 in Central Italy (Fig.1a). CSS 25 was associated with a dataset of 58 earthquakes with $M \geq 4.45$ drawn from the CPTI15 catalog; the 9 strongest events of $M \geq 5.3$ were identified as mainshocks through some clustering analysis. Bayesian analysis of the data was performed showing that the best fit is obtained from the compound model based on the modified Weibull distribution, of which the estimated hazard function is illustrated in Fig. 2.

4 Conclusions and remarks

We analysed four different versions of the SRM, based on the Reid's elastic rebound theory and including the contribution of the tectonic information. Under the hypothesis underlying the SRM, the four models differ from one to the other in the definition of a quantity, i.e. the stress, that drives the mechanism of earthquakes occurrence. We found that scaled energy is the best option in SRM. Moreover, we derived the exact distribution of the waiting times to the next event conditioned on the past history of the SRM; it is a Gompertz distribution whose shape parameter is history dependent. Finally, we proposed a new compound model for earthquake occurrences that captures both long- and short-time features of seismic activity by combining in a hierarchical structure the SRM based on scaled energy and a generalized Weibull model that admit bathtub shaped hazard functions.

Acknowledgements This work was partly funded by the Italian Ministry of Education, University and Research (MIUR) in the framework of the PRIN-2015 project 'Complex space-time modeling and functional analysis for probabilistic forecast of seismic events'.

References

1. Rotondi, R., Varini, E.: Failure models driven by a self-correcting point process in earthquake occurrence modeling. Stochastic Environmental Research and Risk Assessment **33**, 709–724 (2019)
2. Varini, E., Rotondi, R., Basili, R., Barba, S.: Stress release model and proxy measures of earthquake size. Application to Italian seismogenic sources. Tectonophysics, **682**, 147–168 (2016)
3. Varini, E., Rotondi, R.: Probability distribution of the waiting time in the stress release model: the Gompertz distribution. Environmental and Ecological Statistics, **22**, 493–511 (2015)
4. Vere-Jones, D.: Earthquake prediction – A statistician's view. Journal of Physics of the Earth, **26**, 129–146. (1978)
5. Wells, D.L., Coppersmith, K.L.: New relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. Bulletin of Seismological Society of America **84**, 974–1002 (1994)

Cluster Analysis for the Characterization of Residential Personal Exposure to ELF Magnetic Field

Analisi Cluster per la Valutazione dell'Esposizione ai Campi Magnetici ELF nei Bambini

Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Isabelle Magne, Martine Souques, Serena Fiocchi, Marta Parazzini and Paolo Ravazzani

Abstract Cluster analysis was applied to characterize indoor personal exposure to Extremely Low Frequency magnetic field (ELF MF) in children with respect to the type and distance of electric networks near the child home. An association effect analysis was then applied to investigate the possible impact of environmental variables, such as heating type, age and size of the home, and family size on the exposure patterns identified with cluster analysis. Cluster analysis identified three different patterns of exposure: the first one was of children with the highest exposure living near overhead lines of high (63-150 kV), extra-high (225 kV) and ultra-high voltage (400 kV); the second one was of children with mid exposures living near underground networks of low (400V) to mid voltage (20 kV) and substations; the last one was of children with lowest level of exposure living far from electric networks. Electric heating, or living in big buildings or in larger families led to higher levels of exposures.

Abstract *L'analisi Cluster è stata utilizzata per l'identificazione dei pattern di esposizione residenziale ai campi magnetici a bassissima frequenza (ELF MF) nei bambini considerando come variabili di interesse il tipo e la distanza delle linee elettriche dalle abitazioni. Successivamente, è stata fatta anche un'analisi di associazione tra i cluster identificati precedentemente e una serie di variabili di tipo ambientale, quali la tipologia di riscaldamento usata nell'abitazione, l'età e la dimensione dell'abitazione e la dimensione del nucleo famigliare. L'analisi cluster ha identificato 3 diversi pattern di esposizione: il primo è caratterizzato da alti*

[1] Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Serena Fiocchi, Marta Parazzini, Paolo Ravazzani, CNR IEIIT – Consiglio Nazionale delle Ricerche, Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Milan, Italy; email: gabriella.tognola@ieiit.cnr.it.

Isabelle Magne and Martine Souques, EDF Electricité de France, Paris, France; email: isabelle.magne@edf.fr.

Marta Bonato is also with DEIB, Politecnico di Milano, Milan, Italy; email: marta.bonato@ieiit.cnr.it

livelli di esposizione ed è costituito da bambini che abitano nelle vicinanze di linee elettriche aeree ad alta (63-150 kV), extra-alta (225 kV) e ultra-alta potenza (400 kV); il secondo cluster è caratterizzato da livelli di esposizione media ed è costituito da bambini che vivono vicino a linee elettriche interrate di bassa (400 V) e media potenza (20 kV) e alle stazioni di trasformazione elettrica; l'ultimo cluster ha i livelli più bassi di esposizione ed è costituito dai bambini che vivono più lontani dalle linee elettriche. Per quanto riguarda l'effetto delle variabili ambientali, i bambini che vivono in edifici con un elevato numero di unità abitative, oppure in abitazioni ove si fa uso di dispositivi di riscaldamento di tipo elettrico ed infine in famiglie numerose sono tendenzialmente esposti ad un livello di campi magnetici maggiore che negli altri casi.

Key words: Cluster Analysis, residential magnetic field exposure, children, environmental variables

1 Introduction

Studies in the first decade of year 2000 (see e.g., Kheifets et al., (2010)) showed that the risk of childhood leukemia increased for a daily average of Extremely Low Frequency Magnetic Fields (ELF MF, 50/60 Hz) exposure greater than 0.4 μT, without evidencing a causal relationship. Since then, several studies and measurement campaigns have been implemented to characterize ELF MF exposure in children in everyday contexts (Bedja et al., (2010); Magne et al., (2017); Schüz et al., (2016)).

In the present study, we aimed to identify recurrent patterns of indoor ELF MF exposure in children with respect to the presence of electric lines near the child home and then assess with an association effect analysis if other 'secondary' variables, such as the use of electric heating appliances, the age and type of the residence, and family size have an impact on the exposure patterns identified with cluster analysis.

2 Materials and Methods

2.1 Materials

The analyzed dataset consisted of 24h indoor measurements of ELF MF (magnetic flux density B) from 884 children living in France. Measurements were performed

with a personal exposimeter (EMDEX II, Enertech, Campbell, CA, USA; sensitivity: 0.01-300 μT; sampling rate 3 s).

The dataset comes from the EXPERS study (Bedja et al., (2010); Magne et al., (2017)): for each child, it reported the number of electric lines near home for 400 V underground cables within 40 m (UNDlow), 20 kV underground cables within 40 m (UNDmid), 63-150 kV underground cables within 20 m (UNDhigh), 225 kV underground cables within 20 m (UNDextra), 400 V overhead lines within 40 m (OVHDlow), 20 kV overhead lines within 40 m (OVHDmid), 150 or 90 or 63 kV overhead lines within 100 m (for 150 kV lines) or 70 m (for 63 and 90 kV lines) (OVHDhigh), 225 kV overhead lines within 120 m (OVHDextra), 400 kV overhead lines within 200 m (OVHDultra), and 20 kV/400 V substations inside the building within 40 m (Substation). The dataset also reported the type of heating appliances used in the child home, the age and type of the residence, and the family size.

2.2 Methods

Data were clustered with the *K*-means algorithm (Matlab, ver. R2018a, MatWorks Inc., Natick MA, USA) (Kaufman and Rousseeuw, (1990)). More details on the procedure we used to perform clustering can be found in (Tognola, Bonato et al., (2019); Tognola, Chiaramello et al., (2019)). The similarity between the observations in a cluster was measured by the squared Euclidean distance. As clustering variables we considered the geometric mean (GM) of the 50 Hz component of B and the number of electric networks close to the child home, according to the classes of electric networks defined above, i.e., for the variables from 'UNDlow' to 'substation'.

We then applied an association test on a number of 'secondary' variables that did not contribute to differentiate the clusters but might have a potential effect on the clusters previously identified. Specifically, we applied the Chi-square (χ^2) test on the categorical variables house heating, residence age, and residence type and the Kruskal-Wallis test on the numerical variable family size. For all tests, we set the significance level at $p<0.05$. The variable "heating type" was discretized into three categories, namely: non-electric, mixed, and electric heating. The variable "residence type" was discretized into 4 categories, namely: individual home, terraced home, apartment in small building (with two to nine residential units), and apartment in big building (with 10 or more residential units). Finally, the variable "residence age" was discretized into four categories, namely: house built before 1950, from 1950 to 1969, from 1970 to1989, and after 1989.

3 Results

The median GM over the whole dataset was 0.010 μT (1st quartile Q1: 0.003 μT; 3rd quartile Q3: 0.028 μT). For only four children, the exposure level (GM value) was >0.4 μT.

The optimal clustering solution was found by partitioning the data with $K = 3$ clusters; this solution had a silhouette score greater than 0.60 (indicative of a cluster solution of good quality). The results of clustering are displayed in Figure 1, which shows the normalized values of the centroids of the three clusters. The centroid is the 'representative' of the observations belonging to a given cluster and it is calculated as the mean of the observations assigned to that cluster for each variable used in clustering.

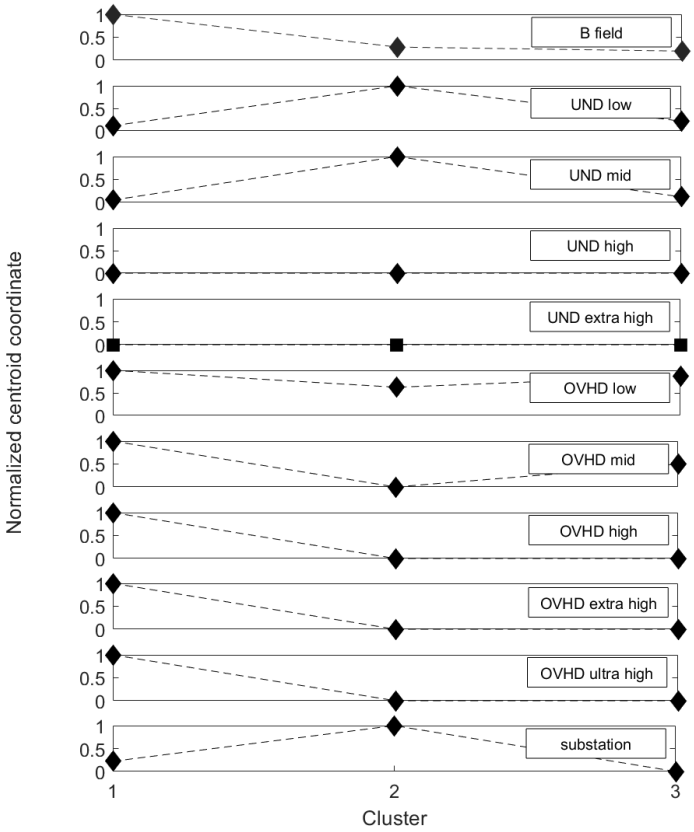


Figure 1: Normalized values of the centroids of the three clusters.

In our analysis each centroid has 11 dimensions, i.e., one for each variable we used to perform the clustering. The centroids in Figure 1 were normalized to the maximum of each of the 11 analyzed variables. The normalization of the centroids allows to better identify which variables were most important in discriminating the clusters. As a matter of fact, it is noted from Figure 1 that the B field, UNDlow,

UNDmid, OVHDmid to OVHD ultra and 'substation' were the variables whose centroids changed more across the clusters, meaning that these latter variables were most important in discriminating cluster characteristics.

As to the clusters, observations were partitioned in three groups: Cluster 1 consisted of children with the highest exposure (average GM: 0.126 μT) living near overhead lines of high (63-150 kV), extra-high (225 kV) and ultra-high voltage (400 kV); Cluster 2 consisted of children with mid exposures (average GM: 0.036 μT) living near underground networks of low (400V) and mid voltage (20 kV) and substations (20kV/400V); Cluster 3 consisted of children with the lowest level of exposure (average GM: 0.025 μT) living far from electric networks.

For what concerns the association effect, we found no statistical significant association of the residence age with the clusters; instead, the type of heating (χ^2: 17.58, p=0.002), the type of residence (χ^2: 138.49 p=2.09x10^{-27}) and the family size (η^2 = 8.63, p<0.02) had a statistical significant association with clusters. For what concerns heating, we found that Cluster 2 (which corresponded to mid levels of indoor exposure) was mainly characterized by children living in homes with electric heating appliances; vice versa, children in Cluster 3 (lowest level of exposure) mainly lived in homes with mixed electric and non-electric heating. As to the residence size, we found that Cluster 3 was characterized by children living in individual and terraced houses, whereas Cluster 2 was mainly characterized by children living in big buildings with more than 10 residential units. Finally, we found that the family size differed significantly across the clusters; in particular, we found that Cluster 2 was characterized by children living in families of bigger size than those in Cluster 3 (Dunn's post-hoc test, p<0.005), whereas no statistically significant differences were found between Cluster 1 and Cluster 2 or 3.

4 Discussion and Conclusions

The current study evidenced how cluster analysis could be helpful in identifying recurrent patterns in personal exposure to indoor ELF MF that were characterized by the type of electric networks close to the child home.

Moreover, this study revealed that secondary environmental variables had an impact on the pattern of ELF MF exposure but of a lesser extent than electric lines. In particular, we found that the use of electric heating appliances or living in big flats or in larger families was generally associated to higher level of personal indoor exposure. This would suggest that to fully characterize ELF MF residential exposure it is important not only to take into account the type and position of electric lines but also the environmental variables such as heating, residence type, family size.

Acknowledgements

This study was supported by ANSES (2015/1/202) Project “ELFSTAT-In depth evaluation of children’s exposure to ELF (40–800 Hz) magnetic fields and implications for health risk of new technologies”. Data come from the EXPERS study database, subsidized by the French Ministry of Health, EDF (Electricite De France) and RTE (Réseau de Transport d'Électricité), and carried out by Supélec, EDF and RTE.

References

1. Bedja, M., Magne, I., Souques, M., Lambrozo, J., Le Brusquet, L., Fleury, G. et al.: Methodology of a study on the French population exposure to 50 Hz magnetic fields. Radiat. Prot. Dosimetry, {142}, 146--152 (2010)
2. Kaufman, L., Rousseeuw, PJ.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., Hoboken (NJ) (1990)
3. Kheifets, L., Ahlbom, A., Crespi, CM., Draper, G., Hagihara, J., Lowenthal, RM., et al.: Pooled analysis of recent studies on magnetic fields and childhood leukemia. Br. J. Cancer, {103}, 1128--1135 (2010)
4. Magne, I., Souques, M., Bureau, I., Duburcq, A., Remy, E., Lambrozo, J.: Exposure of children to extremely low frequency magnetic fields in France: Results of the EXPERS study. J. Expo. Sci. Environ. Epidemiol. {27}, 505--512 (2017)
5. Schüz, J., Dasenbrock, C., Ravazzani, P., Röösli, M., Schär, P., Bounds, PL. et al.: Extremely low-frequency magnetic fields and risk of childhood leukemia: A risk assessment by the ARIMMORA consortium. Bioelectromagnetics, {37}, 183--89 (2016)
6. Tognola, G., Bonato, M., Chiaramello, E., Fiocchi, S., Magne, I., Souques, M, et al.: Use of Machine Learning in the analysis of indoor ELF MF exposure in children. Int. J. Env. Res. Public Health, {16}, 1230--1243 (2019)
7. Tognola, G., Chiaramello, E., Bonato, M., Magne, I., Souques, M., Fiocchi, S. et al.: Cluster Analysis of residential personal exposure to ELF magnetic field in children: effect of environmental variables. Int. J. Env. Res. Public Health, {16}, 4363--4376 (2019)

Statistical Assessment and Validation of Ship Response in High Sea State by Computational Fluid Dynamics

Analisi e Validazione Statistica della Risposta di Navi in Mare Grosso Ottenuta Mediante Fluidodinamica Computazionale

Andrea Serani, Matteo Diez and Frederick Stern

Abstract The study presents the statistical assessment and validation of ship response in heavy weather using the moving block bootstrap method. Computational fluid dynamic results are validated versus experimental fluid dynamics data. The test case is a free-running model of a naval destroyer in heavy weather, namely the 5415M model in irregular stern-quartering waves in sea state 7.

Abstract *Lo studio presenta l'analisi e la validazione statistica della risposta di navi in mare grosso, ottenuta mediante fluidodinamica computazionale. L'approccio per l'analisi e validazione usa un metodo di tipo* moving block bootstrap*. I risultati computazionali sono validati attraverso il confronto con dati sperimentali. Il caso di studio è un modello di cacciatorpediniere, nello specifico il modello 5415M in stato di mare 7.*

Key words: Ship motions, computational fluid dynamics, moving block bootstrap, validation

1 Introduction

The design of air/ground/water-born vehicles requires accurate computational models along with uncertainty quantification and optimization methods for stochastic

Andrea Serani
CNR-INM, National Research Council–Institute of Marine Engineering, Via di Vallerano 139, 00128 Rome, Italy, e-mail: andrea.serani@cnr.it

Matteo Diez
CNR-INM, National Research Council–Institute of Marine Engineering, Via di Vallerano 139, 00128 Rome, Italy, e-mail: matteo.diez@cnr.it

Frederick Stern
The University of Iowa, IIHR–Hydroscience & Engineering, 100 C. Maxwell Stanley Hydraulics Laboratory, Iowa City, Iowa 52242-1585, USA, e-mail: frederick.stern@uiowa.edu

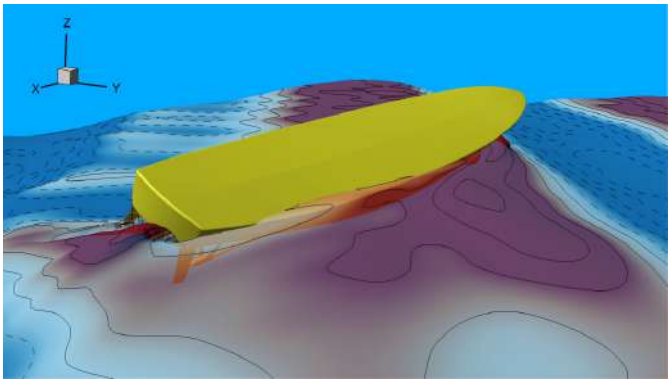


Fig. 1 5415M model in sea state 7 irregular stern-quartering waves by free-running CFD simulations (CFDShip-Iowa).

analysis and optimization in real-world environmental and operating conditions. The validation of stochastic computational responses versus experimental data is an important issue to ensure accurate analysis and informed decision making. In ship hydrodynamics, statistical validation procedures for ship responses were proposed in [2] and applied to computational fluid dynamics (CFD) simulations in irregular waves.

The objective of the current study is to present the extension to high sea states of the statistical validation of stochastic CFD simulations versus experimental fluid dynamics (EFD) data. Namely, free-running CFD simulations of a naval destroyer in heavy-weather conditions are presented. The 5415M model is assessed for course keeping in irregular stern-quartering waves in sea state 7.

The validation is achieved by statistical assessment of EFD data and CFD results for both input waves and ship response time series. Statistical estimators of interest are the expected value, standard deviation, and in general the probability density function. To define validation values and confidence intervals for the statistical estimators a moving block bootstrap (MBB) approach [4] is applied.

The results are part of the activity of the NATO Science and Technology Organization Task Group AVT-280 “Evaluation of prediction methods for ship performance in heavy weather” and have been presented in [7].

2 Test Case Description

The 5415M model is a geosim replica of the DTMB 5415 model with different appendages designed by the Maritime Research Institute of Netherlands (MARIN). EFD data are provided by MARIN and available in [8]. The free-running tests were performed with propeller rate of revolutions fixed to the self-propulsion point of the

model for the envisaged speed. A proportional-derivative controller for the rudder is used taking into account yaw and sway motions.

The selected test case includes seven runs at fixed revolutions per minute RPM = 950 (in model scale), stern-quartering sea (heading of 300 deg) using the JONSWAP spectrum with nominal significant wave height equal to 6 m (full scale), modal period equal to 9.2 s, and 2000 wave components. A total of 130 encounter waves have been recorded, with an average of 20 encounter waves per run. The average run length per run is close to 300 s (in full scale). The model scale data rate is equal to 198.8 Hz.

3 Computational Fluid Dynamics Method and Setup

The unsteady Reynolds-averaged Navier-Stokes equation code CFDShip-Iowa V4.5 [3] is used for CFD computations. Details of equations, numerical implementations, and validation of the numerical solver can be found in [3]. For the current study, absolute inertial earth-fixed coordinates are employed with Menters blended $k-\omega/k-\varepsilon$-BKW) with shear stress transport (SST) using no wall function. The 6 degrees of freedom rigid body equations of motion are solved to calculate linear and angular motions of the ship.

The irregular wave is based on a linear superposition of 80 (regular wave) components. The wave angular frequency values for the wave components are selected evenly distributed within 0.41 and 1.47 rad/s as in the EFD campaign. The phase for each component is randomly selected. Free-running CFD simulations are conducted at Froude number $\mathrm{Fr} = 0.33$ and Reynolds number $\mathrm{Re} = 6.62E+06$ (model scale), with constant propeller RPM. Details on the computational grid and setup can be in found in [7].

4 Analysis and Validation Method

Statistical assessment and validation are studied for time series values of wave elevation, ship motions, rudder angle, ship x- and y-velocities, and immersion probes.

Statistical estimators of interest are the expected value (EV), standard deviation (SD), and in general the probability density function (PDF). EV and SD are evaluated using a sample of M items from time series, $J_i = J(t_i)$, $i = 1,\ldots,M$, as

$$\mathrm{EV}(J) = \frac{1}{M}\sum_{i=1}^{M} J_i \qquad \mathrm{SD}(J) = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left[J_i - \mathrm{EV}(J)\right]^2} \tag{1}$$

The PDF is evaluated via kernel density estimation [5] as

$$\mathrm{PDF}(J,y) = \frac{1}{Mh}\sum_{i=1}^{M}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(y-J_i)^2}{2h^2}\right] \tag{2}$$

where $h = \mathrm{SD}(J)M^{-1/5}$ [6].

To define the validation values and confidence intervals for the statistical estimators of Eqs. 1 and 2 the MBB method [4] is applied to time series, using a number $C = M - l + 1$ moving blocks, each formed by J_i, $i = j,\ldots,j+l-1$, where j is the block index and $l = (2\phi/c)^{2/3}M^{1/3}$ is an optimal block length with

$$\phi = \frac{M\sum_{i=1}^{M-1}\left[J_{i+1}-\mathrm{EV}(J)\right]\left[J_i-\mathrm{EV}(J)\right]}{(M-1)\sum_{i=1}^{M}\left[J_i-\mathrm{EV}(J)\right]^2} \tag{3}$$

and $c = (1-\phi)(1+\phi)$ [1]. From the original set of C blocks, a number of $C' = M/l$ blocks are drawn at random with replacement and concatenated in the order they are picked, forming a new bootstrapped series of size M. $B = 100$ bootstrapped series are used. The validation value for EV and its 95% confidence lower and upper bounds are evaluated as

$$\mathrm{EV} = \mathrm{Median}(\mathrm{EV}_b) = \mathrm{EV}_{\lceil 0.5B\rceil} \quad \mathrm{EV}_l = \mathrm{EV}_{\lceil 0.025B\rceil} \quad \mathrm{EV}_u = \mathrm{EV}_{\lceil 0.975B\rceil} \tag{4}$$

where EV_b represents the EV value of the b-th bootstrapped series, ordered such as $\mathrm{EV}_{b-1} \le \mathrm{EV}_b \le \mathrm{EV}_{b+1}$. Finally, the uncertainty of the estimate is $U_{\mathrm{EV}} = 0.5(\mathrm{EV}_u - \mathrm{EV}_l)$. The validation values and confidence interval for SD are evaluated similarly.

The validation error for EV is evaluated as $E = \mathrm{EV}^{(\mathrm{CFD})} - \mathrm{EV}^{(\mathrm{EFD})}$ and the associated validation uncertainty is defined as

$$U_v = \sqrt{\left[U_{\mathrm{EV}}^{(\mathrm{CFD})}\right]^2 + \left[U_{\mathrm{EV}}^{(\mathrm{EFD})}\right]^2}. \tag{5}$$

Validation is achieved if $|E| \le U_v$. SD validation error and uncertainty are assessed similarly.

5 Results

The statistical validation results are presented for the input wave elevation, ship motions, rudder angle, x- and y-velocities, and two immersion probes. Errors and uncertainties for the statistical estimators provided by the MBB method are normalized with 2SD.

Table 1 shows the input wave validation. EFD and CFD uncertainties of the estimators are close to 4.4 and 9%, respectively. A reduction of the uncertainty would require both EFD and CFD to run longer. Nevertheless, validation is achieved for wave elevation SD (about 5.2% error). Finally, Figure 2a shows the PDF of wave

Table 1 Validation of time series statistics by MBB.

Variable	Units	EV						SD					
		Value			Uncertainty (U)			Value			Uncertainty (U)		
		EFD	CFD	$\lvert E \rvert$	EFD	CFD	U_v	EFD	CFD	$\lvert E \rvert$	EFD	CFD	U_v
Wave elevation	m	0.21	0.00	5.74	1.12	2.54	2.54	1.83	1.64	5.19	4.38	9.06	9.23
Surge	m	5.17	3.56	3.76	8.66	10.86	10.41	21.42	11.40	23.39	6.55	12.04	9.16
Sway	m	5.11	6.23	6.97	17.47	18.52	28.06	8.04	9.53	9.27	9.04	22.20	27.82
Heave	m	-0.24	-0.20	2.17	2.39	2.22	3.25	0.92	0.91	0.54	6.69	9.64	11.65
Roll	deg	6.48	3.02	8.50	2.64	2.33	3.49	20.36	20.01	0.86	5.04	1.97	5.40
Pitch	deg	-0.04	-0.06	0.81	0.99	0.85	1.31	1.23	1.25	0.81	7.00	7.99	10.72
Yaw	deg	1.40	1.03	11.56	6.95	12.51	17.61	1.60	2.07	14.69	4.45	16.00	21.17
Motion avg.				5.63			10.69			8.26			14.32
x-velocity	kn	23.34	23.13	7.34	16.40	9.18	17.73	1.43	1.05	13.29	6.02	6.74	7.79
y-velocity	kn	-0.05	0.04	5.06	2.47	3.66	4.90	0.89	1.03	7.87	3.48	8.55	10.49
Velocity avg.				6.20			11.32			10.58			9.14
Rudder angle	deg	-2.71	4.51	32.49	2.97	7.44	8.58	11.11	12.02	4.10	9.39	7.35	12.30
Immersion probe 3	m	2.78	3.32	17.31	10.76	7.13	13.66	1.56	1.84	8.97	2.44	2.01	3.40
Immersion probe 5	m	3.72	3.72	9.36	12.35	5.22	12.93	2.35	1.72	13.40	6.81	1.75	6.93
Immersion probe avg.				13.33			13.29			11.19			5.17
Average				9.26			10.37			8.53			11.34

Note: E and U are %2SD.

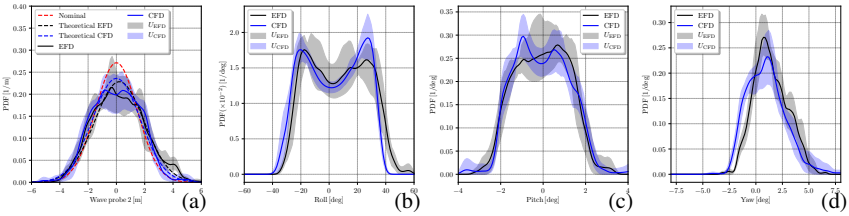


Fig. 2 Comparison of EFD and CFD probability density functions for selected variables.

elevation, using MBB. Both EFD and CFD values are shown along with nominal and theoretical distribution, revealing a good agreement.

Results for ship motions are shown in Fig. 2 and Tab. 1. Specifically, Figs. 2 compare EFD and CFD PDFs of significant ship-motion variables. The results are in good agreement. Ship motions SD is validated on average with an error of 8.3% and an associated uncertainty of 14.3%. Ship velocity EV is validated with an error equal to 6.2% and an uncertainty equal to 11.3%. Rudder angle SD is also validated with error and uncertainty equal to 4.1 and 12.3% respectively. Immersion probes EV and SD show an average error close to 12% with an associated uncertainty close to 9%, likely due to the use of the level set method (in CFD) in combination with rough waves and extreme motions. Variables are validated with an average error of 9.6% and 8.8% for EV and SD, respectively, and the corresponding uncertainties are equal to 11.1% and 11.5%. It is worth noting hot EFD distribution for the roll angle shows an evident bi-modal shape, which is very well captured by CFD.

6 Conclusions

A statistical validation has been presented and assessed for the 5415M model in (sea state 7) irregular stern-quartering waves by free-running CFD simulations. Stochastic input wave and ship response validation has been investigated by statistical assessment of EFD data and CFD results via MBB analysis of time series. On average, validation has been achieved with 9.3 and 8.5% validation errors, along with 10.4 and 11.3% validation uncertainties, for time series EV and SD, respectively. A reduction of the validation uncertainty is desired and require both EFD and CFD to run longer. Current results show that CFD is in a very good agreement with EFD.

Acknowledgments

The University of Iowa (UI) is supported by the Korea Institute of Science and Technology under grant #19026000. CNR-INM is supported by UI under a subaward to the same grant. The ONR grants N00014-17-1-2083 and N00014-17-1-2084 under the administration of Drs. Thomas Fu, Woei-Min Lin and Ki-Han Kim partially sponsored UI research. Dr. Serani is grateful to CNR for support through Short-Term Mobility Program 2018. The research was performed within NATO STO Task Group AVT-280. The availability of experimental data from Cooperative Research Navies is greatly acknowledged. The authors wish to thank Dr. Frans van Walree (MARIN) for the very fruitful discussions and helping with the experimental data.

References

1. E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The annals of statistics*, 14(3):1171–1179, 1986.
2. M. Diez, R. Broglia, D. Durante, A. Olivieri, E. F. Campana, and F. Stern. Statistical assessment and validation of experimental and computational ship response in irregular waves. *Journal of Verification, Validation and Uncertainty Quantification*, 3(2), 2018.
3. J. Huang, P. M. Carrica, and F. Stern. Semi-coupled air/water immersed boundary approach for curvilinear dynamic overset grids with application to ship hydrodynamics. *International Journal for Numerical Methods in Fluids*, 58(6):591–624, 2008.
4. H. R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241, 1989.
5. J. C. Miecznikowski, D. Wang, and A. Hutson. Bootstrap MISE estimators to obtain bandwidth for kernel density estimation. *Communications in Statistics-Simulation and Computation*, 39(7):1455–1469, 2010.
6. B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.
7. F. van Walree, A. Serani, M. Diez, and F. Stern. Prediction of heavy weather seakeeping of a destroyer hull form by means of time domain panel and CFD codes. In *Proceedings of the 33nd Symposium on Naval Hydrodynamics, Osaka, Japan*, 2020.
8. F. van Walree and C. Visser. Analysis and post processing of measured signals vol. 2: Seakeeping water on deck tests. Technical Report 22810-1-SMB, Maritime Research Institute Netherlands, November 2010.

Uncertainty Quantification for PDEs with random data using the Multi-Index Stochastic Collocation method

Quantificazione dell'Incertezza per EDP con dati aleatori tramite il metodo "Multi-Index Stochastic Collocation"

Lorenzo Tamellini and Joakim Beck

Abstract In many real-life applications, one needs to solve partial differential equations (PDEs) to predict the behavior of a system, most often by numerical methods. This goal is often hampered by the fact that the parameters of the equations might be not known exactly, and modeled as random variables; one therefore would like to assess how this uncertainty propagates to the solution of the PDE. To this end, we discuss in this contribution the Multi-Index Stochastic Collocation method, and show its effectiveness with a few numerical tests.

Abstract *In molti contesti applicativi, è necessario risolvere delle equazioni alle derivate parziali (PDE) per ottenere una predizione del comportamento di un sistema, il più delle volte tramite tecniche numeriche. Questo obiettivo è spesso reso difficoltoso dal fatto che i parametri dell'equazione potrebbero non essere noti con esattezza, e descritti comoe variabili aleatorie; di conseguenza, è interessante stimare come questa incertezza si propaga alla soluzione della PDE. A questo scopo, in questo contributo discutiamo il metodo "Multi-Index Stochastic Collocation", e mostriamo la sua efficacia tramite alcuni esempi numerici.*

Key words: Uncertainty Quantification, Partial Differential Equation with random data, Multi-Index Stochastic Collocation Method

Lorenzo Tamellini
Consiglio Nazionale delle Ricerche – Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes" (CNR-IMATI), via Ferrata 1, 27100 Pavia (PV) Italy, e-mail: tamellini@imati.cnr.it

Joakim Beck
King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia e-mail: joakim.beck@kaust.edu.sa

1 Introduction

In many applications, scientists and engineers need to solve ordinary / partial differential equations (ODEs / PDEs respectively) to predict the behavior of a system, either analytically or more commonly by computer simulations (in the remainder of this contribution we focus only on PDEs). While the mathematical structure of these PDEs is often well-understood (at least for classical applications, such as solid mechanics, fluid dynamics, heat exchange, electromagnetism, etc) and numerous well-studied numerical methods exist to approximate their solutions, the predictive power of these simulations is often hampered by the fact that the parameters of the equations might be not known exactly, and they might be considered e.g. as random variables (or random fields). "Parameters" here is used in a broad sense, and includes physical/chemical coefficients (e.g., density, viscosity, thermal conductivity, permeability), shape of the domain, initial conditions, boundary conditions, which can be thoughts as the "inputs" or "data" of the PDE, as opposed to the solution of the PDE, which can be thought as the "output" of the PDE. This uncertainty can be due to multiple reasons, such as limited experimental measurements, or intrinsic randomness of the quantity (e.g. rainfall, earthquakes).

It is therefore crucial to assess how the uncertainty on the data/inputs of the ODEs/PDEs propagates to the outputs of these equations, i.e., to compute mean, variance and higher moments of the solution, and ideally its probability density function. This is the goal of the so-called forward Uncertainty Quantification (UQ) analyses. Other related tasks are inverse UQ and Optimization Under Uncertainty): they can be tackled with similar approaches, but we do not discuss them here. In this context, it is useful to think of the solution of the PDE as a random function, that associates to each realization of the random parameters the corresponding solution of the PDE. In the following, assuming that the PDE depends on N random parameters: we denote by $\boldsymbol{y} \in \mathbb{R}^N$ the vector containing the realizations of the random parameters, by $\Gamma \subset \mathbb{R}^N$ the set in which $\boldsymbol{y}$ can take values (the so-called "parameter space"), and by $u(\boldsymbol{x}, \boldsymbol{y})$ the solution of the PDE, where we have also highlighted the dependence of u on the space/time variables $\boldsymbol{x}$.

Most methods for forward UQ rely on solving the PDE for multiple realizations of the random parameters (i.e., sampling $u(\boldsymbol{y})$ over Γ) and then post-processing the corresponding solutions to obtain the desidered statistical information on the solution. The most trivial of these methods is of course the Monte Carlo method, whereby the statistical information for the solution is obtained by simply averaging the samples of $u(\boldsymbol{y})$ obtained. It is of course possible to replace Monte-Carlo with more effective sampling methods, such as Quasi Monte Carlo or Latin Hypercube Sampling, see e.g. [7]. These methods however do not take full advantage of the possible smoothness of the function to be sampled, i.e. the fact that the map $\boldsymbol{y} \to u(\boldsymbol{x}, \boldsymbol{y})$ could be not only a continuous function but actually a function whose derivatives up to a certain degree might be continuous or square-integrable – even an analytic function at times. These properties derive from the structure of the PDE at hand. This fact can instead been exploited by methods that are more traditional of the numerical analysis background, i.e., numerical quadrature and interpolation methods

(stemming from the fact that computing an expected value / higher moment is nothing but a weighted integral over the parameter space Γ). On the other hand, classical numerical quadrature/interpolation methods scale poorly with the dimension N of the parameter space – in the worst case, the number of required samples grows exponentially with N ("curse of dimensionality"). In addition to this, the function $u(\boldsymbol{y})$ to be sampled is typically expensive to evaluate (since it requires solving numerically a PDE), so that naive approaches become unbearably expensive. A very popular and quite effective approach to reduce the "curse of dimensionality" effect are the so-called "sparse-grids" schemes, i.e., quadrature/interpolation schemes carefully designed to deal with high-dimensional, possibly smooth functions, see e.g. [9, 10].

Another very popular approach to reduce the computational complexity is the so-called multi-level approach, first proposed in the context of numerical finance and then applied to engineering applications [3, 2]. In this approach, a first sampling of the parameters space is performed by solving the required PDEs with a coarse computational mesh (hence, with a cheap method); then, a refinement of the mesh is introduced and a few additional PDEs are solved, to correct the previous estimate. The procedure can be iterated on a hierarchy of increasingly refined meshes where at each level less and less PDEs are solved. In this way, the computational cost of the procedure is substantially reduced without compromising the accuracy of the prediction. In this version (and in the several improvements proposed in literature, see e.g. [5, 8]) the sampling strategy at each level is still a Monte-Carlo / Quasi-Monte-Carlo strategy, i.e., strategies that again do not fully exploit the regularity of the solution u. This can be reached instead combining the multi-level strategy with the sparse-grids schemes mentioned above, [11, 4, 1]. In this contribution we focus on one specific example of such method, the Multi-Index Stochastic Collocation method. In particular, our exposition follows closely [1].

2 Multi-index Stochastic Collocation (MISC)

To fix the ideas, let's consider a specific PDE with random coefficients for which we want to perform a forward UQ analysis. For instance, let's consider the heat equation to compute the pointwise temperature of a metal bar being heated, whose heat-conduction coefficient is uncertain. In this scenario, the parameter $\boldsymbol{y}$ represents the heat conduction coefficient: one parameter $N = 1$ is enough if the bar is homogeneous, while if the bar is composite with pieces with different material, then we will need $N > 1$. Each $y_n, n = 1, \dots N$ can be thought as a uniform random variable over a certain range $y_n \sim \mathscr{U}(a_n, b_n)$, and we can assume y_n to be independent. The parameter space Γ is the hypercube obtained by taking Cartesian products of each $[a_n, b_n]$, and the probability density function of $\boldsymbol{y}$ is simply $\rho(\boldsymbol{y}) = \prod_{n=1}^{N} \frac{1}{b_n - a_n}$. The solution u of the PDE is the pointwise temperature in the bar, and is a function of the spatial coordinate $\boldsymbol{x}$ as well as of the random heat-conductivities $\boldsymbol{y}$, i.e. $u = u(\boldsymbol{x}, \boldsymbol{y})$.

Let us now consider an exahedral mesh to approximate u for a given value of the heat coefficient, by solving the PDE. For simplicity, let us assume that all the

elements have the same size and are allowed to be non-cubic, i.e., their edges have size $h_1 = c_1 2^{-\alpha_1}$, $h_2 = c_2 2^{-\alpha_2}$, $h_3 = c_3 2^{-\alpha_3}$, for some constants c_1, c_2, c_3 and user-defined integer values $\alpha_1, \alpha_2, \alpha_3$. We collect the three values of α_i in a multi-index $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]$; prescribing the multi-index $\boldsymbol{\alpha}$ thus prescribes the computational mesh to be generated. If this flexibility is not allowed by the mesh-generator (or by the problem itself), one can set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$, i.e., controlling the mesh-generation by a single integer value α. Let us denote by $G_{\boldsymbol{\alpha}}$ the quantity of interest of our UQ analysis computed over the mesh specified by $\boldsymbol{\alpha}$; this could be for instance the value of the temperature in a specific point of the bar being heated. Thus, the final goal of the UQ analysis is to compute an approximation of e.g. $\mathbb{E}[G_{\boldsymbol{\alpha}}]$, i.e., of the expected value of $G_{\boldsymbol{\alpha}}$.

The MISC method can be used for this goal. It is based on selecting the values $\boldsymbol{y}_j$ as points of a Cartesian grid obtained by tensorization of univariate quadrature rules (which should be chosen according to $\rho(\boldsymbol{y})$ for computational efficiency). In this work, we use as univariate quadrature rule the Clenshaw–Curtis (CC) univariate quadrature, which is optimal when $y_1, y_2, \ldots, y_N$ are uniform and independent random variables as in our case. Then, for a multi-index $\boldsymbol{\beta} \in \mathbb{N}^N$ and given the function $m(i)$ with $m(0) = 0$, $m(1) = 1$, $m(i) = 2^{i-1} + 1$ for $i \geq 2$, $m(\beta_1)$ CC values are generated for y_1, $m(\beta_2)$ CC values for y_2 etc, and then consider the grid obtained by taking the Cartesian product of the N sets of points thus generated. The quadrature weight of each point of the Cartesian grid is immediately obtained by taking the product of the corresponding univariate weights. The approximation of $\mathbb{E}[G_{\boldsymbol{\alpha}}]$ computed over this grid is denoted as $\mathscr{Q}_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. Clearly, one would like to have both multi-indices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with large components, say $\boldsymbol{\alpha} = \boldsymbol{\alpha}^\star$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$, i.e., to average the values of many PDE solutions over a refined computational mesh. However, as we already mentioned this is typically unfeasible due to computational costs.

Instead, the idea of MISC resorts to the previously mentioned multi-level approach, i.e. the single, highly refined approximation $\mathscr{Q}_{\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star}$ is replaced by a linear combination of many coarser $\mathscr{Q}_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, where whenever one refines the spatial discretization $\boldsymbol{\alpha}$, the quadrature level $\boldsymbol{\beta}$ is kept to a minimum and vice versa (of course, the combined cost of computing the set of coarse discretizations should be smaller than the cost of the highly refined one). In formula,

$$\mathbb{E}[G_{\boldsymbol{\alpha}^\star}] \approx \mathscr{Q}_{\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star} \approx \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}] \in \mathscr{I}} c_{[\boldsymbol{\alpha},\boldsymbol{\beta}]} \mathscr{Q}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \tag{1}$$

where $c_{[\boldsymbol{\alpha},\boldsymbol{\beta}]}$ is a real number and $\mathscr{I}$ is a collection of feasible discretizations, designed with the purpose just explained. For instance,

$$\mathscr{I} = \{\boldsymbol{\alpha} \in \mathbb{R}^3, \boldsymbol{\beta} \in \mathbb{R}^N : |\boldsymbol{\alpha}| + |\boldsymbol{\beta}| \leq L\}$$

for some integer value L. A suitable set $\mathscr{I}$ can be designed can either be designed a-priori, by a careful analysis of the PDE at hand, see e.g. [1], or on-the-run by adaptive algorithms, see e.g. [6]; in this contribution the focus is on the former option.

3 Numerical results

To show the effectiveness of MISC, we briefly comment in this section on a set of results originally reported in [1], to which we refer the interested reader for details. The metal bar is shown in Figure 1 top-left. The pointwise heat-conductivity is modeled as a random field, parametrizes with $N = 3$ i.i.d. uniform random variables, $y_i \sim \mathscr{U}(-1,1)$, i.e., $\Gamma = [-1,1]^3$; one possible realization of this random field is shown in Figure 1 top-right. The quantity of interest (the quantity $G_{\boldsymbol{\alpha}}$) is the integral of the temperature over the metal bar, and we aim at computing its expected value, $\mathbb{E}[G_{\boldsymbol{\alpha}}]$.

Figure 1 bottom-left reports the growth of the computational time required to approximate $\mathbb{E}[G_{\boldsymbol{\alpha}}]$ as we require a smaller and smaller tolerance. The methods considered are MISC, the standard Multi-Level Monte-Carlo method (MLMC) and a certain variant of MLMC, called Multi-Index Monte Carlo (MIMC). We can immediately see that the computational time required by MISC grows significantly slower than for MLMC and MIMC, especially for small tolerances. We also report in dotted lines the theoretical growth of the computational times for all these methods (as well as standard Monte-Carlo), which show good agreement between the theory and the actual computational times.

Finally, Figure 1 bottom-right shows for the number of PDEs solved on each mesh, for some of the tolerances required to obtain the tolerance-vs-time plot. The horizontal axis reports the "total mesh refinement", i.e. the sum $\alpha_1 + \alpha_2 + \alpha_3$: the larger this sum, the more refined the computational mesh, and the more expensive solving the PDE over it. Tolerances are named $TOL_1, TOL_2, \ldots, TOL_5$, with TOL_1 being the largest and TOL_5 the smallest. As expected, if a large tolerance is required, only a few PDE solves are needed, on a few meshes, none of them very refined. As the required tolerance gets smaller, the number of meshes required increases, and more refined meshes are introduced; at the same time, more and more PDEs are being solved on each mesh. Crucially, however, the number of PDEs that need to be solved on the refined meshes is always significantly smaller than those to be solved on coarse meshes: e.g., for TOL_5 we are solving a few thousand PDEs on meshes with total refinement 3 and slightly more than a hundred on the meshes with total refinement 8, which makes MISC a quite effective method for forward UQ purposes.

References

1. Beck, J., Tamellini, L., Tempone, R.: IGA-based Multi-Index Stochastic Collocation for random PDEs on arbitrary domains. Computer Methods in Applied Mechanics and Engineering **351**, 330 – 350 (2019)
2. Cliffe, K., Giles, M., Scheichl, R., Teckentrup, A.: Multilevel monte carlo methods and applications to elliptic pdes with random coefficients. Computing and Visualization in Science **14**(1), 3–15 (2011)
3. Giles, M.B.: Multilevel monte carlo path simulation. Operations Research **56**(3), 607–617 (2008)

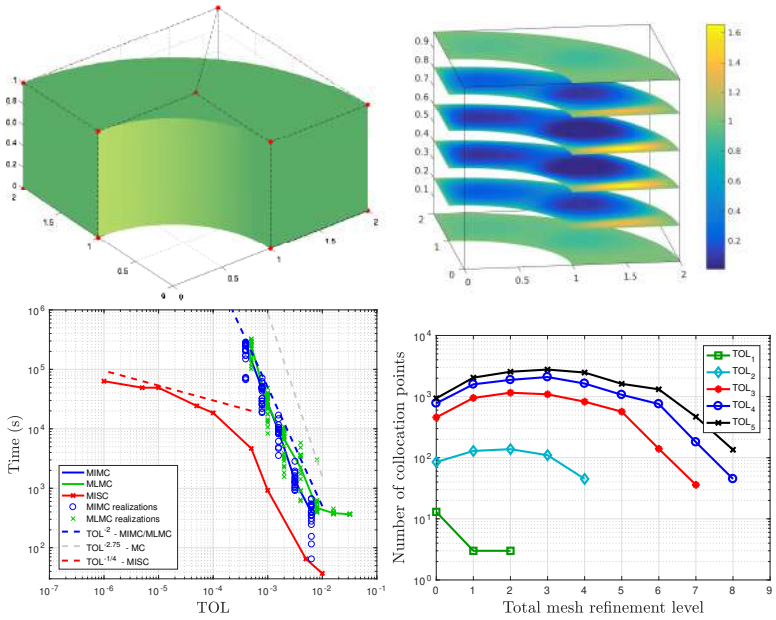


Fig. 1 Test 1 (plots already appeared in [1]). Metal bar (top-left); one realizations of the random heat-capacity (top-right); Tolerance-vs-time plot for MISC and other methods (bottom-left); number of PDEs solved on each grid for each tolerance (bottom-right).

4. Haji-Ali, A., Nobile, F., Tamellini, L., Tempone, R.: Multi-index stochastic collocation for random {PDEs}. Computer Methods in Applied Mechanics and Engineering **306**, 95 – 122 (2016)
5. Haji-Ali, A.L., Nobile, F., Tempone, R.: Multi-index Monte Carlo: when sparsity meets sampling. Numerische Mathematik pp. 1–40 (2015)
6. Jakeman, J.D., Eldred, M., Geraci, G., Gorodetsky, A.: Adaptive multi-index collocation for uncertainty quantification and sensitivity analysis (2019)
7. Kuo, F.Y., Nuyens, D.: Application of quasi-monte carlo methods to elliptic pdes with random diffusion coefficients: A survey of analysis and implementation. Foundations of Computational Mathematics **16**(6), 1631–1696 (2016)
8. Kuo, F.Y., Schwab, C., Sloan, I.: Multi-level Quasi-Monte Carlo Finite Element Methods for a Class of Elliptic PDEs with Random Coefficients. Foundations of Computational Mathematics **15**(2), 411–449 (2015)
9. Nobile, F., Tamellini, L., Tempone, R.: Convergence of quasi-optimal sparse-grid approximation of Hilbert-space-valued functions: application to random elliptic PDEs. Numerische Mathematik **134**(2), 343–388 (2016)
10. Schillings, C., Schwab, C.: Sparse, adaptive Smolyak quadratures for Bayesian inverse problems. Inverse Problems **29**(6) (2013)
11. Teckentrup, A.L., Jantsch, P., Webster, C.G., Gunzburger, M.: A Multilevel Stochastic Collocation Method for Partial Differential Equations with Random Input Data. SIAM/ASA Journal on Uncertainty Quantification **3**(1), 1046–1074 (2015)

Emerging challenges in official statistics: new data sources and methods

Small area poverty indicators adjusted using local spatial price indices

Indicatori di povertà per piccole aree corretti con indici spaziali dei prezzi a livello locale

Stefano Marchetti, Luigi Biggeri, Caterina Giusti and Monica Pratesi

Abstract In this work we focus on estimating monetary poverty indicators at provincial level in Italy taking into account the different price levels within the country. Indeed, the historical Italian North-South divide also correspond to a strong differential in price levels, which can affect the poverty threshold. The local price levels are accounted using two sub-regional spatial price indices: 1. the indices based on retail volumes (units) and prices for food and grocery, and 2. the indices based on housing costs collected through the Household Budget Survey. Sub-regional poverty estimates are then computed using area level small area models, which link direct unreliable estimates to aggregated auxiliary information.

Abstract *In questo lavoro si propone un metodo per ottenere stime di povertà relativa in Italia a livello provinciale, considerando il diverso livello dei prezzi presente tra il nord ed il sud del paese. La linea di povertà nazionale viene opportunamente corretta utilizzando due indici dei prezzi: 1. indice spaziale dei prezzi per alimenti, calcolato utilizzando big data delle vendite al dettaglio nelle catene della grande distribuzione e 2. indice spaziale dei prezzi per l'affitto calcolato utilizzando i dati l'indagine sulle spese delle famiglie. Per la stima della povertà relativa a livello provinciale si utilizzano metodi di stima per piccole aree in quanto le stime tradizionali ("dirette") non risultano affidabili.*

Key words: poverty mapping, retail scanner data, small area estimation.

Stefano Marchetti, University of Pisa and ASESD (Advanced Studies on Equitable and Sustainable Development) Centre "C. Dagum"; email: stefano.marchetti@unipi.it

Luigi Biggeri, University of Florence and ASESD Centre "C. Dagum"; email: biggeri@disia.unifi.it

Caterina Giusti, University of Pisa and Centre "C. Dagum"; email: caterina.giusti@unipi.it

Monica Pratesi, University of Pisa and Centre "C. Dagum"; email: monica.pratesi@unipi.it

Acknowledgement This work has been supported by the H2020 project MAKSWELL G.A. 770643

1 Introduction

Local poverty indicators are essentials for a data driven welfare policy. Among the possible measures, we focus on monetary poverty indicators, and in particular on the Head Count Ratio (HCR), which is the proportion of poor households in a given area or domain.

In Italy, Istat, the National Statistical Office, publishes HCRs estimated from the Household Budget Survey (HBS), which provides good estimates at the national and regional level However, sub-regional estimates, such as provincial estimates, can help the decision makers to plan and implement local welfare policies. Usually, such local estimates are unreliable because of the small sample size of the HBS at local level. For example, for the HBS 2017 the national sample was composed by 16496 households, with provincial sample sizes ranging from 20 to 1036, with a median value of 125. Small area estimation methods (Rao and Molina, 2015) can allow for reliable estimates at local level.

When the focus of the HCR are regions, provinces or municipalities, the national (monetary) poverty threshold can be misleading because the price levels within the country are not equal, as noted by Biggeri et al. (2018). In this work we propose to adjust the relative national poverty line (NPL) at the province level taking into account the spatial distribution of two indexes: the Spatial Food Price Index (SFPI), limited to the prices of food and grocery items as collected by the retail trade scanner data on prices processed by Istat, and the Spatial Housing Cost Index (SHCI), limited to the housing rents as collected by the Italian Household Budget Survey.

To obtain estimates of HCRs at provincial level adjusted taking into account different price levels, we first compute spatial price indexes that are used together to adjust the national poverty line at provincial level, and we finally estimate the provincial HCRs using area-level model-based estimators.

2 Estimation of Spatial Food Price Indexes (SFPI)

The data set used is a selection of price quotations of 73,000 different products (barcodes or GTINs – Global Trade Item Numbers) in food and beverages categories, excluding fresh food, classified in 55 ECOICOP-5-digits, and turning out into about 1,300,000 annual quotations in 2017. In the following they are indicated as the scanner data of grocery products[1].

Retail scanner data on prices are the result of an agreement started in 2013 between ISTAT and the associations of modern distribution, retail trade chains (RTCs) and Nielsen. In 2017 the price scanner data used by Istat were limited to the quotations of a sample of outlets selected by a probability sampling design. Scanner data for 1,781

[1] Retail scanner data were provided by Istat under an agreement with the Centre ASESD "Camilo Dagum" for the research in the framework of the H2020 MAKSWELL project www.makswell.eu

of the main 16 retail trade chains (RTCs) covering the entire national territory were monthly collected by ISTAT on a weekly basis at item code level. Outlets have been stratified according to provinces (107), chains (16) and outlet-types (hypermarket, supermarket) for a total of 867 strata. Probabilities of selection were assigned to each outlet based on the corresponding turnover value. Concerning the selection of the sample of items, a cut off sample of barcodes (GTINs) has been selected within each outlet/aggregate of products (covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). For each GTIN, prices were calculated taking into account turnover and quantities: weekly price is equal to the weekly turnover divided by weekly quantities. Monthly and annual prices are calculated by the arithmetic mean of weekly prices weighted with quantities.
The spatial food price indexes (SFPI) have been computed by using the scanner data and applying the Country Product Dummy (CPD) model according to Laureti and Rao (2018). The CPD is essentially a hedonic regression model where the observed price is related to the country of origin and the characteristics of the product are represented by the commodity itself.

The CPD model can be seen as a simple fixed effects model where country-effects provide estimates of the spatial price index and commodity-specific effects provide estimates of between areas prices. In our work we specified the CPD as follows:

$$\ln p_{jr} = \alpha_0 + \sum_{r=1}^{R} \gamma_r D_r + \sum_{j=1}^{J} \eta_j A_j + \epsilon_{jr} \qquad (1)$$

where p_{jr} is the price of the commodity j in area (province) r, D_r is the same as in equation (1), γ_r is the area r price coefficient, A_js are commodity dummy variables, η_j account for the quality of commodity j and ϵ_{jr} is a normal distributed error with constant variance. As usual, by imposing the constraint $\gamma_1 = 0$, then γ_r is the difference of (fixed) effects connected with the area r compared with the base area 1. To use as a reference Italy instead of area 1, the coefficient γ_r has been adjusted following Suits (1984). In this way, γ_r represent the fixed effect of area r compared to Italy. Thus, the quantity $\exp(\gamma_r)$ represents the spatial price index for food in area r ($SFPI_r$) with respect to Italy, and it is also called purchasing power parity of area r (PPP_r). The SFPI at provincial level are summarized in Table 1 grouped by geographic repartition (North, Centre, South).

Table 1: Distribution of province SFPI grouped by geographical repartition

Repart.	***Min***	***1st Q.***	***Median***	***Mean***	***3rd Q.***	***Max***
North	0.9116	1.0027	1.0365	1.0358	1.0799	1.1216
Centre	0.9339	0.9845	1.0039	1.0025	1.0178	1.0888
South	0.7925	0.9057	0.9423	0.9554	1.0002	1.1350

From Table 1 we can see that food prices in the retail trade chains are higher in the northern and central provinces with respect to the southern ones, as expected (Italy is again the reference).

3 Estimation of Spatial Housing Price Indexes (SHPI)

To estimate the Spatial Housing Price Indexes (SHPI) we used the data coming from the HBS conducted yearly by Istat. Data are collected on the basis of a two-stage sample design where the first stage are the municipalities and the second stage are the households. The regions are the finest geographical level for which direct estimates of the target indicators are usually reliable. The survey collects information on the rent payed by the occupants and on the main characteristics of each house, as well the characteristics of the area where the house is located.

Taking into account the data available, we use a hedonic price method to estimate the SHPI at provincial level. The hedonic price method is basically a regression of the price of the house (rent) against known relevant determinants (characteristics of the unit) that indirectly affect the price. A classical hedonic equation is as follows:

$$\ln p_{ir} = \alpha_0 + \sum_{r=1}^{R} \alpha_r D_r + \sum_{k=1}^{K} \sum_{h=1}^{H} \beta_{kh} C_{kh} + \epsilon_{ir}, \qquad (2)$$

where p_{ir} is the rent cost per square meter of house i in province r, D_r is a vector equal 1 if house i is in area r and 0 otherwise, α_r is the area r price, C_{kh} is the characteristic k and classification h of the house i with β_{kh} its regression coefficient, also called characteristic shadow price and ϵ_{ir} is the error term for house i in area r, which should satisfy the standard assumptions of the multiple linear regression model. As for model (1), α_r represent the fixed effect of area r compared to Italy. Once the α_r are estimated, the SHPI of area r is obtained as $\exp(\alpha_r)$.

To control for the characteristics and classification of the house, we use the following variables: municipality type, kitchen included in the leaving room, presence of: small kitchen, heating, building age, satellite TV, garden, dish-washer, broadband; rent type, number of rooms, surface. Parameters have been estimated using weighted least squares to account for the presence of heteroscedasticity. The estimated SHPIs are summarized in Table 2, where provinces are grouped by geographic repartitions (North, Centre, South).

Table 2: Distribution of province SHPI grouped by geographical repartition

Repart.	*Min*	*1st Q.*	*Median*	*Mean*	*3rd Q.*	*Max*
North	0.7047	1.0436	1.1444	1.1675	1.2818	1.9378
Centre	0.6807	1.0032	1.0595	1.1494	1.3234	1.7712
South	0.5042	0.6390	0.8025	0.8292	0.9919	1.2651

From Table 2 we can see that the distribution of SHPI among provinces of north and central Italy are very similar, even though SHPI in the northern provinces are a little higher than those in the centre. The SHPI values in southern provinces are smaller, with almost 75% of the provinces with an index below 1, the SHPI value of Italy. It can be noted that the values of the SFPI show a reduced geographically dispersion than those of the SHPI (see Table 1), but the two indexes are coherent as they both indicated a higher level of the prices in the south.

4 Adjusting the NPL

The NPL for a household of two components is set by Istat as the mean per-capita consumption expenditure. For different household sizes this threshold is adjusted according to the Carbonaro equivalence scale, which takes the following values: 0.6 for households with one component, 1.33 with three, 1.63 with four, 1.90 with five, 2.16 with six and 2.40 for households with seven or more components. In 2017 the NPL for two components is estimated equal to 1102.52 euros.
We decided to adjust the NPL for each province using both the SFPI and the SHPI values opportunely weighted (extending the idea in Renwick et al. (2014)):

$$NPL_r^* = NPL \times (\phi_r SFPI_r + \lambda_r SHPI_r + 1 - \lambda_r - \phi_r),$$

where NPL_r^* is the adjusted poverty line for province *r*, ϕ_r is the estimated share of expenditure for food and beverages and λ_r is the estimated share of expenditure for the house in province *r*. The quantities ϕ_rs and λ_rs are estimated from the HBS and summarized in table 3.

Table 3: Distribution of province share of expenditure for house and food grouped by geographical repartition

Repart.	*Min*	*1st Q.*	*Median*	*Mean*	*3rd Q.*	*Max*
Estimated share of house expenditure (%)						
North	15.66	17.97	19.37	19.55	20.92	24.57
Centre	14.34	18.84	20.53	20.21	21.77	25.13
South	13.68	17.34	18.71	18.98	20.63	25.48
Estimated share of food expenditure (%)						
North	12.89	16.36	17.74	18.04	18.94	26.83
Centre	13.87	18.05	19.11	19.38	20.29	25.47
South	18.96	21.34	23.71	23.34	25.27	27.74

Results from Table 3 show a similar distribution among provinces in the north, center and south of Italy of the share of expenditure for house rent, while the distribution of the share of expenditure for food is higher in the southern provinces than in the central and northern provinces.
For each province we use the adjusted NPL to estimate the HCR. Direct estimates of HCR have been obtained using Horwitz-Thompson like estimators – where design weights have been adjusted for the population size at provincial level. Direct estimates prove to be unreliable showing high level of coefficient of variation. Therefore, we use small area estimation methods to improve the precision of the estimates. Given the availability of provincial auxiliary data and the absence of spatial correlation in the conditional distribution of the target variable given the auxiliary variables, we used the basic Fay-Herriot area-level model (Fay and Herriot, 1979). As province level auxiliary variables, we used the ratio between number of taxed persons over the population, and the ratios between the number of persons with *i.* income coming from salary, *ii*. income coming from pensions and *iii*. income lower than 10,000 euros per

year, over the number of taxed persons. These data come from the Italian tax agency database. The Fay-Herriot model assumptions are reasonable and model fit is good.

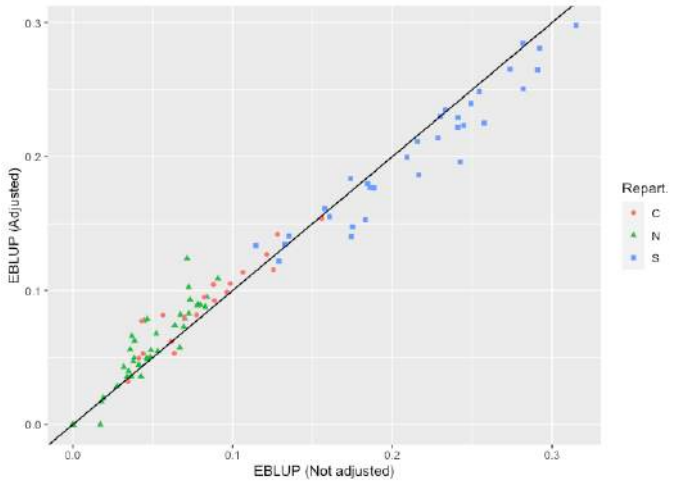


Figure 1: EBLUP estimates of HCR for Italian provinces, Not adjusted poverty line (NPL) and Adjusted poverty line (NPL_r).

From Figure 1 we can see that the HCR of southern provinces estimated using the adjusted poverty line is lower than the HCR obtained using the NPL, while for central and northern provinces is the contrary, as we expected given the perception that in the south of Italy food goods and house rents are cheaper.

References

1. Biggeri, L., Giusti, C., Pratesi, M., Marchetti, S.: Poverty Indicators at Local Level: Definitions, Comparisons in Real Terms and Small Area Estimation Methods. *Statistics and Applications*, **16**, 351-364, ISSN: 2454-7395 (2018)
2. Fay, R.E., Herriot, R.A.: Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, *J. Am. Stat. Assoc.* **74**, 366a, 269-277 (1979)
3. Rao, J.N.K., Molina, I.: Small area estimation. John Wiley & Sons, Inc., Hoboken, New Jersey (2015)
4. Laureti, T., and Rao, D. P.: Measuring Spatial Price Level Differences within a Country: Current Status and Future Developments. *Estudios de economía aplicada*, **36**(1), 119-148 (2018)
5. Renwick, T., Aten, B. Figueroa, E., Martin, T.: Supplemental Poverty Measure: A Comparison of Geographic Adjustments with Regional Price Parities vs. Median Rents from the American Community Survey. BEA Working Papers 0111, Bureau of Economic Analysis (2014)
6. Suits, D. B.: Dummy Variables: Mechanics v. Interpretation. *Rev. Econ. Stat.* **66**, 1, 177-80 (1984)

Smart solutions for trusted smart statistics: the European big data hackathon experience

Soluzioni smart per "trusted smart statistics": l'esperienza del "European big data hackathon"

Francesco Amato, Mauro Bruno, Tania Cappadozzi, Fabrizio De Fausti, Manuela Michelini[1]

Abstract The European Statistical System Committee (ESSC) is promoting innovation in social statistics, recognizing the importance of social statistics to support evidence-based policymaking. Within this context Eurostat promoted the European big data hackathon, with the aim to stimulate research on the use of new data sources for diary-based Time Use survey (TUS) and test new methodologies to improve the quality of statistical outputs. This paper focuses on the characteristics of TUS diary and the challenges for its innovation. In the second part of the paper, we describe SMart Use of TIme Survey (SMUTIS), implemented by Istat during the hackathon. SMUTIS offers a set of innovative functionalities for data processing, such as machine-learning techniques.

Abstract *Lo European Statistical System Committee (ESSC) è impegnato nella promozione dell'innovazione nelle statistiche sociali, riconoscendone l'importanza a supporto della formulazione di politiche sociali. In questo contesto Eurostat ha promosso lo European big data hackathon, con l'obiettivo di stimolare la ricerca nell'uso di nuove fonti di dati per migliorare la qualità dell'indagine Uso del Tempo. Questo articolo si concentra sulle sfide metodologiche per l'innovazione del diario di Uso del Tempo. In particolare, viene descritto il sistema SMart Use of TIme Survey (SMUTIS), realizzato dall'Istat allo European big data hackathon. SMUTIS, offre diverse funzionalità innovative di analisi, come ad esempio tecniche di machine-learning per il processamento di nuove fonti di dati.*

Key words: Time Use survey**,** trusted smart statistics, European big data hackathon, machine learning

[1] Francesco Amato, Istituto Nazionale di Statistica (Istat), framato@istat.it;
Mauro Bruno, Istituto Nazionale di Statistica (Istat), mbruno@istat.it;
Tania Cappadozzi, Istituto Nazionale di Statistica (Istat), cappadoz@istat.it;
Fabrizio De Fausti, Istituto Nazionale di Statistica (Istat), defausti@istat.it
Manuela Michelini, Istituto Nazionale di Statistica (Istat), mamichel@istat.it

1 Introduction

The European Statistical System Committee (ESSC) has been committed to promoting innovation in social statistics for over a decade, recognizing the importance of social statistics to support evidence-based policymaking, in times of resource constraints. Indeed, the high costs of some social surveys[1] often hamper their development in EU Member States [1]. In particular, the collection method based on the traditional paper-and-pencil self-completed diary and the expensive post-coding process are objective obstacles to their introduction as compulsory surveys into the Integrated European Social Statistics (IESS) Regulation. The ESSC states the need for better information on time use and household budgets in terms of coverage and comparability. To this aim, the ESSC promotes the implementation of new capabilities that improve the responsiveness of users, the linkage of TUS or HBS data with other data sources and the design and implementation of new methods and tools.

Eurostat started several projects enabling Member States to develop innovative solutions for diary-based surveys. In particular the Task Force on Innovative Tools and Sources for HBS and TUS [2], in continue cooperation with the Big Data project[2], highlighted how specific mobile applications could be used in the TUS and HBS domains to collect data. In particular the use of smartphones (motion tracker, camera, microphone, etc.), smart watches and activity trackers (e.g. speech recognition for activity labelling, record the time of activity, record location where activity took place, etc.) for collecting TUS data.

Another initiative promoted by Eurostat, for modernizing diary-based social surveys, is the European Big Data Hackathon. The aim of the challenge is to stimulate research on the use of data collected by smartphones and their possible integration with data collected through the respondent by filling in an online prototypal time use diary. The focus of this paper is to briefly describe the characteristics of the TUS diary, the challenge for its innovation and the main outcome achieved by Istat during the hackathon.

1.1 Challenges for the diary-based Time Use survey

In the paper-and-pencil self-completed time use diary, respondents report in free text the activities carried out during a reference day, in a time schedule with minimum slots of 10 minutes [3]. For each daily activity, respondents report the starting and finishing time as well as important contextual information like the place of occurrence and the possible presence of others. This tool allows collecting very detailed and useful information for the formulation of policies related to lifetimes and gender equality. However, the survey is very time and cost consuming for the

[1] Time Use Surveys (TUS) and Household Budgets Surveys (HBS)

[2] More information on the Big Data project is available on CROS portal https://ec.europa.eu/eurostat/cros/content/essnet-big-data_en

NIS and with high burden for the respondents. Therefore, the challenges for this survey to be achieved with the use of new technologies are manifold:

- *reduce respondent's burden*: pre-filling part of the diary using big data collected automatically by smartphone (i.e. travel time using GPS data, the ICT use);
- *improve the diary quality*: introducing immediate consistency checks at the end of the diary;
- *reduce time and costs*: using smart solutions eliminates the registration and coding phases of paper diaries;
- *provide feedback to respondent*: at the end of the diary day, respondents can receive an immediate feedback on their time use compared to the average, as an incentive strategy for participation.

In this challenging context, participation in the Big Data hackathon represented the first concrete attempt to find solutions, albeit experimental, to these innovation needs.

2 The Big Data hackathon experience

The European Big Data Hackathon is an event organised by the European Commission (Eurostat) which gathers teams from all over Europe to compete for the best data product combining official statistics and big data to improve quality of official statistics products. The European Big Data Hackathon has five objectives[1].

1. Solve statistical problems by leveraging algorithms and available data, by engaging with data scientist across Europe to build relevant algorithms and by exposing relevant data sets to participants to come up with ideas.
2. Identify first-class data scientists, by targeting local developers' and data scientists' communities to attend the events, by connecting with the best participants during the events for possible future collaboration.
3. Promote and accelerate big data for statistics initiatives in Europe, by developing prototypes that European countries will be able to integrate, and by generating buzz about big data within communities of scientists' entrepreneurs.
4. Promote partnerships with the research community and the private sector, by raising awareness about big data initiatives in Official Statistics in Europe, by starting partnerships with the private sector and Universities.
5. To produce innovative products and tools, to stimulate the use of open data and public use files and to engage with new audiences.

The teams compete for the best application of big data providing an answer to the following challenge:

"*How can innovative solutions for data collection reduce response burden and enrich or replace the statistical information/data provided by the time use survey?*"

[1] More information on the hackathon is available on CROS portal https://ec.europa.eu/eurostat/cros/EU-BD-Hackathon_en

2.1 The implemented solution

To achieve the hackathon goals and provide an accurate answer to the challenge, Istat team implemented the application SMUTIS (SMart Use of TIme Survey)[1]. The guiding principles in the design and implementation of SMUTIS were the following:

1. *Use of new data sources to enrich traditional output*. The input data provided to the teams combined the collection of data captured by sensors available in smart phones with data entered by the users (e.g. activities performed).
2. *Increase of quality of statistical output by analysing sensor data*. Sensor data provided by the mobile applications included location data and recorded signals from accelerometer, barometer, proximity sensor, etc.
3. *Test the use of machine learning techniques to produce new statistical outputs*. The mobile application used for the hackathon collected pictures of meal and food advertisement. A key goal of the hackathon was to analyse food images as a proxy of nutritional facts and health indicators.
4. *Compliance with official statistics standards.* Implementation of a tool compliant with CSPA principles [4], i.e. software components should be modular, shareable and use standard metadata. Design of a metadata repository according to GSIM [5] information model and classification of metadata according to the GSBPM phases [6].

The hackathon imposed further constraints to the teams, each team had to establish links between the heterogeneous data sets provided and identify and use other data sources beyond those currently at hand. The implemented tools had to integrate or at least had to consider the aspect of data quality for the analysis and visualization.

Given the abovementioned objectives and constraints, SMUTIS architecture fulfills the following requirements:

- *Access and analyse heterogeneous data sources*. SMUTIS needs to access data generated by mobile applications (i-Log and BigO) and the results of the Harmonised European Time Use Survey (HETUS), available on Eurostat website.
- *Analyse and integrate sensor data in a GIS framework*. SMUTIS provides functionalities to process sensor data to derive statistical information on travel/commute activities. The functionality should display data on GIS maps.
- *Use of machine learning algorithms to process food images*. SMUTIS integrates machine-learning algorithms to process food images and produce health indicators, based on nutrition facts. This component is the most challenging and innovative; more details on the deep neural network implemented are provided in the next paragraph.
- *The tool should provide a set of standard services for data processing and analysis.* SMUTIS follows the service-oriented architecture (SOA) design pattern. The server component will provide loose-coupled services with a well-defined interface, e.g. “Sensor data processing service”, “Machine leaning

[1] SMUTIS is an open source project, released under the EUPL 1.1 public license. The source code is available at: https://github.com/mecdcme/smutis

service". The client component should invoke the services and display the outputs on a graphical user interface, accessible via web browser. The main components of SMUTIS architecture are displayed in Figure 1.

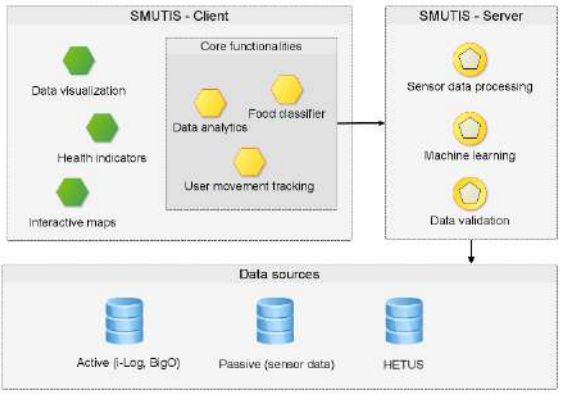


Figure 1: SMUTIS components

SMUTIS, conceived as a proof-of-concept, has confirmed that new data sources and machine learning techniques can be combined to improve traditional output and produce new indicators.

2.2 Machine learning for food classification

With respect to the policy question of the hackathon, we focused on "enrich or replace the statistical information/data provided by the time use survey". We tested the opportunity of creating a new wellness indicator based on the activities of daily life, linking physical activity and healthy eating.

Among the data collected via smart devices, we took advantage of the photos taken at meals through the BigO mobile application[1]. BigO was initially designed to collect and analyze big data on behavior and living environments, to allow public health authorities to plan and execute effective programs to reduce the prevalence of childhood obesity. Further, we created the food classification algorithm with the aim of recognizing the meal eaten by the user and associating the value of the typical average macronutrients for the meal. To implement the food classification algorithm, we used very powerful and versatile machine learning tools such as CNN convolutional neural networks [7], the state of the art in image recognition. For power and relative simplicity, we have chosen to implement the architecture called Incepitonv3 [8]. To train the network we used the Food101 dataset [9], a dataset that collects about one hundred classes of international foods with a cardinality of one thousand images per class. The images collected for the hackathon were classified

[1] More details concerning the "Big data against childhood Obesity" (BigO) project are available at: https://bigoprogram.eu/

for each user according to the Food101 classification and contextually linked to the corresponding typical macronutrients, such as carbohydrates, fats and proteins.

Using the food recognizer as a backend micro service, the SMUTIS application offers the possibility of recognizing in real time the meals eaten and provides information on the ratios of the macronutrients consumed by the user.

3 Conclusion

The European big data hackathon experience proved that new data sources and methodologies could be exploited to achieve challenges related to diary-based surveys. Although such results are experimental, they point the way for both methodological and technological innovation in social surveys.

Indeed Eurostat launched at the beginning of 2020 a new ESSNet project on "Trusted Smart Statistics", with the aim to define the specifications for a European Platform supporting the use of shared smart survey solutions and furthermore to assess the usage of applications for European social surveys, such as the Time Use Survey (TUS) or the Household Budget Survey (HBS).

References

1. DGINS, Wiesbaden memorandum (2011). Available at: https://ec.europa.eu/eurostat/web/ess/about-us/ess-gov-bodies/dgins
2. Cloodt H. Innovative tools and sources for data collection Harmonised European Time Use Survey. In Proceeding of the International Association for Time Use Research Conference 2018. Available at http://www.ksh.hu/iatur2018/iatur40_cloodt.pdf
3. Eurostat, Harmonised European Time Use Surveys (HETUS) 2018 Guidelines (2019)
4. Common Statistical Production Architecture (CSPA) https://statswiki.unece.org/display/CSPA/DRAFT+CSPA+v2.0 (Accessed: 21 February 2020)
5. Generic Statistical Information Model (GSIM) https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model (Accessed: 21 February 2020)
6. Generic Statistical Business Process Model (GSBPM) https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1 (Accessed: 21 February 2020)
7. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361,10 (1995)
8. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2818-2826 (2016)
9. Bossard, Lukas, Matthieu Guillaumin, and Luc Van Gool. "Food-101–mining discriminative components with random forests." European conference on computer vision. Springer, Cham (2014)

The ESSnet Project Smart Surveys: new data sources and tools for Surveys of Official Statistics

Il progetto ESSnet Smart Surveys: nuove fonti di dati e strumenti per le indagini della Statistica Ufficiale

Claudia De Vitiis[1], Francesca Inglese[2]

Abstract Trusted Smart Statistics represent the step forward taken by official statistics as evolution of the Big Data framework and can be seen as the future extended role of official statistics in a world impregnated with smart technologies and new data sources. New data sources represent a unique opportunity to produce new and improve existing statistics within a collective collaborative framework. The ESSnet Smart surveys, in this context, has the objective to conceive a methodological and conceptual framework, define the specifications for a European Platform supporting the use of shared smart survey solutions and, moreover, to assess the usage of smart applications for the European social surveys, implementing privacy by design approaches.

Abstract *Nel contesto europeo delle Trusted Smart Statistics, che costituiscono l'evoluzone del paradigma Big Data per la statistica ufficiale, il paper presenta il progetto ESSNet Smart Surveys, finanziato da Eurostat. Il progetto si propone di testare e implementare soluzioni basate sull'uso di smart devices per la produzione statistica e in particolare per indagini sociali che risultano molto pesanti per il risponente. L'obiettivo principale è definire il framework concettuale per lo sviluppo di una piattaforma europea per la conduzione di smart surveys, insieme alla sperimentazione di strategie e tool già sviluppati da altri progetti Eurostat.*

Key words: Official statistics, Trusted Smart Statistics, Trusted Smart Surveys

[1] Claudia De Vitiis, ISTAT; email: devitiis@istat.it

[2] Francesca Inglese, ISTAT; email: fringles@istat.it

1 Introduction

The last decades have seen important changes in data collection and in their dissemination and use in all sectors of society. This *big data revolution* phenomenon is mainly due to the increased ability to collect and store information automatically through very different sources such as sensors of various kinds, satellites, mobile devises, internet, drones and many others. Big data represented a great opportunity for official statistics in a hyper-connected world dominated by the IoT (Internet of things), but must be used in order to build a reliable information set: they pose new challenges for Statistics, a science that realizes knowledge of reality on empirical-observational data (Arbia, 2019).

The big data revolution, from digitalization to datafication (Ricciato *et al*, 2019), determined for official statistics a totally new situation, together with the massive use of administrative data sources. This state stimulated in recent years the definition of a *new paradigm* for the official statistics: the production of statistics becomes a process in which data from various sources converge, establishing the need not only of methods and techniques to structure and elaborate data coming from each source, but also new reliable methodologies and algorithms to integrate them in order to obtain statistics compliant with the international quality standards (Citro, 2014). In the most recent years a further step forward derived from the transition from the notion of Big data in Official Statistics to the concept of *Trusted Smart Statistics*. This caused a *shift from data sources to data systems*, towards a systemic paradigm change (Ricciato et al 2019).

The paradigm shift requires methodologies, but also an effort to conceptualize, measure, and communicate new sources of error and the statistical accuracy of integrated data. This paradigm needs to be built on new methodological frameworks fit for new data sources (e.g. sensor data), and such frameworks must be designed to enable methodological agility, cooperative development and continuous evolution. Common concepts, as meta-data, data lifecycle, data lineage and provenance, need to be extended from data to methods or equivalently analytics, algorithms, software etc.. New methods and their implementations will not be static or quasi-static but rather dynamic objects (Ricciato et al 2019).

The development of Smart Surveys - surveys in which respondents are asked to employ smart devices (e.g. smartphones, tablets, activity trackers) to collect survey data through active and passive data collection - and Trusted Smart Surveys (TSS) - surveys in which respondents are asked to share existing data collected by trusted third parties, like government authorities and larger, stable enterprises willing to establish data delivery agreements - offers new challenges to improve the quality of social surveys in the NSIs.

The ESSnet on Smart Surveys, which started its activities at the beginning of 2020, fits into this context and represents the focus of this work. The aim of this paper is to present the objectives of the project Smart Surveys and outline the main activities that the ESSnet will undertake to pursue these goals. The paper is organized as follows: in section 2 the background on big data and TSS is outlined, section 3 describes the main

objectives of the ESSnet Smart Surveys and section 4 reports the first ideas regarding the methodological framework for the Smart Survey Platform.

2 Background on Trusted Smart Statistics in the ESS

The European Statistical System Committee (ESSC) adopted in 2013 the Scheveningen Memorandum on Big Data and Official Statistics (European Commision, 2013) and in 2014 an ESS action plan and roadmap (Big Data Action plan and Roadmap - BDAR2). The Bucharest memorandum in 2018 (European Commission, 2018), followed by the Trusted Smart Statistics Action plan and Roadmap, establishes the need for integrating new data sources and data collection methods in the production of official statistics. In particular, the ESS has established privacy-by-design approaches, including the resolution of the legal aspects and putting in place the necessary technical requirements at national and European level.

The definition of TSS will be the eventual result of a process for developing the concept and the respective statistics. Tentatively, we can define TSS as automated, algorithm-based production of official statistics that is trusted by design: it adheres to widely agreed principles and standards of algorithmic transparency, auditability, reproducibility and privacy preservation. The term "Smart Statistics" is inspired by the notion of "smart systems" as systems that are capable of autonomously sensing, processing and interactively analysing information across network infrastructures. Trust is a cornerstone of the current system of official statistics. "Trusted Smart Statistics" from smart systems requires a new coherent set of technical, organisations and legal means respected by all involved stakeholders.

Trusted smart statistics is therefore conceptually integrated in national strategies across the ESS. Its implementation includes prototyping and developing: (i) national and trans-national trusted smart surveys, based on advanced interaction models through smart devices combined with advanced privacy-by design technologies; (ii) use of citizen science data and open data sources; (iii) methodological and quality frameworks integrating the use of multiple data sources (multi-source statistics and multipurpose sources) in the production of official statistics.

Different smart survey solutions were developed within the Eurostat Grants for "Innovative Tools and Sources" for two specific surveys: the Harmonised European Time Use Survey (HTUS) and the Household Budget Survey (HBS)..

In this context, the ESS lauched in 2019 an ESSnet on Smart Surveys which started its activities in 2020. This project will contribute towards many important achievements that are foreseen within the TSSAR: (i) testing and developing (trusted) smart surveys within the ESSnet, (ii) the conceptualisation, development and implementation of a new reference architecture for trusted smart statistics as well as the evolvement of new skills within the ESS. In the ESSnet, the existing solutions for smart surveys will be tested in other European countries and the applicability to other domains will be evaluated. Furthermore, the ESSnet will deliver preparatory work to create a Europeanwide platform to share und re-use smart survey solutions and components.

3 The ESSnet Smart surveys: objectives and planned activities

Mobile devices have become standard tools for communication within the time span of just a decade and have a high population coverage worldwide. Because of their high population coverage and daily life use and because of the availability of sensors, mobile devices have become tools that may supplement surveys with automated data from sensors. Some of these sensor data may also replace survey data (Eurostat Grant Mimod, 2019).

The term *smart surveys* refers to surveys that use smart personal devices, equipped with sensors and mobile applications. The concept of smart surveys goes well beyond the mere use of web-based (online) data collection that essentially transform the paper questionnaire into an electronic version. Smart surveys involve dynamic and continuous interaction with the respondent and with her/his personal device(s). They combine data collection modes based on input from the data subjects with data collected passively by the device sensors (e.g. accelerometer, GPS, microphone, camera, etc.), involving relevant consent issues.

Using "trusted smart surveys" we refer to an augmentation of smart surveys by technological solutions that collectively increase their degree of trustworthiness and hence acceptance by the citizens. Constituent elements of a trusted smart survey are the strong protection of personal data based on privacy-preserving computation solutions, full transparency and auditability of processing algorithms.

The overall goal of the ESSnet Smart Surveys is to define the specifications for a European Platform supporting the use of shared smart survey solutions and furthermore to assess the usage of applications for European social surveys, such as the Time Use Survey (TUS) or the Household Budget Survey (HBS). Both surveys are considered, in fact, to be very burdensome to respondents and to be prone to low recall as well as underreporting errors. Modernization of these surveys and similar ones is therefore urgently needed. In the context of five ESS grants on "Innovative tools and sources for diary-based surveys, the Household Budget Survey and/or the Time Use Survey", members of the ESS are currently developing and testing apps and new data collection methods for these surveys. In the ESSnet Smart Surveys, these solutions will be tested and transferred to other countries and domains.

On the whole the ESSnet Smart Surveys has two main goals: (i) Evaluate existing tools for smart surveys in particular domains, with particular focus on those where tools have already been developed (Time Use Survey, HETUS, and Household Budget Suervey, HBS), with the goal of identifying points for improvement; (ii) Developing technological and methodological solutions for processing the input personal data in a privacy-preserving fashion (e.g., by leveraging Secure Multi-Party Computation), with no need to centralize the data at a single entity (concentration risk) and in a way that provides solid guarantees in terms of privacy protection, full auditability and complete transparency of the processing methods applied over the data.

A conceptual framework will be developed that serves as the basis for a future European platform for trusted smart surveys. This ESSnet is intended to take a first step towards the future development of the platform and will cover the pre-

development stage: collection of requirements, identification of design principles, architecture design and formulation of concrete specifications. The outcome of this ESSnet is intended to serve as a direct input to the future development of such a platform in other follow-up project(s).

The platform will be flexible and implementing a set of common (horizontal) functions and configurable services that can be used to build particular instances of trusted smart surveys for specific application domains and/or target areas. Such a platform should be modular, evolvable, extensible and agnostic to particular application domains. It should provide ready-to-use solutions for horizontal functions. It should allow each platform user (i.e., any ESS member) to instantiate a specific trusted smart survey by selecting and configuring different modules. The platform should include support for secure private computing to avoid data concentration (e.g., secure multiparty computation), full transparency and public auditability.

The activities planned for the ESSnet are structured in three work-packages, the first devoted to coordination and management. The WP2 will carry on the pilot projects using available tools, through the following steps: identification and collection of the functional and technical characteristics of the solutions developed for the HETUS and HBS; establishment and execution of a solutions evaluation protocol for existing data collection applications; developing pilot projects demonstrating the use of scalable secure private computing solutions (e.g. Secure Multiparty Computation) to process individual citizen data without concentrating personal data at a single entity, in combination with appropriate technological solutions to provide full auditability. Test and pilot project will be conducted during 2020-21.

The WP3 will work on the TSS platform through two main tasks: (i) conceptualization and development of a general platform for trusted smart surveys for collecting data for official statistics, following a top-down design approach from abstract framework through architecture down to detailed technical specifications. (ii) development of proofs-of-concept in the form of modular prototype elements for essential aspects of the architecture. A strict connection with the activities of the WP2 will guarantee the coherence and the consistency between the theoretical framework and the feedbacks that will come from the tests and pilots on the existing smart tools.

4 The methodological and conceptual framework for the Smart Surveys: first steps

The first step in the direction of defining a conceptual framework in vision of the platform for the trusted smart survey, is the identification and description of the specificities of the statistical process for a smart survey.

Going through GSBPM, the sub-processes involved/affected especially by the use of smart devices for the data collection will be identified. A methodological scheme of the overall process of a smart survey will be defined, considered as a mixed mode

survey and a multiple data source survey, in which data deriving from different data collection mode are integrated. In particular the points will be highlighted where: (i) smart data are integrated with traditionally collected data; (ii) the impact on the estimates of the use of smart data can be assessed, together with the effect on data quality in terms of selectivity of device usage, nonresponse bias and measurement error (Herzing, 2019).

The specification of methods for data collection and treatment of sensor data through smart devices will be established starting from the Inventory of tools (linking with the WP2 and the Task Force "Innovative Tools and Sources"), together with the identification of the requirements for the APPs for data collection in social surveys.

Relevant steps will be: the review of sensor data sources (starting from the work done in the Eurostat Grant "The use of mixed-mode data collection in social surveys", MIMOD, 2019); the definition of methods for representation and structuring of sensor data to obtain statistical information; the construction of a metadata system that describes the data in terms of concepts, techniques, processes and origins; the definition of new and advanced methodological concepts to link different data sources and to validate data; the definition of technology to obtain secure data access, in conjunction with the appropriate methodology and algorithms to guarantee privacy and confidentiality. These steps are necessary to define a data model that contemplates on the one hand the process of methodological development and the statistics production on the other.

References

1. Arbia, G.: Statistica, nuovo empirismo e società nell'era dei Big Data. Nuova Cultura, Roma (2019)
2. Citro, C.F.. From multiple modes for surveys to multiple data sources for estimates. Survey Methodology, 40(2), 137-161 (2014)
3. ESSnet Mixed-mode Designs for Social Surveys, (2017-2018) Deliverable 3 of Work Package 5 Final methodological report "Discussing the use of mobile device sensors in ESS surveys" LINK
4. European Statistical System Committee. Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics); 2018. http//www.dgins2018.ro/bucharestmemorandum.
5. Eurostat Grant "The use of mixed-mode data collection in social surveys", MIMOD, Deliverable 3 of Work-package 5, "Final methodological report discussing the use of mobile device sensors in ESS surveys", 2019 https://www.istat.it/en/research-activity/international-research-activity/essnet-and-grants
6. Herzing, J.M.E.: Mobile web surveys. FORS Guide No. 01, Version 1.0 (2019) https://forscenter.ch/wp-content/uploads/2019/02/herzing_fg01_mobile_v1.0.pdf
7. Lohr, S.L., Raghunathan, T. E.: Combining survey data with other data sources. Statistical Science, 32(2), 293-312 (2017)
8. Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M.: Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS, 35 (2019) 589–603.
9. Scheveningen Memorandum on Big Data and Official Statistics, 2013. https//ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en.

Factorial and dimensional reduction methods for the construction of indicators for evaluation (SVQS-SIS)

A comparison of MBC with CLV and PCovR methods for dimensional reduction of the soccer players' performance attributes

Un confronto di MBC con i metodi CLV e PCovR per la riduzione dimensionale degli attributi di performance dei calciatori

Maurizio Carpita, Enrico Ciavolino and Paola Pasca

Abstract This work extends the research on soccer performance attributes classification given by EA Sports esperts and available on Kaggle in the KES dataset. Although a first clustering-based inquiry through the Clustering around Latent Variables (CLV) as well as the Principal Covariates Regression (PCovR) produced more detailed and meaningful groupings, no consistent improvement in the predictive performance of match results was observed. Moreover, as the new composite indicators are produced, these dimensional reduction methods does not account for the data uncertainty. In order to overcome this limit, a Model-Based Clustering (MBC) method is considered and tested on players' performance variables.

Abstract *Questo lavoro estende e approfondisce alcuni studi per classificare con metodi statistici gli attributi di performance dei calciatori forniti dagli esperti di EA Sports presente nel dataset KES su Kaggle, piattaforma dedicata alla Data Science. Sebbene l'uso di metodi quali Clustering around Latent Variables (CLV) e Principal Covariates Regression (PCovR) produca classificazioni più interessanti, non si è rilevato un miglioramento nella capacità predittiva dei risultati delle partite. Inoltre, le precedenti modalità di clustering non incorporano l'incertezza dei dati. Per risolvere tale limite, si propone il clustering basato su modelli di mistura (MBC), applicato agli indicatori di performance.*

Key words: sport performance indicators, model-based clustering, clustering around latent variables, principal covariate regression, classification

Maurizio Carpita
University of Brescia, Department of Economics and Management
e-mail: maurizio.carpita@unibs.it

Enrico Ciavolino
Paola Pasca
University of Salento, Department of History, Society and Human Studies
e-mail: enrico.ciavolino@unisalento.it
e-mail: paola.pasca@unisalento.it

1 Introduction

The latest benefits and developments in sports research, specifically that on soccer, come from a "data-scientific" approach. Players' performance on the soccer field has been extensively measured and described by soccer experts: particularly known and detailed are the attribute classification provided by the experts from Electronic Arts (EA)[1]. In their opinion, players' performance can be thought of as a multidimensional construct made up of 7 performance composite indicators (e.g. *defending*), each of which consists of several, more specific skills (e.g. *marking, standing tackle* and *sliding tackle* as elements of the above mentioned *defending* dimension).

Thanks to world's widespread Data Science platforms such as Kaggle, this wide variety of players' performance attributes can be joined to data about real soccer matches and, most of all, is now accessible to everyone [1]. These data lend themselves to both predictive modeling [3], but also to test the soundness of experts' classification of players' performance. A statistical inquiry of experts' performance areas showed strong correlations between variables belonging to different dimensions, thus indicating that experts' theorization might not be statistically supported. If so, a clustering approach might provide an increase in both theoretical and predictive power of players' performance variables.

2 Performance composite indicators of soccer players: a review

Literature about modeling soccer matches is rich and a vivid debate is still going on (see, among the others [11, 12, 2]). In such predictive models, experts' ratings of players are used as predictors to feed machine learning algorithms. To this end, the Kaggle European Soccer (KES) database represents one of the biggest available datasets in the European framework [10]. A more recent version [5] covers a timespan of 7 seasons (from 2009-2010 to 2015-2016) in 10 different countries (Belgium, England, France, Germany, Italy, Netherlands, Portugal, Scotland, Spain and Switzerland) and melts information from two different sources: 20,973 *matches*, which include home and away team goals, as well as players' position on the pitch, and 33 *performance attributes* (variables) reflecting *players' abilities* measured at regular intervals (twice a year until 2013, then just once a month). For what concerns attributes' description, experts of Electronic Arts (EA) Sports state that players' performance can be summarized into 7 *performance composite indicators* made up, in their turn, of specific skills to be mastered by players.

Such procedure, although reflecting experts' opinion, needs a statistical assessment. Indeed, from a statistical point of view, performance indicators aim at measuring the abilities of players that can be considered as latent traits. Moreover, different indicators measure the same latent trait and can be summarized into a composite

[1] EA Sports, a division of Electronic Arts, develops and publishes sports video games, such as the FIFA series. A link to the website: https://www.easports.com/

indicator. [3] created 7 role-based players' composite indicators and used them to predict match outcomes. An alternative method to create composite indicators is based on multivariate data inspection and dimensionality reduction technique such as Principal Components Analysis (PCA) and Cluster Analysis (CA), as proposed in [15] and [8]. In [6] and [4] respectively, performance indicators were clustered via two different methods:

– *Clustering around Latent Variables* (CLV) [13, 14], a method that simultaneously catches variables that occur together (K clusters) and characterizes K latent traits.
– *Principal Covariates Regression* (PCovR) [7], a method that allows to flexibly tune on predictors reconstruction rather than on the predictive power of a regression model and vice versa.

However, such methods do not account for the uncertainty in the data, and in this paper we propose a *Model-Based Clustering* (MBC) method that relies on the definition of a mixture distribution for the players' performance attributes with a random number of mixture components.

2.1 The proposed Model-Based Clustering (MBC) method

As specified in the previous section, in the paper [6] performance indicators are first clustered via Clustering of Variables around Latent Variables (CLV) approach. This procedure simultaneously finds out K clusters of variables as well as K latent components (dimensions) such that variables in each cluster are strongly correlated with the corresponding latent component. Variables that belong to the same cluster show strong association among them and lower association with variables belonging to other cluster. This goal is achieved by decomposing the variance-covariance matrix between variables. In this work, we adopt a different approach, a Model-Based Clustering (MBC) based on a mixture model. Mixture models, widely known in literature [9], can be used to classify variables according to unobserved latent ones. More formally, we say that a distribution f is a mixture of K component distributions $f_1, f_2, ..., f_k$ if:

$$f(x;\pi,\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k)$$

with the π_k being the mixing weights or marginal probabilities of mixture components ($\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$) and θ_k are the parameters characterizing these mixture components. We consider a Gaussian mixture model, so that:

$$f_k(x;\theta_k) = \phi(x;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right).$$

Let X_{pi} be the p-th performance attribute ($p = 1, ..., P$) for the i-th role ($i = 1, .., 4$; attack, defense, mild-player and goal keeper). Following the motivations in [6], we

do not consider the goal keepers, and therefore $P = 28$ and $i = 1,2,3$.
With the MBC method we assume that performance attributes are modeled as a K-component Gaussian mixture model with parameters μ_k and σ_k^2 ($k = 1,...,K$):

$$X_{pi} \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2). \tag{1}$$

The mixture in Equation (1) is a convex combination, or weighted average, of Gaussian distributions. According to the mixture in Equation (1), for each performance indicator we can compute the probability to belong to a mixture component whose mean μ_k defines the latent trait the performance indicator aims at measuring. Indeed, performance indicators measuring the same unobserved latent trait belong to the same mixture component and are thus clustered together.
In order to write the likelihood of the model with a more compact notation, we introduce the variable Z_{pim} that will be equal to k when the corresponding statistical unit belongs to the k-th mixture component. Considering the attributes x_{pim} *observed before* the m-th match ($m = 1,...,20973$), the likelihood of the model is the following:

$$L(X,Z|K,\mu,\sigma^2) = \prod_{m=1}^{20973} \prod_{p=1}^{28} \prod_{i=1}^{3} \sum_{k=1}^{K} \pi_k^{I(Z_{pim}=k)} \phi(x_{pim};\mu_k,\sigma_k^2)^{I(Z_{pim}=k)}. \tag{2}$$

Note that attribute values are *observed before the matches*, so that we may use the obtained performance composite indicators to predict match results. As expected, no closed-form expression can be derived for the model parameters but they are estimated numerically. Notice that the number K of mixture components of the MBC method, interpreted in this context as the number of latent traits the performance indicators aim at modelling, is a model parameter and it is jointly estimated as all the other model components assuming a multinomial distribution. Optimization for MBC is done according to information criteria; as practice in this context, EM algorithm is involved for model estimation. In estimating the model we rely on the package Mclust implemented in R: such a function specifies a mixture distribution with complete variance-covariance matrix whose structure is simplified optimizing the number of mixture component and the BIC criterion. In our case, the best selected model corresponds to VII (spherical, unequal variance structure).

3 Results and Discussion

Preliminary results in Table 1 showed that MBC identifies 5 mixture components with different parameters. $Cluster\_1$ is the largest cluster (the estimate of π_1 is 0.32), with estimated mean 57.4 and variance 13.8, whereas $Cluster\_4$ is the smallest cluster (the estimate of π_1 is 0.11), with lower estimated mean 51 and variance 8.9.

Table 2 showed the 28 performance attributes (variables and long names; the three player roles are considered together), the $sofifa$, CLV and PCovR classifi-

Table 1 Parameter estimate for the clusters obtained with the MBC method

Cluster	μ	σ^2	π
Cluster_1	57.359	13.767	0.321
Cluster_2	67.781	9.817	0.250
Cluster_3	50.641	7.769	0.143
Cluster_4	50.975	8.868	0.107
Cluster_5	64.316	5.820	0.179

cations, and the classification obtained using the MBC method (in this case, each attribute is assigned to the cluster with max π_k).

Table 2 The 28 performance attributes with the *sofifa* and MBC, CLV, PCovR classifications

Attributes			**Dimension classifications with:**		
(variables)	**Long Names**	*sofifa* (LABEL)	MBC	CLV	PCovR
x_1	shot power	power (POW1)	Cluster_1	Group_2	Comp_1
x_2	jumping	power (POW2)	Cluster_2	Group_3	Comp_4
x_3	stamina	power (POW3)	Cluster_5	Group_1	Comp_3
x_4	strength	power (POW4)	Cluster_2	Group_3	Comp_4
x_5	long shots	power (POW5)	Cluster_3	Group_2	Comp_1
x_6	aggression	mentality (MEN1)	Cluster_1	Group_4	Comp_4
x_7	interceptions	mentality (MEN2)	Cluster_4	Group_4	Comp_2
x_8	positioning	mentality (MEN3)	Cluster_4	Group_2	Comp_1
x_9	vision	mentality (MEN4)	Cluster_1	Group_1	Comp_1
x_{10}	penalties	mentality (MEN5)	Cluster_4	Group_2	Comp_1
x_{11}	dribbling	skill (SKI1)	Cluster_2	Group_2	Comp_1
x_{12}	curve	skill (SKI2)	Cluster_1	Group_2	Comp_1
x_{13}	free kick	skill (SKI3)	Cluster_1	Group_2	Comp_1
x_{14}	long passing	skill (SKI4)	Cluster_1	Group_1	Comp_1
x_{15}	ball control	skill (SKI5)	Cluster_2	Group_2	Comp_1
x_{16}	acceleration	movement (MOV1)	Cluster_5	Group_5	Comp_2
x_{17}	sprint speed	movement (MOV2)	Cluster_5	Group_5	Comp_2
x_{18}	agility	movement (MOV3)	Cluster_5	Group_5	Comp_2
x_{19}	reactions	movement (MOV4)	Cluster_2	Group_2	Comp_3
x_{20}	balance	movement (MOV5)	Cluster_5	Group_5	Comp_3
x_{21}	crossing	attacking (ATT1)	Cluster_1	Group_1	Comp_3
x_{22}	finishing	attacking (ATT2)	Cluster_1	Group_2	Comp_1
x_{23}	heading	attacking (ATT3)	Cluster_2	Group_3	Comp_3
x_{24}	short passing	attacking (ATT4)	Cluster_2	Group_1	Comp_1
x_{25}	volleys	attacking (ATT5)	Cluster_1	Group_2	Comp_1
x_{26}	marking	defending (DEF1)	Cluster_3	Group_4	Comp_4
x_{27}	standing tackle	defending (DEF2)	Cluster_3	Group_4	Comp_1
x_{28}	sliding tackle	defending (DEF3)	Cluster_3	Group_4	Comp_1

Classification obtained with the three methods are not the same. The clusters of the MBC solution are a little more balanced than those of the CLV and PCovR solutions: $Cluster\_1$ of MBC incorporates 9 out of the 28 attributes, $Group\_2$ of CLV incorporates 11 attributes and $Comp\_1$ of PCovR incorporates 15 attributes, resulting into lower discrimination of these last two methods. MBC reflect better the players' roles: for example, $Cluster\_1$ are related to attacking capabilities, while

these attributes belong to various groups and components for CLV and PCovR respectively.

For the MBC method, *movement* and *defending* attributes (except MOV4) are in two clusters as for $sofifa$, but this is not the case for CLV and PCovR solutions.

In summary, the MBC clusters result to be good candidates to contruct performance composite indicators to potentially enhance prediction of match results: we will use these and those of the CLV and PCovR into the Skellam model to predict the goal-differences of matches in the KES database.

References

1. Airback: Match outcome prediction in football: www.kaggle.com/airback/match-outcome-prediction-in-football?scriptVersionId=796746 kaggleComp3. (2017)
2. Baboota, R., and Harleen K.: Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting **35**(2): 741–755 (2019)
3. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the Kaggle european soccer database. Statistical Modelling **19**(1):74–101 (2019a)
4. Carpita, M., Ciavolino, E., & Pasca, P.: Exploring the statistical structure of soccer team performance variables using the Principal Covariates Regression (PCovR). In Carpita M. and Fabbris L. (Editors), Book of Abstracts of the Scientific Conference on Statistics for Health and Well-being, ASA Conference 2019, Brescia, September 25-27, 2019. CLEUP Coop. Libraria Editrice . ISBN: 978-88-5495-135-8. (2019b)
5. Carpita, M., Ciavolino, E., Pasca, P.: European Soccer Dataset by Role: www.kaggle.com/paolap86/modified-version-of-the-european-soccer-dataset. (2019c)
6. Carpita, M., Ciavolino, E., Pasca, P.: Role-based Players' Performance Composite Indicators of Soccer Teams: a Statistical Perspective. Accepted by Social Indicators Research (2020)
7. De Jong, S. Kiers, H. A.: Principal covariates regression: Part I. Theory. Chemometrics and Intelligent Laboratory Systems, **14**(1-3): 155–164. (1992)
8. Fin, F., Iannario, M., Piccolo, D., Simone, R.: The effect of uncertainty on the assessment of individual performance: empirical evidence from professional soccer. Electronic Journal of Applied Statistical Analysis **10**(3): 677–692 (2017)
9. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media (2009)
10. Mathien, H.:European Soccer Database (2016) www.kaggle.com/hugomathien/soccer.
11. McHale, I. G., and Szczepański, Ł.:A mixed effects model for identifying goal scoring ability of footballers. Journal of the Royal Statistical Society: Series A. Statistics in Society **177**(2) (2014)
12. Pelechrinis, K., and Wayne W.: Positional value in soccer: Expected league points added above replacement. (2018) arXiv preprint arXiv:1807.07536
13. Vigneau, E., Qannari E.: Clustering of variables around latent components. Communications in Statistics-Simulation and Computation **32**(4):1131–1150 (2003)
14. Vigneau, E. and Chen, M.: Dimensionality reduction by clustering of variables while setting aside atypical variables. Electronic Journal of Applied Statistical Analysis **9**(1): 134–153 (2016)
15. Wibowo, C. P.: Clustering seasonal performances of soccer teams based on situational score line. Communications in Science and Technology **1**(1) (2016)

A framework of cumulated chi-squared type statistics for ordered correspondence analysis. New tools and properties.

Uno schema di statistiche di tipo chi-quadrato cumulato per l'analisi della corrispondenze ordinali. Nuovi strumenti e proprietà.

Antonello D'Ambra, Pietro Amenta and Luigi D'Ambra

Abstract Aim of this contribution is twofold: 1) to show the links between the Cramer-von Mises and the Taguchi statistics with other indices in a single framework, and 2) to suggest a family of correspondence analyses coping with ordinal variables and based on these statistics, providing the basic rationale for an easier single software implementation of all of them.
Abstract *Obiettivo di questo lavoro di mostrare i collegamenti tra la famiglia delle statistiche di Cramer-von Mises e quello di Taguchi con altri indici e, al contempo, di suggerire un framework di analisi della corrispondenze in presenza di variabili ordinali e basate su queste statistiche.*

Key words: Ordinal data, Cramer-von Mises statistics, Taguchi's statistic, Correspondence analysis of cumulative frequencies.

1 Introduction and notation

The Cramer-von Mises family of goodness-of-fit statistics is a well-known group of statistics used to test fit to a continuous distribution. For a single sample of grouped data, this family has been extended to provide tests for discrete distribution when cell probabilities are known [5]. They are based on the empirical distribution func-

Antonello D'Ambra
Department of Economics, University of Campania "Vanvitelli", Italy, Corso Gran Priorato di Malta, Capua, Italy, e-mail: antonello.dambra@unicampania.it

Pietro Amenta
Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Italy, Piazza Arechi II, Benevento, e-mail: amenta@unisannio.it

Luigi D'Ambra
Department of Economics, Management and Institutions, University of Naples "Federico II", Italy, e-mail: dambra@unina.it

tion of the sample, likewise the Kolmogorov-Smirnov statistic [4]. These statistics have been extended to the case where parameters must be estimated from the sample. Extensions of these statistics to test that I samples have the same distribution for grouped data have been also introduced [11]. The Taguchi's statistic [16] is instead a cumulative-sum statistic obtained by assigning a weight to each term that is inversely proportional to its conditional expectation under the null hypothesis of independence. See [6] for its main properties. It has been suggested as a simple alternative to Pearson's statistic for ordered contingency tables. Indeed, the Pearson's statistic can perform poorly when dealing with ordinal structures because it does not have a good power against ordered alternatives [1]. We point out that an interesting difference between the Cramer-von Mises statistics and the Pearson's statistic X^2 is that the formers, as well as the Taguchi's statistic, depend on the order of the cells.

The aim of this paper is twofold: On the one hand, the links between the Cramer-von Mises and the Taguchi statistics with other indices are shown in a single framework and, on the other hand, to suggest a family of correspondence analyses (CA) coping with ordinal variables and based on these statistics. These frameworks can aid the user to better understand the methodological differences between these approaches providing the basic rationale for an easier single software implementation.

Consider a two-way contingency table $\mathbf{N}$ that cross-classifies n statistical units according to I row categories and J ordered column categories of two variables X and Y. Let n_{ij} be the number of observations for the i-th row ($i = 1,\ldots,I$) and the j-th category ($j = 1\ldots,J$). Let $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ be the number of observations of the i-th row (sample) of $\mathbf{N}$. We denote by p_{ij} the proportion of observations that fall in the i-th row and j-th column of the table and the generic element of matrix $\mathbf{P}$. Therefore, denote $\mathbf{D}_I$ and $\mathbf{D}_J$ to be the diagonal matrices of the row and column marginal proportions $p_{i\bullet}$ and $p_{\bullet j}$, respectively, where $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$, $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ and p_j is the probability that an original observation falls into cell j. $p_{\bullet s} = \sum_{j=1}^{s} p_{\bullet j}$ is then the cumulative total of Y evaluated in s ($s = 1,\ldots,J-1$). We consider the null hypothesis (H_0) that p_j is the same for all the I samples or, equivalently, these samples have the same parent population. Under H_0, define $T_{is} = n_{i\bullet} p_{\bullet s}$ and let $Z_{is} = \sum_{j=1}^{s} n_{ij}$ and $Z_{\bullet s} = \sum_{j=1}^{s} n_{\bullet j}$ to be the cumulative count and the cumulative column total up to the s-th column category, respectively. Finally, for a given sample, p_j and $p_{\bullet s}$ can be estimated by using $\hat{p}_j = n_{\bullet j}/n$ and $d_s = Z_{\bullet s}/n$, respectively.

2 A framework of Correspondence Analyses based on a class of weighted cumulative chi squared type statistics

2.1 A class of weighted cumulative chi squared type statistics

Let consider the Kolmogorov-Smirnov (KS) statistic $KS = \sup_x |F_n(x) - F(x)|$ for a given cumulative distribution function $F(x)$ with $F_n(x)$ cumulative empirical distribution function for n i.i.d. ordered observations X_i. It is well known that it can be

modified to serve as a goodness of fit statistic. In a *KS* perspective [4], consider then the following statistic

$$T_{CCS}^{KS} = \sum_{s=1}^{J-1} w_s \sum_{i=1}^{I} \frac{1}{n_{i\bullet}} S_{is}^2 \tag{1}$$

where $S_{is} = Z_{is} - T_{is}$ is the difference between the empirical distribution function of the observations Z_{is} and the cumulative distribution function T_{is} of Y, both evaluated in s. It is possible to show that T_{CCS}^{KS} subsumes several indices according to suitable values of w_s. Indeed, under the hypothesis $H_0 : S_{1s} = \ldots = S_{Is}$ and since $n_{i\bullet}^{-1} S_{is}^2 = n_{i\bullet} (Z_{is}/n_{i\bullet} - d_s)^2$ then we get from T_{CCS}^{KS} the following statistics

w_s	Statistic	Author
$1/J$	$T_N = \sum_{i=1}^{I} n_{i\bullet} \sum_{s=1}^{J-1} (Z_{is}/n_{i\bullet} - d_s)^2/J$	Nair [13]
$1/[d_s(1-d_s)]$	$T = \sum_{i=1}^{I} n_{i\bullet} \sum_{s=1}^{J-1} (Z_{is}/n_{i\bullet} - d_s)^2/[d_s(1-d_s)]$	Taguchi [16]
$\hat{p}_s/[d_s(1-d_s)]$	$A_I^2 = \sum_{i=1}^{I} n_{i\bullet}^{-1} \sum_{s=1}^{J-1} S_{is}^2 \hat{p}_s/[d_s(1-d_s)]$	Anderson-Darling [15]
$\hat{p}_s$	$W_I^2 = \sum_{i=1}^{I} \sum_{s=1}^{J-1} n_{i\bullet} S_{is}^2 \hat{p}_s$	Cramér-von Mises [11]

We point out that the Anderson Darling statistic A_I^2 and the discrete I sample Cramér-Von Mises statistic amount also to $A_I^2 = \sum_{i=1}^{I} n_{i\bullet} \sum_{s=1}^{J-1} (Z_{is}/n_{i\bullet} - d_s)^2 \hat{p}_s/[d_s(1-d_s)]$ and $W_I^2 = \sum_{i=1}^{I} n_{i\bullet} \sum_{s=1}^{J-1} (Z_{is}/n_{i\bullet} - d_s)^2 \hat{p}_s$, respectively. See [15] for further info on the statistic A_I^2.

Let us introduce the new statistic $\tilde{T} = \sum_{s=1}^{J-1} [w_s d_s (1-d_s)] X_s^2$ where X_s^2 is the Pearson's chi-squared statistic for the $I \times 2$ contingency sub-tables $\mathbf{N}_s$ obtained by aggregating the first s column categories ($y_{(1:s)}$) and the remaining categories ($s+1$) to J ($y_{(s+1:J)}$) of table $\mathbf{N}$, respectively. It is possible show that if $w_s = [d_s(1-d_s)]^{-1}$ then $\tilde{T}$ amounts to the Taguchi's statistic T. Nair [13] refers to T as the *cumulative chi-squared statistic* (CCS) because $T = \sum_{s=1}^{J-1} X_s^2$ and introduces then the class of CCS-type tests $T_{CCS} = \sum_{s=1}^{J-1} w_s \sum_{i=1}^{I} n_{i\bullet} (Z_{is}/n_{i\bullet} - d_s)^2$. Since it is possible to show that $X_s^2 = \sum_{i=1}^{I} n_{i\bullet} (Z_{is}/n_{i\bullet} - d_s)^2/[d_s(1-d_s)] = \sum_{i=1}^{I} n_{i\bullet}^{-1} S_{is}^2/[d_s(1-d_s)]$, then statistic $\tilde{T} = \sum_{s=1}^{J-1} [w_s d_s (1-d_s)] X_s^2$ amounts to (1) and so T_{CCS}^{KS} can be also considered as a class of CCS-type statistics (likewise T_{CCS}) because it is a weighted cumulative chi-squared statistics subsuming T, T_N, A_I^2 and W_I^2.

We point out that we can replace the X_s^2 statistics in T_{CCS}^{KS} by likelihood ratio statistics G_s^2 obtaining $T_{CCS}^{KS} \approx \sum_{s=1}^{J-1} [w_s d_s (1-d_s)] G_s^2$. This formulation allows then to achieve approximations of statistics W_I^2, A_I^2 and T_N based on G_s^2. Indeed, Taguchi's statistic can be viewed as a approximate sum of likelihood ratios (with $w_s = 1/[d_s(1-d_s)]$) within the regression model for binary dependent variables following a scaled binomial distribution [6].

We introduce now an alternative interpretation of (1). Indeed, we can rewrite it as $T_{CCS}^{KS} = \frac{n}{n-1} \sum_{s=1}^{J-1} [w_s d_s (1-d_s)] \times C_s$ where C_s is the statistic $C = [(n-1)(J-1)/(n - n^{-1} \sum_{j=1}^{i} n_{\bullet j}^2)] \tau$ [10] computed on the s-th $(I \times 2)$ contingency table obtained by aggregating the first s column categories and the remaining $(s+1)$ to J of $\mathbf{N}$, and $\tau = [\sum_{i=1}^{I} \sum_{j=1}^{J} p_{i\bullet} (p_{ij}/p_{i\bullet} - p_{\bullet j})^2]/(1 - \sum_{j=1}^{J} p_{\bullet j}^2)$ is the Goodman-Kruskal

index [8], which is identical to Light and Margolin's R^2 index [10], and so it has a proportion of explained variation interpretation. Under the zero predictability hypothesis (i.e. $H_0 : p_{ij}/p_{i\bullet} = p_{\bullet j}$) the C-statistic is asymptotically chi-squared distribution with $(I-1)(J-1)$ degree of freedom [10]. This new formulation of T_{CCS}^{KS} highlights that it reflects also a unidirectional association between the categorical variables. Moreover, the numerator of τ index (N_τ) is at hearth of a Non Symmetrical extension of Correspondence Analysis (NSCA) [7]. Also note that $T_{CCS}^{KS} = n\sum_{s=1}^{J-1}[w_s d_s(1-d_s)]\tau_s$.

Let now consider the weighted Leti's index $\bar{D}/2 = \sum_{s=1}^{J-1} w_s p_{\bullet\underline{s}}(1-p_{\bullet s})$ [6] which subsumes the τ denominator with $w_s = 1$. We can decompose $\bar{D}/2$ according to the well-known principle of Between- and Within-group variance decomposition (noted D_B and D_W) of a quantitative variable such that

$$\frac{\bar{D}}{2} = \overbrace{\sum_{i=1}^{I} p_{i\bullet} \sum_{s=1}^{J-1} w_s \frac{p_{is}}{p_{i\bullet}} \left(1 - \frac{p_{is}}{p_{i\bullet}}\right)}^{=D_W} + \overbrace{\underbrace{\sum_{s=1}^{J-1} w_s \sum_{i=1}^{I} \frac{1}{n_{i\bullet}} S_{is}}_{=T_{CCS}^{KS}}}^{=D_B}.$$

T_{CCS}^{KS} then turns out to be a part of the $\bar{D}/2$ index. This result allows to obtain also W_I^2, A_I^2, T and T_N as parts of $\bar{D}/2$ by setting suitable values for w_s (Fig. 1). Moreover, a normalized dependence index for ordinal variables can then be easily defined: $\delta = 2T_{CCS}^{KS}/\bar{D}$. It can be considered as an extension of the τ index for ordinal categorical variables which subsumes several other indices according to suitable w_s values. Its main properties will be highlighted in the extended version of this paper.

According to Nair [12, 13] and under the multinomial law hypothesis for the rows, it is possible to find the asymptotic distribution of T_{CCS}^{KS} that takes into account the highlighted alternative interpretation of (1) based on the statistic C. It is based on a matrix decomposition of T_{CCS}^{KS} into orthogonal components. Let $\mathbf{A} = \mathbf{M} - (\mathbf{d}_{J-1} \times \mathbf{1}^T)$ be a matrix of order $((J-1)\times J)$ where $\mathbf{M}$ is a $((J-1)\times J)$ lower unitriangular matrix and $\mathbf{d}_{J-1} = [d_1, \ldots, d_{J-1}]^T$. The T_{CCS}^{KS} statistic can be also written as $T_{CCS}^{KS} = n \times \mathrm{tr}(\mathbf{D}_I^{1/2}\mathbf{H}\mathbf{A}^T\mathbf{W}\mathbf{A}\mathbf{H}^T\mathbf{D}_I^{1/2})$ where $\mathbf{W}$ is the $((J-1)\times(J-1))$ diagonal matrix of weights w_s and $\mathbf{H} = \boldsymbol{\Xi}_{\mathbf{1}/\mathbf{D}_I}^{\perp}\mathbf{D}_I^{-1}\mathbf{P}$ with $\boldsymbol{\Xi}_{\mathbf{1}/\mathbf{D}_I}^{\perp} = \mathbf{I}_I - \boldsymbol{\Xi}_{\mathbf{1}/\mathbf{D}_I} = \mathbf{I}_I - \mathbf{1}(\mathbf{1}^T\mathbf{D}_I\mathbf{1})^{-1}\mathbf{1}^T\mathbf{D}_I$ that eliminates the row marginal effects from the entire relationship between rows and columns, and $\mathbf{I}_I$ identity matrix of order I. Let $\boldsymbol{\Gamma}$ be the $(J-1)\times(J-1)$ diagonal matrix of the nonzero eigenvalues γ of $\mathbf{A}^T\mathbf{W}\mathbf{A}$ and $\tilde{\mathbf{Q}}$ be the $J\times(J-1)$ matrix of the associated eigenvectors such that $\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}} = \mathbf{I}$ with $\mathrm{SVD}(\mathbf{A}^T\mathbf{W}^{\frac{1}{2}}) \Rightarrow \mathbf{A}^T\mathbf{W}^{\frac{1}{2}} = \tilde{\mathbf{Q}}\boldsymbol{\Lambda}\tilde{\mathbf{S}}^T$ and $\tilde{\mathbf{S}}^T\tilde{\mathbf{S}} = \mathbf{I}$. Using the Satterthwaite's two-moment approximation [14], it is possible to show that the asymptotic distribution of the statistic $\tilde{T}_{CCS}^{KS} = (k/n)T_{CCS}^{KS}$, with $k = [(n-1)(J-1)/(n-n^{-1}\sum_{j=1}^{i} n_{\bullet j}^2)]$, can be approximated by $d(I-1)\times\chi_{(\nu)}^2$ with $\nu = d^{-1}\sum_{s=1}^{J-1}\gamma_s$ degrees of freedom and $d = (I-1)^{-1}\sum_{s=1}^{J-1}\gamma_s^{\,2}/\sum_{s=1}^{J-1}\gamma_s$.

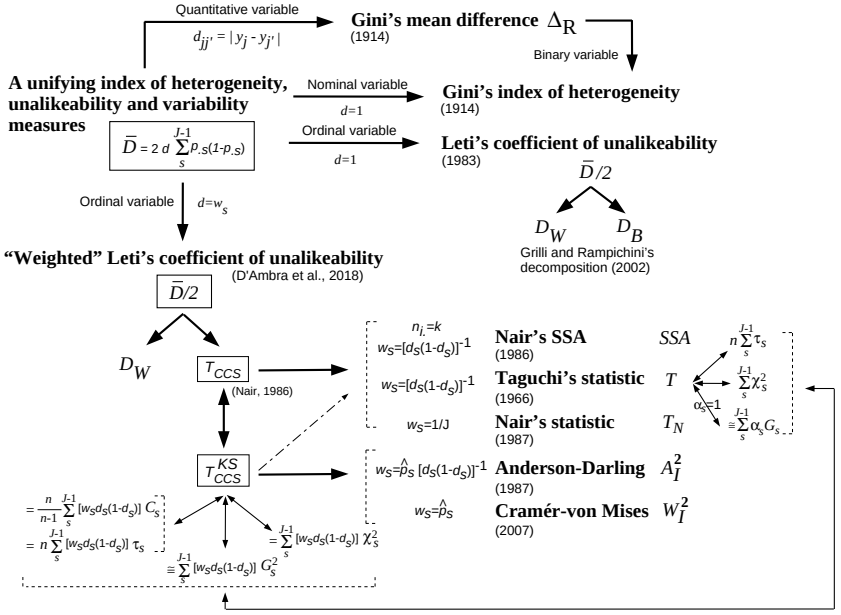


Fig. 1 A recapitulative scheme.

2.2 A family of cumulative correspondence analyses

It is possible to introduce a unified framework of correspondence analyses (CA) coping with ordinal variables and based on the T_{CCS}^{KS} decomposition. Their main goals will be to show how similar cumulative categories are with respect to nominal ones from a graphical point of view, according to a preselected association index.

Let $\tilde{\mathbf{I}}$ be an identity matrix without the last row of order $((J-1)\times J)$. Consider the SVD of matrix $\mathbf{F}=\mathbf{D}_I^{-\frac{1}{2}}(\mathbf{P}-\mathbf{D}_I\mathbf{1}\mathbf{1}^T\mathbf{D}_J)\mathbf{D}_J^{-\frac{1}{2}}\mathbf{R}^T\mathbf{W}^{\frac{1}{2}}$ such that $\mathbf{F}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$ with $\mathbf{U}^T\mathbf{U}=\mathbf{I}$, $\mathbf{V}^T\mathbf{V}=\mathbf{I}$, $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\ldots,\lambda_{min(I,J-1)})$ where λ_i are the singular values of $\mathbf{F}$ arranged in descending order. It is possible to show that $T_{CCS}^{KS}/n=\sum_{i=1}^{min(I,J-1)}\lambda_i$. If the generic element of $\mathbf{W}$ is given by $w_s=1/[d_s(1-d_s)]$, then we achieve a decomposition of the Taguchi's statistic which is at heart of a cumulative extension of correspondence analysis (TA) [2]. Then we can easily get the decomposition of other indices according to figure 1. Table 1 shows the suitable choices of the elements of matrices $\mathbf{W},\mathbf{R}$ and $\mathbf{A}_J$ in order to obtain a CA of cumulative frequencies using a decomposition of a specific statistic: W-CA for W_I^2, A-CA for A_I^2, TA for T and TN-CA for T_N, respectively. Non cumulative methods CA and NSCA are also shown in Table 1 where "Nom" and "Ord" stand for Nominal and Ordinal variable, respectively.

Table 1 A unified framework of CAs coping with ordinal data.

$\text{SVD}[\mathbf{D}_I^{-1/2}(\mathbf{P}-\mathbf{D}_I\mathbf{1}_I\mathbf{1}_J^T\mathbf{D}_J)\mathbf{A}_J^{-1/2}\mathbf{R}^T\mathbf{W}^{\frac{1}{2}}] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$							
Method	Row	Column	$\mathbf{R}$	$\mathbf{W} = diag(w_s)$	$\mathbf{A}_J$	Statistic	Authors
CA	Nom-Ord	Nom-Ord	$\tilde{\mathbf{I}}$	$w_s = 1$	$\mathbf{D}_J$	ϕ^2	[9, 3]
NSCA	Nom-Ord	Nom-Ord	$\tilde{\mathbf{I}}$	$w_s = 1$	$\mathbf{I}$	N_τ	[7]
TA	Nom	Ord	$\mathbf{M}$	$w_s = [d_s(1-d_s)]^{-1}$	$\mathbf{I}$	T/n	[2]
TN-CA	Nom	Ord	$\mathbf{M}$	$w_s = 1/J$	$\mathbf{I}$	T_N	NEW
W-CA	Nom	Ord	$\mathbf{M}$	$w_s = \hat{p}_s$	$\mathbf{I}$	W_I^2	NEW
A-CA	Nom	Ord	$\mathbf{M}$	$w_s = \hat{p}_s[d_s(1-d_s)]^{-1}$	$\mathbf{I}$	A_I^2	NEW

References

1. Agresti, A., 2013. Categorical Data Analysis, John Wiley & Sons
2. Beh, E.J., D'Ambra, L., Simonetti, B., 2011. Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic. Communications in Statistics. Theory and Methods 40, 1620–1632.
3. Benzécri, J.P., 1973. L'Analyse des Données. Volume II. L'Analyse des Correspondances. Paris, France: Dunod.
4. Chakravarti, I.M., Laha, R.G., Roy, J., 1967. Handbook of Methods of Applied Statistics, Volume I, John Wiley.
5. Choulakian, V., Lockhart, R.A., Stephens, M.A., 1994. Cramér- von Mises statistics for discrete distributions, The Canadian Journal of Statistics 22 (1), 125–137.
6. D'Ambra, L., Amenta, P., D'Ambra, A., 2018. Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation, Statistical Methods and Applications 27 (2), 297–318.
7. D'Ambra, L., Lauro, N., 1989. Non symmetrical analysis of three-way contingency tables. In: Coppi, R., Bolasco, S. (eds.) Multiway Data Analysis, pp. 301-315. Amsterdam: Elsevier Science Publishers B.V.
8. Goodman, L.A., Kruskal, W.H., 1954. Measures of association for cross-classifications. Journal of American Statistical Association, 49, 732–764.
9. Hirschfeld, H.O., 1935. A connection between correlation and contingency. Proc. Cambridge Philosophical Society, 31, 520–524.
10. Light, R., Margolin, B., 1971. An analysis of variance for categorical data. Journal of the American Statistical Association, 66 (335), 534–544.
11. Lockhart, R.A., Spinelli, J.J., Stephens, M.A., 2007. Cramér-von Mises statistics for discrete distributions with unknown parameters. Canadian Journal of Statistics, 35 (9), 125–133.
12. Nair, V.N., 1986. Testing in industrial experiments with ordered categorical data. Technometrics 28 (4), 283–291.
13. Nair, V.N., 1987. Chi-squared type tests for ordered alternatives in contingency tables. Journal of American Statistical Association 82, 283–291.
14. Satterthwaite, F., 1946. An approximate distribution of estimates of variance components. Biometrical Bulletin (2), 110–114.
15. Scholz, F.W., Stephens, M.A., 1987. K-sample Anderson-Darling Tests. Journal of the American Statistical Association, 82, 918–924.
16. Taguchi, G., 1974. A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. Saishin Igaku 29, 806–813.

Exploring drug consumption via an ultrametric correlation matrix

Una analisi del consumo di droghe mediante una matrice di correlazione ultrametrica

Giorgia Zaccaria and Maurizio Vichi

Abstract In many real applications, the existence of a general concept (a multidimensional phenomenon) composed of nested specific ones is often theorised. In the specialised literature, different sequential methodologies have been proposed to identify a hierarchy of latent dimensions. In this paper, we investigate drug consumption via an ultrametric correlation matrix which allows to detect different, nonoverlapping groups of drugs and their hierarchical relationships, starting from the correlation matrix of the observed data. Since its social and economic relevance, a model-based approach to drug consumption can provide an in-depth understanding of this challenging phenomenon, which turns out to be fundamental to address policies aimed at reducing it.

Abstract *In molte applicazioni l'ipotesi dell'esistenza di un concetto generale (un fenomeno multidimensionale), definito mediante concetti più specifici, è spesso avvalorata. In letteratura, molteplici metodologie di tipo sequenziale sono state proposte con lo scopo di identificare una gerarchica di dimensioni latenti. In questo articolo indaghiamo il fenomeno del consumo di droghe mediante una matrice di correlazione ultrametrica, che permette di individuare diversi, disgiunti gruppi di droghe e le loro relazioni gerarchiche, a partire dalla matrice di correlazione dei dati osservati. Data la sua rilevanza sociale ed economica, un approccio basato su modello per lo studio del consumo di droghe può fornire una conoscenza più approfondita di tale fenomeno, che a sua volta può risultare fondamentale nella definizione di politiche volte alla sua riduzione.*

Key words: Hierarchical structures, drug consumption, ultrametric correlation matrix, dimensionality reduction

---

Giorgia Zaccaria
University of Rome La Sapienza, P.le Aldo Moro 5 00185, Rome
e-mail: giorgia.zaccaria@uniroma1.it

Maurizio Vichi
University of Rome La Sapienza, P.le Aldo Moro 5 00185, Rome
e-mail: maurizio.vichi@uniroma1.it

1 Introduction

The identification of a hierarchy of nested latent concepts is a considerable aspect in the study of phenomena composed of different facets. Manifold methodologies as higher-order factor models [1, 10] and hierarchical factor models [9, 11] deal with the problem of the construction of a general latent concept via a hierarchy of more specific ones. In order to detect consistent groups of variables and their hierarchical factorial structure, [2] propose a novel exploratory, parsimonious and simultaneous model which is based upon the estimation of an ultrametric correlation matrix to reconstruct the relationships between the observed variables, i.e., within groups of variables and between them.

In many fields as the psychometric and marketing ones, the detection of latent dimensions with different relationship intensities is a crucial need for a correct and all-around understanding of the phenomenon under study, along with the dimensionality reduction of the variable space. In this paper, we demonstrate the large-scale applicability of the model proposed by [2] with its application to the phenomenon of drug consumption. The latter is one of the most challenging problems in the modern societies. Indeed, drug consumption contributes to rise the risk of poor health, crimes, social harm, environmental damage and it has become a social problem over years - especially among young people - governments have to face with. Many studies have been developed to analyse drug consumption, its individual and community effects, e.g., [8]. Therefore, an in-depth analysis of this phenomenon through the aforementioned model - which identifies groups of drugs highly correlated and their (hierarchical) relationships - can contribute to its better understanding and to consequently implement policies aimed at reducing it.

The paper is organised as follows. In Section 2 the methodology used to investigate the phenomenon under study is presented and in Section 3 it is applied on the Drug Consumption data set to stress its usefulness. Finally, in Section 4 some conclusions end the paper.

2 Methodology

The exploratory, parsimonious and simultaneous model, called *Ultrametric Correlation Model* (UCM) and proposed by [2], introduces a novel approach to the identification of hierarchical structures of latent variables (concepts). Indeed, starting from a nonnegative correlation matrix, the UCM estimates highly correlated, nonoverlapping groups of variables and different levels of relationships among them in a least-squares framework. The model is mathematically represented by an ultrametric correlation matrix, whose definition gives rise to a hierarchical structure of variable groups as detailed below. Let us consider the following correlation matrix of order J which is composed of 3 variable groups such that $J_1 + J_2 + J_3 = J$, where J_q, $q = 1,2,3$, is the number of variables in the q^{th} group (g_q).

$$\mathbf{R} = \begin{bmatrix} \begin{matrix} 1 & & \\ & 1 & r_{g_1}^{\text{Within}} \\ & & \ddots \\ & & 1 \end{matrix} & r_{g_1,g_2}^{\text{Between}} & r_{g_1,g_3}^{\text{Between}} \\ & \begin{matrix} 1 & & \\ & 1 & r_{g_2}^{\text{Within}} \\ & & \ddots \\ & & 1 \end{matrix} & r_{g_2,g_3}^{\text{Between}} \\ & & \begin{matrix} 1 & & \\ & 1 & r_{g_3}^{\text{Within}} \\ & & \ddots \\ & & 1 \end{matrix} \end{bmatrix}$$

Within each block of the matrix $\mathbf{R}$, the correlation between variables assumes the same value: $r_{g_q}^{\text{Within}}$, $q = 1,2,3$ represents the correlation within the q^{th} group, whereas $r_{g_q,g_p}^{\text{Between}}$, $q,p = 1,2,3\,(p \neq q)$, represents the correlation between the q^{th} and the p^{th} group. $\mathbf{R}$ is an ultrametric correlation matrix if it satisfies the following conditions[1]:

(i) $r_{g_q,g_p}^{\text{rel}} = r_{g_p,g_q}^{\text{rel}}$ for rel $\in \{\text{Within, Between}\}$, $p,q = 1,2,3$ (symmetry);

(ii) $r_{g_q}^{\text{Within}} \geq \max\{r_{g_q,g_p}^{\text{rel}} : q = 1,2,3\}$ for all $p = 1,2,3$, where rel = Within if $q = p$, rel = Between, otherwise (column pointwise diagonal dominance);

(iii) $r_{g_q,g_p}^{\text{rel}} \geq \min\{r_{g_q,g_h}^{\text{rel}}, r_{g_p,g_h}^{\text{rel}}\}$, for all $q,p,h = 1,2,3$ and rel $\in \{\text{Within, Between}\}$ (ultrametric inequality).

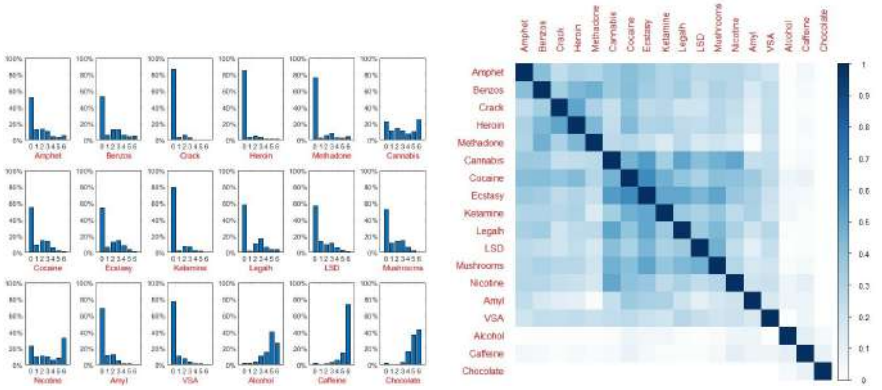
Thus, the ultrametric correlation matrix $\mathbf{R}$ of the UCM model identifies higher correlations within groups of variables and lower (and hierarchically ordered) correlations between those groups by giving rise to a hierarchy of nested latent concepts, each one associated with a variable group. The aforementioned conditions can be easily generalised to Q groups of variables.

3 Drug consumption: evaluation of the hierarchical relationships between groups of drugs

The data set analysed in the paper herein[2] [6] contains information on 1885 respondents, mainly coming from UK (55.58%), USA (29.55%), Canada (4.62%) and Australia (2.86%) and aged from 18 years old, on their drug consumption. Specifically, the use of 18 legal (alcohol, caffeine, chocolate, nicotine) and illegal (am-

[1] The definition of the ultrametric correlation matrix provided herein is based upon [5, pp.58-59].

[2] Drug consumption (quantified) data set available at: `https://archive.ics.uci.edu/ml/datasets/Drug+consumption+\%28quantified\%29`.



(a) Bar chart of the 18 ordinal variables

(b) Correlation matrix of the quantified data via the CatPCA

Fig. 1: Drug consumption data set.

phetamines, amyl nitrite, benzodiazepine, cannabis, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, Volatile Substance Abuse) drugs is investigated in terms of ordinal variables. The response classes are the following: *Never Used, Used over a Decade Ago, Used in Last Decade, Used in Last Year, Used in Last Month, Used in Last Week* and *Used in Last Day*.

In order to apply the methodology described in Section 2 to investigate the correlation structure among drugs, the ordinal variables - each one representing consumption of a specific drug - have to be quantified. This quantification is implemented via the Categorical Principal Component Analysis (CatPCA) [7] and the correlation matrix of the corresponding quantitative variables is computed. Six correlation coefficients assume negative values (not lower than[3] -0.05) which turn out to be statistically nonsignificant; whereas, the variable *Chocolate* has negative correlations with all the other drugs (Figure 1a) - except for *Alcohol* and *Caffeine* - which are not lower than -0.09 and considered nonsignificant in literature [6]. For this reason, in both cases the negative correlations are set to zero such that the non-negativity condition necessary for the UCM holds (Figure 1b). Furthermore, the number of the variable groups necessary to implement the exploratory, parsimonious model described in Section 2 is set according to the scree plot and it is equal to five. It is worthy of remark that hierarchical clustering methods could be implemented to study the correlation between usage of different drugs, but they would not guarantee the correct identification of the underlying hierarchical structure [3].

The application of the model described in Section 2 to the aforementioned data set provides a representation of drug consumption through the identification of different groups of drugs mostly correlated (Figure 1b), and broader ones defined by merging the initial five groups (Figure 2). In this framework, a model-based ap-

[3] In this case, the term *not lower than* refers to small negative correlation coefficients close to zero.

Table 1: Initial five groups identified by the Ultrametric Correlation Model.

Group	Group Name	Variables
Group 1	Depressant and Artificial Drugs	Ampeth, Benzodiazepine, Crack, Heroine, Methadone
Group 2	Stimulant Drugs and Hallucinogens	Cannabis Cocaine, Ecstasy, Ketamine, Legal highs, LSD, Mushrooms, Nicotine
Group 3	Inhalant Drugs	Amyl nitrite, Volatile Substance Abuse
Group 4	Legal Drugs of Daily Use	Alcohol, Caffeine
Group 5	Chocolate	Chocolate

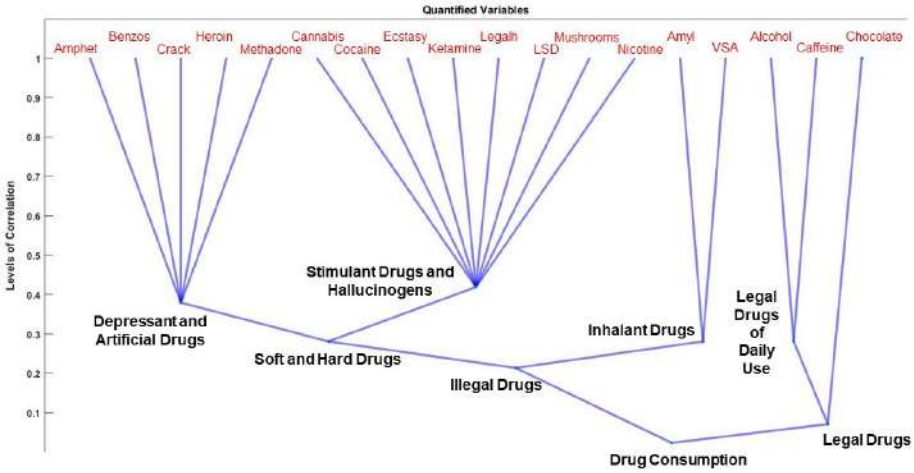


Fig. 2: Path diagram representation of the drug consumption.

proach to analyse correlations between variables can back up the experts' theories on this phenomenon. The initial five groups identified by the model are reported in Table 1. All of them are reliable according to the Cronbach's alpha (α) [4], except for *Inhalant Drugh* and *Legal Drugs of Daily Use*. It is worthy of remark that the Cronbach's alpha of a group is affected by its number of variables.

The hierarchy over the five groups gives rise to broader concepts: *Soft and Hard Drugs* obtained by lumping together Group 1 and Group 2 ($\alpha = 0.87$); *Illegal Drugs* obtained by merging the latter with Group 3 ($\alpha = 0.87$); *Legal Drugs* obtained by lumping together Group 4 and Group 5 ($\alpha < 0.7$). The existence of a general construct representing *Drug Consumption* is assessed through the Cronbach's alpha of the whole data set, which is equal to 0.84. These results turn out to be coherent with the specialised literature on drug consumption (e.g., [6]).

4 Conclusions

In many real applications, the existence of a general concept composed of nested specific ones is often theorised. Manifold sequential methodologies aim at building

a hierarchy of dimensions starting from the observed variables. [2] propose a novel parsimonious and simultaneous model in order to pinpoint latent concepts, each one associated with a variable group, and explore their hierarchical relationships by investigating the correlation matrix of the observed data. In this paper, the aforementioned model is applied to a Drug Consumption data set to study the relationships between groups of drugs. Since its social and economic relevance, a model-based approach to drug consumption analysis can provide a better understanding of the phenomenon which can be fundamental to address policies aimed at reducing it. The results of the application of the model proposed by [2] on the aforementioned data set pinpoint a hierarchy of drug groups which is coherent with the studies on drug consumption (e.g., [6]); this confirms the importance of the UCM in applications where a hierarchical factorial structure can be estimated.

References

1. Cattell, R.B.: The scientific use of factor analysis in behavioral and life sciences. Plenum, New (1978)
2. Cavicchia, C., Vichi, M., Zaccaria, G.: The ultrametric correlation matrix for modelling hierarchical latent concepts. Advances in Data Analysis and Classification, *Accepted*
3. Cavicchia, C., Vichi, M., Zaccaria, G.: Exploring Hierarchical Concepts: Theoretical and Application Comparisons. In: T. Imaizumi, A. Nakayama, S. Yokoyama (eds.) Advanced Studies in Behaviormetrics and Data Science, Springer, Singapore, ISBN: 978-981-15-2699-2, *in press* (2020)
4. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika **16**(3), 297–334 (1951)
5. Dellacherie, C., Martinez, S., San Martin, J.: Inverse M-matrices and ultrametric matrices. Springer International Publishing, Lecture Notes in Mathematics (2014)
6. Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N.: The Five Factor Model of personality and evaluation of drug consumption risk, arXiv (2015) Available at `https://arxiv.org/abs/1506.06297`
7. Gifi, A.: Nonlinear Multivariate Analysis. Wiley, New York (1990)
8. McGinnis, J.M., Foege, W.H.: Actual causes of death in the United States. JAMA **270**(18), 2207—2212 (1993)
9. Schmid, J., Leiman, J.M.: The development of hierarchical factorial solutions. Psychometrika **22**(1), 53–61 (1957)
10. Thompson, G.H.: The factorial analysis of human ability. Houghton Mifflin, New York (1948)
11. Wherry, R.J.: Hierarchical factorial solutions without rotation. Psychometrika **24**(1), 45–51 (1959)

Ranking extraction in ordinal multi-indicator systems

Costruzione di ranking in sistemi multidimensionali di indicatori ordinali

Marco Fattore and Alberto Arcagni

Abstract In this paper, we present a procedure for scoring and ranking statistical units in ordinal multi-indicator systems, by integrating classical dimensionality reduction tools and novel results in Partial Order Theory. Units are ranked based on "dominance" scores, which depend upon both the structure of the partial order and the joint frequency distribution. Dominance scores are complemented with scores of incomparability among units, so to assess the ranking quality. The procedure is computationally light and is here applied to data about financial literacy in Italy.
Abstract In questo articolo, presentiamo una procedura per la costruzione di ranking di unità statistiche, valutate su sistemi multidimensionali di variabili ordinali. La procedura integra algoritmi di riduzione della dimensionalità e recenti risultati della Teoria degli Ordinamenti Parziali, e ordina le unità in base a punteggi di "dominanza", che dipendono dalla struttura dell'ordinamento parziale e dalla distribuzione di frequenze congiunte. Al punteggio di dominanza è affiancato un punteggio di "incomparabilità", per valutare la qualità del ranking. La procedura è computazionalmente leggera ed è qui esemplificata su dati relativi alla competenza finanziaria in Italia.

Key words: Financial literacy; Ordinal data; Partial order; Poset; Ranking.

Marco Fattore
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1 - 20126 - Milano e-mail: marco.fattore@unimib.it

Alberto Arcagni
Sapienza University of Rome, Piazzale Aldo Moro 5 - 00185 - Roma e-mail: alberto.arcagni@uniroma1.it

1 Introduction

Let $\boldsymbol{v}_1,\ldots\boldsymbol{v}_k$ be a system of k ordinal variables, on scales with $m_1,\ldots m_k$ degrees, recorded on n statistical units. Each unit u is associated a score profile $\boldsymbol{u} = (u_1,\ldots,u_k)$, composed of variable scores, among the $m = m_1 \cdot \ldots \cdot m_k$ possible different score configurations. Since the input variables are ordinal, such m score configurations can be partially ordered according to the *product order* rule [1, 2, 4], i.e. given two profiles $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ik})$ and $\boldsymbol{x}_j = (x_{j1},\ldots,x_{jk})$, we put $\boldsymbol{x}_i < \boldsymbol{x}_j$ if and only if $x_{ih} \leq x_{jh} \forall h = 1,\ldots,k$ and there exists an index z such that $x_{iz} < x_{jz}$. Many units can share the same profile and so we are naturally led to consider distributions on product orders, as the typical data structure arising from assessing statistical populations against ordinal multi-indicator systems. The classical problem of ranking units based on their multidimensional profiles is thus the problem of reducing the dimensionality of frequency distributions defined on product orders. In the following, we propose an algorithm to solve this issue.

2 The lexicographic dominance and incomparability matrices

As any finite partially ordered set, a product order π over k ordinal variables can be uniquely reconstructed by means of its set of *linear extensions*, i.e. of linear orders obtained from π, by ordering all non-comparable configuration pairs, in all possible ways. As shown in [6], however, being a product order π can also be reconstructed by using the subset of so-called *lexicographic linear extensions* (LLEs). These are linear orders where score configurations are ordered in "alphabetic fashion", with respect to all possible permutations of the k input variables; thus, there are $k!$ LLEs and these suffice to generate the product order associated to $\boldsymbol{v}_1,\ldots\boldsymbol{v}_k$. LLEs are the key tool for scoring and ranking units, as they bring information on the relative positions of score configurations, in the product order.

Informally speaking, to rank units we want to associate to each of them a dominance score, reflecting to what degree a unit is placed "higher" than the others, in π. As a first step to this goal, we compute the dominance score P_{ij} between two profiles $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, as the fraction of LLEs in which $\boldsymbol{x}_i$ is placed below $\boldsymbol{x}_j$. Scores P_{ij} are called *lexicographic mutual ranking probabilities* (LMRPs) and can be computed analitically, as shown in the following proposition.

Proposition. Let $\boldsymbol{v}_1,\ldots\boldsymbol{v}_k$ be a system of k ordinal variables, on scales with $m_1,\ldots m_k$ degrees, and let $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ik})$ and $\boldsymbol{x}_j = (x_{j1},\ldots,x_{jk})$ be two profiles on it. Let k_1 be the number of indices h such that $x_{ih} < x_{jh}$ and let k_2 be the number of indices h such that $x_{ih} = x_{jh}$. Then:

$$P_{ij} = \frac{k_1}{k} \sum_{s=0}^{k_2} \frac{k_2!}{(k_2-s)!} \frac{(k-s-1)!}{(k-1)!}. \tag{1}$$

Proof. Since the set of LLEs is in bijective correspondence with that of all permutations of variables, we can identify a LLE λ with a permutation $l_1,\ldots,l_k$ of $\{1,\ldots,k\}$. With this notation, the subset G of LLEs such that $\boldsymbol{x}_i < \boldsymbol{x}_j$, can be partitioned as $G = G_0 \cup \ldots \cup G_{k_2}$, where sets G_s and their cardinalities are given by:

1. $G_0 = \{\lambda \,|\, x_{il_1} < x_{jl_1}\} \rightarrow |G_0| = k_1 \cdot (k-1)!$
2. $G_1 = \{\lambda \,|\, x_{il_1} = x_{jl_1}, x_{il_2} < x_{jl_2}\} \rightarrow |G_1| = k_2 \cdot k_1 \cdot (k-2)!$
3. $G_2 = \{\lambda \,|\, x_{il_1} = x_{jl_1}, x_{il_2} = x_{jl_2}, x_{il_3} < x_{jl_3}\} \rightarrow |G_2| = k_2 \cdot (k_2-1) \cdot k_1 \cdot (k-3)!$
4. $\ldots$
5. $G_{k_2} = \{\lambda \,|\, x_{il_1} = x_{jl_1}, \ldots, x_{il_{k_2}} = x_{jl_{k_2}}, x_{il_{k_2+1}} < x_{jl_{k_2+1}}\} \rightarrow |G_{k_2}| = k_2 \cdot (k_2-1) \cdot \ldots \cdot 1 \cdot k_1 \cdot (k-k_2-1)! = k_2! k_1 \cdot (k-k_2-1)!$.

Summing up all of the cardinalities and dividing by the total number of LLEs ($k!$), we get

$$P_{ij} = P(\boldsymbol{x}_i < \boldsymbol{x}_j) = \frac{|G_0| + \ldots + |G_{k_2}|}{k!} = \frac{k_1}{k}\Big[1 + \frac{k_2}{k-1} + \frac{k_2(k_2-1)}{(k-1)(k-2)} + \\ + \frac{k_2(k_2-1)(k_2-2)}{(k-1)(k-2)(k-3)} + \ldots + \frac{k_2!}{(k-1)(k-2)(k-3)\ldots(k-k_2)}\Big] = \\ = \frac{k_1}{k}\sum_{s=0}^{k_2}\frac{k_2!}{(k_2-s)!}\frac{(k-s-1)!}{(k-1)!}.$$

q.e.d.

LMRPs are used as entries of the *lexicographic dominance matrix* P, which is later used to derive the final dominance scores associated to units. Making all profiles comparable, by means of dominance scores, is unavoidably forcing and stretches the data structure, since it destroys the incomparabilities existing in π. In order to keep control on this kind of "distorsion", we complement matrix P with a *lexicographic incomparability matrix* Q, whose entries are defined as $Q_{ij} = \min(P_{ij}, P_{ji})$ $(i \neq j)$ and $Q_{ii} = 0$. The rationale behind this definition is straightforward: the more one profile dominates the other, the less the two are incomparable (the choice of putting the diagonal elements of Q equal to 0 is to assure the maximum incomparability degree between two profiles to be 0.5: this choice has no essential consequences on the subsequent developments). By the above definition, it follows immediately that Q is symmetric.

3 The scoring functions

As in [7], we assign global dominance (*dom*) and incomparability (*inc*) scores to each profile in the input poset, based on the singular value decompositions of P and Q. However, in order to take into account the distribution of units on π, here we weight P and Q with the relative frequencies over poset profiles. Let s_i be the share

of statistical units with profile $\boldsymbol{x}_i$ and let $S = diag(s_1, \ldots, s_m)$. The i-th column of SP comprises the LMRPs of $\boldsymbol{x}_i$, with respect to all of poset profiles, multiplied by the relative frequency corresponding to each of them. In other words, it comprises the probabilities of randomly and independently extracting a statistical unit u and a LLE λ, such that the profile of u is $\boldsymbol{x}_j$ ($j = 1, \ldots, m$) and $\boldsymbol{x}_j$ is not above $\boldsymbol{x}_i$, in λ. Now, let $SP = UD_PV^T$ be the singular value decomposition of matrix SP. From $D_PV^T = U^TSP$, we see that the i-th component dom_i of the first row of D_PV^T is a weighted sum (with non-negative weights, by the Perron-Frobenius Theorem) of such probabilities and can be then interpreted as an overall dominance score of a unit with profile $\boldsymbol{x}_i$, over the other units. As proved in [7][1], the map associating to $\boldsymbol{x}_i$ the dominance score dom_i is *strictly-order preserving* and can serve as a *dominance score function*. Similarly, we take, as global incomparability degrees associated to poset profiles, the components of the first row of D_QB^T, in the singular value decomposition $SQ = AD_QB^T$. In summary, to each profile $\boldsymbol{x}_i$ of the input poset there corresponds a pair (inc_i, dom_i) of incomparability-dominance scores, computed as the i-th components of the first rows of D_QB^T and D_PV^T, respectively. Finally, each statistical unit in the dataset inherits the score pair of its profile and can be mapped into an incomparability-dominance plane, getting both a ranking and a picture of the stretching of the data.

4 Ranking financial literacy in Italy

We have applied the procedure outlined above to data pertaining to financial literacy of Italian adults, collected by Bank of Italy for year 2018 [3]. The survey involved about 2500 respondents, investigating financial knowledge by means of 7 questions, pertaining to different financial notions[2], whose answers are simply classified as *right* or *wrong*. The resulting product order is called 2^7 and has in fact $2^7 = 128$ binary profiles, with $7! = 5040$ lexicographic linear extensions. Dominance and incomparability matrices have been built weighting the LMRPs by the share of respondents in each knowledge profile, also considering the sample weight of each unit. By performing singular value decompositions, each unit has been finally associated the incomparability-dominance score pair. Figure 1 and 2 reproduce the distributions of such pairs for age classes 18-50 and 51+, spatially smoothed to give a more realistic impression of the score patterns. The two plots are scaled to 0-1 on the dominance axis, and the incomparability axis is scaled to 0-0.3 instead due to the small values observed. Units answering correctly all of the questions are scored 1 on the dominance axis, while those providing all wrong answers are scored 0. The value 1 on the incomparability axis corresponds to the theoretical maximum incomparabilty score of a statistical unit, over all possible distributions on the input poset

[1] The pre-multiplication by S does not affect the proof, since all elements of a row are multiplied by the same non-negative quantity.

[2] Questions check knowledge about *inflation*, *simple/compound interest* and *risk management*.

2^7. Consequentially, the small values reported on the incomparability axis mean that the respondents can be reasonably ranked since most of the observed profiles are "almost" comparable. In both age classes the density tends to be concentrated on middle-hìgh dominance scores (more black areas), with evidences of some polarization and the existence of a group of financially illiterate units, separated by the rest of the distribution. Interestingly, older people seem to have more financial competencies than younger subjects; given the increasing relevance of financial products and services for personal lifelong economic sustainability, this appears as a critical feature of the Italian financial literacy distribution. Finally, notice that having similar incomparability degrees need not imply having similar answer profiles, so that care must be taken in interpreting the plot as revealing homogeneous respondent clusters.

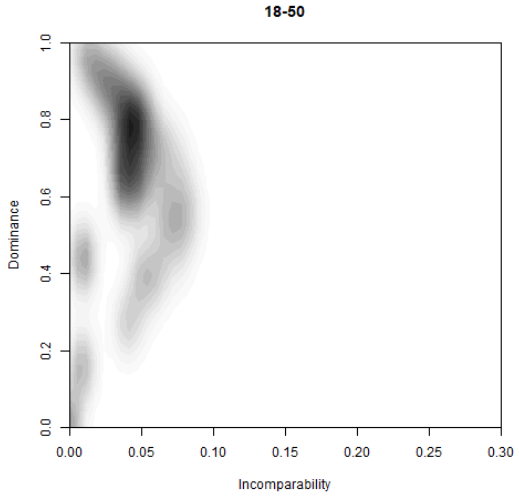


Fig. 1 Smoothed distribution of dominance and incomparability scores related to financial literacy in Italy, for age 18-50. Grey intensity is proportional to frequency density.

5 Conclusions

In this short paper, we have presented a ranking procedure for ordinal multi-indicator systems, which: (i) fully respects the ordinal nature of the data and does not involve any scaling or aggregation of ordinal variables, (ii) is computationally light and can be applied to datasets with a quite large number of variables (possibly, using approximations to the factorial in the computation of LMRPs) and (iii) also provides measures of incomparability among units, so as to keep control on the

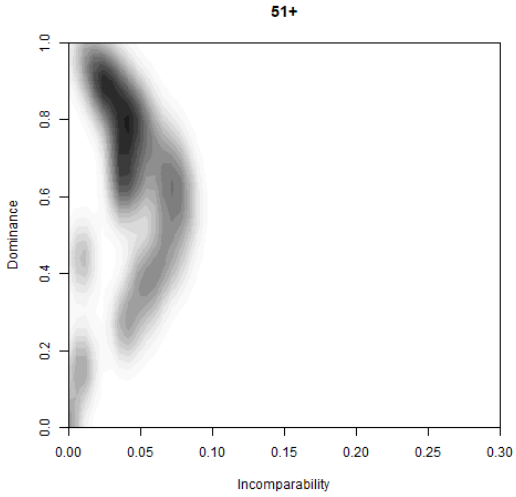


Fig. 2 Smoothed distribution of dominance and incomparability scores related to financial literacy in Italy, for age 51+. Grey intensity is proportional to frequency density.

quality of the ranking process which, as any dimensionality reduction algorithm, unavoidably introduces some distortion into the final data representation. The procedure is freely available in the R package Parsec [5].

References

1. Rainer Brüggemann and Ganapati P Patil. Multicriteria prioritization and partial order in environmental sciences. *Environmental and Ecological Statistics*, 17(4):383–410, 2010.
2. Brian A Davey and Hilary A Priestley. *Introduction to lattices and order*. Cambridge University press, 2002.
3. Antonietta Di Salvatore, Francesco Franceschi, Andrea Neri, and Francesca Zanichelli. Measuring the financial literacy of the adult population: the experience of Banca d'Italia, in "Questioni di Economia e Finanza (Occasional Papers)", n. 435. *Economic Research and International Relations Area*, 2018.
4. Marco Fattore. Partially ordered sets and the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, 128(2):835–858, 2016.
5. Marco Fattore and Alberto Arcagni. Parsec: an R package for poset-based evaluation of multidimensional poverty. In *Multi-indicator systems and modelling in partial order*, pages 317–330. Springer, 2014.
6. Marco Fattore and Alberto Arcagni. A reduced posetic approach to the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, 136(3):1053–1070, 2018.
7. Marco Fattore, Alberto Arcagni, and Filomena Maggino. Optimal scoring of partially ordered data, with an application to the ranking of smart cities. In Giuseppe Arbia, Stefano Peluso, Alessia Pini, and Giulia Rivellini, editors, *Smart Statistics for Smart Applications - Book of Short Papers SIS 2019*, pages 855–860. Pearson, 2019.

Gender statistics

Gender differences in Italian STEM degree courses: a discrete-time competing-risks model

Differenze di genere nelle lauree STEM: un modello a rischi competitivi a tempi discreti

Marco Enea and Massimo Attanasio

Abstract In this paper, we highlight possible gender differences in performance of students attending Italian university STEM degree courses, controlling for socio-demographic and territorial variables. We used the micro-data from the Italian Ministry of Education, Universities and Research database of the 2011-12 freshmen cohort enrolled at an Italian bachelor university degree course. The analysis was dealt with by fitting a discrete-time competing-risks model. Bachelor degree graduation, dropout, and change to non-STEM courses were considered as competing events. The aim is to highlight gender gaps in the student performance, and the STEM courses where this gap is more critical.

Abstract *In questo articolo mettiamo in luce le differenze di genere nelle prestazioni degli studenti italiani che frequentano un corso di laurea STEM, controllando per variabili socio-economiche e territoriali. Abbiamo usato i micro dati del database del Ministero dell'istruzione, dell'universitá e della ricerca della coorte 2011/12 di immatricolati in un corso di laurea triennale. L'analisi é stata effettuata adattando un modello a rischi competitivi e a tempi discreti. Il conseguimento della laurea triennale, l'abbandono e il passaggio di corso verso un non STEM sono stati considerati come eventi competitivi. L'obiettivo é quello di mettere in luce differenze di genere nella performance degli studenti e il corso di laurea STEM in cui tale gap é piú critico.*

Key words: Gender differences, student performances, STEM degree courses

Marco Enea
Department of Health Promotion, Mother and Child Care, Internal Medicine and Medical Specialties - University of Palermo e-mail: marco.enea@unipa.it

Massimo Attanasio
Dipartimento di Scienze Economiche, Aziendali e Statistiche - University of Palermo e-mail: massimo.attanasio@unipa.it

1 Introduction

STEM (Science, Technology, Engineering and Mathematics) is the acronym commonly used to indicate a group of disciplines related to the scientific and technological fields. The Program for International Student Assessment (PISA), an international survey promoted by OECD countries with the objective to periodically assess the levels of education of teenagers of the most industrialized countries, highlights that in most countries, boys outperform girls in science and mathematics. This gap was recently confirmed for the Italian case [1, 2].

In this paper, we investigate about the gender gap in STEM at university following a different approach from PISA survey. PISA consists of scores given to assessment tests, while our research compares gender performances in Italian university courses through dropout and degree completion rates. Our research questions are: is it true that males outperform females at university in STEM courses? Is it true for all STEM courses?

2 Data and preliminary analysis

The data we analyzed were directly provided by the Italian Ministry of Education, Universities and Research (MIUR) through an agreement between the Ministry and some Italian universities [3]. The database consists of around 280.000 records for each cohort from 2008-09 to present. The records are longitudinal micro-level data reporting the students career since his/her enrolment. For the current analysis, we have chosen the 2011-12 freshmen cohort which allows a longer follow-up. We have restricted our analysis to the STEM bachelor (BA) degree courses, excluding the online degree courses, for a total of 61982 freshmen. Because there is not an exact correspondence between 3-year STEM degree courses for all countries, we considered our STEM classification as listed in Table 1, were the percentage of males and females is also reported, as well as the one of the 2015/16 cohort, to have a comparison with a more recent cohort. In that list, we included all the basic scientific courses (following the "Piano Nazionale Lauree Scientific" classification) plus the engineering courses. Among the STEM courses of the 2015/16 cohort, females are under-represented in computer science, engineering, physics, geology and statistics, while they are almost equally represented in chemistry and mathematics. Instead, biology, biotechnology, and environmental and natural sciences are STEM courses were females are numerically more represented than males. The larger difference between the two cohorts concerns a seven points decrease of females enrolments in physics. To detect the effect of socio-demographics and territorial factors, a discrete-time competing-risks model was fitted. Bachelor degree graduation, dropout, and change to non-STEM courses were used as outcomes. Table 2 reports the students' follow up flows by gender and type of competing event. At the end of the first A.Y. the student can dropout but he/she cannot change degree course because early changes are not recordered. Regular BA graduations are, by definition, observed within the end of the third A.Y.

Table 1 Percentage distribution of Italian Bachelor STEM freshmen by course typology and gender. Cohorts 2011-12 and 2015-16. Source: data elaborated from database MIUR

	2011-12 Freshmen		2015-16 Freshmen	
Bachelor Degree Course	Females	Males	Females	Males
Computer Science	13%	87%	13%	87%
Information Engineering	20%	80%	21%	79%
Industrial Engineering	21%	79%	23%	77%
Civil And Environmental Engineering	31%	69%	30%	70%
Physics	37%	63%	30%	70%
Geology	37%	63%	35%	65%
Statistics	45%	55%	43%	57%
Chemistry	52%	48%	50%	50%
Mathematics	56%	44%	52%	48%
Environmental and Natural Sciences	59%	41%	55%	45%
Biotechnology	63%	37%	66%	34%
Biology	73%	27%	71%	29%
Total STEM	36%	64%	35%	65%
Total non-STEM	64%	36%	62%	38%

Table 2 STEM students' follow up for the 2011/12 cohort by gender and type of outcome

		Academic Year						
		11/12	12/13	13/13	14/15	15/16	16/17	17/18*
Females	At risk to event	21216	15006	9144	5481	3443	2292	1899
	Changes to non-STEM	0	5606	783	241	83	41	31
	BA dropouts	1402	604	616	311	241	199	173
	BA graduations	0	0	4463	3111	1714	911	189
	Total	**22618**	**21216**	**15006**	**9144**	**5481**	**3443**	**2292**
Males	At risk to event	36290	29226	18463	11605	7607	5179	4202
	Changes to non-STEM	0	4856	1084	362	138	73	40
	BA dropouts	3074	2208	2024	1122	750	576	590
	BA graduations	0	0	7655	5374	3110	1779	347
	Total	**39364**	**36290**	**29226**	**18463**	**11605**	**7607**	**5179**

*A.Y. censored at 31th July 2018

3 The discrete-time competing-risks model

The model selected is a discrete-time multinomial logistic regression model. Data is censored after 6 years since enrolment, as the seventh year follow-up is incomplete. By indicating with $p(t|z)$ the probability of BA graduation at time t, conditioned to a covariates vector z, the model is

$$\log\left(\frac{p^{(j)}(t|x;z)}{p^{(0)}(t|x;z)}\right) = \sum_{t=1}^{6} \alpha_t^{(j)} + \delta^{(j)\prime} X^{(j)}, \tag{1}$$

with $j = 1,2,3$, corresponding to the event "BA graduation", "BA dropout" or "change to a non-STEM degree course", respectively. $p^{(0)}(t|x;z)$ is the probability

for the reference category, that is the probability of being "still at risk" at time t, for the student with covariate profile given by the vector x. At the same way, $p^{(j)}(t|x;z)$ is the probability of the j-th event at time t, for the student's profile specified by x. The parameters $\alpha^{(j)}(t=k), \quad k=1,2,\ldots,6$, are the time-dependent NPO intercepts, $\delta^{(j)\prime}X^{(j)}$ is the NPO linear predictor for the j-th multinomial equation. The intercepts are assumed to vary along with the time, while the covariates effects are assumed time-independent for simplicity. Parameter interpretation is done in terms of *relative risk ratio* i.e. $\exp(\delta^{(j)})$ is the multiplicative effect of a 1-unit increase in x on the risk of event type j versus the risk that no event occurs [4]. Table 3 reports a description of the covariates used in the model.

Table 3 Description of the covariates used in the model for the BA.

Variable	type	Description
Time	discrete	Time (in years) since BA enrolment
Macropath	categorical	Macro-Regional (MR) Path to BA enrolment: <MR before BA enrolment> - <MR of BA enrolment>
STEM	categorical	STEM degree course typology
Irregular_age	binary	1= Student's year of birth $<$ 1992
Gender_M	binary	1= male gender
HighSchool_type	categorical	Classical, Scientific, Technical, Vocational, Other
High School_mark	categorical	High School final mark with three levels [60-70), [70-90),[90,102]

The linear predictor includes several two-way interactions of gender with all covariates, but Irregular_age which was not significant, plus the interaction High-School_mark and HighSchool_type. The baseline student profile is: female, Irregular_age=0, Macropath="North-North", STEM="Biology", HighSchool_mark="$[60,70)$", HighSchool_type="Other liceo". Estimates of this model are reported in Table 4.

Females have more success in terms of BA graduation in geology, biology, biotechnology and statistics, while they seem to suffer all the remaining STEM courses, especially mathematics, where males outperform females. Math has also the highest dropout rates for both males and females, and where females also do decide to change towards non-STEM degrees, while males do not. In general, males and females have similar behaviours about dropping out, while females are more prone than males to change course, except biology and biotechnology, which are the courses where both males and female are highly probable change to non-STEM degrees. However, this latter result is mainly due to a consolidated practice of students to change towards the medicine degree course, because the advantage that biology and biotechnology have similar first-year subjects than can be validated in the new course. HighSchool_mark, HighSchool_type and Macropath play an important role in detecting possible gender differences in students' performance.

Figure 1 highlights macro-regional differences by showing, for a regular-age student coming from the scientific high school with a $[70,90)$ mark, the estimated hazard of the three outcomes distinguishing by Southern and Northern students and by gender and by few courses.

Table 4 Estimates from the discrete-time competing-risks model fitted on the 2011/12 data cohort.

	BA graduation		**BA dropout**		**Change to non-STEM**	
	coef	pval	coef	pval	coef	pval
Time 1	-	-	-1.84	0.00	-	-
Time 2	-	-	-2.01	0.00	-0.02	0.79
Time 3	-1.33	0.00	-1.63	0.00	-1.24	0.00
Time 4	-1.00	0.00	-1.85	0.00	-1.90	0.00
Time 5	-1.03	0.00	-1.84	0.00	-2.47	0.00
Time 6	-1.18	0.00	-1.72	0.00	-2.73	0.00
Irregular_age TRUE	-0.57	0.00	0.51	0.00	-0.16	0.00
Gender M	0.05	0.60	0.04	0.73	-0.06	0.54
Macropath						
Center-Center	-0.65	0.00	0.18	0.00	-0.00	0.93
South-Center	-1.08	0.00	-0.04	0.71	-0.06	0.33
South-North	-0.63	0.00	0.09	0.39	-0.14	0.07
South-South	-1.11	0.00	0.15	0.00	-0.26	0.00
Other path	-0.65	0.00	-0.02	0.85	-0.13	0.16
STEM						
Biotechnology	0.24	0.00	0.03	0.71	0.41	0.00
Chemistry	-0.36	0.00	0.01	0.85	-0.24	0.00
Phisics	-0.26	0.00	0.19	0.03	-0.70	0.00
Geology	0.32	0.00	-0.07	0.54	-1.04	0.00
Computer Science	-0.29	0.00	0.02	0.83	-0.76	0.00
Civil & Envir. Engineering	-0.48	0.00	-0.31	0.00	-0.96	0.00
Information Engineering	-0.37	0.00	-0.20	0.01	-0.98	0.00
Industrial Engineering	-0.35	0.00	-0.47	0.00	-1.11	0.00
Mathematics	-0.56	0.00	0.27	0.00	-0.89	0.00
Envir. & Natural Sciences	-0.10	0.10	0.26	0.00	-0.06	0.24
Statistics	0.68	0.00	-0.29	0.05	-1.18	0.00
HighSchool_mark						
[70,90)	0.89	0.00	-0.54	0.00	-0.05	0.57
[90,101)	1.53	0.00	-1.24	0.00	-0.19	0.08
HighSchool_type						
Scientific	0.38	0.00	-0.74	0.00	-0.52	0.00
Classical	0.29	0.04	-0.79	0.00	-0.14	0.18
Technical	-0.20	0.13	0.16	0.06	-0.50	0.00
Vocational	-0.66	0.01	0.19	0.11	-0.56	0.00
Gender*Macropath						
M:Center-Center	0.04	0.36	-0.08	0.17	0.22	0.00
M:South-Center	0.24	0.00	-0.00	0.99	0.37	0.00
M:South-North	0.14	0.06	-0.20	0.10	0.34	0.00
M:South-South	0.19	0.00	0.07	0.15	0.46	0.00
M:Other path	0.38	0.00	-0.03	0.87	0.21	0.11
Gender*STEM						
M:Biotechnology	-0.01	0.95	0.00	0.98	-0.20	0.01
M:Chemistry	0.31	0.00	-0.07	0.49	-0.59	0.00
M:Phisics	0.02	0.82	-0.11	0.35	-0.64	0.00
M:Geology	0.12	0.34	0.04	0.81	-0.53	0.00
M:Computer Science	0.14	0.22	0.11	0.36	-0.78	0.00
M:Civil & Envir. Engineering	0.24	0.00	0.17	0.09	-0.27	0.00
M:Information Engineering	0.05	0.51	0.18	0.07	-0.58	0.00
M:Industrial Engineering	0.21	0.00	0.33	0.00	-0.39	0.00
M:Mathematics	0.48	0.00	-0.09	0.46	-0.55	0.00
M:Envir. & Natural Sciences	0.20	0.05	-0.08	0.45	-0.52	0.00
M: Statistics	-0.00	0.98	0.44	0.02	0.05	0.78
Gender*HighSchool_mark						
M:[70,90)	-0.24	0.00	0.01	0.78	-0.21	0.00
M:[90,101)	-0.09	0.14	0.06	0.47	-0.53	0.00
Gender*HighSchool						
M:Scientific	-0.24	0.00	0.11	0.15	0.28	0.00
M:Classical	-0.21	0.01	0.09	0.40	0.45	0.00
M:Technical	0.05	0.56	-0.00	0.98	-0.03	0.74
M:Vocational	0.29	0.06	0.19	0.09	0.21	0.17
HighSchool_mark*HighSchool_type						
[70,90):Scientific	0.09	0.45	-0.17	0.04	-0.05	0.59
[90,101):Scientific	0.26	0.04	-0.57	0.00	-0.14	0.23
[70,90):Classical	0.01	0.94	-0.08	0.47	0.05	0.67
[90,101):Classical	0.21	0.17	-0.19	0.27	0.18	0.18
[70,90):Technical	-0.04	0.74	0.16	0.05	0.04	0.73
[90,101):Technical	-0.06	0.68	0.27	0.03	-0.03	0.84
[70,90):Vocational	-0.00	0.98	0.34	0.00	0.06	0.76
[90,101):Vocational	0.31	0.20	0.63	0.00	0.07	0.77

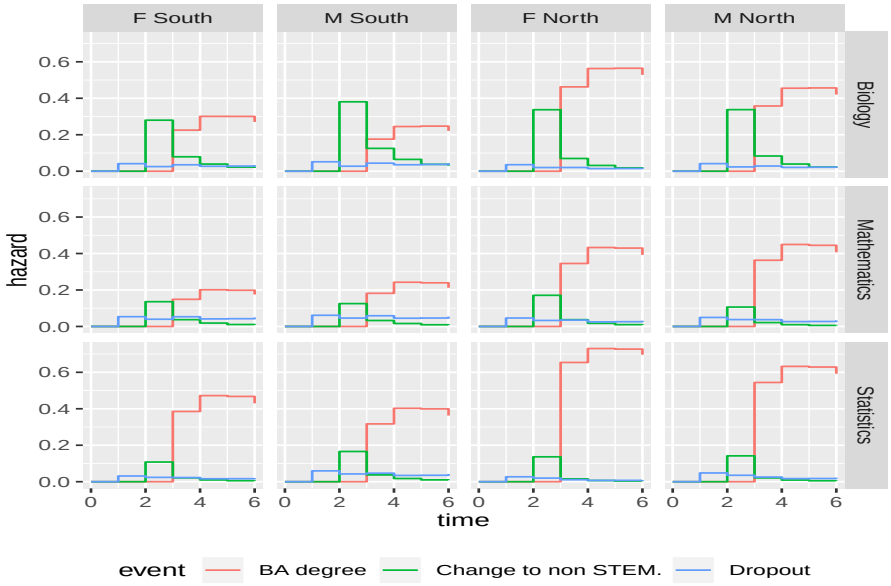


Fig. 1 Probabilities of either BA graduation, BA dropout, or Change to non-STEM degree courses, by macro-region and gender in three STEM courses, for selected students' profiles (regular-age student coming from the scientific high school with a $[70,90)$ mark) .

4 Conclusions

We aimed at detecting possible gender differences in performances of students attending a STEM degree course in an Italian university. Preliminary results highlight the females are under-represented in STEM courses, except for biology, biotechnology, and environmental and natural sciences. Geology, biology, biotechnology and statistics are courses where females have more success in terms of BA graduation are, while they seem to suffer all the remaining STEM courses, especially mathematics, where males outperform males. In general, males and females have similar behaviours about dropping out, while females are more prone to change course.

Acknowledgements This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P

References

1. Contini, D., Salza, G., and Scagni, A.: Dropout and Time to Degree in Italian Universities Around the Economic Crisis. WP16/2017 del Dipartimento di Economia e Statistica "Cognetti de Martiis" (2017).
2. INVALSI: Rapporto prove INVALSI 2018. Roma: INVALSI (2018).
3. Database MOBYSU.IT: Mobilitá degli Studi Universitari in Italia. Protocollo di ricerca MIUR - Universitá degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II, Fonte dei dati ANS-MIUR/CINECA (2016).
4. Steele F.: Event history analysis. NCRM Methods Review Papers, NCRM\004, (2005) URL http://eprints.ncrm.ac.uk/88/1/MethodsReviewPaperNCRM-004.pdf

Some Challenges and Results in Measuring Gender Inequality

Alcune sfide nella misura della disuguaglianza di genere

Fabio Crescenzi, Francesco Di Pede

Abstract In this paper, we show some challenges and results on gender statistics at national and international level. The geo-referencing of statistical information will allow Istat, through the integrated system of registers (SIR), to release statistical and geographical information with a strong territorial detail and will allow users to know more on the longitudinal evolution of data and this will have a potential strong impact also on gender statistics.
Considering the importance given to name of urban places to show the disequilibrium in representing women who have contributed, in all fields, to improving society, we will show some preliminary results obtained analyzing the gender of personalities whose name is in the streets of the National archive of streets.

.

Abstract *In questo articolo mostriamo alcune tra le principali sfide e risultati sulle statistiche di genere a livello internazionale e nazionale. La georeferenziazione delle informazioni statistiche consentirà all'Istat, attraverso il sistema integrato di registri (SIR), di rilasciare informazioni statistiche e geografiche con un forte dettaglio territoriale e consentirà agli utenti di saperne di più sull'evoluzione longitudinale dei dati e questo avrà un potenziale forte impatto anche sulle statistiche di genere.*
Considerando l'importanza data al nome dei luoghi urbani per mostrare lo squilibrio della presenza di donne che hanno contribuito, in tutti i campi, al miglioramento della società, mostreremo alcuni risultati preliminari ottenuti analizzando il genere delle personalità il cui nome è nelle strade dell'archivio nazionale di strade.

Key words: Gender statistics, Geo-referencing, Streets Names, Place Names.

[1] Fabio Crescenzi, Istat, the National Institute of Statistics of Italy; email: fabio.crescenzi@istat.it
Francesco Di Pede, Istat, the National Institute of Statistics of Italy; email: dipede@istat.it

1 Introduction

The Global Gender Statistics Programme is mandated by the United Nations Statistical Commission, implemented by the United Nations Statistics Division (UNSD) and coordinated by the Inter-Agency and Expert Group on Gender Statistics IAEG-GS [4]. The Programme encompasses:

- improving coherence among existing initiatives on gender statistics through international coordination.
- developing and promoting methodological guidelines in existing domains as well as in emerging areas of gender concern.
- strengthening national statistical and technical capacity for the production, dissemination and use of gender relevant data.
- facilitating access to gender relevant data and metadata through a newly developed data portal.

UNSD serves as Secretariat of the Inter-Agency and Expert Group on Gender Statistics (IAEG-GS), the coordinating and guiding body of the Global Gender Statistics Programme. The IAEG-GS was first convened in 2006, meets annually and functions through advisory groups. Presently, the main advisory group's work concentrates on examining emerging and unaddressed key gender issues and related data gap with the aim to develop proposals on how to fill these gaps.

This website[1] serves as a platform for the dissemination of developments in the field of gender statistics and promotes the inclusion of gender statistics into all fields of statistical activities at both the national and international levels[2].

One of the emerging data field collection concern SDGs indicators[3], the main results are presented in the national SDGs report [5]

In the other sections of this paper, we will consider some of the main challenges faced in the field of gender statistics at national level. The geo-referencing of statistical information will allow Istat, through the integrated system of registers (SIR), to release statistical and geographical information with a strong territorial detail and will allow users to know more on the longitudinal evolution of data and this will have a potential strong impact also on gender statistics.

Considering the importance given to name of urban places to show the disequilibrium in representing women who have contributed, in all fields, to improving society, we will show some preliminary results obtained analysing the gender of personalities whose name is in the streets of the National archive of streets.

The use of big data in gender violence analysis was presented and discussed in a Workshop "Big Data e Violenza di Genere, Metodologie a supporto dell'analisi del sentiment " which was held at the National Institute of Statistics (Istat) on February 7, 2020. [1,4,6,7]

---

[1] https://unstats.un.org/unsd/demographic-social/gender/index.cshtml

[2] Concerning data, Gender Equality Data and Statistics (UN and World Bank) provide country level observations in six thematic areas: Economic structures and participation; Education; Health and related services; Public life and decision making; Human rights of women and girls, and Demographic indicators. Data is sourced from the regional commissions of the UN and the World Bank.
The European Institute for Gender Equality publishes the annual Gender Equality Index, allowing researchers to compare index variables for all EU member states, by work domain and pay level.
https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/UNWomen

[3] 1. https://unstats.un.org/sdgs/indicators/database/

2 The Integrated System of Registers (SIR), a formidable source of data for future gender statistics.

In the modernization plan of Istat data from administrative sources and statistical surveys are integrated in the SIR covering the demographic, social, economic and environmental domains. The Base registers are connected by codes and are maintained updated over time using administrative sources and statistical surveys [2 8 9 10]. A greater use of administrative and statistical data sources will allow minimizing the burden on respondents and costs of field data collection. RSBL, (the Italian Statistical Base Register of Territorial Entities), is one of the pillars of SIR which integrates information from many different sources of geographical data [3], with the aim to improve the geo referencing of data. The aim is to build the register only once, to keep it updated, and to use it in all statistical processes.
The geo referencing of statistical information will allow Istat, through the integrated system of registers, to release statistical and geographical information with a strong territorial detail and will allow users to know more on the structure and composition of the territory of Italy. RSBL is a multidimensional register integrating several components: addresses, regular grids and micro zones, buildings and housing units, administrative zones, statistical and functional zones. In the panorama of statistical registers, it represents an innovative way of considering in an integrated way all the most relevant components of territorial data. Moreover, Istat is moving from a decennial census to a permanent census, which will produce census data every year.
SIR represent a formidable source of data also for future gender statistics. Respecting the need to protect privacy, the potential availability of new data to be studied in their longitudinal evolution and territorial distribution is very rich.

3 What street names tell us about gender inequality?

In Italy an associations named *Toponomastica femminile,* founded on Facebook in January 2012, works with the objective of setting up research, publishing data and putting pressure on the society and local authorities to dedicate to women names of streets, squares, gardens and urban places to compensate for the evident disequilibrium that characterizes the current state of the names of places.
The main objective of the associations, which today includes more than 10,000 members, is to spread gender culture, to give visibility to women who have contributed, in all fields, to improving society. Links to conferences and papers on this topic can be found on the association's website[1].
The address component is one of the most important component of RSBL. Addresses in RSBL comes mainly from the National archive of addresses and urban streets (ANNCSU). The Law of May 30, 1989, n. 223 establishes that each municipality compile and update the list of streets and the list of addresses according to the rules issued by the National Statistics Institute. In addition, the Law of 17 December 2012, n. 221, provides for the institution of the

1. [1] https://www.toponomasticafemminile.com/sito/

ANNCSU, made and managed by Istat and the cadastral data of the Land Property Registry. The ANNCSU is the reference database on streets and addresses, which contains information for the entire country in digital format. Each address is geocoded to the Census Enumeration Areas of the Census Mapping. The National Digital Agency recognizes ANNCSU as a core database for the great impact on the transition to the National Register of Resident Population (ANPR) as well as on the many other uses of public interest.
As part of the data processing of the 2011 population census activities, municipal offices verified the misalignment of data and, when necessary, corrected, integrated and validated them.
In the data processing, the National Statistical Office checked the names of the streets and organized them into a complete dictionary. The dictionary contains names in standardized form, each of which can occur in one or multiple cases in different Italian municipalities.
Although work is still in progress, here we present some preliminary results concerning the extraction from the dictionary of the street names that are names of personalities.
Out of 847 names of personalities, 107 refers to females, 737 refers males with an overall gender ratio (number of females/number of males x100) equal to 14,7%. A really poor result, considering that in order to have gender equality this ratio should be equal to 100%.
We continued the analysis by forming groups in relation to the type of personalities found in street names. We got the ten groups listed in **Table 1**:

Table 1: *Groups of personality types found in street names*

Groups	***Examples***
1. Nations and Churches Leaders	Popes, Kings, Emperors, Presidents
2. State Personalities	Politicians, Patriots, Diplomats and Trade unionists
3. Dynastic and Literary Personalities	Dynastic, Aristocrats, Mythological and Literary personalities
4. Religious Figures	Religious, Theologians, Saints and Blessed, Divinities
5. Explorers, Inventors, Heroes, Soldiers	Explorers, Inventors, Heroes, Soldiers
6. Arts and Literature Personalities	Painters, Poets, Writers, Musicians, Historians, Philosophers, Archaeologists
7. Professionals and Business Workers	Lawyers, Magistrates, Doctors, Engineers, Architects, Entrepreneurs, Economists
8. Scientists	Mathematicians, Physicists, Geologists, Astronomers, etc.
9. Metiers Workers	Jewellers, Engravers, Archivists, Educators, Pharmacists, Insurers, etc.
10. Entertainment, Media, Sports Workers	Actresses / Actors, Publishers, Producers, Directors, Singers, Journalists, Photographers, Athletes

We computed the gender ratio by group shown in **Table 2.** The gender ratio is over the overall figure in the following case: Group 3, Dynastic and Literary Personalities (37,5%), Group 10, Entertainment, Media, Sports Workers (37,0), Group 4, Religious Figures (34,5%), Group 9, Metier Workers (16,7%).
The gender ratio is under the overall figure in the following case: Group 7, Professionals and Business Workers (1,4%), Group 5, Explorers, Inventors, Heroes, Soldiers (5,9%), Group 8,

Scientists (6,7%), Group 1, Nations and Churches Leaders (7,7%), Group 6, Arts and Literature Personalities (8,3%), Group 2, State Personalities (10,4%)..

Table 2: Gender ratio by groups

Type	*Females*	*Males*	*Ratios (x100)*
1. Nations and Churches Leaders	3	39	**7,7**
2. State Personalities	14	135	**10,4**
3. Dynastic and Literary Personalities	6	16	**37,5**
4. Religious Figures	49	142	**34,5**
5. Explorers, Inventors, Heroes, Soldiers	3	51	**5,9**
6. Arts and Literature Personalities	17	204	**8,3**
7. Professionals and Business Workers	1	71	**1,4**
8. Scientists	2	30	**6,7**
9. Metiers Workers	2	12	**16,7**
10. Entertainment, Media, Sports Workers	10	27	**37,0**
Total	**107**	**727**	**14,7**

This is only a preliminary result, in the future it will be useful to map this data and analyse the territorial distribution of the gender ratio.
Recently, an example of a geo referencing of this kind of data showed the aerial photos of Italian cities with the streets dedicated to women highlighted in red. The author of this work, the photographer Alessandro Scotti, called it "The Atlas of Misogyny", because the result of the survey is impressive: in the national map of the name of streets, women are marginal, small, sparse red dashes that just they glimpse in a sea of streets. "It is the testimony of the degree of social recognition attributed to the female gender in the cities where we live"[1].

1 https://www.ilmessaggero.it/mind_the_gap/nomi_strade_citta_donne_uomini_mappa_misoginia_censis-4877458.html

4 Conclusions

Many results have already been achieved on gender statistics, however further impetus will have to be given to the study of longitudinal evolutions and the territorial distribution of data.
The use of new sources such as registers, aerial images, big data will help greatly in this direction.

References

1. Almaviva: La metodologia proposta per l'analisi dei BIG DATA e i risultati attesi. Workshop Big Data e Violenza di Genere, Metodologie a supporto dell'analisi del sentiment, Istat, Roma, (2020)
2. Bakker, B.F.M., Rooijen, J. van & Toor: The system of social statistical datasets of Statistics Netherlands: an integral approach to the production of register-based social statistics. Statistical journal of the United Nations ECE, 30(4), p.411-424 (2014).
3. Crescenzi F., Lipizzi F.: The Integration of Geographic and Territorial Data Sources into the Base Register of Territorial and Geographical Entities, Journal of the International Association for Official Statistics, (to appear, 2020)
4. Deriu F.: Il vocabolario della violenza. Il progetto e le metodologie, Workshop Big Data e Violenza di Genere, Metodologie a supporto dell'analisi del sentiment, Istat, Roma, (2020)
5. Istat: SDGS Report. Statistical Information for 2030 Agenda in Italy, pp.93-105. Rome (2019) https://www.istat.it/en/archivio/232745
6. Istat: Le ipotesi di lavoro per l'analisi del sentiment sulla violenza di genere. Workshop Big Data e Violenza di Genere, Metodologie a supporto dell'analisi del sentiment, Istat, Roma, (2020)
7. Pavone P.: Presentazione dei primi risultati e ulteriori sviluppi del lavoro. Workshop Big Data e Violenza di Genere, Metodologie a supporto dell'analisi del sentiment, Istat, Roma, (2020)
8. Schulte Nordholt, E. The Dutch Virtual Census 2001: A new approach by combining different Sources, Statistical Journal of the United Nations Economic Commission for Europe, 2005, 22, p.25–37.
9. UNECE: Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics, United Nations Publication, ISBN 978-92-1-116963-8, (2007)
10. Zhang, L-C.: Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica, (2012)

How Deep is Your Plot? Young SIS and deep statistical learning (ySIS)

A modal approach for clustering matrices

Un approccio modale al clustering di matrici

Federico Ferraccioli and Giovanna Menardi

Abstract In this work we propose a modal approach to density-based clustering for matrix-valued data. We introduce appropriate kernels for this type of data and define a kernel estimator of matrix-variate density functions. Additionally, we propose an extension of the mean-shift procedure for the identification of the modes of the estimated density. Given the intrinsic high dimensionality of matrix-variate data and the resulting computational complexity of the algorithm, we discuss a possible solution to handle the problem. We finally present the performance of the proposed method through an application to real world data, also with respect to a plausible competitor.
Abstract In questo lavoro viene proposto un approccio modale al *clustering* basato su densità per dati in forma di matrici. Vengono introdotti *kernel* appropriati per questo tipo di dati e definito uno stimatore di densità *kernel* per dati matrice-variati. Viene inoltre proposta un'estensione dell'algoritmo *mean-shift* per l'identificazione delle mode della densità stimata. Data l'elevata dimensionalità dei dati considerati e la conseguente complessità computazionale, si discute una possibile soluzione per gestire questo problema. Vengono infine presentate le prestazioni della metodologia proposta attraverso un'applicazione ad dati reali, confrontandola con un possibile metodo concorrente.

Key words: matrix-variate distributions, modal clustering, high dimensional data, mean shift.

Federico Ferraccioli
Department of Statistical Sciences, University of Padova e-mail: ferraccioli@stat.unipd.it

Giovanna Menardi
Department of Statistical Sciences, University of Padova e-mail: menardi@stat.unipd.it

1 Introduction

The analysis of complex data in the form of matrices represents an active area of research. Examples are image data, connectivity maps, as well as data arising from classical longitudinal studies, or modern scientific applications such as brain scans or multi-tissue genome expressions. In the former cases, the data of interest are represented by a collection of matrices, one for each subject. In the latter cases, the statistical observations are represented by vectors of variables, each of them measured on a subject over different times, thus giving a matrix-valued observation for each subject.

While a vast body of literature has focused on the development of supervised methods for matrix-valued data, especially when multivariate observations are gathered over time or in the specific class of semi-definite matrices, far less attention has been devoted to the unsupervised case. For the aim of clustering, scattered contributions which are worth to mention refer to [5, 4, 6].

In this work we extend the *modal*, or *nonparametric* clustering formulation to the framework of matrix-variate data. Here, clusters are identified as the "domains of attraction" of the modes of the true density underlying the data. The issue of density estimation, usually addressed via nonparametric methods, assumes a key role in order to approximate the ideal population goal of modal clustering, along with the operational search of the modal regions.

After providing an overview on modal clustering ideas, we introduce a kernel estimator for matrix-variate density functions, with a special focus on Normal kernels. We then study its asymptotic properties also with reference to the problem of optimal bandwidth selection. Given the intrinsic high dimensionality of matrix-variate data, impacting on both the accuracy of the estimate and the computational complexity, we explore some solution to handle the problem. Additionally, we propose an extension of the mean-shift procedure for the identification of the modes of the estimated density. Finally we illustrate the performance of the proposed method on a set of real data, also with respect to some plausible competitors.

2 An overview on modal clustering

In a standard multivariate setting, *modal*, or *nonparametric*, clustering relies on the assumption that the observed data $\mathscr{X} = (x_1, \dots, x_n)$, are realizations of a multidimensional random variable $x \in \mathbb{R}^p$ with (unknown) probability density function f. The modes of f are regarded to as representatives of the clusters, which are in turn represented by their domains of attraction. Operationally, clustering involves two main choices, which are overviewed in the following. See [3] and references therein for a comprehensive review.

The first choice concerns the estimation of the density function, which determines the high density regions and, hence, governs the final clustering. A standard choice, within the class of nonparametric methods, is the kernel density estimator

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i),$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ is a probability density on $\mathbb{R}^p$ symmetric around x_i, and with scale parameter the bandwidth $h > 0$, which defines the degree of smoothing. While the choice of the function K is known not to have a strong impact in the performance of the estimate, choosing h properly turns out to be crucial. A small value of h leads to an undersmoothed density estimate, with the possible appearance of spurious modes, while a value too large results in an oversmoothed density estimate, possibly hiding relevant features.

A further aspect to account for in modal clustering, is to operationally characterize the clusters as the domain of attraction of the density modes. Most of the contributions in this direction take their steps from the *mean-shift* algorithm which, starting from a generic point $y^{(0)}$, recursively shifts it to a local weighted mean, until convergence. Denoted by $w_i(y^{(s)})$ the vector of weights of x_i at step s, at the next step $s+1$,

$$y^{(s+1)} = \sum_{i=1}^{n} w_{i,h}(y^{(s)})x_i \quad \text{with} \quad w_{i,h}(y) = \frac{\nabla K_h(y - x_i)}{\sum_{i=1}^{n} K_h(y - x_i)},$$

namely the weights $w_{i,h}(\cdot)$ are specified as normalized components of the gradient of the kernel function. These arguments motivate the mean shift as a gradient ascent algorithm for some specific kernel functions as, e.g., the Normal kernel.

Operationally a partition of the data points is obtained by grouping in the same cluster those observations ascending to the same mode of the density.

3 Matrix-variate extension of modal clustering

3.1 Kernel density estimation of matrix-variate data

Consider a sample $X_1, \ldots, X_n$, of *i.i.d.* realizations of a $p \times t$ random matrix X, which we shall assume to be defined on the vector space $\mathbb{R}^{p,t}$. The (unknown) distribution underlying X is naturally described by some probability density function $f : \mathbb{R}^{p,t} \mapsto \mathbb{R}^+$, with $\int_{\mathbb{R}^{p,t}} f(X)\mathrm{d}X = 1$, being the component-wise integral of $f(X)$ on its support.

Given a symmetric smooth kernel K_h supported on (a subset of) $\mathbb{R}^{p,t}$, scaled by h and with symmetry to be intended along all the $p \times t$ dimensions, we define the *kernel density estimator* for matrix-variate data as

$$\hat{f}_h(X) = \frac{1}{n}\sum_{i=1}^{n} K_h(X - X_i). \tag{1}$$

Asymptotic theory about kernel density estimation can be shown to extend naturally to the matrix-variate setting. With the established convention on defining matrix-variate integrals, for instance, we may define the Mean Integrated Square Error as in the standard multivariate framework, and show that the integrated variance plus squared bias decomposition still holds, so that under some regularity conditions, and as $n \to \infty$, the Asymptotic Mean Integrated Squared Error (AMISE) for estimator (1) is

$$\text{AMISE}(\hat{f}_h(\cdot)) = n^{-1}h^{-(p\times t)}R(K) + \frac{1}{4}h^4 m_2(K)^2 R(\Delta f).$$

where $m_2(K) = \int_{\mathbb{R}^d} z_i^2 K(z)dz$ and $R(a) = \int_{\mathbb{R}^{p,t}} a(X)^2 dX$, The optimal bandwidth which minimizes the AMISE is

$$h_{\text{AMISE}} = \left(\frac{(p\times t)R(K)}{m_2(K)^2 R(\Delta f)}\right)^{\frac{1}{(p\times t)+4}} n^{-\frac{1}{(p\times t)+4}}.$$

As for the univariate and multivariate settings, in the matrix-variate framework a pivotal role is played by the matrix Normal density [1], which represents a natural candidate for the kernel function. In the general unconstrained covariance case, which allows full flexibility, the summands of (1) take the form

$$K_{U,V}(X - X_i) = (2\pi)^{-\frac{pt}{2}} |V|^{-\frac{p}{2}} |U|^{-\frac{t}{2}} \exp\left(-\tfrac{1}{2}\text{tr}(V^{-1}(X - X_i)^\top U^{-1}(X - X_i))\right) \quad (2)$$

The $U_{p\times p}$ and $V_{t\times t}$ matrices identify a separable covariance structure of the rows and columns of the variable X. In the simplest case, where $U = h_U \mathbb{I}_p$ and $V = h_V \mathbb{I}_t$, the kernel assumes the form

$$K_{h_U,h_V}(X - X_i) = (2\pi)^{-\frac{pt}{2}} h_U^{-p} h_V^{-t} \exp\left(-\frac{1}{2(h_U h_V)^2}\text{tr}((X - X_i)^\top (X - X_i))\right).$$

Due to the separability of the covariance structure, the choice of two distinct smoothing parameters for rows and columns hence results unnecessary.

3.1.1 Adaptive kernel

As an overall problem shared by nonparametric tools, kernel estimators are known to strongly suffer from the curse of dimensionality. On one side, the required sample size to achieve an acceptable accuracy becomes disproportionately large as the dimensions increases, leading to intractable problems, even computationally. On the other side, in high dimensions, the sparsity of data induce much of the probability mass to flow to the tails of the density. This may give rise to the birth of spurious modes and average away features in the highest density regions, thus possibly affecting severely a clustering goal. While these arguments could discourage from the application of modal clustering on matrix-variate data, which are intrinsically

high-dimensional, non-parametric estimates can still be useful to coarsely describe the data structure, and often, allowing different amounts of smoothing to capture local structures of the data is advisable. One example is the *k−nearest neighbours* estimator, defined as

$$\hat{f}_k(X) = \frac{1}{n}\frac{1}{\delta_{(k)}(X)^{p\times t}}\sum_{i=1}^{n} K\left(\frac{X-X_i}{\delta_{(k)}(X)}\right) \tag{3}$$

where $\delta_{(k)}(X)$ is the k-th nearest neighbour distance to X. This is a special case of adaptive kernel, where the bandwidth $\delta_{(k)}(X)$ varies with X, thus determining a larger amount of smoothing in the low density regions. The (3), albeit not smooth for visualization purposes, is shown to be very efficient, since density estimator at a point X reduces essentially to computing $\delta_{(k)}(X)$, and no explicit kernel function evaluations are required.

3.2 Mean shift for matrix-variate data

Once that the density has been estimated via the matrix-variate extension discussed so far, in order to cluster observations, a proper way for the observations to climb the modes is required.

To this aim, the mean-shift algorithm lends itself quite naturally to the generalization to matrix-valued data. Consider the kernel density estimator (1), and, for the sake of simplicity, a Normal kernel (2). Define

$$w_{i,U,V}(Y) = \frac{\exp(\mathrm{tr}(V^{-1}(Y-X_i)^\top U^{-1}(Y-X_i)))}{\sum_{j=1}^{n}\exp(\mathrm{tr}(V^{-1}(Y-X_)^\top U^{-1}(Y-X_i)))}. \tag{4}$$

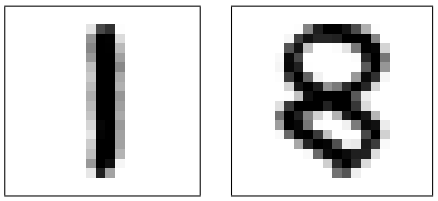
We can now state the following proposition.

Proposition 1. *Starting from a generic $Y^{(0)} \in \mathbb{R}^p \times t$, the sequence $\{Y^{(0)},\dots,Y^{(S)}\}$, defined by the mean-shift update*

$$Y_i^{sj+1)} = \sum_{i=1}^{n} w_{i,U,V}(Y_s^{(j)})X_i$$

converge to a mode of $\hat{f}$.

The main argument justifying this result is that, due to the separability of the covariance matrices U and V, equation (1) boils down to the mean-shift based on a standard multivariate Normal kernel with covariance matrix $U \otimes V$ and built on the vectorization of the involved matrix variables. It can be shown that the mean-shift convergence result to the local modes of $\hat{f}$ applies to a wider set of kernel functions, among which the nearest neighbour estimator (3), that correspond to the special case where $U = \delta_{(k)}(X)I_p$ and $V = \delta_{(k)}(X)I_t$.



Method	ME
Mixture of matrix-variate Normals (2 clusters)	0.09
Mixture of matrix-variate Normals (# of clusters defined via BIC)	0.39
fixed h mean-shift	0.62
NN-based mean-shift	0.12

Fig. 1 Two samples of handwritten digits data(left and middle panel), and the summarizing results of the proposed methods with respect to the competitor in term of misclassification error(right).

4 Application

We applied the proposed methodology to the well known handwritten ZIP codes data [2], described by samples of handwritten digits $(0,\dots,9)$ collected from the ZIP codes on envelopes from US postal mail. Each observation is represented by a matrix of 16×16 pixels, ranging in intensity from 0 to 255. We restrict our attention to the subset of digits (1, 8), with a total of 430 observations (Figure 1).

Two alternative matrix-variate kernel estimators are considered: the (1) with fixed bandwidth selected for optimal estimation of a matrix-variate Normal density, and the (3) with $k = 50$. The procedure is compared with its parametric counterpart, where clusters are identified by the components of a mixture of matrix-variate Normal densities [4]. Results, presented in the right panel of Figure 1, are evaluated in term of misclassification error (ME). While the parametric matrix-normal mixture [4] has the best performance when the number of clusters is correctly specified, in the case of automatic selection via BIC, the procedure is unable to identify different clusters. In the nonparametric case, a fixed bandwidth produce, as expected, multiple spurious clusters, due to the complex geometry of the underlying density and the intrinsic high-dimensionality of the data. The NN-mean shift has far better performances, possibly due to the higher flexibility of the adaptive mean-shift, that results in a higher accuracy in capturing the shape of the underlying density.

References

1. Gupta, A. K., & Nagar, D. K. (2018). Matrix variate distributions. Chapman and Hall/CRC.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
3. Menardi, G. (2016). A review on modal clustering. Int. Stat. Rev., 84(3), 413-433.
4. Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comp., 21(4), 511-522.
5. Viroli, C. (2011). Model based clustering for three-way data structures. Bayes. An., 6(4), 573-602.
6. Wang, M., Fischer, J., & Song, Y. S. (2019). Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. Ann. Appl. Stat., 13(2), 1103-1127.

A Note on Detection of Perturbations in Biological Networks

Una nota sul rilevamento di perturbazioni in reti biologiche

Vera Djordjilović

Abstract Griffin et al. (2018, *Biometrics* **74**, 1351-1361) have recently addressed the problem of detecting perturbations in multi-omics networks by applying the method of network filtering. They have also introduced two novel testing procedures for detecting the origin of perturbation in Gaussian graphical models. In this work, we argue that the proposed testing procedures, although useful in the exploratory data analysis, lack some desirable statistical properties, including type I error control. We illustrate this issue on a small example, and propose a solution in the form of a modified testing procedure.

Abstract Griffin et al. (2018, *Biometrics* **74**, 1351-1361) hanno recentemente affrontato il problema della rilevazione di perturbazioni nelle reti multi-omiche mediante l'applicazione del metodo del filtro di rete, introducendo due nuove procedure di verifica di ipotesi per rilevare l'origine della perturbazione nei modelli grafici Gaussiani. In questo lavoro dimostriamo che le procedure proposte, sebbene utili nell'analisi esplorativa dei dati, non soddisfano alcune proprietà statistiche desiderabili, compreso il controllo dell'errore di I tipo. Considereremo un semplice esempio per illustrare il problema e verrà proposta una soluzione che modifica la procedura originale.

Key words: FWER control, Gaussian graphical models, likelihood ratio tests, network filtering.

1 Introduction

Comparing biological networks between two conditions is an active area of research in statistics. After identifying all network components that show differing behavior,

Vera Djordjilović
Department of Biostatistics, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway e-mail: vera.djordjilovic@medisin.uio.no

the interest is usually in detecting the *origin* of perturbation. A promising tool for addressing this task, *network filtering* [1, 3], has been recently applied to multi-omics networks [2]. The proposed solution and the assumed perturbation model are briefly described in the following.

Data in the control condition are assumed to come from a multivariate normal distribution that is Markov with respect to an unknown graph. The perturbation acts on its target(s) and changes its(their) mean. The effect of perturbation is then propagated through network connections so that further nodes result perturbed. The aim is to detect the site of the original perturbation. This is achieved in two steps: in the first step, data from the control condition are used to estimate the covariance matrix and the graphical structure; in the second step, data from the perturbed condition are transformed in the process of network filtering, and a testing procedure is used to identify the most likely site of the original perturbation.

The focus of the present work is on the testing procedures proposed in [2] and described in Section 2. In Section 3, we show that the single-target procedure suffers from the fact that it is not accompanied by an assessment of uncertainty of the produced ranking. In Section 4, we illustrate how this may cause the multi-target procedure to overestimate the number of targets. In Section 5, we propose a modified procedure that addresses these issues, and, in addition, avoids making any prior assumptions on the number of targets. In Section 6, we draw some general conclusions and recommendations.

2 Testing procedures for perturbation detection

Our focus is on the testing procedure; to this end, we adopt a setting in which the covariance matrix Σ for the considered set of variables is known. Let $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ be a random sample from the perturbed condition, i.e. a random sample of size n from a p-variate normal distribution $\mathsf{N}(\Sigma\boldsymbol{\mu},\Sigma)$, where $\boldsymbol{\mu}\in\mathbb{R}^p$ is a perturbation vector whose only non-zero components are perturbation targets. [2] consider transformed data $\boldsymbol{Z}_i=\Omega\boldsymbol{Y}_i$, $i=1,\dots,n$, where $\Omega=\Sigma^{-1}$. Since $\boldsymbol{Z}_i\sim\mathsf{N}(\boldsymbol{\mu},\Omega)$, the transformed data can be used to make inference on $\boldsymbol{\mu}$.

2.1 Single-target perturbations

The Authors first consider single-target perturbations that affect only one component of $\boldsymbol{\mu}$. To identify the affected component, they propose testing a hypothesis that a node has been perturbed, conditional on there being only one target. More formally, they propose testing the null hypothesis $H:\boldsymbol{\mu}=0$ against a series of alternatives $H_i:\mu_i\neq 0,\boldsymbol{\mu}_{-(i)}=0$, where $\boldsymbol{\mu}_{-(i)}$ denotes a subvector of $\boldsymbol{\mu}$ obtained by omitting the i-th component representing node i. The components of $\boldsymbol{\mu}$ are then ranked according to the associated log-likelihood ratio test statistic T_i. They show

that if node i has been perturbed, the test statistic T_i will stochastically dominate test statistics of the remaining nodes, so that node i is likely to be top-ranked.

2.2 Multi-target perturbations

Test statistics T_i, $i = 1, \ldots, p$, are different test statistics for the global null hypothesis H, and when the power to reject H is sufficiently high, many p-values, including the one associated to the true perturbation target, will be low. When this is the case, it might be difficult to conclude whether the nodes with low p-values represent multiple perturbation targets or are only significant due to their proximity to the perturbation target.

To resolve this issue, the Authors propose a sequential testing procedure that at each step takes into account the perturbation targets identified in the previous steps. In particular, at step $s+1$, they consider the problem of testing

$$H_S : \boldsymbol{\mu}_{-(S)} = 0 \quad \text{against} \quad H_{(S,i)} : \mu_i \neq 0, \boldsymbol{\mu}_{-(i,S)} = 0, \tag{1}$$

where S is the set of nodes detected up to step s. The test is performed for all nodes not already identified, i.e. for all $i \in \bar{S}$, and the node with the lowest p-value is declared the likely perturbation target at step s. The Authors suggest that the false discovery rate can be controlled by applying the method of Benjamini and Hochberg to the resulting sequence of p-values.

3 Ranking uncertainty

The main aim of the proposed testing procedures is to produce a ranking of likely perturbation targets. Unfortunately, the obtained ranking is not accompanied by an assessment of uncertainty.

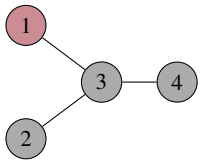


Fig. 1 The graph G used in the example. The perturbation target is highlighted in red.

To illustrate this issue we use a small graph G shown in Figure 1. Let $\Sigma = \left\{\sigma_{ij}\right\}_{i,j=1}^{4}$ be such that $\sigma_{ii} = 1$, $\sigma_{ij} = 0.9$ when nodes i and j are connected in G, and $\sigma_{ij} = 0.81$ otherwise. Then $\Omega = \Sigma^{-1}$ has zeros corresponding to the missing edges of G. Data from the reference condition are assumed to follow a zero-mean

multivariate normal distribution with the covariance matrix Σ. Let the perturbation affect node 1 by shifting its mean from 0 to $\mu_1 = 0.5$, so that μ_1 becomes the only non-zero element of $\boldsymbol{\mu}$. The effect of the perturbation propagates through G, so that after the perturbation data follow a multivariate normal distribution with mean $\Sigma\boldsymbol{\mu}$ and unchanged covariance matrix Σ. We assessed the performance of the single-target testing procedure by simulating 1000 random samples of size $n = 100$ from the perturbed distribution, i.e. from $\mathsf{N}(\Sigma\boldsymbol{\mu}, \Sigma)$. Following the single-target procedure described in Section 2.1, we computed, for each sample, the four p-values for the hypotheses that node i has been perturbed conditional on there being only one perturbation, i.e. the p-values for testing H against H_i.

Figure 2 reports the proportion of Monte Carlo runs in which the p-value of the single-target procedure associated to node i was below the Bonferroni threshold of $\alpha/4$ (wide gray bars; red bars can be ignored for now), where α was set to 0.05. Narrow gray bars show the proportion of Monte Carlo runs in which a node was first ranked. We see that the true perturbation target was first ranked around 80% of times. As a consequence, the first ranked node was not the perturbation target approximately 20% of times.

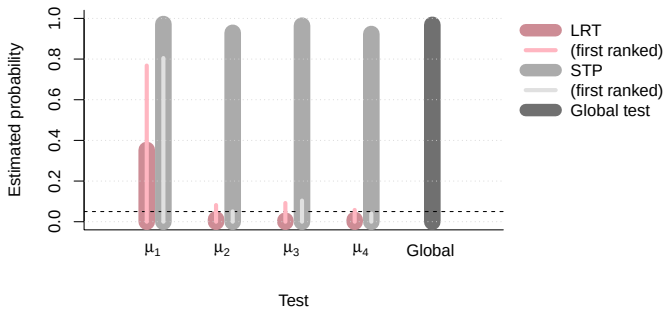


Fig. 2 Estimated probability that a) a node-wise hypothesis is rejected (wide bars); b) a node is first ranked (narrow bars); c) the global null hypothesis H is rejected (Global). STP stands for the single-target procedure; LRT for the unconditional node-wise likelihood ratio test. Horizontal dashed line $y = 0.05$ represents the considered significance level.

4 Type I error control

The lack of statistical guarantees for the obtained ranking affects the multi-target procedure as well. Namely, when the true perturbation target is not ranked first with high probability, the proposed sequential procedure will overestimate the number of targets and correcting the p-values for multiplicity will not resolve this issue.

Let us use again the perturbation of node 1 in G to illustrate this point. We again generated 1000 random samples of 100 observations from the post-perturbation distribution, but this time we applied the multi-target (sequential) procedure. We considered a range of values for $\mu_1 = 0.1, 0.3, 0.5, 0.9$, emulating perturbations of different strength. We investigated the power to detect the true target and the type I error control. For simplicity, we considered the familywise error rate (FWER), and applied the Bonferroni correction to the p-values obtained by the sequential procedure. The power was estimated as the proportion of Monte Carlo runs in which the p-value for node 1 resulted significant; the FWER as the proportion of runs in which at least one p-value of the remaining nodes resulted significant. Hypotheses were tested at level $\alpha = 0.05$. The results are reported in the first column of Table 1. Although the power of the multi-target procedure appears to be very good even

Table 1 Estimated power and FWER multiplied by 10^2 for the multi-target and the node-wise likelihood ratio testing procedure for multiple perturbation strengths. Estimated FWER exceeding the nominal level is highlighted.

	Multi-target		Node-wise LRT	
μ_1	Power	FWER	Power	FWER
0.1	5.1	**6.7**	1.8	2.5
0.3	46.7	**29.1**	12.5	3.2
0.5	82.5	**18.2**	36.4	3.4
0.7	93.3	**8.3**	72.1	3.8
0.9	97.6	**5.4**	92.2	3.7

for mild perturbations (for $\mu_1 \geq 0.3$), applying the Bonferroni correction does not ensure FWER control. In fact, the obtained FWER exceeds the nominal level (0.05) in all considered scenarios, and the result is particularly worrisome in the presence of mild perturbations: for $\mu_1 = 0.3$, the FWER is almost 0.3.

5 Unconditional node-wise likelihood ratio tests

Both issues can be solved with a slight modification of the single-target procedure. It is sufficient to replace the conditional null hypothesis by unconditional node-wise hypotheses

$$H_{0i} : \mu_i = 0, \boldsymbol{\mu}_{-(i)} \in \mathbb{R}^{p-1}, \tag{2}$$

and test them against general alternatives $H_{0i}^c : \mu_i \neq 0, \boldsymbol{\mu}_{-(i)} \in \mathbb{R}^{p-1}$. The modified procedure addresses directly the question of what nodes have been perturbed without making any prior assumptions on the number of targets. Additionally, it allows type I error control by standard multiple testing procedures.

Hypothesis H_{0i} can be tested with a likelihood ratio test. The maximum likelihood estimator of $\boldsymbol{\mu}$ under the null hypothesis is $\hat{\boldsymbol{\mu}}_0$, with $(\hat{\boldsymbol{\mu}}_0)_i = 0$, and $(\hat{\boldsymbol{\mu}}_0)_{(-i)} =$

$\bar{\boldsymbol{Z}}_{(-i)} + \Sigma^{-1}_{(-i,-i)}\Sigma_{(-i,i)}\bar{\boldsymbol{Z}}_i$, where $\Sigma_{(-i,-i)}$ and $\Sigma_{(-i,i)}$ are submatrices of Σ obtained by removing the appropriate rows and columns. The maximum likelihood estimator of $\boldsymbol{\mu}$ under the alternative hypothesis is the unrestricted mean $\bar{\boldsymbol{Z}}$.

Figure 2 compares the single-target procedure with the modified procedure in our small example from Section 3. The original procedure achieves significantly higher power (1 against 0.37 for μ_1) which is not surprising: it is testing the global null hypothesis $H : \boldsymbol{\mu} = 0$. Indeed, it behaves as the likelihood ratio test for H against a general alternative stating that at least one component of $\boldsymbol{\mu}$ is different from zero (the rightmost dark gray bar). Rejecting H and concluding that at lest one component of $\boldsymbol{\mu}$ is non-zero is easier than rejecting H_{01} and concluding that it is the first component of $\boldsymbol{\mu}$ that is different from zero. On the other hand, the modified procedure controls the type I error for nodes 2,3 and 4.

We also considered the modified procedure in the example of Section 4. The second column of Table 1 reports its estimated power and FWER, and shows that it resolves the type I error issue of the multi-target procedure. Namely, the power of the multi-target procedure is higher, but, as already noted, at the expense of type I error control. On the other hand, the modified procedure has lower power but controls FWER across all values of μ_1 (somewhat conservatively: the estimated FWER is around 3.3). When the perturbation is strong, i.e. for $\mu_1 = 0.9$, the performance of the two procedures is very similar, indicating that the conclusions drawn will coincide whenever there is sufficient statistical power.

6 Conclusion

The two testing procedures proposed in [2] appear to be well-suited for the exploratory data analysis concerning the ranking of likely perturbation targets; however, they suffer from the lack of statistical guarantees. In particular, it is the lack of an assessment of uncertainty for a single target procedure and the lack of the type I error control for the multi-target procedure. We proposed a modified testing procedure, based on unconditional node-wise tests, that provides a simple solution to these issues.

References

1. Cosgrove, Elissa J and Zhou, Yingchun and Gardner, Timothy S and Kolaczyk, Eric D: Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. Bioinformatics **24(21)**, 2482–2490 (2008).
2. Griffin, Paula J and Zhang, Yuqing and Johnson, William Evan and Kolaczyk, Eric D: Detection of multiple perturbations in multi-omics biological networks. Biometrics **74(4)**, 351–1361 (2018).
3. Yang, Shu and Kolaczyk, Eric D: Target detection via network filtering. IEEE Trans Inf Theory **56(5)**, 2502–2515 (2010)

Bayesian inference for DAG-probit models

Inferenza bayesiana per modelli DAG-probit

Federico Castelletti

Abstract We consider a binary response variable, e.g. a disease status, together with a set of real-valued variables, such as clinical features of a patient, whereas the interest is in discovering which of these can predict the response. To implicitly account for dependence relationships between all variables, we model the joint distribution of the observables through a particular Directed Acyclic Graph (DAG) structure that we name DAG-probit. We propose a Bayesian methodology for structural learning and parameter estimation of DAG-probit models.

Abstract *Si considera una variabile risposta dicotomica, per esempio una condizione di salute, insieme ad una collezione di variabili reali, come un insieme di caratteristiche cliniche, con l'obiettivo di determinare quali variabili hanno un impatto sulla risposta. Al fine di considerare possibili relazioni di dipendenza tra tutte le variabili di interesse, se ne modella la distribuzione congiunta attraverso un particolare tipo di grafo aciclico direzionato (DAG) denominato DAG-probit. Si propone quindi una metodologia bayesiana per la selezione di modelli DAG-probit e la stima dei relativi parameteri.*

Key words: Graphical model; Directed Acyclic Graph; Probit regression; Bayesian inference

1 Introduction

We consider a collection of real-valued variables which potentially affect a binary response. The interest is in discovering which variables are relevant to predict the

Federico Castelletti
Universitá Cattolica del Sacro Cuore, Dipartimento di Scienze Statistiche, Largo Gemelli 1, Milan,
e-mail: federico.castelletti@unicatt.it

response, an issue which can be formalized as a *variable selection* problem for a dichotomous response. Variable selection in binary response models is typically performed using statistical methods based on logistic or probit regression which are widely employed to provide answers to many scientific queries. For instance, in medical sciences one might be interested in discovering which clinical features can predict a disease status and also how much each feature affects the probability to observe the disease. However, standard regression models do not directly account for dependencies among variables which, if neglected, can produce unreliable estimates of regression coefficients or related quantities of interest, such as odds or *causal effects*. In this paper we approach this issue from a graphical model perspective. We adopt Directed Acyclic Graphs (DAGs) which represent a well established tool to investigate dependence relationships among variables, especially in medicine and genomics [4] and in particular *DAG-probit* models to account for the presence of a (binary) response. In particular, we consider a Gaussian framework where the covariance matrix is Markov w.r.t. an unknown DAG and assume that the binary response is obtained by discretization of a continuous (latent) counterpart [1]. Under a Bayesian setting, the normality assumption allows for efficient posterior inference on the covariance matrix, also accounting for model uncertainty, which is gathered from an available posterior distribution on the space of DAG-probit models. DAGs are particularly suited for causal reasoning as well, which is predicated using the notion of *intervention* and the allied *interventional distribution* [9]. Accordingly, the causal effect on the response due to an intervention on a specific node depends on the underlying DAG and can be expressed as a function of the covariance matrix. Therefore, the proposed methodology can be naturally adopted for causal effect estimation under model uncertainty; see for instance [7] and [3].

2 DAG-probit models

Let $\mathscr{D} = (V,E)$ be a DAG, where $V = \{1,\dots,q\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ a set of edges. The elements of E are $(u,v) \equiv u \to v$ and such that if $(u,v) \in E$ then $(v,u) \notin E$. In addition, $\mathscr{D}$ contains no cycles, that is paths of the form $u_0 \to u_1 \to \cdots \to u_k$ where $u_0 \equiv u_k$. For a given node v, if $u \to v \in E$ we say that u is a *parent* of v (conversely v is a *child* of u). The parent set of v in $\mathscr{D}$ is denoted by $\mathrm{pa}_{\mathscr{D}}(v)$, the set of children by $\mathrm{ch}_{\mathscr{D}}(v)$. We further assume that $\mathrm{ch}_{\mathscr{D}}(1) = \emptyset$ which prescribes that node $v = 1$, which will represent the response variable, has no children.

We now consider a collection of variables $(X_1,\dots,X_q)$ and assume that the joint probability density function $f(x_1,\dots,x_q)$ obeys the Markov property of $\mathscr{D}$, so that we can write the factorization

$$f(x_1,\dots,x_q) = \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathscr{D}}(j)}), \tag{1}$$

Consider now the Gaussian setting,

$$X_1,\ldots,X_q \,|\, \boldsymbol{\Omega} \sim \mathscr{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1}), \tag{2}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix and $\boldsymbol{\Omega} \in \mathscr{P}_{\mathscr{D}}$, the space of symmetric and positive definite (s.p.d.) precision matrices Markov w.r.t. $\mathscr{D}$. For a Gaussian DAG-model factorization (1) becomes

$$f(x_1,\ldots,x_q \,|\, \boldsymbol{\Omega}) = \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}(j)}, \boldsymbol{\Omega}). \tag{3}$$

Equation (3) can be also written as a *structural equation model*. Specifically, assuming a parent ordering of the nodes so that $i > j$ whenever j is a child of i we can write

$$\boldsymbol{L}^\top \boldsymbol{x} = \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{L}$ is a (q,q) lower-triangular matrix of coefficients, $\boldsymbol{L} = \{\boldsymbol{L}_{ij}, i \geq j\}$, such that $\boldsymbol{L}_{ij} \neq 0$ if and only if $i \to j$ and $\boldsymbol{L}_{ii} = 1$. Moreover, $\boldsymbol{\varepsilon}$ is a $(q,1)$ vector of error terms, $\boldsymbol{\varepsilon} \sim \mathscr{N}_q(\mathbf{0}, \boldsymbol{D})$, where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\sigma}^2)$ and $\boldsymbol{\sigma}^2$ is the $(q,1)$ vector of *conditional* variances whose j-th element is $\sigma_j^2 = \mathbb{V}\mathrm{ar}(X_j \,|\, \boldsymbol{x}_{\mathrm{pa}(j)}, \boldsymbol{\Sigma})$. Therefore

$$\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^\top. \tag{5}$$

Equation (5) is called the *modified Cholesky decomposition* of $\boldsymbol{\Omega}$ [2]. Let now $\prec j \succ = \mathrm{pa}(j)$ and $\prec j] = \mathrm{pa}(j) \times j$. The modified Cholesky decomposition (5) induces a re-parametrization of $\boldsymbol{\Omega}$ (equivalently, $\boldsymbol{\Sigma}$) in terms of the node-parameters $(\boldsymbol{L}_{\prec j]}, \sigma_j^2)$, $j = 1,\ldots,q$, where

$$\boldsymbol{L}_{\prec j]} = -\boldsymbol{\Sigma}_{\prec j \succ} \boldsymbol{\Sigma}_{\prec j]}, \quad \sigma_j^2 = \boldsymbol{\Sigma}_{jj \,|\, \mathrm{pa}(j)}. \tag{6}$$

Accordingly, Equation (3) can be written as

$$\begin{aligned} f(x_1,\ldots,x_q \,|\, \boldsymbol{\Omega}) &= \prod_{j=1}^{q} f(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}(j)}, \boldsymbol{L}_{\prec j]}, \sigma_j^2) \\ &= \prod_{j=1}^{q} d\mathscr{N}(x_j \,|\, -\boldsymbol{L}_{\prec j]}^\top \boldsymbol{x}_{\mathrm{pa}(j)}, \sigma_j^2). \end{aligned} \tag{7}$$

Let now $Y \in \{0,1\}$ be the binary outcome. Given the threshold $\theta_0 \in (-\infty, +\infty)$, we assume that Y is generated from X_1 (its latent counterpart) as

$$Y = \begin{cases} 1 & \text{if } X_1 \in [\,\theta_0, +\infty), \\ 0 & \text{if } X_1 \in (-\infty, \theta_0); \end{cases} \tag{8}$$

so that $P(Y=1) = P(Z \geq \theta_0)$; see also [1]. Therefore, letting $f(x_1)$ be the p.d.f. of X_1 and using the convention $\theta_{-1} = -\infty, \theta_1 = +\infty$, the joint density of (Y, X_1) can be

written as

$$f(y,x_1\,|\,\theta_0) = f(x_1)\cdot \mathbb{1}(\theta_{y-1} < x_1 \leq \theta_y). \tag{9}$$

We now assume that $(X_1,\ldots,X_q)$ follows the Gaussian DAG-model (2) which implies the factorization in (7). Accordingly the joint density of $(Y,X_1,\ldots,X_q)$ in (9) becomes

$$f(y,x_1,\ldots,x_q\,|\,\boldsymbol{\Omega},\theta_0) = f(x_1,\ldots,x_q\,|\,\boldsymbol{\Omega})\cdot \mathbb{1}(\theta_{y-1} < x_1 \leq \theta_y). \tag{10}$$

Equation (10) defines a Gaussian *DAG-probit model*. Recall now from Equation (7) that the conditional distribution of the latent variable X_1 is $\mathcal{N}(-\boldsymbol{L}_{\prec 1]}^{\top}\boldsymbol{x}_{\mathrm{pa}(1)},\sigma_1^2)$. To guarantee the identifiability of our DAG-probit model we then set $\sigma_1^2 = 1$ as in standard probit regression.

Finally, by considering n independent samples $(y_i,x_{i,2},\ldots,x_{i,q})$, $i=1,\ldots,n$, the augmented likelihood can be written as

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_q\,|\,\boldsymbol{\Omega},\theta_0) &= \prod_{i=1}^{n} f(y_i,x_{i,1},\ldots,x_{i,q}\,|\,\boldsymbol{\Omega},\theta_0) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{q} f(x_{i,j}\,|\,\boldsymbol{x}_{i,\mathrm{pa}(j)},\boldsymbol{L}_{\prec j]},\sigma_j^2)\cdot \mathbb{1}(\theta_{y_i-1} < x_{i,1} \leq \theta_{y_i}), \end{aligned} \tag{11}$$

where $\boldsymbol{y} = (y_1\ldots,y_n)^{\top}$ and $\boldsymbol{x}_j = (x_{1,j},\ldots,x_{n,j})^{\top}$ for $j=1,\ldots,q$.

3 Bayesian inference

Under a Bayesian framework, we complete our DAG-probit model by specifying prior distributions for precision matrix $\boldsymbol{\Omega}\in\mathscr{P}_{\mathscr{D}}$, DAG $\mathscr{D}$ and cut-off θ_0. For a *complete* DAG $\mathscr{D}$, a prior on $\boldsymbol{\Omega}$ (unconstrained) can be assigned through a Wishart distribution $\mathscr{W}_q(a,\boldsymbol{U})$, where $a > q-1$ and $\boldsymbol{U}$ is a s.p.d. matrix. This induces a prior on $(\boldsymbol{L},\boldsymbol{D})$ such that the node-parameters $(\boldsymbol{L}_{\prec j]},\sigma_j^2)$ are independent and Normal-Gamma distributed. Accordingly, the priors $p(\boldsymbol{L}_{\prec j]},\sigma_j^2)$ are conjugate with the corresponding normal densities in (2), which guarantees closed-form results for the posterior distribution of the model parameters as well as for the marginal likelihood of a complete DAG $\mathscr{D}$.

Consider now the case in which $\mathscr{D}$ is not complete. To assign a prior on $\boldsymbol{\Omega}\in\mathscr{P}_{\mathscr{D}}$ we adopt the procedure of [5] which is based on a set of assumptions which are naturally satisfied under the Gaussian setting (2). Accordingly, we assign a prior to the parameters of each DAG model starting from a unique prior on the parameters of an unconstrained (complete) DAG. In turns, closed-form expressions for the posterior distribution of each DAG model parameters and marginal likelihoods can be also obtained; see also the original paper for details.

Next, a prior on DAG $\mathscr{D}$, $p(\mathscr{D})$, can be assigned through a Bernoulli prior on the elements of the adjacency matrix of $\mathscr{D}$; see for instance [2]. This can be also used to regulate sparsity in the DAG-probit model space by choosing a prior probability of inclusion for each (possible) edge. Finally, for the unknown threshold θ_0 we assume a uniform prior, so that $p(\theta_0) \propto 1$.

4 Computational details

Given the results resumed in the previous sections, an efficient MCMC scheme for posterior inference and model selection of DAG-probit models can be introduced. Specifically, our proposal is based on a collapsed MCMC scheme in which at each iteration we propose and update DAG $\mathscr{D}$ and precision matrix $\boldsymbol{\Omega}$ given the observed data $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_q$ and the latent $\boldsymbol{x}_1$. At each step, the latent $x_{1,i}$, for $i = 1, \ldots, n$, is sampled from its conditional distribution, which corresponds to a normal density truncated at $(\theta_{y_i-1}, \theta_{y_i}]$. The update of θ_0 is performed through a Metropolis Hastings step where a cut-off g_0 is proposed from a $\mathscr{N}(\theta_0, \sigma_0^2)$.

Our algorithm results in a collection of T DAGs and precision matrices visited by the MCMC chain. In particular, we can use the output $\{\mathscr{D}^{(1)}, \ldots, \mathscr{D}^{(T)}\}$ to approximate the posterior distribution of the DAG-probit space or related quantities of interest, such as the probability of inclusion for each edge $u \to v$ which can be estimated as

$$\hat{p}_{u \to v}(\boldsymbol{y}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{u \to v}\{\mathscr{D}^{(t)}\}, \tag{12}$$

where $\mathbb{1}_{u \to v}(\cdot)$ is the indicator function taking value 1 if and only if $\mathscr{D}$ contains $u \to v$. In addition, for each individual i we can provide an estimate of the probability $P(Y_i = 1 \,|\, x_{i,2}, \ldots, x_{i,q}; \boldsymbol{\Omega}, \theta_0) = P(Z_i \geq \theta_0 \,|\, x_{i,2}, \ldots, x_{i,q}; \boldsymbol{\Omega})$ e.g. by computing the expected value of the predictive distribution of Y_i.

5 Application to breast cancer data

In this section we apply the proposed methodology to the breast cancer dataset presented in [8]. The objective of the study was to develop a prediction model which can be used as a biomarker of breast cancer, based on a collection of anthropometric data and parameters gathered in routine blood analysis and measured on $n = 116$ individuals. Accordingly, each individual i is classified as $y_i = 1$ (type 1) in case of breast cancer, $y_i = 0$ (type 0) otherwise. Data are provided as a supplement to the original paper. We implement our methodology to approximate the posterior distribution over the DAG-probit space and the probability of disease for each subject i. As the results, we report in Figure 1 the heat map with estimated posterior probabilities of edge inclusion computed as in (12) and the posterior probabilities of $Y_i = 1 \,|\, \cdot$

computed across the n individuals. Comparisons with frequentist probit regression is also provided.

References

1. Albert, J. H., Chib, S.: Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association **303**(5659), 799–805 (2004)
2. Cao, X., Khare, K., Ghosh, M.: Posterior graph selection and estimation consistency for high-dimensional Bayesian CAG models. The Annals of Statistics **1**(47), 319–348 (2019)
3. Castelletti, F., Consonni, G.: Bayesian inference of causal effects from observational data in Gaussian graphical models. Submitted (2020)
4. Friedman, N.: Inferring Cellular Networks Using Probabilistic Graphical Models. Science **88**(422), 669–679 (1993)
5. Geiger, D., Heckerman, D.: Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. The Annals of Statistics **30**(5), 1412–1440 (2002)
6. Lauritzen, S. L.: Graphical Models. Oxford University Press, Oxford (1996)
7. Maathuis, M. H., Kalisch, M., Bühlmann, P.: Estimating high-dimensional intervention effects from observational data. The Annals of Statistics **37**(6A), 3133–3164 (2009)
8. Patrício, M. Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., Caramelo, F.: Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer **18**, 29 (2018)
9. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)

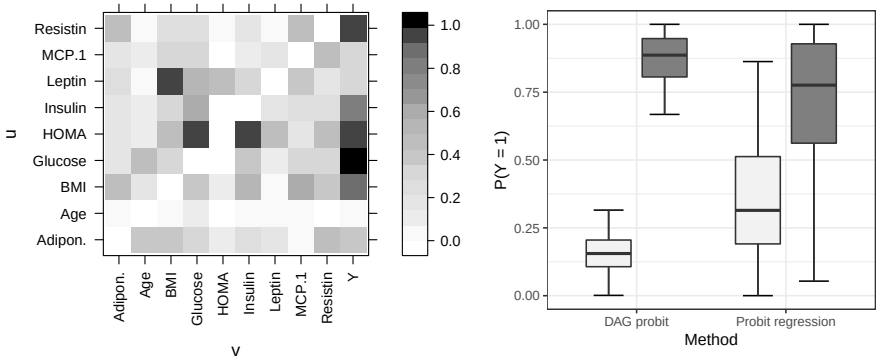


Fig. 1: Heat map with posterior probabilities of inclusion for each edge $u \to v$ (left side). Box plots with estimated posterior probabilities $P(Y_i = 1 \mid \cdot)$ for the two groups of individuals, type 0 [ ] and type 1 [■]. Comparison is between our DAG-probit method and frequentist probit regression.

Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process

Inferenza Bayesiana Variazionale per Modelli Fattoriali Gaussiani dotati di Processo a Contrazione Cumulativa

Sirio Legramanti

Abstract The cumulative shrinkage process is an increasing shrinkage prior that can be employed within models in which additional terms are supposed to play a progressively negligible role. A natural application is to Gaussian factor models, where such a process has proved effective in inducing parsimonious representations while providing accurate inference on the data covariance matrix. The cumulative shrinkage process came with an adaptive Gibbs sampler that tunes the number of latent factors throughout iterations, which makes it faster than the non-adaptive Gibbs sampler. In this work we propose a variational algorithm for Gaussian factor models endowed with a cumulative shrinkage process. Such a strategy provides comparable inference with respect to the adaptive Gibbs sampler and further reduces runtime.
Abstract *Il processo a contrazione cumulativa è una distribuzione a priori a contrazione crescente che può essere impiegata in modelli per i quali si suppone che termini aggiuntivi giochino un ruolo via via più trascurabile. Una sua naturale applicazione è nei modelli fattoriali gaussiani, dove tale processo si è dimostrato efficace nell'indurre rappresentazioni parsimoniose fornendo allo stesso tempo un'inferenza accurata sulla matrice di covarianza dei dati. Il processo a contrazione cumulativa è stato proposto insieme a un Gibbs sampler adattivo che modula il numero di fattori latenti nel corso delle iterazioni, il che lo rende più veloce del Gibbs sampler non adattivo. In questo lavoro proponiamo un algoritmo variazionale per modelli fattoriali gaussiani dotati di processo a contrazione cumulativa. Tale soluzione fornisce un'inferenza paragonabile a quella del Gibbs sampler adattivo e riduce ulteriormente i tempi di calcolo.*

Key words: Shrinkage prior, Spike and slab, Stick-breaking representation

Sirio Legramanti
Bocconi University, Milan, Italy, e-mail: sirio.legramanti@phd.unibocconi.it

1 Introduction

The cumulative shrinkage process [5] is an increasing shrinkage prior based on a sequence of spike-and-slab distributions, with growing mass assigned to the spike. It can be defined for both countable and finite sequences. A definition for the countable case can be found in [5], while the following is a definition for finite sequences.

Definition 1. We say that $\{\theta_h \in \Theta \subseteq \mathbb{R} : h = 1,\dots,H\}$ is distributed according to a cumulative shrinkage process with shrinkage parameter $\alpha > 0$, slab P_0 and spike P_∞ if, conditionally on $\{\pi_h \in (0,1) : h = 1,\dots,H\}$, each θ_h is independent and

$$(\theta_h \mid \pi_h) \sim (1-\pi_h)P_0 + \pi_h P_\infty, \qquad (h = 1,\dots,H), \tag{1}$$

where $\pi_h = \sum_{l=1}^{h} \omega_l$ for $h = 1,\dots,H$ and $\omega_l = v_l \prod_{m=1}^{l-1}(1-v_m)$ for $l = 1,\dots,H$, with $v_1,\dots,v_{H-1}$ being independent Beta$(1,\alpha)$ random variables and $v_H = 1$.

This construction, based on the stick-breaking representation of the Dirichlet process [4], implies that the sequence π_h is non-decreasing and that $\pi_H = 1$.

The cumulative shrinkage process can be used in a variety of models, e.g. Poisson factorization [3], but here we focus on Gaussian factor models, which are ubiquitous in statistics and have been used in [5] as illustrative example. In [5] posterior inference for this model under the cumulative shrinkage process is carried out through an adaptive Gibbs sampler which tunes H as it progresses. This algorithm, together with the ability of the prior to favor the recovery of the number of active latent factors, allows for reduced runtime with respect to the non-adaptive Gibbs sampler. However, the increasing availability of large datasets demands for even faster algorithms. This need for scalability has pushed Bayesian statisticians towards approximate methods for posterior inference, including Laplace approximation, variational Bayes and expectation propagation [2].

In this work we employ mean-field variational Bayes, which is straightforward to derive for Gaussian factor models under a convenient specification of the cumulative shrinkage process which slightly differs from the one in [5]. Such a specification is detailed in § 2, while the variational approximation is described in § 3. Finally, in § 4 we illustrate the performance of the variational algorithm on real data.

2 Model and Prior

We focus on learning the structure of the $p \times p$ covariance matrix $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$ of the data $y_i \in \mathbb{R}^p$ from the Gaussian factor model $y_i = \Lambda\eta_i + \varepsilon_i$ $(i = 1,\dots,n)$, where $\Lambda = [\lambda_{jh}] \in \mathbb{R}^{p\times H}$, $\eta_i \sim N_H(0, I_H)$, $\varepsilon_i \sim N_p(0,\Sigma)$ and $\Sigma = \mathrm{diag}(\sigma_1^2,\dots,\sigma_p^2)$. As priors, we let $\sigma_j^2 \sim \mathrm{InvGa}(a_\sigma, b_\sigma)$ for $j = 1,\dots,p$ and, differently from [5], we place a cumulative shrinkage process directly on the loadings, with π_h as in Def. 1:

$$(\lambda_{jh} \mid \pi_h) \sim (1-\pi_h)N(0,\theta_0) + \pi_h N(0,\theta_\infty), \qquad (j = 1,\dots,p;\ h = 1,\dots,H). \tag{2}$$

This simpler specification facilitates the derivation of the variational algorithm, while preserving the increasing shrinkage property. In fact, setting $\theta_0 > \theta_\infty$, the loadings are increasingly shrunk towards zero in probability, i.e. $\text{pr}\{|\lambda_{j,h+1}| < \varepsilon\} \geq \text{pr}\{|\lambda_{jh}| < \varepsilon\}$ for any $\varepsilon > 0$, encoding the prior assumption that additional factors provide a decreasing contribution to the model. However, setting both the spike and the slab to Gaussians is suboptimal to the specification in [5], where the Student-t slab is more differentiated from the Gaussian spike, thus facilitating the separation of active and inactive factors.

The derivation of the variational algorithm is further facilitated by the introduction of the augmented data $z_h = (z_{h1}, \ldots, z_{hH}) \sim \text{Mult}\{1, (\omega_1, \ldots, \omega_H)\}$, which exploits the fact that equation (2) can be obtained by marginalizing out z_h from

$$(\lambda_{jh} \mid z_h) \sim \{1 - \textstyle\sum_{l=1}^{h} z_{hl}\} N(0, \theta_0) + \sum_{l=1}^{h} z_{hl} N(0, \theta_\infty), \quad (j = 1, \ldots, p;\ h = 1, \ldots, H).$$

3 Variational Inference

Variational Bayes approximates the posterior density with the density q^* that is closest to it, in Kullback-Leibler (KL) divergence, within a family Q of tractable densities (see [1] for a review). The ideal variational family Q should combine flexibility, that allows for a good approximation, and tractability. Here we choose the mean-field variational family, whose elements factorize as follows:

$$q(\lambda, \eta, \sigma, z, v) = q(\lambda) q(\eta) q(\sigma) q(z) q(v). \tag{3}$$

The KL divergence between such a q and the intractable posterior cannot be computed or minimized directly. Equivalently, we maximize the evidence lower bound

$$\begin{aligned} ELBO(q) &= \log p(y) - KL(q(\lambda, \eta, \sigma, z, v) || p(\lambda, \eta, \sigma, z, v \mid y)) = & (4) \\ &= E_q[\log p(y, \lambda, \eta, \sigma, z, v)] - E_q[\log q(\lambda, \eta, \sigma, z, v)]. & (5) \end{aligned}$$

Equation (4) highlights that, since the KL divergence is always non-negative, the ELBO lower-bounds the log-evidence, thus justifying its name. Moreover, since $\log p(y)$ does not depend on q, maximizing the ELBO is equivalent to minimizing the KL divergence with respect to q. Since (4) involves the intractable posterior, the equivalent expression (5) is used to actually compute the ELBO. The optimization is solved through coordinate ascent, iteratively maximizing the ELBO with respect to each factor on the right-hand side of (3). Following [2, Ch. 10], each factor update is derived as follows (we report only the loadings term for illustrative purposes):

$$\log q^*(\lambda) = E_{\neq \lambda}[\log p(y, \lambda, \eta, \sigma, z, v)] + const,$$

where $E_{\neq \lambda}$ denotes the expectation under q with respect to all variables other than the loadings. With no parametric assumption on the factors in (3), we obtain:

1 **for** *j from 1 to p* **do**
  set $V_j^{(\lambda)} = \{diag(\theta_1^*,\dots,\theta_H^*) + (A^{(\sigma)}/B_j^{(\sigma)})(\mu^{(\eta)\mathrm{T}}\mu^{(\eta)} + nV^{(\eta)})\}^{-1}$,
  where $\theta_h^* = (1-\sum_{l=1}^h \kappa_{hl})\theta_0^{-1} + (\sum_{l=1}^h \kappa_{hl})\theta_\infty^{-1}$, and $\mu_j^{(\lambda)} = (A^{(\sigma)}/B_j^{(\sigma)})V_j^{(\lambda)}\mu^{(\eta)\mathrm{T}}y_{\cdot j}$;

2 Set $A^{(\sigma)} = a_\sigma + n/2$ and **for** *j from 1 to p* **do**
  set $B_j^{(\sigma)} = b_\sigma + 0.5\cdot\sum_{i=1}^n\{y_{ij}^2 - 2y_{ij}\mu_i^{(\eta)\mathrm{T}}\mu_j^{(\lambda)} + \sum_{h=1}^H\sum_{k=1}^H(\mu_{ih}^{(\eta)}\mu_{ik}^{(\eta)} + V_{hk}^{(\eta)})(\mu_{jh}^{(\lambda)}\mu_{jk}^{(\lambda)} + V_{j;hk}^{(\lambda)})\}$;

3 Set $V^{(\eta)} = (I_H + \mu^{(\lambda)\mathrm{T}}diag(A^{(\sigma)}/B^{(\sigma)})\mu^{(\lambda)} + \sum_{j=1}^p(A^{(\sigma)}/B_j^{(\sigma)})V_j^{(\lambda)})^{-1}$;
for *i from 1 to n* **do**
  set $\mu_i^{(\eta)} = V^{(\eta)}\mu^{(\lambda)\mathrm{T}}diag(A^{(\sigma)}/B^{(\sigma)})y_{i\cdot}$;

4 **for** *h from 1 to H* **do**
  for *l from 1 to h* **do** set $\kappa_{hl} \propto \exp\{E(\log\omega_l) - 0.5\cdot p\log\theta_\infty - 0.5\cdot\theta_\infty^{-1}E[\lambda_{\cdot h}^T\lambda_{\cdot h}]\}$;
  for *l from h+1 to H* **do** set $\kappa_{hl} \propto \exp\{E(\log\omega_l) - 0.5\cdot p\log\theta_0 - 0.5\cdot\theta_0^{-1}E[\lambda_{\cdot h}^T\lambda_{\cdot h}]\}$;
  where $E[\lambda_{\cdot h}^T\lambda_{\cdot h}] = \sum_{j=1}^p(\mu_{jh}^{(\lambda)2} + V_{j;hh}^{(\lambda)})$ and, with Ψ being the digamma function,
  $E(\log\omega_l) = \mathbb{1}\{l<H\}\{\Psi(A_l^{(v)}) - \Psi(A_l^{(v)} + B_l^{(v)})\} + \mathbb{1}\{l>1\}\sum_{m=1}^{l-1}\{\Psi(B_m^{(v)}) - \Psi(A_m^{(v)} + B_m^{(v)})\}$;

5 **for** *h from 1 to* $(H-1)$ **do**
  set $A_h^{(v)} = 1 + \sum_{l=1}^H \kappa_{lh}$ and $B_h^{(v)} = \alpha + \sum_{l=1}^H\sum_{m=h+1}^H \kappa_{lm}$.

Algorithm 1: One cycle of the variational algorithm for Gaussian factor models

$$q^*(\lambda,\dots,v) = \prod_{j=1}^p N_H(\lambda_{j\cdot};\mu_j^{(\lambda)},V_j^{(\lambda)})\prod_{i=1}^n N_H(\eta_{i\cdot};\mu_i^{(\eta)},V^{(\eta)})\prod_{j=1}^p \mathrm{InvGa}(\sigma_j^2;A^{(\sigma)},B_j^{(\sigma)})\cdot$$
$$\cdot\prod_{h=1}^H \mathrm{Mult}(z_h;1,\kappa_h)\prod_{h=1}^{H-1}\mathrm{Beta}(v_h;A_h^{(v)},B_h^{(v)}).$$

Notice that each factor further factorizes into exponential-family distributions, thus facilitating computations. The update equations for the parameters are coupled, meaning that each factor update involves expectations with respect to other factors. We then proceed iteratively cycling over the steps of Algorithm 1. This routine converges to a local maximum, hence should be run from several initializations [1]. Convergence of each run can be assessed by monitoring the monotone growth of the ELBO. From the optimal variational parameters we can also compute the variational expectation of the number H^* of factors that are active, in the sense that they are modeled by the slab: $E_{q^*}[H^*] = \sum_{h=1}^H\sum_{l=h+1}^H \kappa_{hl}$.

4 Application to Personality Data

We compare our variational algorithm for the model in § 2 to the adaptive Gibbs sampler for the model proposed in [5], on the same real dataset considered there. Namely, we consider a subset of the dataset `bfi` from the `R` package `psych`, containing the six-point-scale answers of $n = 126$ individuals older than fifty years to

Table 1 Performance of adaptive Gibbs sampler and variational algorithm on the `bfi` dataset

Method	MSE	E[H*]	Running time (s)
Adaptive Gibbs sampler	0.01	2.7	340
Variational inference	0.01	3.0	63

$p = 25$ questions about five personality traits. As in [5], we center the 25 items and, to have coherent answers within each personality trait, we change sign to answers $1, 9, 10, 11, 12, 22$ and 25, as suggested in the documentation of the `bfi` dataset.

For the adaptive Gibbs sampler, the model and the hyperparameters are specified as in [5]. For our variational algorithm, we set $\alpha = 5$, $\theta_0 = 1$, $\theta_\infty = 10^{-6}$ and we conservatively let $H = p + 1$, which coincides with the initial value of H for the adaptive Gibbs sampler and corresponds to at most p latent factors.

We run the variational algorithm from 20 random initializations, stopping each run when the ELBO grows less than 0.05. We then pick the run reaching the highest ELBO. Using the optimal variational parameters of this run, we get a sample of size 2000 for Ω, from which we derive a sample for the correlation matrix $\Omega^* = (\Omega \odot I_p)^{-1/2} \Omega (\Omega \odot I_p)^{-1/2}$, with $\odot$ denoting the element-wise product. From this sample we compute a Monte Carlo estimate of the mean squared deviations $\sum_{j=1}^{p} \sum_{q=j}^{p} E(\Omega^*_{jq} - S_{jq})^2 / \{p(p+1)/2\}$ between Ω^* and the sample correlation matrix S. The same quantity is computed from a posterior sample of equal size obtained running the adaptive Gibbs sampler in [5] for 10000 iterations after a burn-in of 5000 and then thinning every five. The two quantities are reported as MSE (Mean Square Error) in Table 1, together with the expected number of active factors and the total running time for each of the two methods.

With respect to the adaptive Gibbs sampler, the proposed variational algorithm provides the same MSE (rounded off to the second decimal digit) and a similar expected number of active factors, but is more than five times faster.

Acknowledgements The author is grateful to Daniele Durante for his helpful comments, and acknowledges the support from MIUR-PRIN 2017 project 20177BRJXS.

References

1. Blei, D. M., Kucukelbir, A., McAuliffe, J. D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**, 859–877 (2017)
2. Bishop, C. M.: Pattern recognition and machine learning. Springer (2006)
3. Dunson, D. B., Herring, A. H.: Bayesian latent variable models for mixed discrete outcomes. Biostatistics **6.1**, 11–25 (2005)
4. Ishwaran, H., James, L. F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)
5. Legramanti, S., Durante, D., Dunson, D. B.: Bayesian cumulative shrinkage for infinite factorizations. Biometrika, in press (2020+)

Measuring poverty and vulnerability

Choosing the vulnerability threshold using the ROC curve

Determinare la soglia di vulnerabilità: un approccio basato sulla curva ROC

Chiara Gigliarano and Conchita D'Ambrosio

Abstract When estimating vulnerability one usually estimates a probability of being poor in the future, which varies between 0 and 1. This leads to the problem of choosing which threshold to use in order to separate the estimated vulnerable from the estimated non-vulnerable individuals. This threshold is usually set at 50 percent. However, this choice appears to be clearly arbitrary. Aim of this paper is to study the optimal choice of vulnerability threshold that is based on the ROC curve, by providing a comparison among three different strategies.

Abstract *Nell'analisi della vulnerabilità si stima la probabilità di essere poveri in futuro, che varia tra 0 e 1. Ciò implica la necessità di individuare una soglia per separare gli individui vulnerabili da quelli non vulnerabili. Questa soglia è di solito fissata pari al 50 percento, e tale scelta appare arbitraria. Scopo del lavoro è individuare una soglia ottimale di vulnerabilità basata sulla curva ROC, mediante un confronto tra tre diverse metodologie.*

Key words: Vulnerability measurement, ROC curve, optimal threshold

1 Introduction

Poverty is an ex-post measure of household welfare. However, the current poverty status may not necessarily be a good indicator of being poor tomorrow. Non-poor households today may have a high probability of becoming poor tomorrow, while some households that are currently poor may be only transiting poverty because of

Chiara Gigliarano
Department of Economics, Università degli Studi dell'Insubria, Varese, Italy. e-mail: chiara.gigliarano@uninsubria.it

Conchita D'Ambrosio
Faculty of Humanities, Education and Social Sciences, Universitè du Luxembourg, Esch-sur-Alzette, Luxembourg. e-mail: conchita.dambrosio@uni.lu

an unexpected temporary shock. Therefore, it seems important not only alleviating the economic condition of those who are poor today, but also preventing people from becoming poor in the future, adopting a forward looking perspective. The economic literature has recently introduced the concept of vulnerability to poverty, referring to the uncertainty of being poor in the future.

The concept of economic vulnerability introduces the notion of uncertainty: future distributions of incomes are indeed unknown and the vulnerability of households is unobservable to the policy-maker. Due to this complexity, there is still lack of consensus about the definition of economic vulnerability to poverty. In this paper we will use the definition of vulnerability of expected poverty (VEP) approach, which defines vulnerability as the ex-ante risk that an individual will be poor in the future. In contrast with poverty, which is an ex-post measure of a household welfare, vulnerability is a forward-looking measure of household welfare. As suggested by Chaudhuri et al. (2002), whereas the status of poor is observable, the status of vulnerable can only be estimated or inferred.

We include among the vulnerable both families that are non-poor and have a high probability of falling below the poverty line in the future, and families who are currently poor and have a high probability of remaining poor in the future.

When estimating vulnerability we estimate a probability of being poor in the future, which varies between 0 and 1, and not a vulnerability status that can take only the values of 0 or 1. This leads to the problem of choosing which threshold to use in order to separate the estimated vulnerable from the estimated non-vulnerable individuals. This threshold is usually set at 50 percent, so that values above 50 percent indicate that the household has a higher probability of being poor than not. However, a household with 51 percent probability of being poor is not that different from a case with a probability of 49 percent. Aim of this note is, therefore, to determine the optimal vulnerability threshold that simultaneously maximizes true positive rate and minimizes false positive rate. For this purpose, we will discuss the criteria based on the ROC curve.

2 Optimal threshold selection based on ROC curve

The Receiver Operating Characteristic (ROC) curve is one of the most common statistical tools useful to assess classifier performance (Lusted, 1971; Krzanowski and Hand, 2009). This curve is generated by plotting the fraction of true positives out of the positives (true positive rate) versus the fraction of false positives out of the negatives (false positive rate), at various threshold settings.

We are interested in classifying individuals (or households) into two categories: vulnerable or not, according to their estimated vulnerability, that is the estimated probability of being poor tomorrow, given a vector of covariates.

Let the continuous random variable X represent the vulnerability (i.e. the probability of being poor tomorrow), with cumulative distribution function F. The random variable X_{NP} refers to the vulnerability of those who will be actually non-poor to-

morrow with c.d.f. F_P while X_P refers to the vulnerability of those who will be poor tomorrow, with c.d.f F_{NP}. The vulnerability is defined in such a way that individuals who will be poor tomorrow tend to have higher estimated vulnerability, while the individuals non-poor tomorrow tend to have lower vulnerability, with a threshold that divides the scores into the two groups. For a pre-specified cut-off c, individual i will be labelled as poor tomorrow if $x_i \geq c$ and as non-poor tomorrow otherwise.

The performance of a vulnerability model can be summarized in a table (confusion matrix) which compares actual and predicted classifications for a fixed cut-off level c. From the confusion matrix four conditional probabilities can be computed: true positive rate (TP), or *sensitivity*, which corresponds to the proportion of those who will be non-poor tomorrow and predicted today as non-vulnerable; true negative rate, or *specificity*, that is the proportion of the poor tomorrow who are predicted today as vulnerable; false positives rate (FP), or 1- *specificity*, which is the proportion of the poor tomorrow who are predicted as non-vulnerable; false negatives rate, i.e. the proportion of the non-poor tomorrow who are predicted as vulnerable.

Given a set of cut-off points and the corresponding confusion matrices, one of the most common methods to select the best model is the ROC curve. The ROC curve is obtained representing, for any fixed cut-off value, a point in the cartesian plane having as x-value the false positive rate and as y-value the true positive rate.

More formally, the true positive rate is $F_{NP}(c) = Pr(X \leq c|\text{Non-poor tomorrow})$, while the false positive rate is $F_P(c) = Pr(X \leq c|\text{Poor tomorrow})$. The ROC curve is defined as a plot of $\{(u, ROC_X(u)), u \in (0,1)\}$, where

$$ROC_X(u) = F_P(F_{NP}^{-1}(u)),$$

that is, it is a plot of the true positive rate, on the vertical axes, versus its corresponding false positive rate on the horizontal axes, as c varies from $+\infty$ to $-\infty$.

The best curve is the one that is leftmost, the ideal one coinciding with the y-axis. Perfect discrimination corresponds to a ROC curve that passes through the point (0,1). At the other extreme, a model that discriminates not better than random, so that $F_{NP}(c) = F_P(c)$ for any c, will give rise to the diagonal line.

The ROC curve shows the overall performance of a classifier (the vulnerability model) across all possible choices of the cut-off. However, policy interventions aimed at targeting anti-poverty strategies require a unique threshold to be chosen that classifies each individual/household into vulnerable or non-vulnerable.

The most common methods used in the literature for fixing the vulnerability line c are: (i) $c = 0.5$: if the individual probability of being poor tomorrow (vulnerability) is greater than his probability of not-being poor tomorrow, then he will be considered vulnerable; (ii) c equal to the observed poverty rate today: if the individual probability of being poor tomorrow is greater than the average, then the individual will be considered vulnerable; (iii) other authors propose to fix a desirable level of true positive rate (e.g. 80% or 90%) and choose the threshold that minimizes the false positive rate (Celidoni 2015, Landau et al. 2012).

However, these choices appear to be clearly arbitrary, and we propose here to consider methods that are aimed at choosing the optimal threshold based on the

ROC curve. Here we will focus on the Youden index and the Distance index to test which is the optimal threshold to be used in order to separate those who will be poor tomorrow from those non-poor and in order to maximize true positive rate while minimizing false positive rate.

1. *Youden index:* The optimal cut-off is the one that maximizes the Youden index YI
$$YI = \max_{c \in \mathbb{R}} (TP - FP),$$
i.e. the probability of correctly classifying the vulnerable individuals. Graphically, the Youden index is the longest vertical distance between the ROC curve and the chance line (i.e. the positive diagonal).
2. *Distance index.* The optimal cut-off is the one that minimizes the distance d from the scenario of perfect discrimination:
$$d = \left(\sqrt{(1 - S_n)^2 + (1 - S_p)^2} \right),$$
where S_n and S_p are Sensitivity and Specificity, respectively. The optimal cut-off point is the one that minimizes the distance d from the scenario of perfect discrimination (that is the point (0,1)). Graphically, it is the shortest distance between errorless classification (the point (0,1)) and the ROC curve. See Wodon (1997) and López-Ratón et al. (2014).

3 An empirical illustration

The empirical application is based on EU-SILC (European Union Statistics on Income and Living Conditions) 2012 longitudinal dataset. EU-SILC collects comparable longitudinal micro data on households income and living conditions referred to years 2009, 2010, 2011, 2012, for 25 EU member states plus Norway and Iceland. For the analysis we consider a sample of representative countries: Belgium, Denmark, Greece, Spain, France, Italy, Luxembourg, United Kingdom.

As income variable we choose the household equivalent disposable income using the modified OECD scale[1]. We then deflate incomes using the harmonised consumer price indices (base year 2005) provided by Eurostat, and we drop all non-positive incomes. The unit of analysis is the household and we refer to the one-year-ahead vulnerability, that is the probability of being poor in the next year.

As empirical strategy we allow for an autoregressive structure in our prediction model for log of income at time t based on information at time $t-1$ (see Calvo and Dercon, 2013). We consider rolling balanced 3-year panels, keeping in the dataset only the households that remain in the panel for three consecutive years (that is 2009-2011 and 2010-2012). For estimating future income we require two waves

[1] The OECD modified scale assigns value of 1 to the household head, of 0.5 to each additional adult member and of 0.3 to each child in the household.

to forecast household income in the third year and estimate its vulnerability, and a third wave to compute ROC curve and assess the estimate's accuracy. Information available for the first two years of each panel are used to forecast household income in the third year. The observed incomes in the third year are then used to determine which households are poor in that year and which are not. We have to note that in SILC information on income refer to the previous year. The logarithm of income was regressed on lagged log income and a number of explanatory variables. In particular, for forecasting income for year $t+1 = 2011$ and $t+1 = 2010$, we

$$\begin{aligned} \text{estimate } \alpha \text{ and } \beta \text{ using} \quad y_t &= \alpha y_{t-1} + \beta X_t + \varepsilon_t, \\ \text{predict using} \quad \hat{y}_{t+1} &= \hat{\alpha} y_t + \hat{\beta} X_{t+1}. \end{aligned}$$

The covariates included are both at household and at head of the household[2] level: log income of the previous year, number of children (0-17 years old), number of elderly (65+), number of member 18-34 years old and number of member 35-64 years old tenure status, number of rooms, number of fully employed and number of part-time workers in the household, gender, marital status and education of the householder. Using these predictions, one-year-ahead household vulnerability can be estimated. Due to space constraints, in this paper we will show only results related to vulnerability estimation for the year 2011.

The vulnerability is here estimated using a parametric approach proposed in the literature: the vulnerability, $p(\tilde{y} < z)$, where z is the poverty line, is estimated using a normal distribution with mean corresponding to the predicted future income and standard deviation put equal to 1 for each household (Celidoni, 2015). As poverty line we consider the 60% of the median income in each country.

The ROC curve is obtained by comparing the true positive and false positive rates in the third year of the panel for each possible cut-off applied to the predicted income for the third year. Next, we compute the Youden index and the Distance index illustrated in the methodological section. In general, the two indicators may suggest different optimal thresholds. In our case, the Youden and Distance index provide exactly the same values for all countries but the UK. Since the ROC curve is symmetric around the negative diagonal of the unit square, the two decision rules point to the same optimal threshold. That means that, if we consider as poor the cases whose predicted probability is greater than the optimal threshold and as non-poor the cases below this optimal cut-off point, we are minimizing both the distance from the perfect discrimination scenario and maximizing the trade-off between true alarm rate and false alarm rate. For each country considered, Table 1 provides the optimal cut-off point, together with the corresponding true positive rate and the false positive rate.

[2] The householder is here defined as the person responsible for the rent or mortgage interest.

References

1. Calvo, C., Dercon, S.: Vulnerability to individual and aggregate poverty. Soc. Choice Welf. **41**, 721-740 (2013)
2. Chaudhuri, S., Jalan, J., Suryahadi, A.: Assessing household vulnerability to poverty from cross-sectional data: A methodology and estimates from Indonesia. Discussion Paper Series 0102-52, Columbia University, Department of Economics (2002)
3. Celidoni, M.: Decomposing vulnerability to poverty. Review Income Wealth **61**, 1493-1506 (2015)
4. Krzanowski, W.J., Hand, D.J.: ROC curves for continuous data. Chapman and Hall, London (2009)
5. Landau, K., Klasen, S., Zucchini, W.: Poverty, equity and growth in developing and transition countries: Statistical methods and empirical analysis. Discussions Papers 118, Georg-August-Universitaet-Goettingen (2012)
6. Lusted, L.B.: Signal detectability and medical decision-making. Science **171**, 1217-1219 (1971)
7. López-Ratón, M., Rodríguez-Álvarez, M., Cadarso-Suárez, C., Gude-Sampedro, F.: Optimal-Cutpoints: an R package for selecting optimal cutpoints in diagnostic tests. J. Stat. Soft. **61**, 1-36 (2014)
8. Wodon, Q.: Targeting the poor using ROC curves. World Development **25**, 2083-2092 (1997)

Table 1 Comparison of the optimal thresholds vs. the 50% threshold, for several EU countries

Country	Method	True Positive	False Positive	Optimal threshold	% vulnerable (opt.)	% vulnerable (50%)	% poor % in 2012
BE	Youden	0.75	0.21	0.39	27.11	17.52	12.03
	Distance	0.75	0.21	0.39			
	threshold 50%	0.61	0.12				
DK	Youden	0.80	0.12	0.25	19.52	6.73	10.51
	Distance	0.80	0.12	0.25			
	threshold 50%	0.49	0.02				
EL	Youden	0.69	0.21	0.35	29.06	16.54	17.96
	Distance	0.69	0.21	0.35			
	threshold 50%	0.47	0.10				
ES	Youden	0.77	0.23	0.24	32.82	10.29	17.5
	Distance	0.77	0.23	0.24			
	threshold 50%	0.36	0.05				
FR	Youden	0.82	0.22	0.28	30.08	10.46	12.8
	Distance	0.82	0.22	0.28			
	threshold 50%	0.49	0.05				
IT	Youden	0.81	0.19	0.35	30.7	17.52	18.3
	Distance	0.81	0.19	0.35			
	threshold 50%	0.63	0.07				
LU	Youden	0.82	0.15	0.30	24.05	9.26	13.89
	Distance	0.82	0.15	0.30			
	threshold 50%	0.47	0.03				
UK	Youden	0.65	0.17	0.36	23.97	12.12	14.66
	Distance	0.73	0.26	0.27			
	threshold 50%	0.43	0.07				

New advances in applications, a Bayesian nonparametric perspective

Bayesian Mixture Models for Latent Class Analysis

Modelli mistura bayesiani per l'analisi delle classi latenti

Raffaele Argiento, Bruno Bodin and Maria De Iorio

Abstract Bayesian mixture models provide rich and flexible class tools which are particularly useful when there is unobserved heterogeneity in the data. When the number of subpopulations, called components, is assumed random, we allow the data to determine the complexity of the model. The latter property allows us to include a finite mixture model with a random number of components into the wider class of Bayesian nonparametric models. In this paper we consider multivariate discrete data, so that the class of mixtures is also referred to as *latent class* models. In particular, we let the number of latent classes to be random, and resort to Bayesian nonparametric techniques to devise a MCMC algorithm. The model is illustrated on an benchmark application dealing with *role conflict*.

Abstract *Le misture costituiscono una famiglia estramamente ricca e flessibile di modelli per analizzare dati provenienti da fonti etorogenee. In ambito bayesiano, quando il numero di sottopopolazioni (ovvero di componenti) è assunto aleatorio, sono i dati a determinare la complessità del modello. Quest'ultima proprietà rende una mistura con numero di componenti aleatorio un modello non parametrico. Un* modello a classi latenti *è un modello mistura per dati mutivariati e discreti. In questo lavoro ricorriamo a delle tecniche bayesiane non parametriche per costruire un algoritmo MCMC per l'analisi a posteriori di un modello a classi latenti bayesiano. Le prestazioni del modello sono illustrate mediante un applicazione riguardante lo studio di diversi comportamenti in caso di* conflitto fra ruoli.

Key words: Mixture Models, Latent Class Analysis, Bayesian Inference

Raffaele Argiento
Università Cattolica del Sacro Cuore e-mail: raffaele.argiento@unicatt.it

Bruno Bodin, Maria De Iorio
Yale-NUS College, Singapore e-mail: bruno.bodin@yale-nus.edu.sg, maria@yale-nus.edu.sg

1 Introduction

Mixture models are a very powerful and natural statistical tool to model data from heterogeneous populations. Under this class of models, observations $Y_1, \ldots, Y_n$ are assumed to have arisen from one of M (finite or infinite) groups, each group being suitably modelled by a density typically from a parametric family. A Bayesian mixture model for a set of observation in $\mathbb{R}^d$ with $d \geq 1$ is usually expressed as follow

$$\begin{aligned} &Y_i|\tau_1, \ldots, \tau_M, \mathbf{w}, M \overset{\text{iid}}{\sim} \sum_{m=1}^{M} w_m f(y \mid \tau_m), \qquad i = 1, \ldots, n \\ &\tau_m \mid M \overset{\text{iid}}{\sim} p_0(\tau_m) \qquad m = 1, \ldots, M \\ &\mathbf{w} \mid M \sim p_W(\mathbf{w}) \\ &M \sim q_M(m), \qquad m = 1, 2, \ldots \end{aligned} \tag{1}$$

where the *kernel* $f(\cdot \mid \tau_m)$ belongs to a parametric family of density on $\mathbb{R}^d$, with finite dimensional parameter τ_m, while the vector $\mathbf{w} = (w_1, \ldots, w_M)$ is such that $\sum_{m=1}^{M} w_m = 1$ and is referred as vector of the *weights* of the mixture, w_m being the weight of the $m-$th component in the population density. The vector of parameters τ_m assumes values in $\Theta \subset \mathbb{R}^s$, $s \geq 1$ and is assigned a prior density p_0. In a fully Bayesian approach the number of components of the mixture is assumed random with a prior here denoted by q_M. Moreover, conditionally on M a prior distribution with the support on the $M-1$ dimensional simplex should be assumed to the vector of weights $\mathbf{w}$, to ensure that its components represent the probability that an observation belongs to the corresponding component. Data $\{Y_i\}$ generated from a mixture can be univariate or multivariate, continuous, discrete-valued or mixed-type, outcomes of a regression model, or even time series data; see [3] for a comprehensive review of finite mixture distributions. When data are multivariate discrete a mixture model is also known as *latent class analysis* [5].

Recently in [2] we have developed a new theoretical framework for finite mixture model (i.e. $M < \infty$ almost surely). The latter work is based on the key observation that the hierarchical parametric distribution q_M, p_W, p_0 introduced in Eq. (1) define the law of an almost surely (a.s.) finite-dimensional random probability measure on the parameter space Θ. Indeed, given a realization $M, \boldsymbol{w}, \boldsymbol{\tau}$ we can define

$$P(d\theta) = \sum_{m=1}^{M} w_m \delta_{\tau_m}(d\theta) \tag{2}$$

This implies that the joint probability distribution on M, $\mathbf{w}$ and $\boldsymbol{\tau}$ induces a distribution on P defined in Eq. (2), whose support is the space of the a.s. finite-dimensional random probability measures on Θ. From this observation, the link between infinite (nonparametric) and finite mixture models becomes evident as the model in Eq. (1) can be easily rewritten as

$$\begin{aligned} &Y_1,\dots,Y_n|\theta_1,\dots,\theta_n \overset{ind}{\sim} f(y;\theta_i) \\ &\theta_1,\dots,\theta_n|P \overset{iid}{\sim} P \\ &P \sim \mathscr{P} \end{aligned} \tag{3}$$

where P is defined in Eq. (2) and $\mathscr{P}$ is the law of P defined via q_M, P_W, P_0. In [2] a constructive definition of $\mathscr{P}$ is provided by considering the weights $\{w_m\}$ as the normalised jumps of a *finite point process* and the parameters $\{\tau_m\}$ are defined in terms of realisations of the same point process. As in any mixture, the parameter θ_i's in Model (3) are equal to one of the τ_m in Eq. (2), depending on which component the i-th observation is assigned to.

2 Normalized independent finite point processes

Let Θ be the space of parameter of the family of the kernel components $f(\cdot,\tau)$. We consider the collection of random variables $\{(S_1,\tau_1),\dots,(S_M,\tau_M)\}$ such that $(S_m,\tau_m) \in \mathbb{R}^+ \times \Theta$ for any m and $M \geq 1$. The law of the random variables is assigned hierarchically so that

$$\begin{aligned} \mathscr{L}((S_1,\tau_1),\dots,(S_M,\tau_M)) &= \mathscr{L}((S_1,\tau_1),\dots,(S_m,\tau_m)|M=m)P(M=m) \\ &= \prod_{l=1}^{m} \mathscr{L}(S_l,\tau_l|M=m)P(M=m), \quad m=1,2,\dots \end{aligned}$$

Moreover we assume that S_l and τ_l are conditionally independent with density $h(s)$ and p_0 respectively, while we denote with q_M the probability mass function of M. Since we are assuming $q_M(0)=0$, the random variable $T := \sum_{j=1}^{M} S_j$ is almost surely different from 0, so that we can give the following definition.

Definition 1. A normalized independent finite point process (Norm-IFPP) with parameter h and q_M and p_0 is the discrete probability measure on Θ defined by

$$P(\cdot) = \sum_{m=1}^{M} P_m \delta_{\tau_m}(\cdot) \overset{d}{=} \sum_{m=1}^{M} \frac{S_m}{T} \delta_{\tau_m}(\cdot) \tag{4}$$

where $T = \sum_{m=1}^{M} S_m$. We will use the notation $P \sim Norm-IFPP(h, q_M, p_0)$.

Let $(\theta_1,\dots,\theta_n)$ be a sample from a Norm-IFPP; we denote by $\theta_1^*,\dots,\theta_K^*$ the unique values of this sample (observed with probability greater than 0). In this way we define a random partition by letting $\boldsymbol{\rho}_n := \{C_1,\dots,C_K\}$ on the set $\mathbb{N}_n := \{1,\dots,n\}$ where $C_j = \{i : \theta_i = \theta_j^*\}$ for $j = 1,\dots,K$.

3 Latent Class analysis

In this paper we consider the case when data $Y_i = (Y_{i1}, \cdots, Y_{id})$ are vectors of dichotomous responses, i.e. $Y_i \in \{0,1\}^d$ and the kernels for the components of the mixture are assumed to be multivariate Bernoulli distribution. Therefore, the component specific parameter vector is $\tau_m = (\tau_{m1}, \cdots, \tau_{md}) \in [0,1]^d$, where τ_{mj} denotes the probability that $Y_{ij} = 1$, for $j = \{1, \ldots, d\}$ in the m-th component (subpopulation).

When data are multivariate discrete, the mixture is referred to as *Latent Class model* [5, 4]. In the Latent Class analysis there are d possible characteristics, and τ_{mj} represents the probability (in the m-th subpopulation) that an individual endorses the $j-$th characteristic.

In what follows we set the kernels of the mixture as

$$f(y_i \mid \tau_m) = \prod_{j=1}^{d} \tau_{mj}^{y_{ij}} (1-\tau_{mj})^{1-y_{ij}}$$

Moreover, as mixing distribution, we assume $P \sim Norm - IFPP(h, q_M, p_0)$ where the prior for the locations, p_0, is the density of independent beta priors with parameters $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)$:

$$p_0(\theta \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{d} \pi_{\theta_j}(\theta_j \mid \alpha_j, \beta_j) = \prod_{j=1}^{d} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j - 1} (1-\theta_j)^{\beta_j - 1}, \quad 0 \le \theta_j \le 1,\ j = 1, \ldots, d$$

This prior is conjugate to the multivariate Bernoulli density and, as such, greatly simplifies computations. The prior on the weights is obtained by normalizing M conditionally iid gamma variables with same shape parameter γ and rate 1, so that

$$h(s;\gamma) = \frac{s^{\gamma-1} e^{-x}}{\Gamma(\gamma)} \qquad s > 0, \ \text{ for } \gamma > 0 \tag{5}$$

Finally, for the number of components we assume a Negative Binomial density with parameters $0 \le p \le 1$ and $r > 0$ and support on $\{1, 2, \ldots\}$, i.e.

$$q_M(m) = \frac{\Gamma(r+m-1)}{(m-1)!\Gamma(r)} p^{m-1} (1-p)^r, \quad m = 1, 2, \ldots \tag{6}$$

The posterior distribution of our latent class model can be approximated using a blocked Gibbs sampling scheme by a straightforward modification of the *conditional* algorithm described in [2]. The algorithm has been implemented `cpp` and is included the R package `AnTMAN` [1].

Latent Class Analysis to study *role of conflict*. The *role conflict* data consist of four observed dichotomous variables. These variables describe the response patterns for $n = 216$ individuals to four questionnaire items in which four different situations (denoted by A, B, C,and D) of role conflict are considered. The respon-

dents are cross-classified (see Table 1) with respect to whether they tend toward "universalistic" values (coded by + or 1) or "particularistic" values (coded by − or 0) when confronted by situations of role conflict.

A	B	C	D	Obs. Freq.	A	B	C	D	Obs. Freq.
+	+	+	+	42	-	+	+	+	1
+	+	+	-	23	-	+	+	-	4
+	+	-	+	6	-	+	-	+	1
+	+	-	-	25	-	+	-	-	6
+	-	+	+	6	-	-	+	+	2
+	-	+	-	24	-	-	+	-	9
+	-	-	+	7	-	-	-	+	2
+	-	-	-	38	-	-	-	-	20

Table 1 The data collect responses of individuals to a questionnaire concerning four different situations of role conflict (situations A, B, C, and D). The respondent tends either towards universalistic values (+) or toward particularistic values (-) when responding to the situation with which he or she is confronted. Source of the data: [4].

The dataset has been initially analysed by [5] who employ a particular latent class model that has five latent classes. Further analysis of the dataset is also discussed in [4], which highlight the importance of inferring the number of classes for the data, since the interpretation of the latent classes can drastically change. In this work we allow the number of possible classes M to be random, and let the data drive the inference on this parameter. To fit the data we use the R package `AntMAN` [1]. We set $\boldsymbol{\alpha} = \{1,...,1\}$, $\boldsymbol{\beta} = \{1,...,1\}$, and we add an extra level to the hierarchy by assuming $\gamma \sim Gamma(1,1)$, $p \sim Beta(1,1)$ while $r = 1$ (see Eq. (5) and (6)). All the results discussed here are based on a posterior sample of size 5000 obtained after a burn-in period of 5000 iterations and with a thinning of 50 iterations.

The posterior distribution on the number of K clusters (i.e. active classes) is displayed in the first panel of Figure 1. The posterior mode of the latter distribution is 2. The estimated clustering configuration is obtained by minimizing the Binder loss function, obtaining two active classes which cover 32.4% and 67.6% of the data respectively. We refer to [6] and [7] for more details on the Binder loss function and on the computational strategies to minimize it.

To better interpret the groups obtained under the mixture model, we estimate cluster specific parameters, i.e. the parameters $\theta_k^\star$ for $k = 1,\dots,\hat{K}$ that represent the probability of an individual to be universalistic in the four situation A, B, C and D given that it belongs to cluster k. These estimates are obtained by *conditioning* to the estimated clusters, the results are displayed in the last two panels of Figure 1.

We can interpret the two classes as a "universalistically inclined" latent class and a "particularistically inclined" latent class. The probability of a universalistic response in situation A is 0.90 for those who are in the universalistically inclined latent class; for situations B, C, and D, the corresponding probabilities are 0.77, 0.76, and 0.59, respectively. In addition, the probability of a universalistic response in situation A is 0.716 for those who are in the particularistically inclined latent class;

for situations B, C, and D, the corresponding probabilities are 0.34, 0.37, and 0.15, respectively. In situation A, the modal response is always universalistic, but the corresponding probability of a universalistic response is reduced in the particularistic latent class from 0.90 to 0.71.

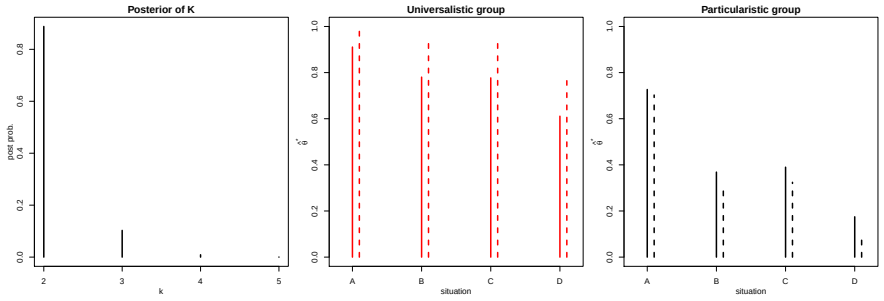


Fig. 1 Posterior inference on clustering. The first panel displays the posterior distribution of the number of active classes. The second and third panels report the posterior probabilities (continuous lines) of observing a 1 for the four situations and the corresponding empirical relative frequencies (dashed lines).

In Figure 1, the observed frequencies of 1 in each class are reported as dashed lines. We can see how, especially for the "universalistic group", the Bayesian parameter estimates are lower than the observed frequencies, due to the correlation induced by the latent class model, the uncertainty on cluster estimation as well as the prior specification. Further investigation is needed to better understand the latter issue. In fact, the problem of estimating a partition of the data using the posterior samples of a MCMC algorithm in a mixture model is still object of active research. We refer to [6] and [7] for recent discussions of this problem.

References

1. R. Argiento, B. Bodin, and M. De Iorio. *AntMAN: Anthology of Mixture Analysis Tools*, 2019. R package version 1.0.
2. R. Argiento and M. De Iorio. Is Infinity that far? a Bayesian nonparametric perspective of finite mixture models. *arXiv preprint arXiv:1904.09733*, 2019.
3. S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert. *Handbook of Mixture Analysis*. Chapman and Hall/CRC, 2019.
4. J. A. Hagenaars and A. L. McCutcheon. *Applied latent class analysis*. Cambridge University Press, 2002.
5. P. F. Lazarsfeld and N. W. Henry. *Latent Structure Analysis*. Houghton Mifflin Co., 1968.
6. R. Rastelli and N. Friel. Optimal Bayesian estimators for latent variable cluster models. *Statistics and computing*, pages 1–18, 2017.
7. S. Wade and Z. Ghahramani. Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*, 13(2):559–626, 2018.

Non-Parametric Inference and Forecasting of Functional and Object Data

An interpretable estimator for the function-on-function linear regression model with application to the *Canadian weather* data

Uno stimatore interpretabile per il modello di regressione lineare con regressore e risposta funzionale con un'applicazione ai Canadian weather data

Fabio Centofanti and Matteo Fontana

Abstract An interpretable estimator for the functional linear regression model, where values of the response function depend on the whole trajectory of the covariate function, is described. It is able to increase the interpretability of the model by better identifying domain regions where the covariate does not influence the response. This property is ensured by a functional LASSO penalty and two roughness penalties. The application of the proposed estimator to the well known *Canadian weather* dataset shows its practical advantages.

Abstract *In questo lavoro viene introdotto uno stimatore del modello di regressione lineare con regressore e risposta funzionali. Lo stimatore proposto è in grado di migliorare l'interpretabilità del modello identificando in maniera più accurata le parti del dominio dove la covariata non influenza (significativamente) la risposta. Tale proprietà viene ottenuta mediante un opportuno termine di penalizzazione LASSO funzionale e due termini di penalizzazione della rugosità. I vantaggi pratici dello stimatore proposto sono evidenziati mediante applicazione ai dati Canadian weather.*

Key words: functional data analysis, linear regression, LASSO, B-splines

Fabio Centofanti
Dept. of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125, Naples, Italy
e-mail: fabio.centofanti@unina.it

Matteo Fontana
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy
e-mail: matteo.fontana@polimi.it

1 Methods

The development in data acquisition systems allow massive amounts of data to be recorded at high-rate and modelled as functions defined on compact domains, i.e., *functional data* . Functional data analysis (FDA) is the set of tools to deal with this type of data [12, 7, 8, 9]. In this paper, we consider the generalization of the classical linear regression analysis to the case where either the covariate or the response or both are functional data, i.e. functional linear regression (FLR) [11, 12, 7, 5, 3, 10, 6, 2]. In particular, we study the Function-on-Function (FoF) linear regression model, where both the predictor and the response variable are functions and each value of the latter, for any domain point, depends on the full trajectory of the former that is, for $i = 1, \ldots, n$,

$$Y_i(t) = \int_{\mathscr{S}} X_i(s)\,\beta(s,t)\,ds + \varepsilon_i(t) \quad t \in \mathscr{T}. \tag{1}$$

The pairs (X_i, Y_i), $i = 1, \ldots, n$, are independent realizations of the covariate X and response Y, that are smooth random functions with mean zero defined on the compact domains $\mathscr{T}$ and $\mathscr{S}$, and ε_i are independent random functional errors with mean zero and variance function σ^2 that are independent of X_i. The bivariate function β is the *coefficient function*.

In this work, we consider an interpretable estimator of the coefficient function β, named S-LASSO (Smooth plus LASSO), that has been proposed in [4]. It is locally sparse, i.e., is zero on the region where β is zero (*null region*), and, smooth on the region where β is different from zero (*non-null region*). The S-LASSO estimator $\hat{\beta}_{SL}$ is obtained as the minimum of an objective function, composed by the usual sum of squared errors added to two smoothness penalties and a functional LASSO penalty, where the latter ensures sparseness of the estimator.

From a computational point of view, $\hat{\beta}_{SL}$ is computed by means of a modified version of the *orthant-wise limited-memory quasi-Newton* (OWL-QN) algorithm [1], after the introduction of a suitable finite dimensional approximation of the optimization problem.

2 Real-Data Examples: Canadian Weather Data

In this section, we apply the S-LASSO estimator to the Canadian Weather data. The S-LASSO estimator is compared with a the estimator proposed by [12], referred to as SMOOTH, with regularization obtained by introducing two roughness penalties. We aim to demonstrate that the S-LASSO estimator has advantages in terms of both prediction accuracy and interpretability, over the SMOOTH estimator. The Canadian weather data have been studied by [12] and [13]. The data set contains the daily mean temperature curves, measured in Celsius degree, and the log-scale of the daily rainfall profiles, measured in millimeter, recorded at 35 cities in Canada

(Figure 1). The aim is to predict the log-daily rainfall based on the daily temper-

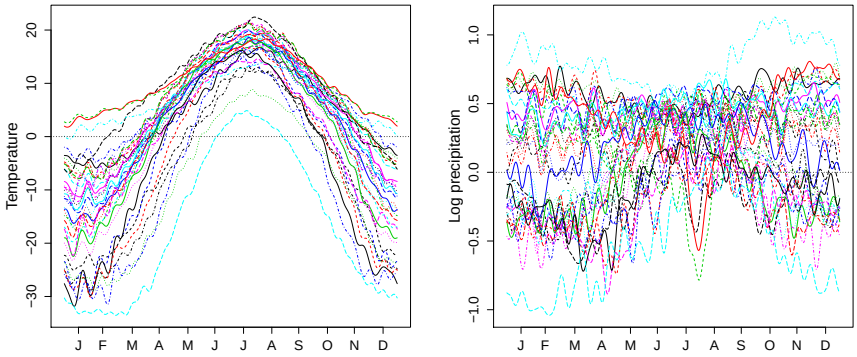


Fig. 1 Daily mean temperature and log-daily rainfall profiles at 35 cities in Canada over the year.

ature using the model reported in Equation (1). Figure 2 shows the S-LASSO and SMOOTH estimates of the coefficient function β. The SMOOTH estimate is ob-

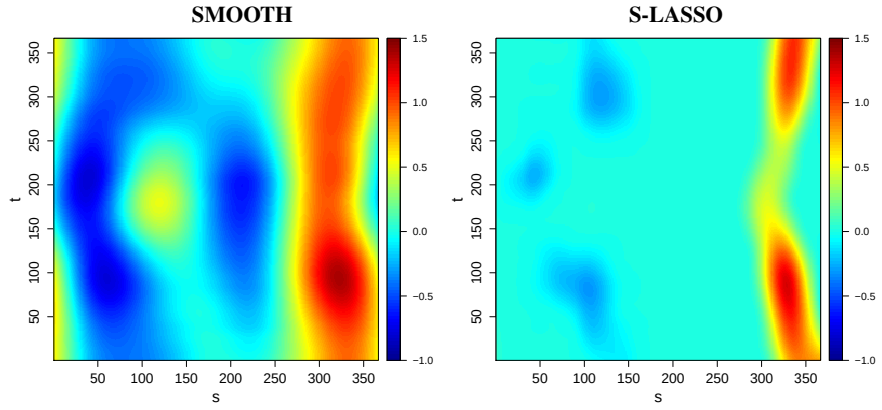


Fig. 2 SMOOTH (on the right) and S-LASSO (on the left) estimates of the coefficient functions β at different days for the Canadian weather data.

tained using a Fourier basis and roughness penalties were chosen by using 10-fold cross-validation over an opportune grid of values. The S-LASSO estimates is zero over large domain portions. In particular, except for values from July through August ($t \in [210,240]$), it is always zero in summer months ($s \in [180,270]$) and in January and February ($s \in [1,58]$). Thus, in those months rainfalls are not significantly affected by daily temperature during the year. In addition, temperature in

October, November and December ($s \in [300, 365]$) has a positive effect on the daily rainfalls. That is, the higher (the lower) the temperature is in October, November and December, the heavier (the lighter) the precipitations are during the year. The S-LASSO estimate in spring months ($s \in [90, 150]$) is negative form January through April ($t \in [1, 120]$), and from October through December ($t \in [305, 365]$). Thus, the higher (the lower) the temperature is in the spring the lighter (the heavier) the daily rainfalls are from October through April. Finally, the S-LASSO have better prediction performance than the SMOOTH one, that is, 10-fold cross-validation mean squared errors are 22.314 and 22.365, respectively.

References

1. Andrew, G., Gao, J.: Scalable training of l 1-regularized log-linear models. In: Proceedings of the 24th international conference on Machine learning, pp. 33–40. ACM (2007)
2. Besse, P.C., Cardot, H.: Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. Canadian Journal of Statistics **24**(4), 467–487 (1996)
3. Cardot, H., Ferraty, F., Sarda, P.: Spline estimators for the functional linear model. Statistica Sinica pp. 571–591 (2003)
4. Centofanti, F., Fontana, M., Lepore, A., Vantini, S.: Smooth lasso estimator for function-on-function linear regression model. Manuscript - MOX report (2019)
5. Cuevas, A.: A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference **147**, 1–23 (2014)
6. Hall, P., Horowitz, J.L., et al.: Methodology and convergence rates for functional linear regression. The Annals of Statistics **35**(1), 70–91 (2007)
7. Horváth, L., Kokoszka, P.: Inference for functional data with applications, vol. 200. Springer Science & Business Media (2012)
8. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons (2015)
9. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
10. Li, Y., Hsing, T., et al.: On rates of convergence in functional linear regression. Journal of Multivariate Analysis **98**(9), 1782–1804 (2007)
11. Morris, J.S.: Functional regression. Annual Review of Statistics and Its Application **2**, 321–359 (2015)
12. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer (2005)
13. Sun, X., Du, P., Wang, X., Ma, P.: Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. Journal of the American Statistical Association pp. 1–11 (2018)

Statistical process monitoring of multivariate profiles from ship operating conditions

Monitoraggio statistico di processo delle condizioni operative di una nave basato su analisi di dati funzionali

Christian Capezza

Abstract Motivated by CO_2 emissions monitoring from maritime transportation, we propose and apply statistical methods for industrial process monitoring of scalar quality characteristic in presence of multivariate functional covariates. The two main objectives are (i) the prediction of CO_2 emissions based on covariates' profiles describing the ship operating conditions at each voyage and (ii) the statistical process monitoring of ship operating conditions and CO_2 emissions based on functional control charts. We aim to provide industrial practitioners with interpretable tools that give clear indications of anomalies by identifying the related causes, possibly in real-time. We show the evolution of the methods proposed in our research from multivariate methods, to feature-oriented and functional data analysis approaches.

Abstract *Motivati dal problema del monitoraggio e la previsione di emissioni di CO_2 nel trasporto marittimo, vengono proposti metodi statistici per il monitoraggio di processi industriali. I due obiettivi principali possono essere ricondotti (i) alla previsione di emissioni di CO_2 basata sui profili di covariate che descrivono le condizioni operative della nave ad ogni viaggio e (ii) al monitoraggio statistico di processo delle condizioni operative e delle emissioni di CO_2 basato su carte di controllo funzionali. L'obiettivo del lavoro è la definizione di carte di controllo interpretabili dagli ingegneri navali riguardo le possibili anomalie e l'identifiazione delle relative cause. I metodi di analisi statistica affrontati vengono presentati per livello di complessità crescente: a partire da metodi di analisi multivariata, approcci feature-oriented, fino all'analisi di dati funzionali.*

Key words: Functional control charts, profile-driven features, real-time monitoring.

Christian Capezza
Department of Industrial Engineering, University of Naples Federico II, Naples, e-mail: christian.capezza@unina.it

1 Introduction

Moving from the multivariate data analysis setting [1, 3, 6, 10, 12, 13, 15], the aim of this work is to provide statistical methods for industrial process monitoring of a scalar quality characteristic in presence of multivariate functional covariates. The industrial context is the prediction and monitoring of CO_2 emissions from maritime transportation. Specifically, the two main objectives are (i) the prediction of CO_2 emissions on the basis of covariates' profiles describing the ship operating conditions at each voyage and (ii) the statistical process monitoring of ship operating conditions and CO_2 emissions based on functional control charts.

The relevance of the industrial scenario is highlighted at international level by the EU regulation 2015/757 [7], which urges shipping operators to set up systems for the monitoring, reporting and verification of CO_2 emissions. The aim is to give greater transparency to operations and public access to CO_2 emissions data in the shipping sector. On the other hand, modern ships allow the continuous acquisition of operational data, which calls for the application of opportune statistical methods for high-dimensional data. In fact, multi-sensor systems installed on board of modern ships stream massive amounts of observational data with high frequency. For each voyage, these data can be considered to be varying over a continuous domain, then ship operating conditions can be described by sensor signals and stored as profiles. However, currently there is no standard solution or method available in the shipping industry that can be robustly adopted in real environments.

In the following sections, we show the evolution of the methods proposed in our research from multivariate methods, to feature-oriented and functional data analysis approaches. In particular, in Section 2 we describe the dataset used in the proposed approaches. In Section 3 we describe methods for the prediction of ship CO_2 emissions, while in Section 4 we describe methods for statistical process monitoring of ship performance.

2 Data description

In the following, we briefly describe the navigation data set analyzed using the proposed approaches. The data are acquired from a Ro-pax ship sailing over different route. Each model is built for a specific route, and each observation corresponds to a voyage. The response variable is the total CO_2 emissions in the navigation phase per each voyage, measured in tons. The covariates are stored as profiles since observations are available at five minute frequency, we list them as follows. The *cumulative sailing time* variable measures the voyage navigation time. The *speed over ground* (SOG) variable is the ratio between the distance travelled by the ship and the cumulative sailing time. The *acceleration* variable is obtained as the derivative of SOG. The *power difference between port and starboard propeller shafts* indicates possible anomalies or malfunctioning in the main engines, when one of the main engines is turned off. The *Distance from the nominal route* variable is the distance

from the nominal route of the current position of the vessel. The *longitudinal wind component* and *transverse wind component* variables are obtained on the basis of the wind speed direction relative to the ship acquired by the anemometer sensor. The *air temperature* variable is the average of the temperatures measured from the sensors installed on the four main engines. The *Trim* variable is obtained through the inclinometer sensor measurements. Additional information about the variables can be found in [1, 5, 6].

3 Methods for the prediction of ship CO_2 emissions

The marine engineering literature mainly relies on the use of white-box models for prediction of CO_2 emissions. The most common example is the speed-power curve [18, 20], which describes an ideal univariate relationship between the engine power, which is related to the CO_2 emissions, and the vessel speed and is usually calibrated through dedicated tests. However, this curve overlooks other factors affecting the vessel during navigation and leads to poor predictions of CO_2 emissions. Nowadays, the statistical and data science domains offer potentially interesting tools to circumvent this limitation.

3.1 Feature-oriented methods

The profile of variables automatically acquired at each voyage by multi-sensor systems is typically complex, unstructured, intrinsically collinear and with non-stationary behavior. In this scenario, classical approaches used in the naval literature may fail or are at least suboptimal, since they are limited to the analysis of averages per voyage. However, in spite of the easier interpretability, compressing a variable profile into a single average value may lead to significant information loss and to discarding most of the relevant dynamic patterns.

In the opposite spectrum of complexity, multivariate statistical methods commonly used for monitoring batch processes usually require the implementation of data pre-processing techniques that constitute an additional challenge for practitioners and may hamper their practical usability. For example, data needs to be correctly unfolded to handle its three-way structure, resulting in a very large number of pseudo-variables and model parameters. Furthermore, complex synchronization methods are required in order to ensure that the voyages' major landmarks are aligned and that all voyages have the same number of observations.

The batch process monitoring literature is vast and another class of approaches that is growing in importance is the class of feature-oriented methods. These methods are simpler to apply because they do not require synchronization and tend to be more parsimonious since the number of model parameters is smaller. Examples of feature-oriented methods include profile-driven features, recently proposed by

[17], and statistical pattern analysis, proposed by [8]. These techniques compress each variable into a small number of features that can be utilized for data-driven model building. We showed the successful implementation of these techniques in a shipping industry application in [16].

3.2 Scalar-on-function regression

The most complex approach to the prediction problem avoids the simplification of feature-oriented methods that extract relevant features from the profiles, but it treats the data describing the operating conditions of a ship over each voyage as unique, complex mathematical objects, such as functions or vectors of functions. In this context, functional data analysis techniques, in particular scalar-on-function regression, can be used to predict the CO_2 emissions at each voyage.

Denote by y the scalar response variable representing the CO_2 emissions at each voyage, let $\{(\boldsymbol{X}_i, y_i)\}_{i=1,\ldots,n}$ be a random sample from $(\boldsymbol{X}, y)$, with $\boldsymbol{X}_i = \boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{iP}(t))$ being a vector of functional covariates. The conditional distribution of y_i given the corresponding observation of the functional covariates $\boldsymbol{X}_i$ can be modeled by means of the following scalar-on-function regression model

$$y_i = \beta_0 + \sum_{p=1}^{P} \int_{\mathscr{T}} X_{ip}(t)\beta_p(t)dt + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$, with $\beta_p \in L^2(\mathscr{T})$, the space of square integrable functions, are the coefficient to be estimated, and $\varepsilon_1, \ldots, \varepsilon_n$ are the error terms, which are assumed to be independent identically distributed. Estimation of these types of models requires to deal with the infinite dimensionality of the dataset and is based on multivariate functional principal component analysis. More details on the methodology can be found on [4, 5].

4 Methods for statistical process monitoring of ship operating conditions and CO_2 emissions

There are two relevant aims related to statistical process monitoring of ship performance and CO_2 emissions: (*i*) statistical monitoring of ship fuel consumption and CO_2 emissions for anomaly detection and (*ii*) quantifying CO_2 emissions reduction consequent to energy efficiency initiatives or dry-dock operations. Many of the available methods for prediction and monitoring have strong limitations when applied to high-dimensional and correlated data, or they do not fully exploit all of the available information. There is a common framework among all the methods proposed in this work for statistical process monitoring, which can be summarized in the following steps:

1. the starting point is a regression model where the ship CO_2 emissions are the response variable to be predicted using one of the approaches described in the previous section;
2. since the predictor variable space is characterized by high or infinite dimensionality, dimensionality reduction is applied to both stabilize the estimation of regression coefficients in the prediction problem and to describe covariates in a more efficient and interpretable way, in a lower dimensional subspace; functional principal component analysis and partial-least squares methods are used for this purpose;
3. the dimension reduction provides a nice split of the covariate space into two complementary subspaces: correspondingly, two control charts are used. Usually the Hotelling T^2 statistic is calculated on the variables obtained from the projection of covariates onto the subspace of the components retained in the model, while a squared prediction error statistic monitors the squared distance of the covariates in the original space from the projection subspace; finally a third monitoring statistic looks at the prediction error on the response variable based on the regression model;
4. on the basis of the monitoring statistics and a reference sample of in-control observations, control charts are built to monitor future observations
5. when some control charts detect out-of-control observations, contribution plots are built in order to decompose the monitoring statistics as sums over the covariates, in order to identify the variable(s) responsible of anomalies.

The main advantage of considering profiles instead of single observations per each voyage is the possibility to give real-time predictions and indications on possible anomalies in ship operating conditions during a voyage and on the instant at which anomalies may have occurred. Two main approaches are presented in this work to profile monitoring. Both of them are able to monitor profiles with different length at different voyages (due to the different duration). The first approach is presented in [11, 9] and uses multi-way partial least-squares (PLS) regression [14] and a multilinear version of PLS proposed by [2, 19], which is called three-way PLS. They require a three-dimensional array that contains ship operational reference profiles at given domain points with the following three dimensions: the number of replications, the number of variables, and the number of evaluation points. The second approach is presented in [5] and is based on multivariate functional principal component analysis.

References

1. Bocchetti, D., Lepore, A., Palumbo, B., Vitiello, L.: A statistical approach to ship fuel consumption monitoring. Journal of Ship Research **59**(3), 162–171 (2015)
2. Bro, R.: Multiway calibration. multilinear pls. Journal of chemometrics **10**(1), 47–61 (1996)
3. Capezza, C., Coleman, S., Lepore, A., Palumbo, B., Vitiello, L.: Ship fuel consumption monitoring and fault detection via partial least squares and control charts of navigation data. Transportation Research Part D: Transport and Environment **67**, 375–387 (2019)

4. Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional control charts for monitoring ship operating conditions and CO_2 emissions based on scalar-on-function linear model. In: SIS2019 Statistical Conference—Smart Statistics for Smart Applications. Università Cattolica, Milan, Italy (2019)
5. Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Control charts for monitoring ship operating conditions and CO_2 emissions based on scalar-on-function regression. Applied Stochastic Models in Business and Industry (2020)
6. Erto, P., Lepore, A., Palumbo, B., Vitiello, L.: A procedure for predicting and controlling the ship fuel consumption: Its implementation and test. Quality and Reliability Engineering International **31**(7), 1177–1184 (2015)
7. European Commission: Regulation (EU) 2015/757 of the European Parliament and of the Council of 29 April 2015 on the monitoring, reporting and verification of carbon dioxide emissions from maritime transport, and amending directive 2009/16/EC. Official Journal of the European Union **L123**, 55–76 (2015)
8. He, Q.P., Wang, J.: Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. AIChE journal **57**(1), 107–121 (2011)
9. Lepore, A., Palumbo, B., Capezza, C.: Monitoring ship performance via multi-way partial least-squares analysis of functional data. In: SIS2017 Statistical Conference—Statistics and Data Science: new challenges, new generations. University of Florence, Italy (2017)
10. Lepore, A., Palumbo, B., Capezza, C.: An Empirical Approach to Monitoring Ship CO_2 Emissions via Partial Least-Squares Regression. In: C. Perna, M. Pratesi, A. Ruiz-Gazen (eds.) Studies in Theoretical and Applied Statistics, p. 219. Springer (2018)
11. Lepore, A., Palumbo, B., Capezza, C.: Analysis of profiles for monitoring of modern ship performance via partial least squares methods. Quality and Reliability Engineering International **34**(7), 1424–1436 (2018)
12. Lepore, A., Palumbo, B., Capezza, C.: Orthogonal LS-PLS approach to ship fuel-speed curves for supporting decisions based on operational data. Quality Engineering **31**(3), 386–400 (2019)
13. Lepore, A., Reis, M.S., Palumbo, B., Rendall, R., Capezza, C.: A comparison of advanced regression techniques for predicting ship CO_2 emissions. Quality and Reliability Engineering International **33**(6), 1281–1292 (2017)
14. Nomikos, P., MacGregor, J.F.: Multi-way partial least squares in monitoring batch processes. Chemometrics and intelligent laboratory systems **30**(1), 97–108 (1995)
15. Reis, M.S., Palumbo, B., Lepore, A., Rendall, R., Capezza, C.: On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data. In: SIS2017 Statistical Conference—Statistics and Data Science: new challenges, new generations. University of Florence, Italy (2017)
16. Reis, M.S., Rendall, R., Palumbo, B., Lepore, A., Capezza, C.: Predicting ships' CO_2 emissions using feature-oriented methods. Applied Stochastic Models in Business and Industry **36**(1), 110–123 (2020). DOI 10.1002/asmb.2477
17. Rendall, R., Lu, B., Castillo, I., Chin, S.T., Chiang, L.H., Reis, M.S.: A unifying and integrated framework for feature oriented analysis of batch processes. Industrial & Engineering Chemistry Research **56**(30), 8590–8605 (2017)
18. Schrady, D.A., Smyth, G.K., Vassian, R.B.: Predicting Ship Fuel Consumption: Update. Tech. rep., Naval Postgraduate School, Monterey, California, Department of Operations Research (1996)
19. Smilde, A.K.: Comments on multilinear pls. Journal of Chemometrics **11**(5), 367–377 (1997)
20. Van Manen, J., Van Ossanen, P., Lewis, E.: Principles of naval architecture, second revision, volume II: resistance, propulsion, and vibration. Society of Naval Architects and Marine Engineers (1988)

Prior choice in Bayesian Modelling (SISbayes)

Bayesian Learning of Multiple Essential Graphs

Apprendimento Bayesiano di Grafi Essenziali Multipli

L. La Rocca, F. Castelletti, S. Peluso, F.C. Stingo and G. Consonni

Abstract Structural learning of graphical models is a well-established approach to the identification of complex dependencies in biological networks. We here present a Bayesian methodology for learning directed networks from observational data when distinct subgroups of a population are observed.

Abstract *L'apprendimento strutturale di modelli grafici è un approccio consolidato all'identificazione di dipendenze complesse in reti biologiche. Presentiamo qui una metodologia bayesiana per l'apprendimento di reti orientate da dati osservazionali quando si osservino sottogruppi distinti di una popolazione.*

Key words: Markov equivalence, Markov random field, objective Bayes

Luca La Rocca
Department of Physics, Informatics and Mathematics, Università degli Studi di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy, e-mail: luca.larocca@unimore.it

Federico Castelletti
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Edificio Lanzone 18, 20123 Milano, Italy, e-mail: federico.castelletti@unicatt.it

Stefano Peluso
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Edificio Lanzone 18, 20123 Milano, Italy, e-mail: stefano.peluso@unicatt.it

Francesco Claudio Stingo
Department of Statistics, Computer Science, Applications "G. Parenti", Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy, e-mail: francescoclaudio.stingo@unifi.it

Guido Consonni
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Edificio Lanzone 18, 20123 Milano, Italy, e-mail: guido.consonni@unicatt.it

1 Introduction

Understanding the genome, in terms of structure and function, is at the foundation of genomic medicine [5, 13]. The identification of complex dependencies in biological networks, such as co-expression, gene regulatory and protein interaction networks, plays a major role in this effort and graphical models have been successfully applied to infer the dependence structure of *omics* variables [8, 18, 20]. An important case is that of observations collected from a population divided into groups. For instance, gene expression measurements can be available both for cancer and normal tissue samples. In this case, both differences and similarities are expected between the distinct groups and their networks should be learned together. A few proposals to tackle this problem are available in the literature, including the joint lasso [7], direct estimation of differential networks [23] and Bayesian methods [15, 19].

The above proposals learn undirected networks. However, in genetic analyses, there is a special interest for directed pathways. Furthermore, in graphical modeling, acyclic directed graphs provide an effective data generating mechanism. Motivated by these reasons, we contribute to the literature on learning multiple directed structures [14, 22] a Bayesian method that explicitly deals with *Markov equivalence*: the fact that some directed structures cannot be distinguished through observational data, at least in the Gaussian setting we consider; see Wang & Drton [21] for non-Gaussian observations. Our method learns partially directed structures: if possibile it detects the direction of an association, otherwise it acknowledges that a direction cannot be detected. Sect. 2 provides the necessary background on graphical models and formalizes our structural learning problem; we refer to Drton & Maathuis [9] for a broader presentation of graphical models and structural learning.

2 Structural Learning of Graphical Models

A *simple graph* $\mathbb{G} = (V, \mathcal{E})$ consists of a finite set of vertices $V = \{1, \ldots, q\}$ together with a set of edges $\mathcal{E} \subseteq V \times V$ such that $u \neq v$ for all $(u, v) \in \mathcal{E}$ (no loops). We say that $\mathbb{G}$ contains the *line* $u - v$ if both $(u, v) \in \mathcal{E}$ and $(v, u) \in \mathcal{E}$, while we say that $\mathbb{G}$ contains the *arrow* $u \to v$ if $(u, v) \in E$ but $(v, u) \notin E$. If $u \to v$ is contained in $\mathbb{G}$, we say that u is a *parent* of v. Now let A be a nonempty subset of V. The *parent set* of A is defined as $\mathrm{pa}(A) = \{u \in V | u \text{ is a parent of some } v \in A\}$. The *subgraph* of $\mathbb{G}$ induced by A is defined as $\mathbb{G}_A = (A, \mathcal{E}_A)$, where $\mathcal{E}_A = \mathcal{E} \cap (A \times A)$. An *immorality* is a subgraph of the form $u \to v \leftarrow w$, while a *flag* is a subgraph of the form $u \to v - w$. A sequence of distinct vertices $v_0, v_1, \ldots, v_k$ is a *path* of length k from v_0 to v_k in $\mathbb{G}$ if $\mathbb{G}$ contains $v_{j-1} - v_j$ or $v_{j-1} \to v_j$ for all $j = 1, \ldots, k$. A *cycle* is defined like a path, but with $v_0 = v_k$. An *undirected path* is a path where there are only lines, a *directed cycle* is a cycle where there is at least one arrow.

An *Undirected Graph* (UG) is a simple graph containing only lines. The *skeleton* of $\mathbb{G}$ is the UG obtained from $\mathbb{G}$ by replacing all its arrows with lines. An UG is *chordal* if every cycle of length $k \geq 4$ admits a *chord*, that is, two nonconsecutive

vertices joined by a line. A *directed graph* is a simple graph containing only arrows. A directed graph without cycles is called a *Directed Acyclic Graph* (DAG). A *Chain Graph* (CG) is a simple graph without directed cycles. For a CG $\mathbb{G}$, we call *chain component* an inclusion maximal set of vertices $\tau \subseteq V$ such that all its pairs of vertices are joined by an undirected path. The set of all chain components of a CG, denoted by $\mathcal{T}$, is a partition of V, and all arrows between two chain components share the same direction, so that every CG can be seen as a DAG containing *meta-arrows* between chain components [16]. Both UGs and DAGs are special cases of CGs: UGs have no meta-arrows, while DAGs have singleton chain components.

A *graphical model* is a multivariate statistical model specified by conditional independence constraints that can be read from a graph through a *Markov property* suitable for that kind of graph [9]. We are interested in DAG models, as we explain in Sect. 1, but it turns out that we need a broader class of CG models to deal with observational data: any two DAGs with the same skeleton and the same immoralities specify the same model as their *essential graph*, which is a CG with no flags, whose chain components induce chordal subgraphs and whose meta-arrows have weakly sunsubstantial arrowheads, using the characterization of Roverato [16], so that such a CG is all that can be learned from data collected without interventions. Details and historical information can be found in [16], while interventional data are dealt with by Hauser and Bühlmann [10]; see also Castelletti & Consonni [2].

Now let $Y_1, \ldots, Y_q$ be our observables of interest, in a given population, stacked in a column vector $\mathbf{Y}_{1:q}$ with joint probability density function $f(\mathbf{y}_{1:q})$, $\mathbf{y}_{1:q} \in \mathbb{R}^q$. We assume that f is Gaussian and Markov with respect to an essential graph $\mathbb{G}$:

$$f(\mathbf{y}_{1:q}) = \prod_{\tau \in \mathcal{T}} f_{\mathbb{G}}(\mathbf{y}_\tau | \mathbf{y}_{\mathrm{pa}(\tau)}, \mathbf{B}_{[\tau]}, \boldsymbol{\Omega}_{[\tau]}), \tag{1}$$

where $\mathbf{y}_\tau$ denotes the subvector of $\mathbf{y}$ indexed by τ, $\mathbf{B}_{[\tau]}$ is a $\{|\mathrm{pa}(\tau)|+1\} \times |\tau|$ matrix, $\boldsymbol{\Omega}_{[\tau]}$ is a $|\tau| \times |\tau|$ positive definite matrix, whose entries associated to lines that are not contained in $\mathbb{G}_\tau$ are equal to zero, and $f_{\mathbb{G}}(\mathbf{y}_\tau | \mathbf{y}_{\mathrm{pa}(\tau)}, \mathbf{B}_{[\tau]}, \boldsymbol{\Omega}_{[\tau]})$ denotes the $|\tau|$-dimensional Gaussian density having $\mathbf{B}_{[\tau]}^\top (1, \mathbf{y}_{\mathrm{pa}(\tau)}^\top)^\top$ as mean vector and $\boldsymbol{\Omega}_{[\tau]}^{-1}$ as covariance matrix. Anderson et al. [1] discuss the underlying Markov property and data generating mechanism.

We observe $\mathbf{Y}_{1:q}$ in K subgroups of a population. Let $\mathbf{Y}_{[k]}$ be the $n_k \times q$ data matrix from group k, for $k = 1, \ldots, K$, and denote by $\mathbf{Y}_{[1:K]}$ the full data matrix, obtained by stacking $\mathbf{Y}_{[1]}, \ldots, \mathbf{Y}_{[K]}$ on top of one another. Assuming independent sampling, within and across subgroups, the i-th row of $\mathbf{Y}_{[k]}$, denoted by $\mathbf{y}_{[k]i}^\top$, for $i = 1, \ldots, n_k$, contributes to the likelihood a factor like (1), but with a group specific graph $\mathbb{G}_{[k]}$, as well as group specific matrices $\mathbf{B}_{[k,\tau]}$ and $\boldsymbol{\Omega}_{[k,\tau]}$. We aim at learning the graphs, whose ensemble we denote by $\mathbb{G}_{[1:K]}$ and call a multiple essential graph, while we regard the matrices $\mathbf{B}_{[k,\tau]}$ and $\boldsymbol{\Omega}_{[k,\tau]}$ as nuisance parameters.

We adopt a Bayesian approach and assign a *parameter prior* $\pi_{\mathbb{G}_{[k]}}(\boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]})$ to the nuisance parameters, independently over $k = 1, \ldots, K$ and $\tau \in \mathcal{T}_k$, where $\mathcal{T}_k$ denotes the set of all chain components of $\mathbb{G}_{[k]}$. We can thus eliminate the nuisance parameters by computing the *marginal likelihood*

$$m(\mathbf{Y}_{[1:K]}|\mathbb{G}_{[1:K]}) = \prod_{k=1}^{K} \prod_{\tau \in \mathcal{T}_k} m_{\mathbb{G}_k}(\mathbf{Y}_{[k]\tau}|\mathbf{Y}_{[k]\mathrm{pa}(\tau)}), \tag{2}$$

where $\mathbf{Y}_{[k]\tau}$ is the submatrix of $\mathbf{Y}_{[k]}$ formed by the columns of $\mathbf{Y}_{[k]}$ indexed by τ, while $m_{\mathbb{G}_k}(\mathbf{Y}_{[k]\tau}|\mathbf{Y}_{[k]\mathrm{pa}(\tau)})$ is obtained by integrating out $\boldsymbol{B}_{[k,\tau]}$ and $\boldsymbol{\Omega}_{[k,\tau]}$ from the expression $\pi_{\mathbb{G}_{[k]}}(\boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]}) \prod_{i=1}^{n_k} f_{\mathbb{G}_{[k]}}(\mathbf{y}_{[k]i\tau}|\mathbf{y}_{[k]i\mathrm{pa}(\tau)}, \boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]})$. We also assign a prior distribution $\Pr(\mathbb{G}_{[1:K]})$ on the set of all multiple essential graphs, which can be updated with the evidence gauged by (2), using a version of Bayes' theorem, to the posterior distribution $\Pr(\mathbb{G}_{[1:K]}|\mathbf{Y}_{[1:K]}) \propto m(\mathbf{Y}_{[1:K]}|\mathbb{G}_{[1:K]}) \Pr(\mathbb{G}_{[1:K]})$. The latter can eventually be summarized through the *projected median probability graph model* described in [3]. Sect. 3 presents our priors and their computational implications.

3 Prior Choice and Posterior Computation

We deal with nuisance parameters using an objective Bayes approach. Specifically, we assign the parameter prior $\pi_{\mathbb{G}_{[k]}}(\boldsymbol{B}_{[k,\tau]}, \boldsymbol{\Omega}_{[k,\tau]})$ as suggested by Consonni et al. [6], which produces a closed-form expression for the term $m_{\mathbb{G}_k}(\mathbf{Y}_{[k]\tau}|\mathbf{Y}_{[k]\mathrm{pa}(\tau)})$ in (2) and thus prevents a significant computational issue; see also [4]. As for the multiple essential graph $\mathbb{G}_{[1:K]}$, we first specify a prior for the skeletons of $\mathbb{G}_1, \ldots, \mathbb{G}_K$, whose ensemble we represent by a collection of K-dimensional vectors $\mathbf{s}_{ij}$, $1 \le i < j \le q$, such that $s_{ijk} = 1$ if $i - j$ is contained in the skeleton of $\mathbb{G}_{[k]}$ and $s_{ijk} = 0$ otherwise. We then assume that all multiple essential graphs with given skeletons are equally probable. In this way, we can encourage similarity in the skeletons as suggested by Peterson et al. [15] for UGs and thus borrow strength across groups.

The prior in [15] is a hierarchical Markov random field specified as follows:

$$\pi(\mathbf{s}_{ij}|\nu_{ij}, \boldsymbol{\Theta}) \propto \exp(\nu_{ij} \mathbf{1}_K \mathbf{s}_{ij} + \mathbf{s}_{ij}^\top \boldsymbol{\Theta} \mathbf{s}_{ij}), \tag{3}$$

independently over $1 \le i < j \le q$, where ν_{ij} is a scalar sparsity parameter, $\boldsymbol{\Theta}$ is a $K \times K$ symmetric matrix, whose entries measure pairwise group associations, $\mathbf{1}_K$ is a K-dimensional vector of ones, and the normalizing constant can be analytically calculated, as a function of ν_{ij} and $\boldsymbol{\Theta}$, for reasonably small values of K. We denote by $\boldsymbol{\nu}$ the vector of all sparsity parameters and by $\boldsymbol{\theta}_{km}$ the generic entry of $\boldsymbol{\Theta}$.

The entries of $\boldsymbol{\Theta}$ follow a spike and slab prior, independently of the sparsity parameters, because $\boldsymbol{\theta}_{km} = 0$ identifies conditional independence between group k and group m, given the other groups, which is an interesting lack of association. The spike $\boldsymbol{\theta}_{km} = 0$ has probability w, independently over $1 \le k < m \le K$, while the slab distribution (a gamma distribution) has positive support to encourage similarity. The sparsity parameters are independent and $e^{\nu_{ij}}/(1 + e^{\nu_{ij}})$, which can be interpreted as a baseline inclusion probability for the line $i - j$, follows a beta distribution, whose parameters express prior information on sparsity.

We compute a Monte Carlo approximation of the posterior distribution on the set of all multiple essential graphs by marginalizing a Markov chain that targets the joint

posterior distribution of $\mathbb{G}_{[1:K]}$, $\boldsymbol{\Theta}$ and ν, using a Metropolis-Hastings algorithm. Following He et al. [11], when we propose a new essential graph, we make local changes to its structure through *perfect operators*, so as to guarantee that the chain has good theoretical properties. We update the essential graphs one at a time and the same we do with the entries of $\boldsymbol{\Theta}$, toggling between the spike and the slab, as well as with the sparsity parameters. The acceptance ratio for a new essential graph depends on the ratio between the number of DAGs that share its skeleton and the same quantity for the current skeleton; these quantities are hard to compute and we resorted to a simple heuristic approximation [4]. Sect. 4 highlights our results.

4 Results

We first validated our approach in the same simulation setting used by Castelletti et al. [3] for single essential graphs. Our simulations showed that the proposed method compares favorably to state-of-the-art methods for learning single essential graphs: we achieve comparable performance when the simulated networks are independent and a gain in reconstruction accuracy when the simulated networks are close to each other. Details of our simulations and their outcomes are reported in [4].

We then analyzed two datasets from multiple protein networks [12, 17]. We found networks consistent with the literature, but a few alternative regulatory mechanisms emerged, which warrant further investigation. Details of our analyses and their outcomes are reported in [4]. We here only comment on sparseness and overlap for the networks reconstructed from the data of Kornblau et al. [12], which consist of the levels of $q = 18$ proteins in $K = 4$ subtypes of leukemia patients. Table 1 reports the number of edges shared between pairs of networks, with the number of edges in each network on the diagonal. Out of $q(q-1) = 306$ possible edges (UG with all edges) a maximum of 15 is selected (for the subtype $M2$). The sparser networks show a significant amount of overlap with the denser ones, with a minimum of 2 edges in the network $M1$ that are not present in the network $M2$. These results show the value of formulating this learning problem in terms of multiple structures.

Table 1 Shared edge counts for the reconstructed leukemia protein networks

Subtype	$M0$	$M1$	$M2$	$M4$
$M0$	9	6	7	6
$M1$		9	6	5
$M2$			15	6
$M4$				12

Acknowledgements The work of Federico Castelletti, Guido Consonni and Stefano Peluso was partially supported by grants from UCSC (projects D1) and the EU COSTNET project (CA15109).

References

1. Andersson, S.A., Madigan, D., Perlman, M.D.: Alternative Markov properties for chain graphs. Scand. J. Stat **28**, 33–85 (2001)
2. Castelletti, F., Consonni, G.: Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. Ann. Appl. Stat. **13**, 2289–2311 (2019)
3. Castelletti, F., Consonni, G., Della Vedova, M., Peluso, S.: Learning Markov Equivalence Classes of Directed Acyclic Graphs: an Objective Bayes Approach. Bayesian Anal. **13**, 1231–1256 (2018)
4. Castelletti, F., La Rocca, L., Peluso, S., Stingo, F.C., Consonni, G.: Bayesian learning of multiple directed networks from observational data (2019). Submitted
5. Chin, L., Andersen, J.N., Futreal, P.A.: Cancer genomics: from discovery science to personalized medicine. Nat. Med. **17**, 297–303 (2011)
6. Consonni, G., La Rocca, L., Peluso, S.: Objective Bayes covariate-adjusted sparse graphical model selection. Scand. J. Stat. **44**, 741–764 (2017)
7. Danaher, P., Wang, P., Witten, D.M.: The joint graphical lasso for inverse covariance estimation across multiple classes. J. Roy. Stat. Soc. B **76**, 373–397 (2014)
8. Dobra, A., Hans, C., Jones, B., Nevins, J.R., West, M.: Sparse graphical models for exploring gene expression data. J. Multivariate Anal. **90**, 196–212 (2004)
9. Drton, M., Maathuis, M.H.: Structure learning in graphical modeling. Annu. Rev. Stat. Appl. **4**, 365–393 (2017)
10. Hauser, A., Bühlmann, P.: Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. J. Roy. Stat. Soc. B **77**, 291–318 (2015)
11. He, Y., Jia, J., Yu, B.: Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. Ann. Stat. **41**, 1742–1779 (2013)
12. Kornblau, S.M., Tibes, R., Qiu, Y.H., Chen, W., Kantarjian, H.M., Andreeff, M., Coombes, K.R., Mills, G.B.: Functional proteomic profiling of AML predicts response and survival. Blood **113**, 154–164 (2009)
13. Kristensen, V.N., Lingjærde, O.C., Russnes, H.G., Vollan, H.K.M., Frigessi, A., Børresen-Dale, A.L.: Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer **14**, 299–313 (2014)
14. Oates, C., Smith, J., Mukherjee, S., Cussens, J.: Exact estimation of multiple directed acyclic graphs. Stat. Comput. **26**, 797–811 (2016)
15. Peterson, C., Stingo, F.C., Vannucci, M.: Bayesian inference of multiple Gaussian graphical models. J. Am. Stat. Assoc. **110**, 159–174 (2015)
16. Roverato, A.: A unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs. Scand. J. Stat. **32**, 295–312 (2005)
17. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science **308**, 523–529 (2005)
18. Stingo, F.C., Chen, Y.A., Vannucci, M., Barrier, M., Mirkes, P.E.: A Bayesian graphical modeling approach to microRNA regulatory network inference. Ann. Appl. Stat. **4**, 2024–2048 (2010)
19. Tan, L.S.L., Jasra, A., De Iorio, M., Ebbels, T.M.D.: Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. Ann. Appl. Stat. **11**, 2222–2251 (2017)
20. Telesca, D., Müller, P., Parmigiani, G., Freedman, R.S.: Modeling dependent gene expression. Ann. Appl. Stat. **6**, 542–560 (2012)
21. Wang, Y.S., Drton, M.: High-dimensional causal discovery under non-Gaussianity. Biometrika **107**, 41–59 (2020)
22. Yajima, M., Telesca, D., Ji, Y., Müller, P.: Detecting differential patterns of interaction in molecular pathways. Biostatistics **16**, 240–251 (2014)
23. Zhao, S.D., Cai, T.T., Li, H.: Direct estimation of differential networks. Biometrika **101**, 253–268 (2014)

Bayesian post-processing of Gibbs sampling output for variable selection

Analisi bayesiana del Gibbs sampling per la selezione di variabili.

Stefano Cabras

Abstract We consider the problem of variable selection in general, but with specific focus on ordinary linear regression model where the number of covariates is larger and big enough to prevent exploration of all possible models. In this context, a stochastic model exploration with Gibbs-sampling is usually needed. With the aim of estimating the model inclusion probability (or including other interesting feature), the empirical estimator is consistent for a large number of steps with respect to the size of model space. However, we found that there is still space to improve such estimation in even a low number of steps with respect to the size of the model space. The focus is then here on post-processing Gibbs-sampling output using within a proper Bayesian analysis of such output. In particular a Dirichlet Process (DP) prior is used on the space of all possible models. The probability distribution on all model space is thus estimated combining the DP prior and the output of the Gibbs-sampling.

Abstract *Consideriamo il problema della selezione di variabili in generale e con particolare riferimento al modello lineare quando il numero delle covariate è sufficientemente grande da impedire l'esatto confronto tra tutti i possibili modelli. In questo contesto, lo spazio dei modelli è esplorato mediante il Gibbs-sampling. L'obiettivo è ottenere una stima della probabilità di inclusione a posteriori delle covariate. La stima empirica è consistente quando il numero di passi del Gibbs-sampling è sufficientemente elevato rispetto alla dimensione dello spazio dei modelli. Abbiamo osservato che è possibile migliorare le stime rispetto alla versione empirica. L'obiettivo del lavoro è analizzare l'output del Gibbs-sampling mediante procedure bayesiane non-parametriche utilizzando il processo di Dirichet come distribuzione a priori su tutto lo spazio dei modelli. La probabilità di inclusione è quindi stimata combinando tale a priori con l'output del Gibbs-sampling.*

Stefano Cabras
Department of Statistics, Universidad Carlos III de Madrid, C/Madrid, 126 - 09042 Getafe (Spain),
e-mail: stefano.cabras@uc3m.es

Key words: Conventional priors, Dirichlet Process, Ordinary linear regression, Variable selection

1 Introduction

Consider the usual problem in regression analysis consisting in finding a suitable set of predictors for a response variable y chosen among a fixed design matrix X with p columns. In a Bayesian setting we have to evaluated the posterior probability $\pi(\mathscr{M}_\gamma|y)$ of each regression model $\mathscr{M}_\gamma$, where $\gamma = (\gamma_1, \ldots, \gamma_p)$, with $\gamma_j = 1$ if the jth column of X_p is in $\mathscr{M}_\gamma$, namely $\pi(\mathscr{M}_\gamma|y)$ for all $\gamma \in \Gamma$ and Γ has cardinality 2^p.

Each $\pi(\mathscr{M}_\gamma|y)$ is obtained, exactly, by means of all Bayes Factors (BF) $BF_{\gamma 0}$ in which each possible model $\mathscr{M}_\gamma$ is compared against a common null nested model $\mathscr{M}_0$, usually the regression with only the intercept. To obtain $\pi(\mathscr{M}_\gamma|y)$ a prior on model space, $\pi(\mathscr{M}_\gamma)$ must be chosen. We consider the uniform $\pi(\mathscr{M}_\gamma) = 2^{-p}$, although other alternatives are possible. Different definitions of $BF_{\gamma 0}$ can be employed depending on the prior on model parameters. For instance, the conventional prior approach in [5] provides a closed form expression of $BF_{\gamma 0}$ with suitable properties. An important one is the predictive matching property which assures that $BF_{\gamma 0} = 1$ if there is not enough information in the data to distinguish between $\mathscr{M}_\gamma$ and $\mathscr{M}_0$. Closed form expression of the $BF_{\gamma 0}$ can be also achieved by using non-local priors on model parameters as detailed in [11] which again as other interesting properties. Beyond choices of model parameters prior and/or model prior, the point is that even under closed form expression of BFs an exhaustive exploration of Γ is not possible and thus a stochastic model search must be employed in order to obtain an estimation of $\pi(\mathscr{M}_\gamma|y)$. Along with such estimation, other interesting features can be derived from $\pi(\mathscr{M}_\gamma|y)$. The one of main interest here is the inclusion probability of a covariate j, τ defined on marginalizing $\pi(\mathscr{M}_\gamma|y)$ over all γ such that $\gamma_j = 1$. τ is necessary in order to define the median probability [4] model defined as the model such that $\tau > 0.5$ for all covariates in it. If this model exists, it is proved to be very near to the true model even under strong collinearity [3]. Under the scenario in which Γ cannot be fully explored, stochastic model search is usually employed in lieu of heuristic algorithms. In the case of normal linear regression model the stochastic approach is done by means of the well known the Gibbs-sampling algorithm discussed in [9]. Such an algorithm provides a dependent sequence of $\mathscr{M}^{(S)} = \left(\gamma^{(1)} \in \Gamma, \ldots, \gamma^{(S)} \in \Gamma\right)$ supposed to be a sample of size S from $\pi(\mathscr{M}_\gamma|y)$.

The aim of this paper is to use such a sample to obtain an estimation of τ, namely $\widehat{\tau}$. Different definition of $\widehat{\tau}$ derived from the Gibbs-sampling algorithm have been studied in [8] applying known concept of sampling theory. In particular, two estimators are compared:

i) the intuitive *empirical* proportion of the sampled models containing covariate j, $\widehat{\tau}_{\mathrm{e}} = S^{-1} \sum_{\gamma \in \mathscr{M}^{(S)}}^{S} \mathbf{1}(\gamma)_{\gamma_j=1}$, called the empirical estimator;

ii) the *renormalized* proportion of the sampled models containing covariate j, $\widehat{\tau}_{\mathrm{r}} = \frac{\sum_{\gamma \in \mathscr{M}^{(S)}} \mathbf{1}(\gamma)_{\gamma_j=1} \pi(\mathscr{M}_\gamma) BF_{\gamma^{(s)}0}}{\sum_{\gamma \in \mathscr{M}^{(S)}} \pi(\mathscr{M}_\gamma) BF_{\gamma^{(s)}0}}$, called the renormalized estimator.

It results that both are consistent estimators of τ for $S \to \infty$ and that the error of $\widehat{\tau}_{\mathrm{r}}$ is the sum of two components: the error of $\widehat{\tau}_{\mathrm{e}}$ plus (or minus) a term that depends on the correlation between posterior probability of a $\mathscr{M}_\gamma$ and its occurrence as a visited model. Basically, if the Gibbs-sampling moves very little around a model $\mathscr{M}_\gamma$ just because it has highest posterior probability (not necessarily a large one), $\widehat{\tau}_{\mathrm{r}}$ will be more biased than $\widehat{\tau}_{\mathrm{e}}$.

The idea of this paper is to rethink about all these estimators of τ and use a proper Bayesian approach to analyze the sequence $\mathscr{M}^{(S)}$. In practice, $\mathscr{M}^{(S)}$ can be viewed as a sequence of non ordinal categorical variables with two levels which gives raise to a *very sparse* contingency table of dimension 2^p whose cells probabilities are $\pi(\mathscr{M}_\gamma|y)$. This is the same setup as in [7]. Deriving the posterior distribution of cells probability is the same as deriving an estimation of $\pi(\mathscr{M}_\gamma|y)$ and thus the τ. That is, from $\pi(\tau|\mathscr{M}^{(S)})$ we define $\widehat{\tau}_{\mathrm{b}} = E_{\pi(\tau|\mathscr{M}^{(S)})}\tau$ and this is the estimator we will compare against $\widehat{\tau}_{\mathrm{e}}$ and $\widehat{\tau}_{\mathrm{r}}$ allowing also for properly account the uncertainty around the obtained value of $\widehat{\tau}$.

The following sections will illustrate the Bayesian model used to obtain $\widehat{\tau}_{\mathrm{b}}$ along with examples on simulated and real data. Some final remarks are left in the last section.

2 Post processing with the Dirichlet process prior

Using the notation in [7] let $\pi = \left\{\pi_{\gamma_1\gamma_2\cdots\gamma_p}, \gamma_j = 0,1, j = 1,\ldots,p\right\} \in \Pi$ be the set of all probabilities tensor of on Γ of size 2^p (all joint probabilities), where $\|\pi\|_1 = \sum_{\gamma_1=0}^{2} \cdots \sum_{\gamma_p=1}^{d_p} \left|\pi_{\gamma_1\gamma_2\cdots\gamma_p}\right| = 1$ and every $0 \le \pi_{\gamma_1\gamma_2\cdots\gamma_p} \le 1$. In this notation and without loss of generality we denote the two categories of $\gamma_j^{(s)}$, by $c_j = 0,1$. Probability π can be decomposed as an addictive mixture of k sets of probabilities,

$$\pi = \sum_{h=1}^{k} v_h \Psi_h, \quad \Psi_h = \psi_h^{(1)} \otimes \psi_h^{(2)} \otimes \cdots \otimes \psi_h^{(p)}$$

where $\mathbf{v} = (v_1,\ldots,v_k)'$ is the probability vector of the k sets, $\Psi_h \in \Pi_{1\ldots p}$, with $\psi_h^{(j)}$ being the probability for covariate j to be in the suitable set of predictors given a certain probability distribution over all Γ labeled by a specific value of h. The likelihood of $\mathbf{v}$ and ψ is

$$\Pr\left(\gamma_1^s = c_1,\ldots,\gamma_p^s = c_p|\mathbf{v},\psi\right) = \pi_{c_1\cdots c_p} = \sum_{h=1}^{k} v_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}.$$

Introducing the latent class indicator $z_s \in \{1,\dots,k\}$, the conditional probability to a specific set h is $\Pr\left(\gamma_j^s = c_j | z_s = h\right) = \psi_{hc_j}^{(j)}$ we have that ψs are just the inclusion probabilities of covariates for a given distribution on all model space. Thus the marginal distribution of $\tau = \psi^{(j)} = \sum_{h=1}^{k} \Pr\left(\gamma_j^s = 1 | z_s = h\right) \Pr(z_s = h)$, is the distribution of interest which leads to the definition of our estimator

$$\widehat{\tau}_{\mathrm{b}} = E_{\pi(\tau|\mathscr{M}^{(S)})}\tau = E(\psi^{(j)}|\mathscr{M}^{(S)}) = \sum_{h=1}^{k} \Pr\left(\psi_{hc_j}^{(j)}|\mathscr{M}^{(S)}\right) \Pr\left(v_h|\mathscr{M}^{(S)}\right).$$

The larger the k is the better is the representation of π, thus we allow for $k = \infty$ by using the following non-parametric prior:

$$\begin{aligned} \pi &= \sum_{h=1}^{\infty} v_h \Psi_h, \quad \Psi_h = \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)} \\ \psi_h^{(j)} &\sim P_{0j}, \quad \text{independently for } j = 1,\dots,p \text{ and } h = 1,\dots,\infty \\ \mathbf{v} &\sim Q, \end{aligned} \tag{1}$$

where P_{0j} corresponds to a Dirichlet measure and Q to a Dirichlet process.
The usual stick-breaking stochastic representation of the model is

$$\begin{aligned} \gamma_j^s &\sim \text{Multinomial}\left(\{0,1\}, \psi_{z_i}^{(j)}, \dots, \psi_{z_i d_j}^{(j)}\right) \\ z_i &\sim \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l)\, \delta_h, \quad V_h \sim \text{beta}(1,\alpha) \\ \psi_h^{(j)} &\sim \text{Dirichlet}\left(a_{j1},\dots,a_{jc_j}\right) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), \end{aligned}$$

where the Multinomial notation is actually over-engineered as observations are Bernoulli r.v.. Parameters $a_{j1} = \ldots = a_{jc_j} = 1$ induce non informative *a priori* information about the probabilities of each covariate of being in the model. α is the usual concentration parameter on the space of latent classes of model probability distributions. That is for small values of α the probability of having many classes of different probability distributions decreases. Hiperparameters are thus chosen such that the prior sample size is low, $a_\alpha + b_\alpha = 1/2$ and $a_\alpha = 1/4$.

Such a stochastic representation suggests the Gibbs sampling exposed in [7] to obtain simulations from the posterior distribution of ψs and vs. However, this algorithm can be very slow for large p and the benefits of $\widehat{\tau}_{\mathrm{b}}$ can be compensated by just calculating $\widehat{\tau}_{\mathrm{e}}$ or $\widehat{\tau}_{\mathrm{r}}$ on larger value of S. To avoid this we make use of a recent and faster variational algorithm proposed in [1] and also implemented in the R package `mixdir`. The algorithm relies on using approximated distribution, for the posterior of $\mathbf{v}$ and ψ, derived by applying the mean field theory to variational inference

(see [10]). These distributions lay down to be mixtures of Dirichlet distributions (for more details see [1]).

3 Examples and Simulations

In what follows we will consider the Riboflavin dataset (see [6]) related to the riboflavin production by Bacillus subtilis. We have 71 observations and $p = 4088$ predictors (gene expressions) and a one-dimensional response (riboflavin production), y. We assume the normal linear model where $BF_{\gamma 0}$ are obtained from the conventional prior approach (R package: `BayesVarSel`). $\pi(\mathscr{M}_\gamma)$ is the Uniform prior. The Gibbs algorithm starts at the null with $S = 1000$ steps after a burn in of 100 steps. In the first scenario we analyze the data set in its original version in order to compare $\widehat{\tau}_{\mathrm{e}}$, $\widehat{\tau}_{\mathrm{r}}$ and $\widehat{\tau}_{\mathrm{b}}$ from a biological perspective. In the second scenario we keep comparing by simulating many times y making depending it on only 2 out of the p predictors.

The following table 1 reports the median probability models with specific genes and their corresponding τ according to the three estimators.

Table 1 Estimated of τ larger than 0.5 (definition of the median probability model) along with estimated posterior distribution of τ (right hand side).

Gene	$\widehat{\tau}_{\mathrm{e}}$	$\widehat{\tau}_{\mathrm{r}}$	$\widehat{\tau}_{\mathrm{b}}$	Mentioned$^{(*)}$	v_h 0.8737	0.126	1e-04
YOAB_at	1.00	1.00	1.00	*	0.996	1.000	1.000
YXLD_at	0.44	0.69	0.55	*	0.486	1.000	1.000
YXLE_at	0.54	0.31	0.56	*	0.496	1.000	1.000
ARGH_at	0.17	0.59	0.29				
GYRA_at	0.04	0.54	0.17				
YFII_at	0.17	0.54	0.30				
YLXQ_at	0.08	0.71	0.20				

$^{(*)}$ Whether or not genes have been mentioned in [2].

The proposed approach provides smoothed results than $\widehat{\tau}_{\mathrm{e}}$ and apparently less false discoveries than $\widehat{\tau}_{\mathrm{r}}$. However, we can also provide an entirely probability distribution of τ as reported in the right-hand side of Table 1.

We simulated 100 data sets where $y = \sum_{i=1}^{2} i \times X_{p_i} + \varepsilon$, $\varepsilon \sim N(0,1)$ and p_1, p_2 are 2 columns of X_p picked at random in each data set. Results are showed in Figure 1

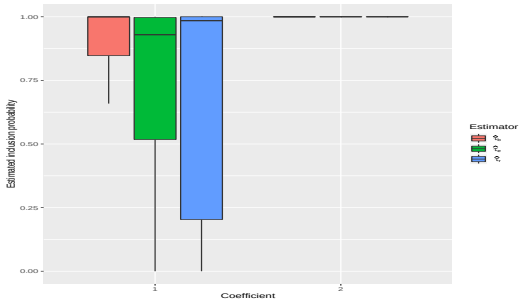


Fig. 1 Simulation results. Distributions of $\widehat{\tau}_{\mathrm{e}}$, $\widehat{\tau}_{\mathrm{r}}$ and $\widehat{\tau}_{\mathrm{b}}$ over the 100 datasets for covariates inside the model and ordered by the magnitude of their corresponding coefficients.

4 Remarks

We have presented a post processing approach of Gibbs-sampling output. However, the Bayesian NP model suggested here to derive an estimation of $\pi(\mathscr{M}_\gamma|y)$ could be further used by directly incorporating the NP prior into the Gibbs-sampling in order to match π in (1) with $\pi(\mathscr{M}_\gamma)$ as also done with $\widehat{\tau}_{\mathrm{r}}$ in [8]. GitHub code of this paper is reported at `https://github.com/scabras/varseldmmp`.

References

1. Ahlmann-Eltze, C., Yau, C.: Mixdir: Scalable bayesian clustering for high-dimensional categorical data. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 526–539. IEEE (2018)
2. Bar, H.Y., Booth, J.G., Wells, M.T.: A scalable empirical bayes approach to variable selection in generalized linear models. Journal of Computational and Graphical Statistics **0**(ja), 1–31 (2019). DOI 10.1080/10618600.2019.1706542
3. Barbieri, M., Berger, J.O., George, E.I., Rockova, V.: The median probability model and correlated variables (2018)
4. Barbieri, M.M., Berger, J.O.: Optimal predictive model selection. The Annals of Statistics **32**(3), 870–897 (2004)
5. Bayarri, M.J., Berger, J.O., Forte, A., García-Donato, G.: Criteria for Bayesian model choice with application to variable selection. The Annals of Statistics **40**, 1550–1577 (2012)
6. Bühlmann, P., Kalisch, M., Meier, L.: High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application **1**(1), 255–278 (2014)
7. Dunson, D.B., Xing, C.: Nonparametric bayes modeling of multivariate categorical data. Journal of the American Statistical Association **104**(487), 1042–1051 (2009)
8. Garcia-Donato, G., Martinez-Beneito, M.A.: On Sampling strategies in Bayesian variable selection problems with large model spaces. Journal of the American Statistical Association **108**(501), 340–352 (2013)
9. George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Statistica Sinica **7**, 339–373 (1997)
10. Jaakkola, T.S., Jordan, M.I.: Improving the Mean Field Approximation Via the Use of Mixture Distributions, pp. 163–173. Springer Netherlands, Dordrecht (1998). DOI 10.1007/978-94-011-5014-9-6
11. Nikooienejad, A., Johnson, V.E.: BVSNLP: Bayesian variable selection in high dimensional settings using nonlocal priors (2018)

Priors on precision parameters of IGRMF models

Distribuzioni a priori sui parametri di precisione nei modelli IGRMF

A. Gardini, F. Greco and C. Trivisano

Abstract When intrinsic Gaussian random Markov field (IGRMF) priors are assumed for random effects of a latent Gaussian model, a notable issue concerns prior elicitation for the precision hyperparameters. In fact, the structure of the precision matrix could lead to the undesired feature that the same prior for different precisions imply different marginal priors for the random effects. The work is aimed at investigating this problem following a rigorous mathematical procedure, in order to propose a new strategy and compare it to a widespread solution based on matrix scaling. Finally, an application of the proposed method to a real data problem is presented.

Abstract *Nell'ambito della stima di modelli Gaussiani latenti, vengono spesso impiegate distribuzioni a priori appartenenti alla famiglia dei campi Gaussiani Markoviani intrinseci. In questo contesto, l'elicitazione della distribuzione a priori sul parametro di scala rappresenta una criticità ben nota in letteratura. La struttura della matrice di precisione può portare a strutture di variabilità a priori differenti per gli effetti casuali a parità di prior sul parametro di precisione. Il lavoro è finalizzato allo studio di tale problema attraverso un rigoroso approccio matematico: si propone una nuova strategia per l'elicitazione della prior. In conclusione, viene presentato un caso di studio e un confronto con un approccio molto diffuso in letteratura.*

Key words: BYM model, INLA, Disease Mapping

Aldo Gardini
Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna, e-mail: aldo.gardini2@unibo.it

Fedele Greco and Carlo Trivisano
Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna

1 Introduction

Several popular statistical models could be expressed as *latent Gaussian models* [7]. In fact, they are able to express additive regression models incorporating flexible relations with covariates (e.g. penalised splines), structured random effects (such as spatial or temporal ones) and unstructured error terms.

A latent Gaussian model is based on the assumption that the response variable $\mathbf{y} \in R^n$ follows an exponential family distribution with the i-th ($i = 1, ..., n$) conditional expectation μ_i that is related to the linear predictor η_i through a link function $g(\mu_i) = \eta_i$. The most general structure of the predictor is:

$$\eta_i = \beta_0 + \sum_{l=1}^{p} \beta_l c_{li} + \sum_{j=1}^{q} f_j(z_{ji}) + \varepsilon_i, \tag{1}$$

where β_0 is the intercept, β_j are the fixed effect coefficients related to the covariate values $\mathbf{c}_l$, $f_j(\cdot)$ are smooth functions that allow to model the dependence between the response and a set of covariates $\mathbf{z}_j$. Finally, ε_i is an unstructured error term.

If the estimation of such a model is carried out into the Bayesian framework, a prior distribution for the parameters must be specified. In this paper we consider the case in which intrinsic Gaussian random Markov field (IGRMF) [6] is set as prior for the structured effects. Labelling with $\mathbf{x} \in R^n$ the IGRMF specified as prior for an effect $f_j(\cdot)$, then it is defined as a random vector distributed according to a zero-mean improper Gaussian density, i.e. having a precision $\tau_x \mathbf{K}$, where $\mathbf{K}$ is a sparse matrix not full rank and τ_x is a further scalar parameter that represents a precision. The rank of $\mathbf{K}$ is $n-k$, and k is the order of the IGRMF, whose density can be expressed as proportional to:

$$\pi(\mathbf{x}|\tau_x) \propto \exp\left\{-\frac{\tau_x}{2}\mathbf{x}^T\mathbf{K}\mathbf{x}\right\}. \tag{2}$$

The prior specification for $\mathbf{x}$ must be completed with the choice of a prior for the precision parameter τ_x. An issue that arises when an IGRMF prior is specified is that the structure of $\mathbf{K}$ influences the total variability expressed by the conditional prior $\pi(\mathbf{x}|\tau_x)$. As a direct consequence, the same prior distribution for the precision parameter $\pi(\tau_x)$ could lead to different marginal priors $\pi(\mathbf{x})$ for diverse $\mathbf{K}$ precision matrices. Moreover, the relation between the sparseness pattern of the matrix and the marginal prior variability is not straightforward. An empirical strategy used to overcome the problem is to scale the prior of τ_x with the geometric mean of the diagonal entries of the generalized inverse $\mathbf{K}^-$ [4, 8].

This work is aimed at proposing a new prior specification strategy for models involving IGRMF priors based on the solution of an integral equation that allows the user to have full control on the prior. From a technical point of view, the equation is solved by means of the Mellin transform and a brief overview is provided in section 2. The popular Besag, York and Mollié (BYM) model [1] has been considered in

section 3 to show how the proposed method works, whereas results of a real data application are shown in section 4.

2 Some mathematical tools

To develop the theory that is presented in the following sections, the distribution of a linear combination of r Chi-squared random variables with 1 degree of freedom and positive weights l_j $(j = 1, ..., r)$ is of interest. The density function of the variable $X = \sum_{j=1}^{r} l_j \chi_1^2$ has been derived by Ruben [5] as:

$$f_X(x) = \sum_{k=0}^{\infty} a_k \frac{x^{r/2+k-1} \exp\left\{-\frac{x}{2\beta}\right\}}{(2\beta)^{r/2+k} \Gamma(r/2+k)}, \tag{3}$$

where the coefficients a_k are defined recursively and are functions of the weights, and β is a parameter that rules the speed of convergence of the series.

Another relevant tool used in the work is the Mellin transform of a function $f(x)$ having positive support. It is an integral transform strictly related to the Fourier transform and is defined for $z \in \mathbf{C}$ as:

$$\hat{F}(z) = \int_0^{+\infty} f(x) x^{z-1} \mathrm{d}x. \tag{4}$$

It has been used in statistics and in probability for its appealing properties, particularly to deal with the product of random variables [2].

3 The BYM model

The BYM model is a widely used model in disease mapping applications. Moreover, it has been frequently taken as example for the development of prior specification strategies useful when both structured and unstructured random components are considered.

A study region that is partitioned into n areas is considered. For each area $i = 1, ..., n$ the observed count Y_i and the expected count E_i of cases for a given disease are available. The BYM model is defined as:

$$Y_i | \mu_i \sim \mathscr{P}(\mu_i E_i), \; i = 1, ..., n; \tag{5}$$

$$\log(\mu_i) = \eta_i = \mathbf{x}_i^T \beta + s_i + u_i, \tag{6}$$

where $\mathbf{x}_i$ is a vector of observed covariates, β are the corresponding coefficients, $\mathbf{s} = (s_1, ..., s_n)$ is a vector of spatially structured random effects and $\mathbf{u} = (u_1, ..., u_n)$ is the vector of unstructured random effects. The chosen priors for the random effects

are:

$$\mathbf{s}|\tau_s \sim \mathscr{N}_n\left(\mathbf{0}, \tau_s^{-1}\mathbf{Q}^-\right), \tag{7}$$

$$\mathbf{u}|\tau_u \sim \mathscr{N}_n\left(\mathbf{0}, \tau_u^{-1}\mathbf{I}_n\right), \tag{8}$$

where $\mathbf{Q}^-$ is the generalized inverse of the structure matrix $\mathbf{Q} = \mathbf{D} - \mathbf{W}$. The matrix $\mathbf{D}$ is diagonal with generic entry equal to n_i, that is the number of neighbourhoods for area i, whereas the elements of the adjacency matrix $\mathbf{W}$ are defined as $w_{ij} = 1$ if i and j are neighbour areas and $w_{ij} = 0$ otherwise.

The prior (7) for vector $\mathbf{s}$ is a first order IGRMF on an irregular grid and the precision matrix $\mathbf{Q}$ has rank $n-1$.

To complete the model specification, the hyperpriors $\pi_{\tau_s}(t_s)$ and $\pi_{\tau_u}(t_u)$ are required. For the precision of the structured effect, it is common to assume a gamma prior $\tau_u \sim \mathscr{G}(a,b)$, whereas the prior on τ_s is the subject of the following discussion.

Sørbye and Rue [8] underline that the precision matrix $\mathbf{Q}$ of the IGRMF prior should be accounted in specifying the prior for the precision hyperparameter τ_s. They suggest to assign the prior on $\tau_s/\sigma^2_{\text{ref}}(\mathbf{s})$ instead of τ_s, assuming that the proposed distribution can be directly compared to the one on τ_u. The quantity $\sigma^2_{\text{ref}}(\mathbf{s})$ is a reference standard deviation for $\mathbf{u}$, computed for the IGRMF prior assuming $\tau_s = 1$ and is the geometric mean of the diagonal entries of $\mathbf{Q}^-$.

3.1 Our proposal: balanced priors

The proposed prior specification strategy is based on the prior independence of the random effects and it allows to control the total prior variance set for the linear predictor η and its partition between the structured and unstructured effect:

$$\frac{\eta^T\eta}{n} = \frac{\mathbf{u}^T\mathbf{u}}{n} + \frac{\mathbf{s}^T\mathbf{M}\mathbf{s}}{n} = V_u + V_s; \tag{9}$$

where $\mathbf{M}$ is a centering matrix that accounts for the rank deficiency of $\mathbf{Q}$. Recalling the prior of the random effects (7) and (8), it can be shown that the distribution of the vectors variances conditioned with respect to the respective scalar precisions are:

$$V_u|\tau_u \sim \mathscr{G}\left(\frac{n}{2}, \frac{n\tau_u}{2}\right); \;\; V_s|\tau_s \sim \frac{1}{\tau_s}\sum_{i=1}^{n-1}\frac{\lambda_i}{n}\chi^2_1, \tag{10}$$

where λ_i are the reciprocals of the non-zero eigenvalues of $\mathbf{Q}$.

Under the proposed $\mathscr{G}(a,b)$ prior for τ_u, the marginal prior on V_u is a generalized beta distribution of the 2^{nd} type (GB2) that has parameters $V_u \sim GB2(1,n/2,2b/n,a)$.

On the other hand, due to the structure of $\mathbf{Q}$, if a gamma prior on the precision is assumed, then the marginal prior of V_s is not GB2. In this work we assume that the user wish to achieve the complete prior balance by specifying the same marginal prior for V_u and V_s: the consequent underlying prior for τ_s must be computed. To

reach this target, the proposed strategy is to solve a Fredholm integral equation of the first kind. By exploiting the particular structure of the involved function, it is possible to find the desired prior density by inverting a Mellin transform [3].

4 An application

To show an application of the proposed prior specification procedure and compare it to the widely used INLA's scaled priors, the counts of deaths registered in the $n = 341$ municipalities of the Emilia-Romagna Region due to a particular cause are analysed. The total amount of cases is 2095. The BYM model under the compared prior strategies was fitted in R using INLA [7].

For the unstructured precision τ_u a gamma prior with $a = 1$ and $b = 1/2$ is chosen. As a consequence, the scaled prior on τ_s is a gamma with $a = 1$ and $b = \sigma^2_{\text{ref}}(\mathbf{s})/2$, with $\sigma^2_{\text{ref}}(\mathbf{s}) = 0.79$ for the map of the Emilia-Romagna Region. On the other hand, our prior is obtained numerically by solving the integral equation aimed at obtaining the prior that mixed with the IGRMF prior implies the same marginal GB2 prior for V_s and V_u. The deduced priors are compared in figure 1, and it can be noted that the distance from our exact balanced prior is larger for the scaled prior with respect to the $\mathscr{G}(1, 0.5)$ prior. In particular, posterior balancing between structured and unstructured random effects is affected by prior specification. Some interesting quantities fitted through the BYM model under the two prior strategies are compared in figure 2.

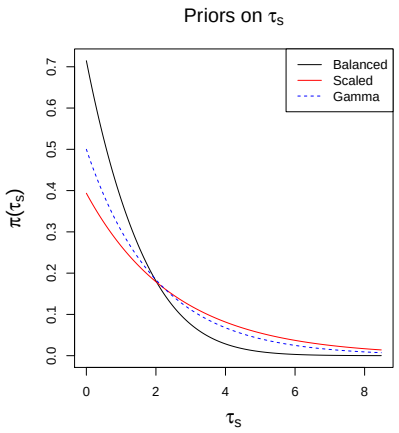


Fig. 1 Hyperpriors for the precision parameter τ_s: the black line reports the proposed balanced prior obtained by solving the Fredholm integral equation; the red lined is the prior used by INLA when the scalded matrix is considered. Finally, the dashed blue line is the gamma prior with $a = 1$ and $b = 0.5$.

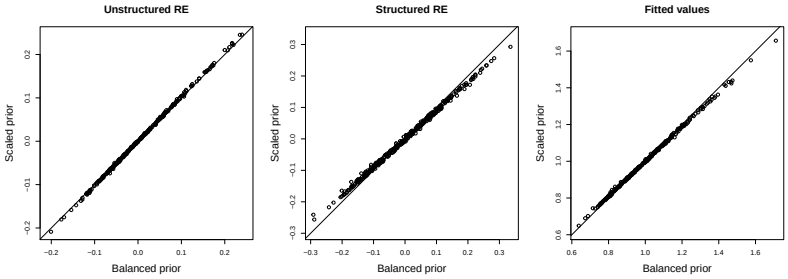


Fig. 2 The BYM model estimates obtained using INLA's scaled prior and our balanced priors approach are compared. From left to right the following quantities are shown: unstructured random effects (**u**), structured random effects (**s**) and fitted relative risks.

References

[1] Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics **43**(1), 1–20 (1991)

[2] Epstein, B.: Some applications of the Mellin transform in statistics. The Annals of Mathematical Statistics **19**(3), 370–379 (1948)

[3] Polyanin, A.D., Manzhirov, A.V.: Handbook of integral equations. CRC press (1998)

[4] Riebler, A., Sørbye, S.H., Simpson, D., Rue, H.: An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Statistical Methods in Medical Research **25**(4), 1145–1165 (2016)

[5] Ruben, H.: Probability content of regions under spherical normal distributions, IV: The distribution of homogeneous and non-homogeneous quadratic functions of normal variables. The Annals of Mathematical Statistics **33**(2), 542–570 (1962)

[6] Rue, H., Held, L.: Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC (2005)

[7] Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **71**(2), 319–392 (2009)

[8] Sørbye, S.H., Rue, H.: Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spatial Statistics **8**, 39–51 (2014)

Sequence Analysis: methods and applications

Internal migration, family formation and social stratification in Europe. A life course approach.

Migrazioni interne, formazione familiare e stratificazione sociale in Europa attraverso un'analisi dei corsi di vita.

Roberto Impicciatore, Gabriele Ballarino and Nazareno Panichella

Abstract Geographical mobility trajectories generally intersects other life course patterns such as student career, job experiences and family choices. Life course approach emphasizes that different events are not separate experiences, but are linked to each other. Focusing on 11 European countries (Austria, Czech Republic, Denmark, France, Germany, Greece, Italy, Poland, Spain, Sweden, and Switzerland), this analysis has three main aims: describing the life patterns experienced by internal migrants, analyzing the selection into different migration trajectories, and identifying the association with different occupational achievements and social mobility pathways. To achieve these goals, we applied Sequence and Cluster Analysis to SHARELIFE data (2008-09 and 2017). Preliminary results reveal that different migration patterns are characterized by a marked selectivity of movers (mainly based on education), particularly in contexts where migration works well as an escalator strategy. In addition, specific life trajectories are associated to better occupational returns.

Abstract *Le traiettorie di mobilità geografica generalmente intersecano altri percorsi di corsi di vita come l'istruzione, la carriera lavorativa e le scelte familiari. L'approccio del corso di vita pone l'attenzione sul fatto che eventi diversi non sono esperienze separate ma collegate tra loro. Concentrandosi su 11 paesi europei (Austria, Repubblica Ceca, Danimarca, Francia, Germania, Grecia, Italia, Polonia, Spagna, Svezia e Svizzera), questa analisi ha tre obiettivi principali: descrivere i modelli di vita sperimentati dai migranti interni, analizzare il processo di selezione sulle diverse traiettorie di migrazioni e identificare l'associazione con specifici*

[1] Roberto Impicciatore, University of Bologna; email: roberto.impicciatore@unibo.it
Gabriele Ballarino, University of Milan; email: gabriele.ballarino@unimi.it
Nazareno Panichella, University of Milan; email: nazareno.panichella@unimi.it

risultati lavorativi e con le tratiettorie di mobilità sociale. Per raggiungere questi obiettivi, abbiamo applicato una Sequence and Cluster Analysis ai dati SHARELIFE (2008-09 e 2017). I risultati preliminari rivelano che i diversi percorsi migratori risultano caratterizzati da una marcata selettività (basata principalmente sull'istruzione), partciolarmentenei contesti dove la migrazione funziona bene come strategia di ascesa sociale. Inoltre, alcune specifiche traiettorie di vita risultnao associate a migliori opportunità in campo lavorativo.

Key words: internal migration, social stratification, life course, sequence analysis

1 Introduction

Spatial mobility and family are intertwined from early through later life. Life course theory emphasizes that life events are not separate experiences, but are linked to one another (Giele & Elder, 1998). On this respect, the analysis of geographical mobility is crucial because it intersects with other life course patterns such as student career, job experiences and family choices. The positive impact on occupation given by internal migration has already underlined in the literature mainly suggesting that chances of obtaining higher socio-economic status later in life than those who did not move (van Ham, 2003), in particular when migration is directed to areas with higher educational and labour-market opportunities such as large cities (Fielding, 1992). However, the timing of migration is also relevant since the advantages tend to be higher if the migration takes place in the early stages of the labour career. In other words, it is likely that the migration, as an investment in human capital, will pay off in the long run (Mulder and van Ham, 2005). Furthermore, the decision to move is related to other life events such as family events (including family reunification), childbearing and change in family size (Kulu & Milewski, 2007). This creates substantial heterogeneity in migration experiences and trajectories, which are based on a huge interrelation of biographical events, and which may have different effects on the integration of migrants.

Our approach is based on the idea that mobility is a crucial step in the life pattern as embedded into the process of the transition to adulthood and thus intrinsically linked with the achievement of residential autonomy, economic independence and family formation (Impicciatore and Panichella 2019). We mainly focus on the following three research questions: a) which are the life patterns experienced by internal migrant in Europe? b) Does the selection into migration change when it occurs at different phases of life-course events? c) Do different types of life trajectories are associated with specific outcomes in the labour market?

However, there is a substantial lack of literature of comparative analyses on the effect of life trajectories on occupational outcomes. We aim at contributing to the existing literature by analyzing the relations between geographical mobility, education, occupational career and family behavior focusing on the internal migration in 11 European countries. The comparative perspective used, ensured by using standardized data, allows to evaluate whether the life patterns of internal tend

to be similar in different European countries or, alternatively, there are peculiarities at country-level.

2 Data and method

The analysis is based on SHARELIFE, i.e. third and the seventh wave of SHARE (Survey on Health, Ageing and Retirement in Europe) held, respectively, on 2008-09 and 2017. Selected data provide life-history information about a representative sample of about 45,000 respondents aged 50 and over living in Europe (21,800 for Wave 3 and 23,500 for wave 7). The selected subsample is composed by individuals born between 1920 and 1960 living at the interview in the following countries: Austria, Czech Republic, Denmark, France, Germany, Greece, Italy, Poland, Spain, Sweden, and Switzerland. Respondents are asked to report all the changes in accommodation (at regional level for internal moves) they had throughout their lives. We focus on interregional moves, namely among NUTS-2 level regions. The domains of interests also include family relationships history, housing, educational career and working history.

We applied a method called sequence analysis for social sciences (Abbott, 1983, 1995; Abbott and Forrest, 1986). The basic idea is to represent each life course trajectory as a string of characters (one for a specific time unit), resembling the one used to code DNA molecules in biological sciences. After having defined the set of sequences, a dissimilarity measure between each pair of sequences can be computed. Among different possible methods, we used the Longest Common Subsequence metric (LCS) proposed by Elzinga (2006) and based on the length of common distinct sub-sequences between life course trajectories. Compared to the commonly used Optimal Matching strategy (OM), the LCS metric does not require the definition of costs. The following step is to group similar sequences (according to their dissimilarities) through a hierarchical cluster analysis. In our case, we used the Ward linkage and we fixed the number of cluster by observing the hierarchical tree diagram also known as dendrogram.

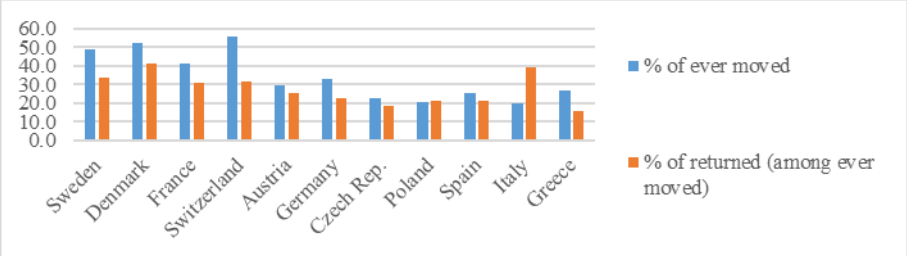
Our window of observation is the age range between 15 and 40 years. For each of this age, an individual can assume one specific state. The set of all possible states (state space) is given by the combination of the following characteristics: in education (yes/no); working (yes/no), in couple (yes/no), with children (yes/no), moved to another region (no/yes/returned), resulting in potentially states. This dimension requires a simplification of the complexity. One efficient strategy is the so-called multichannel sequence analysis approach (Gauthier, 2010; Pollock, 2007), which makes it possible to specify multiple domains in order to construct a single matrix of dissimilarities by locally aligning distinct life trajectories simultaneously. In our analysis, we distinguished between transitions in the study-work domain and transitions in the family domain. The sequence analysis was performed using the TraMineR package, which is available in the open source statistical environment R (Gabadinho et al., 2011).

Once the cluster has been identified, they can be used an additional variable in multivariate regression. In particular, we focused on the social selectivity of internal migrants, estimating the probability of being a mover using a binomial logistic regression model and the probability of being included in each cluster of mobility (conditional on the probability of being a mover) by means of multinomial logit. Finally, we develop binomial logistic models by considering the probability of having a specific labour market position (distinguishing among service, middle and lower class) at 50 years of age according to the previous life pattern.

3 Preliminary findings

In our subsample, one every three individuals experienced at least an interregional move between 15 and 40 years of age. Across countries (Fig. 1), it emerges a quite high heterogeneity with higher values in Switzerland and the Scandinavian area (around 50%) and a lower propensity to move in the Eastern and Mediterranean area (around 23%). Even though this rough comparison is somewhat dependent to the number and the dimensions of regions within each countries, the differences among European countries in the propensity to move internally tend to confirm what have been underlined in the existing literature (Bell et al. 2018). The propensity to move back to their region of origin (among ever moved) is higher in the countries with a higher mobility even though differences across countries are less marked. The most evident exception is Italy where the (few) migrants often returned to their region of origin.

Figure 1: Percentage of individual who experienced at least one inter-regional move between 15 and 40 years of age and percentage of those returned to their region of origin before 40 years of age (among ever moved). Individuals born between 1920 and 1960.



Source: Own elaboration on SHARE wave 3 (2008-09) and wave 7 (2017).

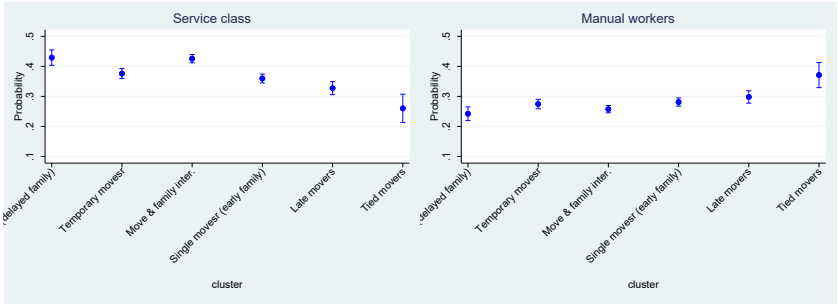
Through the cluster analysis applied to sequence of states, we found six clusters that can be briefly described as follows: *Single movers* (cluster 1: individual moved as single with union formation occurred relatively late and often childless; 9% of cases); *Temporary movers* (cluster 2: migrants returned to their region of origin and family formation after the return; 19% of cases); *Move & family interrelation* (cluster 3: delayed move and a strong interrelation between migration, union and childbirth; 29% of cases); *Single movers (early family formation)* (cluster 4: early

mover as a single and then a rapid family formation with children; 25% of cases): *Late movers* (cluster 5: rapid family transition and a slow migration not connected to other, i.e. movers as a family unit; 11% of cases); *Tied movers* (cluster 6: late movers as inactive (in few cases returned), family formation usually after the move but with no children; 7% of cases).

Our preliminary suggests that the overall different migration patterns are characterized by a marked selectivity of movers and the 'positive' selection of internal migrants found by other research may hide important differences based on the type of life course pattern.

Figure 2 considers inclusion in the occupational structure at 50 years of age. *Tied movers* (cluster 6) and, to a lesser extent, *late movers* (cluster 5) show a clear disadvantage in the labor market given the lower probability of being in the service class and, at the same time, a higher probability of being a manual worker. Conversely, two patterns emerge as the most advantaged, namely cluster 1 (moved as a single and delayed family formation) and cluster 3 (interrelated events).

Figure 2: Probability of entering the service class and the lower class (manual workers) among workers at 50 years of age. Logit models, predictive margins and 95% confidence intervals.



Source: Own elaboration on SHARE wave 3 (2008-09) and wave 7 (2017).
Note: control variables in the model: country; gender; birth cohort; level of education; social class of origin according to the job position of the main breadwinner when 10 years of age; international migrant (born abroad or arrived before 15 years of age); area of origin.

These first results suggest that mobility can have good occupation returns following two very distinct patterns. Firstly, when migration is experienced early in the life course but independently by family formation. This pattern is mobility experienced mainly as a human capital investment both in education and in order to better exploit the potentialities of the labor market in the area of arrival. Given that it includes costly geographical movements, this cluster identifies a mechanism of social closure, since it allows the reproduction of the upper class from the region of origin to the region of destination. Secondly, good outcomes can be obtained through a general postponement. Mobility and family formation can be experienced later in the life course after having obtained a good position in the labor market in the context of departure.

In conclusion, provisional results suggest that the adoption of a family viewpoint on internal migration represents a powerful empirical strategy for contributing to the existing literature. However, further analysis should be added in order to shed light on the heterogeneity across European countries. For example, in some countries, such as those in the Northern Europe, the internal migration may works well as an escalator strategy. Additional analysis will also consider the inter-generational social mobility and emphasize the gender roles in migration.

References

1. Abbott, A. (1995). Sequence Analysis: New Methods for Old Ideas. Annual Review of Sociology, 21(1), 93–113.
2. Abbott, A. (1983). ‘‘Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes.’’ Historical Methods 16:129-47.
3. Abbott, A. and J. Forrest (1986). ‘‘Optimal Matching Methods for Historical Sequences.’’ Journal of Interdisciplinary History 16:471-94.
4. Bell, M., Charles-Edwards, & E., Bernard, A. (2018). Global Trends in Internal Migration. In I. S. Tony Champion, Thomas Cooke (Ed.), Internal Migration in the Developed World Are we becoming less mobile?, 1st Edition. Routledge.
5. Blau, P. M., & Duncan, O. D. (1967). The American occupational structure. New York: Wiley.
6. Elzinga, C.H. (2006). Sequence analysis: Metric representation of categorical time series. Technical report. Department of Social Science Research Methods, Vrije Universiteit Amsterdam, The Netherlands.
7. Fielding, A. (1992). Migration and social mobility: South East England as an escalator region. Regional Studies, 26(1), 1–15.
8. Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. Journal of Statistical Software, 40(4), 1–37.
9. Gauthier, J.-A. (2010). Multichannel sequence analysis applied to social science data. Sociological Methodology, 40(1), 1-38.
10. Giele, J. Z., & Elder, G. H., Jr. (Eds.). (1998). Methods of life course research: Qualitative and quantitative approaches. Sage Publications, Inc.
11. Impicciatore, R., Panichella, N., (2019). Internal Migration Trajectories, Occupational Achievement and Social Mobility in Contemporary Italy. A Life Course Perspective. Population, Space and Place, 25(6).
12. Kulu, H., & Milewski, N. (2007). Family change and migration in the life course: An introduction. Demographic Research, 17(19), 567–590.
13. Mulder, C. H., & van Ham, M. (2005). Migration histories and occupational achievement. Population, Space and Place, 11(3), 173–186.
14. Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(1), 167-183.
15. van Ham, M. (2003). Job access at labour market entry and occupational achievement in the life course. International Journal of Population Geography, 9(5), 387–398.

Socio economic integration of migrants

A study on the characteristics of spouses who intermarry in Italy

Uno studio delle caratteristiche dei matrimoni misti in Italia

Agnese Vitali and Romina Fraboni

Abstract This contribution compares the characteristics of spouses in endogamous vs. exogamous marriages in Italy using marriage registers for 2017. We model the probability of observing an endogamous vs. exogamous marriage between a native and foreign spouse coming from different migrant origins. The marriage patterns that we find are consistent with a (mismatched) marriage market where men and women's expectations are changing at different rhythms.

Abstract *Questo contributo confronta le caratteristiche degli sposi nei matrimoni endogami e in quelli esogami osservati in Italia nel 2017. Studiamo la probabilità di osservare un matrimonio endogamo rispetto ad uno esogamo tra uno sposo/a italiano/a e una sposa/o straniera/o con diverse origini migratorie. Emergono dei modelli matrimoniali coerenti con un mercato matrimoniale squilibrato in cui le aspettative di uomini e donne cambiano in modo differente.*

Key words: Intermarriage, Marriage, homogamy, assortative mating, marriage market

1 Background and Hypothesis

The abundance of high-educated women and the scarcity of high-educated men may create mismatches in the marriage market (Grow and Van Bavel, 2015). In the Italian context, where the male breadwinner family type is persistent, the emergence of a reversal of the gender gap in education contrasts sharply with a still prevailing

[1] Agnese Vitali, Dipartimento di Sociologia e Ricerca Sociale, Università di Trento; email: agnese.vitali@unitn.it

Romina Fraboni, Istituto Nazionale di Statistica; email: fraboni@istat.it

gendered ideology permeating family life. Contrary to what found for other settings, in Southern Europe high-educated women have the lowest likelihood of being in a union (Kalmijn, 2013) and, together with low-educated men, are most exposed to the risk of lifelong singlehood (Bellani et al., 2017).

The growing migrant population may improve the marriageability of those groups whose marriage opportunities are being shrunk, i.e. high-educated women and low-educated men (Van Bavel, 2012). Low-educated men and high-educated women, we argue, will be more likely to intermarry (vs. marrying a native spouse) if compared to better-educated men and lower-educated women, respectively. Indeed, if the preference for homogamy cannot be satisfied in the national marriage market (e.g. there are too few low-educated women and too few high-educated men), native spouses may be willing to marry out of their group in exchange of a certain attribute viewed as a 'compensating differential' (Grossbard-Shechtman and Fu, 2002). We distinguish between different migrant groups so to account for differences in their cultural and economic characteristics. Following Medrano et al. (2014), we consider spouses coming from Western Europe, North America and Oceania as having a prevailing gender-egalitarian ideology, while we consider spouses coming from Eastern Europe, Africa, Asia and Latin America as having a prevailing traditional gender ideology.

We expect low-educated native men to be willing to marry foreign wives from traditional contexts, in exchange of their traditional gender ideology, while we expect high-educated native women to be willing to marry foreign husbands from gender-egalitarian contexts.

2 Data and Methods

We study the population of Italian men and women who married for the first time in 2017, using marriage register data (Rilevazione dei Matrimoni, ISTAT).

We are interested in comparing the characteristics of spouses in exogamous marriages with those in endogamous marriages. To do so, using a multinomial logistic regression model, we model the probability that a native man or woman with

certain socio-demographic characteristics will marry a native vs. a foreign spouse from a given country of origin. Our variable of interest is constituted of seven categories: a native woman (man) marrying a native spouse (ref.), or intermarrying with a spouse from North America and Oceania, Western Europe, Eastern Europe, Africa, Asia, or Latin America. As independent variables we consider: age of the native spouse (linear and quadratic term); age difference between native and foreign spouse (same age –ref.–, native spouse is older, and younger); education of the native spouse (low, medium, high), educational difference between native and foreign spouse (same education –ref.–, native spouse is more educated and native spouse is less educated), employment status of the couple (both employed, jobless,

only man employed, only woman employed) and order of marriage of the foreign spouse (first –ref.–, second or higher order). We run separate models by gender of the native spouse.

3 Results

In this section we identify the profiles of native spouses who intermarry. Such profiles differ between men and women and vary with the origin of the foreign spouse.

About one third of Italian women intermarry with Western-European husbands, whereas about half of Italian men intermarry with Eastern-European wives (Table 1). Results of the multinomial regression show that native men and women who marry spouses coming from areas characterized by a gender-egalitarian ideology tend to be highly educated, whereas those who marry spouses coming from areas characterized by a traditional gender-egalitarian ideology tend to be low educated (Tables 2 and 3). The educational differential among spouses also matters: whereby marriages among natives are associated with homogamy, intermarriage is associated with hypergamy when native men marry wives from traditional areas and with hypogamy when they marry wives from gender-egalitarian areas. Men who intermarry tend to be older than men who marry a native wife. Native women who intermarry with husbands from gender-egalitarian areas are more likely to intermarry with a spouse who has achieved at least the same educational level as they did, and tend to be older than women who married a native husband. As for low-educated men, high-educated women might have spent some time searching for a similarly high-educated spouse before turning to foreign spouses. Instead, native women who intermarry with husbands from traditional areas tend to be less educated and younger than women who marry a native husband. Here, the gradient in the educational gap between spouses is less clear, except for a departure from homogamy.

Table 1: Intermarriages by gender of the native spouse and origin of the foreign spouse. First marriages of Italian husbands, 2017

Origin of foreign spouse	***Italian Husband***	***Italian Wife***
Traditional:		
Latin America	15.47	9.90
Eastern Europe	47.3	28.7
Africa	8.6	20.0
Asia	6.6	6.4
Gender egalitarian:		
Western Europe	19.1	30.7
North America and Oceania	2.9	4.3
Total N.	19,865	12,574

Table 2: Relative Risk Ratios from multinomial logistic regression model on the probability of marrying a wife of a given origin for Italian native men who married for the first time in 2017

	Native Italian husband at first marriage					
	RRR		Std. Err.	RRR		Std. Err.
	Western Europe			*North America, Oceania*		
Age of husband	1.111	***	0.024	0.947		0.039
Age of husband^2	0.999	**	0.000	1.001		0.000
Order of marriage of wife (First ref.)						
Second or higher order	0.683	**	0.097	1.659		0.434
Husband education (Primary ref.)						
Tertiary edu.	2.446	***	0.239	2.962	***	0.598
Secondary edu.	1.064		0.094	1.461	**	0.259
Age difference (Same Age ref.)						
Husband older	0.899		0.075	1.129		0.186
Wife older	1.485	***	0.157	1.369		0.289
Educational Homogamy (Equally ref.)						
Husband more educated	0.607	***	0.081	0.281	**	0.103
Wife more educated	1.344	**	0.123	1.415		0.251
	Eastern Europe			*Africa*		
Age of husband	1.066	***	0.008	1.008		0.019
Age of husband^2	1.000	***	0.000	1.000		0.000
Order of marriage of wife (First ref.)						
Second or higher order	2.933	***	0.110	2.005	***	0.223
Husband education (Primary ref.)						
Tertiary edu.	0.807	***	0.032	0.428	***	0.049
Secondary edu.	0.849	***	0.025	0.530	***	0.044
Age difference (Same Age ref.)						
Husband older	1.970	***	0.072	2.372	***	0.260
Wife older	1.114	**	0.058	1.676	***	0.239
Educational Homogamy (Equally ref.)						
Husband more educated	1.335	***	0.060	2.329	***	0.263
Wife more educated	1.239	***	0.039	0.784	**	0.075
	Asia			*Latina America*		
Age of husband	1.060	**	0.022	1.036	**	0.012
Age of husband^2	1.000		0.000	1.000		0.000
Order of marriage of wife (First ref.)						
Second or higher order	1.964	***	0.209	1.840	***	0.119
Husband education (Primary ref.)						
Tertiary edu.	1.502	***	0.150	0.766	***	0.049
Secondary edu.	1.142		0.095	0.880	**	0.042
Age difference (Same Age ref.)						
Husband older	1.679	***	0.158	1.646	***	0.095
Wife older	1.912	***	0.221	1.894	***	0.135
Educational Homogamy (Equally ref.)						
Husband more educated	1.060		0.120	1.501	***	0.102
Wife more educated	1.119		0.099	1.022		0.054

*p < .05; **p < .01; ***p < .001.

Table 3: Relative Risk Ratios from multinomial logistic regression model on the probability of marrying a husband of a given origin for Italian native women who married for the first time in 2017

	Native Italian wife at first marriage					
	RRR		Std. Err.	RRR		Std. Err.
	Western Europe			*North America, Oceania*		
Age of wife	1.100	**	0.032	0.835	***	0.039
Age of wife^2	0.999	**	0.000	1.002	**	0.001
Order of marriage of housband (First ref.)						
Second or higher order	1.057		0.150	3.124	***	0.859
Wife education (Primary ref.)						
Tertiary edu.	2.778	***	0.288	2.841	***	0.678
Secondary edu.	1.128		0.115	1.555	**	0.349
Age difference (Same Age ref.)						
Husband older	0.918		0.076	0.591	**	0.105
Wife older	1.773	***	0.182	1.910	**	0.425
Educational Homogamy (Equally ref.)						
Husband more educated	1.336	**	0.173	1.221		0.357
Wife more educated	0.436	***	0.045	0.543	**	0.115
	Eastern Europe			*Africa*		
Age of wife	0.695	***	0.012	0.736	***	0.010
Age of wife^2	1.004	***	0.000	1.003	***	0.000
Order of marriage of housband (First ref.)						
Second or higher order	1.571	**	0.254	1.705	***	0.201
Wife education (Primary ref.)						
Tertiary edu.	0.786	**	0.075	0.484	***	0.041
Secondary edu.	0.757	***	0.058	0.603	***	0.038
Age difference (Same Age ref.)						
Husband older	0.697	***	0.053	1.104		0.081
Wife older	3.759	***	0.342	6.017	***	0.495
Educational Homogamy (Equally ref.)						
Husband more educated	1.339	**	0.138	1.281	**	0.108
Wife more educated	1.385	***	0.112	1.395	***	0.102
	Asia			*Latina America*		
Age of wife	0.736	***	0.021	0.764	***	0.016
Age of wife^2	1.003	***	0.000	1.003	***	0.000
Order of marriage of housband (First ref.)						
Second or higher order	1.643		0.431	2.264	***	0.370
Wife education (Primary ref.)						
Tertiary edu.	1.070		0.182	1.249		0.149
Secondary edu.	0.896		0.128	0.899		0.092
Age difference (Same Age ref.)						
Husband older	0.807		0.117	0.866		0.090
Wife older	4.295	***	0.707	4.375	***	0.515
Educational Homogamy (Equally ref.)						
Husband more educated	1.366		0.258	1.604	***	0.207
Wife more educated	0.988		0.151	0.922		0.100

*p < .05; **p < .01; ***p < .001.

Our initial hypotheses are therefore confirmed: for men, the probability of marrying a wife from traditional areas vs. a native wife increases as his educational attainment decreases; for women, the probability of marrying a husband from gender-egalitarian areas vs. a native husband increases as her educational attainment increases. This contribution confirms results in Gabrielli and Paterno (2016) and Maffioli, Paterno and Gabrielli (2014) on different data sources and extends their findings highlighting the heterogeneous patterns that exist across different migrant origins. Furthermore, the perspective we propose in this contribution shows that intermarriage not only can be used, as traditionally done, as an indicator of integration of migrants, but can also shed light on the marriage patterns of the native population.

References

1. Bellani, D., Esping-Andersen, G. and Nedoluzhko, L. (2017). Never partnered: A multilevel analysis of lifelong singlehood. Demogr Res, 37: 53–100.
2. Gabrielli, G., & Paterno, A. (2016). Selection criteria of partner: comparison between transnational and homogamous couples in Italy. Genus, 71(2-3).
3. Grossbard-Shechtman, S., & Fu, X. (2002). Women's Labor Force Participation and Status Exchange in Intermarriage: A Model and Evidence for Hawaii. J Bioeconomics, 4(3): 241–268.
4. Grow, A., & Van Bavel, J. (2015). Assortative mating and the reversal of gender inequality in education in Europe: An agent-based model. PloS one, 10(6), e0127806.
5. Guetto, R., & Azzolini, D. (2015). An empirical study of status exchange through migrant/native marriages in Italy. J Ethn Migr Stud, 41(13): 2149–2172.
6. Kalmijn, M. (2013). The Educational Gradient in Marriage: A Comparison of 25 European Countries. Demography, 50: 1499–1520.
7. Maffioli, D., Paterno, A., & Gabrielli, G. (2014). International married and unmarried unions in Italy: Criteria of mate selection. Int Migr, 52(3): 160–176.
8. Medrano, J. D., Cortina, C., Safranoff, A., & Castro-Martín, T. (2014). Euromarriages in Spain: Recent trends and patterns in the context of European integration. Popul, Space and Place, 20(2): 157–176.
9. Van Bavel, J. (2012). The reversal of gender inequality in education, union formation, and fertility in Europe. Vienna Yearb Popul Res, 10: 127–154.

Statistical Analysis for mobility and transportation

A multilevel Analysis of University attractiveness in the network flows from Bachelor to Master's degree

Un'analisi multilivello dell'attrattività delle Università nel flusso di passaggio dalla Laurea Triennale alla Magistrale

Silvia Columbu and Ilaria Primerano

Abstract In this work we aim to study the mobility choices of Italian students in the transition from bachelor to masters degree in order to assess the role played by the field of study. We consider micro-data from the Italian National Student Archive on a cohort of students enrolled for the first time at the university in a.y. 2011-12 who enrolled to a master degree program in the a.y. 2014-15 or 2015-16. We study the incoming and outgoing flows of students moving from bachelor to master's degree between provinces and universities. We then assess the effects on mover choices of network centrality measures in terms of hub and authorities adopting a multilevel multinomial logit model.

Abstract *Con questo contributo ci proponiamo di studiare le scelte di mobilità per gli studenti italiani nel passaggio tra Laurea Triennale e Magistrale in base all'indirizzo di studi. Si considerano micro-dati provenienti dall'Anagrafe Nazionale Studenti relativi alla coorte di studenti immatricolati in una triennale nell'a.a. 2011-12 e iscritti a una Laurea Magistrale nell'a.a. 2014-15 o 2015-16. Studiamo i flussi di studenti in entrata ed uscita tra province e università italiane nel passaggio alla Laurea Magistrale. Inoltre, analizziamo l'influenza delle misure di centralità in termini di hub e authority applicando un modello multilivello multinomiale.*

Key words: Second level Student Mobility, Multinomial Multilevel, Fields of study

Silvia Columbu
Dept. of Mathematics and Computer Sciences, University of Cagliari, Via Ospedale, 72, Cagliari e-mail: silvia.columbu@unica.it

Ilaria Primerano
Dept. of Computer Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano e-mail: iprimerano@unisa.it

1 Introduction

In the Italian University System there is a growing interest in understanding the dynamics of internal mobility for students. This aspect is relevant given that the attractiveness of university students coming from other territories and/or institutions is considered one of the criteria for the distribution of public financial resources.

Exploratory studies on internal student mobility have been carried out mainly considering macro-geographical and interregional flows of students in order to analyze the attractiveness of different territories [5] [2] rather than those of universities. If we consider the students' internal migration in the transition from the first (Bachelor) to the second (Master) level of university studies, the literature is quite limited. A recent contribution [4] focuses on the flows from the South to the North of Italy.

In this work we analyze students' mobility flows to detect and understand the main migration patterns in the transition from Bachelor's to Master's degree between Italian provinces accross the country and the universities. We adopt a two step procedure of analysis. At first, we use social network centrality measures to investigate the role and position of each University/Province in the network. Then, we define a multilevel multinomial logit model to assess the effects of these measures on students' mobility choices conditionally to the specificity of the field of study undertaken. The paper is structured as follows. The data we use are presented in section 2, while the methodological approach is described in section 3. First results and concluding remarks are given in section 4.

2 Data

The micro-data used to analyze the Italian mobility student flows are provided by the Italian National Student Archive (NSA) [1]. We focus on a longitudinal cohort of 46,127 students, clustered in 53 fields of study, enrolled for the first time at university in the a.y. 2011-12. We use these data to study the incoming and the outgoing flows of students in the transition from bachelor's to master's degree between provinces and universities. We, therefore, consider only the students who have completed an undergraduate program and enrolled to a master degree program in the a.y. 2014-15 or 2015-16.

Looking at the main regions and provinces of destination of graduates, it arises that about 84% of regional *mover* students are spread in six regions (namely, Lombardy, Emilia Romagna, Lazio, Tuscany, Piedmont, and Veneto) and about 54.13% are concentrated in four provinces (Milan, Rome, Bologna, and Turin). We defined a regional *mover master* variable by taking into consideration all the possible alternative paths in the student mobility choices at bachelor and master degree. We distinguished among students who migrated only for their bachelor, only for the

[1] The micro-data at student level are available by the NSA archive for the universities involved in the Italian Ministerial grant PRIN 2017 CUP: B78D19000180001.

master or for both levels of study. In particular, we classified students with respect to their mover history in five categories: 1) Stayer I (Bachelor) & II Level degree (Master); 2) Mover I & Stayer II; 3) Stayer I & Mover II; 4) Mover I & II (same region) and 5) Mover I & II (different region). We observe that overall 11% of students changes region for the first time in the transition from bachelor's to master's degree, with losses in terms of outgoing bachelor graduates for each region (with respect to the number of graduated) that vary from the 23% and 25% of Sardinia and Sicily to the 6% and 5% of Lombardy and Piedmont. The proportion of first level mover graduates enrolling for the master in their region of origin is trascurable (1.37% of the total).

3 A two-step strategy of analysis

3.1 First step: Network data and centrality measures

Students mobility flows in the transition from Bachelor's to Master's degree can be represented as a network data structure, where the universities and the provinces are the nodes and the students exchanges between them are the edges. The number of students involved in these exchanges define the weight of each edge. The corresponding adjacency matrix **A** is directed and weighted (with a link from the origin to the destination). Formally, we are dealing with a valued directed graph $\mathscr{G}_v(\mathscr{N},\mathscr{L},\mathscr{V})$ consisting of three sets of information: a set of nodes, $\mathscr{N} = \{n_1,n_2,\ldots,n_g\}$, a set of edges (or arcs), $\mathscr{L} = \{l_1,l_2,\ldots,l_L\}$, and a set of values, $\mathscr{V} = \{v_1,v_2,\ldots,v_L\}$ attached to the edges (or arcs). In particular, to explore the second level mobility flows, we consider as origin the provinces (or universities) where students achieved the Bachelor's degree and as destination the provinces (or universities) where students enrolled for their Master's degree.

Among the network centrality measures, we adopt the Hub and Authority scores [3] to shade some lights on the role played by the territories and the universities in the master students' mobility network. Specifically, the authority score is here perceived as an indicator of the reputation of a given province or university, thus of its attractiveness; whereas the hub score is interpreted as an indicator of awareness of students' mobility choices.

In this context, the classification of nodes in hub and authority allow us to define as good hubs the universities/provinces which are pointing to many good authorities (i.e. good exporter). At the same time, good authorities are the universities/provinces which are pointed to by many good hubs (i.e. good importers).

Descriptive statistics on hub and authority measures, as discussed in [1], highlight that there is an exchange of students across few leading provinces mainly located in the North of Italy. Similar observations can be make looking at the flows among universities. This result is confirmed also from a medium correlation among the measures of hub and authority in the province network (0.49) as well as in the

university one (0.58), i.e. the exchange of students at second level is driven by a limited number of provinces and/or universities.

3.2 Second Step: Multilevel approach

We apply a two level random intercept multinomial logistic model [6] [7] to study the probability of bachelor graduates, clustered in fields of study, to be in a specific *mover* status. Let us denote with Y_{ij} a categorical variable which indicates the mobility status k ($k = 1,\dots,K$) of student i ($i = 1,\dots,n$) in the field of study j ($j = 1,\dots,J$). We assume for the response Y_{ij}, conditional on the random effects, a multinomial distribution of parameters $(\pi_{ij}^{(1)},\dots,\pi_{ij}^{(K)})$. The estimated model can be specified using a multinomial logistic function as follow

$$\log\left(\frac{\pi_{ij}^{(k)}}{\pi_{ij}^{(1)}}\right) = \tau^{(k)} + \boldsymbol{X}_{ij}^{T}\boldsymbol{\beta}^{(k)} + \boldsymbol{Z}_{ij}^{T}\boldsymbol{\gamma}^{(k)} + \theta_{j}^{(k)} = \eta_{ij}^{(k)} \quad k = 2,\dots,K \qquad (1)$$

where the probability of the response to belong to one of the k categories is specified as

$$\pi_{ij}^{(k)} = P(Y_{ij} = k) = \frac{\exp(\eta_{ij}^{(k)})}{1 + \sum_{k=2}^{K}\exp(\eta_{ij}^{(k)})}.$$

In equation (1), $\boldsymbol{X}_{ij}$ is a vector of individual explanatory variables, $\boldsymbol{Z}_{ij}$ is the vector of the network measures, $\theta_{j}^{(k)}$ is the 2-level random component for the contrast of response category k with reference category $k = 1$. The parameters in equation (1) are specific to the category k. We assume that the random effects $(\boldsymbol{\theta_j}^{(2)},\dots,\boldsymbol{\theta_j}^{(K)})$ follow a multivariate normal distribution with mean $\mathbf{0}$ and matrix of covariance Σ. The model has been estimated with the runmlwin routine which calls MLwiN scripts from Stata [8] by adopting Monte Carlo Markov Chain algorithm.

The parametrization adopted aims to assess the effect of network measures on the different typologies of *mover* students enrolled in different fields of study. We therefore consider as explanatory variables the measures of authority and hub at University and Province level. We consider also some demographic characteristics of students such as age, gender, whether students had a highschool diploma from a Lyceum or from another school, the region where is located the University where they enrolled for their master, the geographical macro-area from where they originally come. In the model estimation we consider $K = 4$ (the group of Mover I & Stayer II was not considered in the estimation). We compare all the possible mobility choices with the only one representing the non-mobility (the reference category was Stayer I & II Level degree).

4 First results and concluding remarks

Model estimates, reported in Table 1, suggest that there is a field of study effect in the student mobility conditions, as the values of the between field of study variability display for all mover category status (k=2,..,4) with respect to students not in mobility (k=1). In particular, this variability is higher in those groups of students that make the stronger decision (k=2 and k=4). We observe also that there are some associations among students in mobility enrolled in the same field of study. In particular, based on random effects variances and covariances estimates, we find that there is a high positive correlation (0.97) among students in mobility at the second level for the first time (k=2) and those already in mobility that choose to make another migration (k=4). It is moderate also the sharing between the two groups of always mover students (k=3 and k=4), with a correlation of 0.41.
Estimates of fixed effects coefficients confirm that measures of network centrality can contribute to explain the mobility choices of master students conditional upon the field of study they enrolled. In particular, we observe that the authority of the territory (Authority Province) is one of the major factors of attractivity for students in mobility. Once we have controlled for the specific field of study, it arises that the power of authoritative universities (Authority University) in attracting master students affects with a negative sign the log-odds of being in mobility at both levels of study inside the same region, whereas it vanishes in the other two groups of movers. Furthermore, bachelors graduated in universities that proved to be good exporters (Hub University) are less inclined to make a mobility choice at the new enrollement, except for movers at both levels changing region. On the other hand, for this last group of students, coming from an exporter province decreases the propensity to be in mobility. This last result suggests that, for students in the same field, the decision to undertake a new mobility choice is strictly linked to the University of origin rather than to the Province of origin. The demographic students characteristics do not reveal unexpected results. The estimates show that older students sharing similar choices in their study path are less inclined to the mobility, there is a gender effect and it is confirmed the typical italian North-South divide.

Acknowledgements. This contribution is financially supported by the Italian Ministerial grant PRIN 2017 From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide, Principal Investigator Massimo Attanasio n. 2017HBTK5P. CUP: B78D19000180001.

References

1. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.: Exploring the Italian student mobility flows in higher education. In: Bini M., Amenta P., D'Ambra A., Camminatiello I. (eds). Statistical Methods for Service Quality Evaluation. Book of short papers of IES 2019 Rome, Italy, July 4-5, pag. 46-49. Cuzzolin Editore, Napoli (2019).
2. D'Agostino A., Ghellini G., Longobardi S.: Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. Electronic Journal of Applied Statistical Analysis. **12 (4)** (2019).

PREDICTORS	BASELINE CATEGORY: STAYER I & II $k=1$		
	STAYER I & MOVER II $k=2$	MOVER I & MOVER II (SR) $k=3$	MOVER I & MOVER II (DR) $k=4$
FIXED EFFECTS			
Intercept	-4.857***(0.4)	-3.912***(0.238)	-7.309*** (0.606)
Authority University	-0.042(0.118)	-0.613***(0.101)	-0.287(0.181)
Authority Province	0.673***(0.111)	1.056***(0.089)	1.023*** (0.186)
Hub University	-0.370***(0.090)	-0.767***(0.076)	0.589***(0.142)
Hub Province	-0.709***(0.104)	0.025(0.098)	-1.830***(0.159)
Age	-0.057***(0.018)	-0.009 (0.01)	-0.035(0.027)
Gender (F vs M)	0.095**(0.039)	-0.031(0.035)	0.009(0.062)
Lyceum (vs Other)	0.318***(0.052)	0.230***(0.042)	0.495***(0.089)
DESTINATION REGION (vs ABRUZZO)			
TRENTINO A.A.	7.289***(0.200)	5.533***(0.165)	7.896***(0.295)
EM. ROMAGNA	6.401***(0.183)	4.771***(0.146)	6.610***(0.259)
FRIULI VEN. GIU.	6.341***(0.196)	4.684***(0.153)	6.886***(0.291)
VENETO	6.108***(0.170)	4.144***(0.135)	6.298*** (0.244)
PIEDMONT	5.812***(0.176)	3.038***(0.148)	5.935***(0.260)
LIGURIA	5.444*** (0.223)	3.673***(0.168)	5.683***(0.367)
LOMBARDY	5.044***(0.181)	3.470***(0.146)	5.380***(0.273)
TUSCANY	3.769***(0.164)	2.543***(0.127)	4.084*** (0.235)
LAZIO	2.985***(0.168)	1.397***(0.130)	3.450***(0.244)
MARCHE	2.968***(0.188)	1.702***(0.144)	2.993***(0.289)
UMBRIA	2.884***(0.253)	2.329***(0.182)	3.830***(0.327)
MOLISE	-0.998*(0.510)	-0.567**(0.235)	-267.3*(147.9)
CAMPANIA	-2.678***(0.215)	-3.792***(0.128)	-2.349*** (0.355)
CALABRIA	-2.741***(0.393)	-3.846***(0.249)	-49.15*(27.43)
BASILICATA	-3.204** (1.293)	-1.938***(0.329)	-139.8(129.4)
APULIA	-3.711***(0.420)	-3.206***(0.143)	-2.338***(0.436)
SARDINIA	-5.110***(1.478)	-5.857***(0.761)	-3.733***(1.388)
SICILY	-4.146***(0.540)	-2.947***(0.136)	-2.155***(0.413)
AOSTA VALLEY	-390.4**(191.7)	-189.0*(107.2)	-623.5*(377.8)
ORIGIN MACRO AREA (vs CENTER)			
SOUTH	4.095***(0.091)	4.487***(0.086)	4.639***(0.114)
NORTH	-2.273*** (0.070)	-1.656***(0.067)	-2.450***(0.110)
RANDOM EFFECTS: FIELD OF STUDY			
$SD(\theta_j^{(k)})$	0.332***(0.076)	0.079***(0.023)	0.567***(0.140)
$COV(\theta_j^{(2)},\theta_j^{(3)})$	0.042(0.029)		
$COV(\theta_j^{(3)},\theta_j^{(4)})$		0.086**(0.04)	
$COV(\theta_j^{(2)},\theta_j^{(4)})$		0.421***(0.097)	

Table 1 Multinomial Logistic model estimates

3. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), **46(5)**, 604–632 (1999)
4. Enea, M.: From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree. In: Perna C., Pratesi M., Ruiz-Gazen A. (eds) Studies in Theoretical and Applied Statistics. SIS 2016. Springer Proceedings in Mathematics & Statistics, **227**. Springer, Cham (2018)
5. Giambona, F., Porcu, M., and Sulis, I.: Students mobility: assessing the determinants of attractiveness across competing territorial areas. Social Indicators Research, **133(3)**:1105-1132 (2017).
6. Grilli L., Rampichini, C.: A multilevel multinomial logit model for the analysis of graduates' skills. Stat. Meth. & Appl. **16**, 381–393 (2007)
7. Leckie, G.: Three-Level Multilevel Models-Concepts. LEMMA VLE Module 11,1–47 (2013)
8. Leckie, G., Charlton, C.: runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. Journal of Statistical Software, **52(11)**,1–40 (2013)

Analysis of mobility data through a novel Cheng and Church algorithm for functional data

Analisi di dati di mobilità estendendo il modello di bi-clustering di Cheng e Church ai dati funzionali

Marta Galvani[1] , Agostino Torti[2,3] and Alessandra Menafoglio[3]

Abstract The aim of this work is to discover spatio-temporal patterns on the usage of a generic mobility system through bi-clustering algorithms. Since mobility data are continuously dependent on time, it seems natural to treat them as continuous functions over a temporal domain. Many different algorithms have been developed with the aim of clustering functional data; however, a lack of techniques occurs when these functional data are intrinsically ordered in a matrix structure and the interest is to simultaneously group the rows and the columns of the functional data matrix. In this work, we developed a new bi-clustering technique for functional data, extending the well known Cheng and Church algorithm for multivariate data. This new algorithm has been applied on a real dataset describing the flows of users on a mobility network. As results we are able to discover groups of stations and days with similar behavior, thus characterizing the mobility demand and identifying the presence of eventual issues in the system.

Abstract *Lo scopo di questo lavoro è di studiare i pattern spazio-temporali in un sistema di mobilità utlizzando tecniche di bi-clustering. Poichè i dati di mobilità hanno una dipendenza continua nel tempo, appare naturale trattarli come funzioni su un dominio continuo. Diversi algoritmi sono stati proposti allo scopo di clusterizzare dati funzionali; tuttavia la letteratura relativa a tecniche di bi-clustering per dati funzionali risulta ancora carente. In questo lavoro viene proposto un nuovo algoritmo di bi-clustering per dati funzionali che estende l'algoritmo introdotto da Cheng e Church per dati multivariati. Questo nuovo modello viene applicato su un dato reale relativo ai flussi di mobilità. In questo modo siamo in grado di raggruppare blocchi di stazioni e giorni con coportamenti simili, caratterizzando così la domanda di mobilità e evidenziando eventuali criticità nel sistema.*

Key words: Mobility, Bi-clustering, Functional Data

---

[1]Department of Mathematics, University of Pavia, Italy
[2]Center for Analysis Decisions and Society, Human Technopole, Milano
[3]MOX - Department of Mathematics, Politecnico di Milano, Italy

1 Introduction

In the last years, due to urbanization and globalization, the demand for transportation has been increased like never before. In order to design the mobility of the future, companies operating in the transportation field have started to collect a huge quantity of data monitoring the evolution of mobility flows. The analysis and managing of these data represent a new challenge in the development of new statistical methods as these data are characterize by a high spatio-temporal dependence.
The recent increase in the collection of high frequency data represents a new challenge in the development of novel statistical methods. In this context, the statistics community developed a wide number of algorithms considering these new data as functional data, i.e. continuous functions over a domain, see [5].
Many different algorithms have been developed with the aim of clustering this type of data. However, a more particular case, with respect to standard clustering problems, appears when data are intrinsically ordered in a matrix structure and the interest is to simultaneously group the rows and the columns of a data matrix. This problem is knew as bi-clustering and was first introduced in [2] for the bi-clustering of expression data. For a complete review of bi-clustering methods for multivariate data see [4].
The aim of this work is to discover spatio-temporal patterns on the usage of a bike sharing mobility system identifying areas and days with a similar behavior. In addition, since a mobility datum is continuously dependent on time and we would like to characterize its within day variability, we decide to see it as continuous function of time. To this end we developed a new bi-clustering technique for functional data, grouping simultaneously locations and days with the same daily function.
Bi-clustering models need to be extended when considering a functional data matrix, i.e. where each element is a function. New proposals of bi-clustering methods for functional data are introduced by [6] and [1]. These two works both proposed a procedure which generalizes the classical latent block model ([3]) for multivariate data. These procedures are model-based and assume the existence of a latent-block structure in the data-matrix, therefore they are both semi-parametric. In addition the output of these models is an exhaustive bi-clustering of the data matrix with a check-board structure.
In this work, we would like to introduce a novel flexible bi-clustering algorithm for functional data extending the method first introduced by [2] for multivariate data. The concept of bi-clustering and the measure of goodness of a bi-cluster are here extended to functional data. The new obtained model is non parametric, thus it does not assume any prior modeling assumptions on the data and obtains as final result a functional non exclusive bi-cluster, thus allowing some elements not to belong to any bi-cluster.
The functional bi-clustering model has been applied on a real dataset about the Bike sharing system of Lyon called Vélo'v, with the aim of discovering spatio-temporal patterns in the daily usage of the system. Launched in 2005, Vélo'v is the first bicycle-sharing system in France, with a network of more than 3000 bikes spread over 345 stations across the city. Specifically, we analysed data for one week

in March 2014, discovering groups of stations and days with similar behavior, thus characterizing the mobility demand.
Notice that, in this work, we briefly report the methodology and the results that have been deeply presented in [7].
In Section 2 the methodological framework is presented coupled with obtained results on the analysed data. Conclusions are presented in Section 3.

2 Methods and Analysis

In this work, we extend the definition of bi-cluster and the well know Cheng and Church algorithm [2] to the case of a functional data matrix A. The aim of the Cheng and Church bi-clustering algorithm is to find a submatrix $A' \in A$, corresponding to a subset of rows I and a subset of columns J, with a high similarity score. In particular, in the Cheng and Church algorithm, an *ideal bi-cluster* is a set of rows I and a set of columns J such that each element in the bi-cluster can be expressed by the average value in the bi-cluster plus the residue of rows and columns average value and the total average value. A particular measure of goodness is evaluated for a bi-cluster $A'(I,J)$ considering a score H which is the *Mean Squared Residue* between all the real values in $A'(I,J)$ and their representative values in the bi-cluster.
Extending these concepts to FDA, each element of the dataset matrix A is a function $f_{ij}(t)$ defined on a continuous domain T. In such framework we define an ideal bi-cluster A' as a subset of rows I and columns J such that each function belonging to the bi-cluster $A'(I,J)$ can be defined as follows:

$$f_{ij}(t) = f_{IJ}(t) \quad \forall i \in I \quad and \quad \forall j \in J$$

where $f_{IJ}(t) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} f_{ij}(t)$.
For easiness of interpretation we define the bi-cluster template observing only the average function in the bi-cluster.
Consequently, the extended H-score of a bi-cluster $A'(I,J)$ is:

$$H_{IJ} = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} \int_T \left\{ f_{ij}(t) - [f_{IJ}(t)] \right\}^2 dt$$

The algorithm starts considering the whole dataset and following an iterative procedure tries to find the biggest bi-cluster with a H-score value lower then a given threshold δ. In deteils to find a bi-cluster an iterative procedure is performed first removing the rows and the columns that contribute more to the H-score and then adding the rows and columns that contribute less. These steps are a natural extension of the node addition and deletion phases contained in the original Cheng and Church algorithm [2] to the functional framework. Each time a bi-cluster is found the algorithm proceeds looking for new bi-clusters in the remaining part of the data

matrix. The procedure stops when no more bi-clusters are found. The output of the algorithm is a non exhaustive bi-clustering, thus some elements do not belong to any bi-cluster.
We apply this algorithm on a real dataset about the Bike sharing system of Lyon called Vélo'v with the aim of discovering spatio-temporal patterns in the daily usage of the system. Notice that the system is a fixed stations service where bikes are picked up and dropped off in specific bike docks located around the city. For each station and for each day we now the hourly loading value (number of available bike over the number of available docks). Firstly we treat the available raw data as continuous functions, thus we model the bike station loading profile during the entire day by means of a point belonging to a space of continuous functions representing the number of available bikes divided by the total number of bike docks at each timestamp on the time domain (0,24). In this way we obtain 2415 curves, i.e. 345 stations per 7 days, representing all the elements of a dataset matrix A with the same dimensions (345x7).
After obtaining a functional data matrix of 2415 elements we apply the functional Cheng and Church algorithm obtaining 53 bi-clusters.

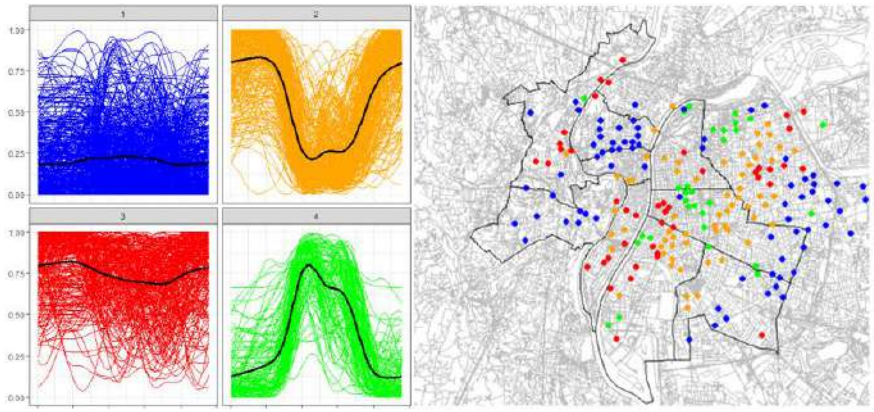


Fig. 1 The first four bi-clusters with in black the representative function of the bi-cluster (left) and the position on the map of Lyon of bike stations belonging to the first four bi-clusters (right)

In Figure 1 the first four bi-clusters, covering almost the 50% of the dataset, are reported. In particular, bi-cluster 1 and 3 represent constant profile which are respectively always empty or always full bike stations during the whole week. Bi-cluster 1 is composed by bike-stations located in the suburbs where people could not find a bike, while bi-cluster 3 represents bike-stations where people could not leave a bike. They both underlay an issue in the system identifying not used bike-stations. Bi-cluster 2 and 4, instead, represent two opposite users interactions with the system. They cover only working days and represent, respectively, bike stations always empty during the day and full during the night, and bike stations always full during the day and empty during the night. These particular behaviour could be explained

observing that bike-stations in bi-cluster 2 are almost all located in residential areas while, as opposite, bike stations in the bi-cluster 4 are mainly located in industrial areas.

3 Conclusions

In this paper, we analysed a real dataset concerning the daily loading profile of bike stations in the city of Lyon with the aim of discovering spatio-temporal patterns. A novel functional bi-clustering method is introduced extending the Cheng and Church algorithm for multivariate data. The application of this model to the data at hand allows us do discover sub-groups of stations following the same loading profile in a subgroup of days, underlying particular users behaviour and identifying some issue in the system.

References

1. Bouveyron, C. and Bozzi, L. and Jacques, J. and Jollois, F.: The functional latent block model for the co-clustering of electricity consumption curves. Journal of the Royal Statistical Society: Series C (Applied Statistics **67(4)**, 897–915 (2018)
2. Cheng, Y. and Church, G.M.: Biclustering of expression data. In Proceedings of Ismb **8**, 93–103 (2000)
3. Govaert, G. and Nadif, M.: Co-clustering: models, algorithms and applications. John Wiley & Sons (2013)
4. Pontes, B. and Giráldez, R. and Aguilar-Ruiz, J.S.: Biclustering on expression data: A review. Journal of biomedical informatics **57**, 163–180 (2015)
5. Ramsay, J.O.: Functional data analysis. Encyclopedia of Statistical Sciences (4), Wiley Online Library. (2004)
6. Slimen, Y.B. and Allio, S. and Jacques, J.: Model-based co-clustering for functional data. Neurocomputing **291**, 97–108 (2018)
7. Galvani, M., Torti, A., Menafoglio, A., Vantini, S. (2019) A novel bi-clustering method for functional data with misalignment, Manuscript.

Bridge closures in a transportation network: analysis of the impacts in the region of Lombardy

Chiusura di ponti in una rete stradale: analisi degli impatti in Lombardia

Agostino Torti[1,2], Marika Arena[3], Giovanni Azzone[1,3] and Piercesare Secchi[1,2]

Abstract This paper introduces a methodology to evaluate the socio-economic impacts of closures for maintenance of one or more infrastructures of a large and complex road network. Motivated by a collaboration with Regione Lombardia, we focus on a subset of bridges in the region, although we aim at developing a method scalable to all road infrastructures of the regional network, consisting of more than 10000 tunnels, bridges and overpasses. The final aim of the endeavor is to help decision-makers in prioritizing their interventions for maintaining, repairing and extending infrastructure segments. We develop two different levels of impact assessment, both providing a unique global score for each bridge closure and investigating its spatio-temporal effects on mobility. We take advantage of a functional data analysis approach enhanced by a complex network theory perspective, thus modelling the roads of Lombardy as a network in which weights attributed to the edges are functional data. Results reveal the most critical bridges of Lombardy; moreover, for each bridge closure, the most impactful hours of the day and the most impacted municipalities of the region are identified. The proposed approach develops a flexible and scalable method for monitoring infrastructures of large and complex road networks.

Abstract *In questo lavoro viene analizzata e discussa una nuova metodologia per la stima degli impatti socio-economici dovuti alla chiusura al traffico di una o più infrastrutture in una rete stradale complessa. Motivati da una collaborazione con Regione Lombardia, vengono condotte delle analisi su un caso studio di più ponti della regione, il cui obbiettivo a lungo termine è quello di sviluppare una metodologia scalabile e applicabile a tutte le infrastrutture stradali lombarde, ovvero più di 10000 fra tunnel, ponti e cavalcavia. Lo scopo ultimo è fornire degli strumenti che*

[1]Center for Analysis Decisions and Society, Human Technopole, Milano
[2]MOX - Department of Mathematics, Politecnico di Milano
[3]Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy

permettano di pianificare al meglio la gestione della rete stradale. Nel lavoro svolto vengono forniti più indici di impatto, definendo sia un indice globale univoco per la chiusura di ogni ponte sia analizzandone gli effetti spazio-temporali sulla mobilità. A tal fine, vengono adottati allo stesso tempo strumenti modellistici dell'analisi di dati funzionali e della teoria dei network, modellizzando quindi la rete stradale della Lombardia come un network, il cui peso sugli archi è un dato funzionale. I risultati rivelano i ponti più critici della Lombardia individuando per ogni ponte le ore più impattanti della giornata e i comuni più colpiti. La metodologia sviluppata, grazie alla sua flessibilità e scalabilità, è pronta ad essere applicata per il monitoraggio di qualsiasi tipo di infrastrutture in altre reti stradali.

Key words: Bridge Closure, Spatio-Temporal, Network, Functional Data.

1 Introduction

Civil road infrastructures play a pivotal role in the modern society and are generally considered one of the main drivers of socio-economic development. However, these structures are naturally subject to deterioration caused by the effects of natural hazards, environmental conditions and aging ([1], [2]). In order to allow these infrastructures to be safe over their entire life, monitoring and maintenance activities are necessary. In a situation where resources are limited, it is clear the importance of planning maintenance interventions based on some priorities, for defining which facilities should be visited, when and how maintenance should be carried out, respecting budget and other resource constraints ([3]. This paper aims to address the issue of determining the socio-economic impacts of closures of one or more road infrastructures so to handle and prioritize maintenance activities. The problem of evaluating the impacts of road closures in transportation networks is not new in the literature. Over the years, several measures of impact assessment have been introduced looking at different indicators (e.g. [4]; [5]). However, all these works do not address two issues which we consider to be important for the correct management of human mobility in a damaged network: the temporal and the spatial aspects. Hence, this paper proposes a novel two-way approach which investigates the impacts of a bridge closure from a spatio-temporal perspective: from the temporal point of view, we analyze how the impact of a road closure varies according to the hour of the day; from the spatial point of view, we identify the areas most affected by the closure. Specifically, we provide two different levels of impact assessment, investigating the spatio-temporal effects of a bridge closure and providing, for each bridge, a unique global score which can be used for ranking. The case study illustrated in the paper refers to the region of Lombardy in Italy, following a research project financed by Regione Lombardia, the regional government, to Politecnico di Milano. The project has the final aim of developing a scalable method for the assessment of the impacts of the closure of all the about 10000 bridges of the region. The results reported in this paper, however, relate a pilot study focusing on a sample of 290 bridges. Notice

that in this work we briefly report the methodology and the results that have been deeply presented in [6].
In Section 2 the methodological framework is presented coupled with analyses and results on the case study of Lombardy. Conclusions are presented in Section 3.

2 Methods and Analyses

Lombardy is one of the twenty administrative regions of Italy with an area of 23,844 square kilometres and a resident population of about 10 million people, forming one-sixth of Italy's population. Every day in Lombardy there are about 17 million trips across the region with different means of transport, 75% of which involves a motor vehicle on the road. For our analyses, we use two main complementary information sources both provided to us by Regione Lombardia: a modeling of the road network of Lombardy and origin-destination (OD) matrices. The modeled road network is a spatial network made of about 37000 nodes and 82000 directional links, i.e. segments of road between two intersections, which model all types of roads of the real network with a major simplification for the municipal roads. For each directed edge of the network, the length (km), the typical velocity without traffic (km/hour) and the typical travel time (hour) are known. Moreover, for all the 290 bridges under examination the corresponding edge to which each bridge belongs is known. The OD matrices of Lombardy contain the number of trips between all the municipalities of the region, which is split in 1450 mobility areas, for each hour of a typical working day in the time span between February and May 2016. Since it is well known that the daily traffic profiles are characterised by a within-day variability, we model the OD trips making use of tools from Functional Data Analysis (FDA), the branch of statistics dealing with curves, surfaces or anything else varying over a continuum [7]. Specifically, we model each mobility flow between an origin and a destination during the 24 hours by means of a point belonging to a space of continuous functions defined on the time domain (0,24). For more details about the OD trips modelling see [6].

For the scope of our analyses, we first pass through a data fusion step aggregating our sources of information. Using the modeled road network as base network we assign each mobility area of the OD matrices, whose GPS coordinate is known, to the closest node of the modeled road network. In this way we are able to estimate how the OD trips move across the modeled road network by searching the shortest paths. Then, we asses the impact of each bridge according to its importance in maintaining a proper connectivity between all origin and destination couples of the OD matrices ([4], [8]). For each bridge, we remove from the road network the edge to which it belongs and we measure its impact in terms of total-trip cost variation on all the affected trips. The first step is to evaluate the cost for each OD trip caused by the removal of a edge of the network. In our case each cost is equivalent to the increase in terms of travel time. Then, to obtain a global index of impact for the closure of the bridge belonging to the removed edge, we weight this cost by the number of trips

it affects during the day and we sum over all the the OD trips. The unit measure of the estimated global index is man-hour, that is representative of the socio-economic impacts of the bridge closure. Applying this procedure on our sample of bridges, we obtain for each one of them a unique global index that can be used to obtain a clear ranking of the bridges, so to help the regional government in the prioritization of maintenance investments. In Figure 1 all the 290 analysed bridges are reported according to their global index. It appears that the most critical bridges are located in the north side of the region, which is mainly in a mountain area with a sparse road network made of few main roads. Moreover, these bridges are often over a river.

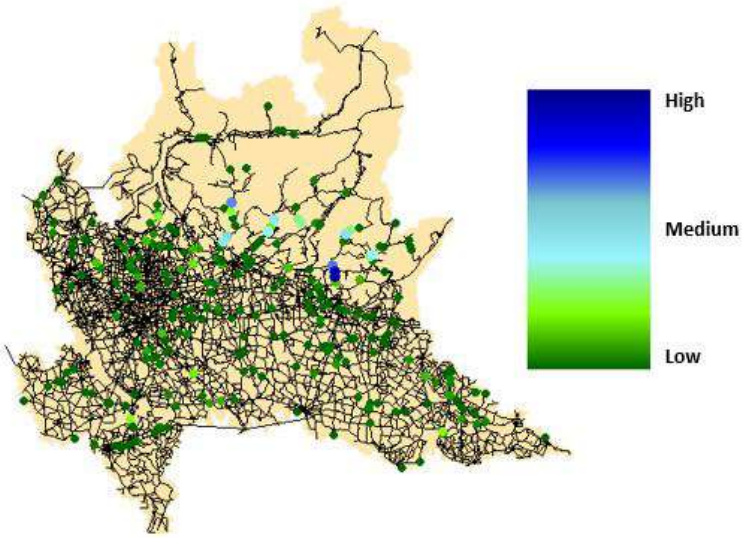


Fig. 1 Map of Lombardy highlighting the 290 analysed bridges according to their global index.

In addition to the just estimated global index, we provide a novel two-way spatio-temporal approach to explain how this impact spreads across the region and during the day. From a spatial perspective, we estimate how the impact of a bridge closure is distributed across the region, namely highlighting the most impacted areas of Lombardy. Hence, we apply again the same procedure used above for the construction of the global index, but now, when summing over all the the OD trips, we fix the origins or the destinations, respectively. For each origin we measure the cumulated impact of the closure of a bridge summing over all the trips starting from that origin. Analogously, for each destination we sum over all the trips directed to that destination. As result, for each bridge, we obtain a global index for each origin and for each destination. Obviously, summing up this spatial impact over all the origins or other all the destinations respectively we obtain the global index of the bridge closure. From a temporal perspective, we estimate how the impact of a bridge closure changes along time. To this end, we apply the same procedure used above

to estimate the global index but instead of weighting each OD cost by the total daily number of trips we weight it by the functional datum describing the number of trips from an origin to a destination at any time t of the day. These spatio-temporal analyses can be of great importance to better organize the road maintenance schedule. For example, the information about the most impacted origins or destinations by a bridge closure can be passed on to the affected travelers, with the aim of letting them optimize their trips. In addition, knowing the most impactful hours of the day for each bridge closure, road works can be planned by opening and closing the lanes of the road accordingly. Results related to this part are reported in [6].

3 Conclusions

In this work we discussed methodologies to evaluate the socio-economic impacts of the closure of critical infrastructures in a road network, stimulated by the analysis of a subset of bridges in the region of Lombardy in Italy. Specifically, we developed two different levels of impact assessment, both providing a unique global score for each bridge closure and investigating its spatio-temporal effects on mobility. The application of this methodology on the case of Lombardy revealed the most critical bridges of the region, which are mainly located in a mountain area, providing useful information to prioritize maintenance investments and for the correct allocation of available resources. The proposed approach is flexible and repeatable, thus ready to be applied to other realities in other countries.

References

1. Wang, C., Zhang, H., Li, Q. (2017). Reliability assessment of aging structures subjected to gradual and shock deteriorations. Reliability Engineering System Safety, 161, 78-86.
2. Ellingwood, B. R. (2005). Risk-informed condition assessment of civil infrastructure: state of practice and research issues. Structure and infrastructure engineering, 1(1), 7-18.
3. Shah, Y. U., Jain, S. S., Parida, M. (2014). Evaluation of prioritization methods for effective pavement maintenance of urban roads. International Journal of Pavement Engineering, 15(3), 238-250.
4. Taylor, M. A., Sekhar, S. V., D'Este, G. M. (2006). Application of accessibility based methods for vulnerability analysis of strategic road networks. Networks and Spatial Economics, 6(3-4), 267-291.
5. Rupi, F., Bernardi, S., Rossi, G. and Danesi, A. (2015) The evaluation of road network vulnerability in mountainous areas: a case study. Networks and Spatial Economics, 15(2):397–411.
6. Torti, A., Arena, M., Azzone. G, Secchi, P. and Vantini, S. (2019) Bridge closures in the transportation network of Lombardy: analysis of the socio-economic impacts from a spatio-temporal perspective, Manuscript.
7. Ramsay, J.O.: Functional data analysis. Encyclopedia of Statistical Sciences (4), Wiley Online Library. (2004)
8. Cantillo, V., Macea, L. F., Jaller, M. (2019). Assessing vulnerability of transportation networks for disaster response operations. Networks and Spatial Economics, 19(1), 243-273.

Statistical Methods and Applications in Social Network Analysis

A clustering procedure for ego-network data: an application to Italian elders living in couple

Una procedura di clustering per l'analisi di reti ego-centrate: applicazione ai dati italiani sugli anziani che vivono in coppia

Elvira Pelle and Roberta Pappadà

Abstract The analysis of ego-network characteristics (especially size and composition) has become crucial in studying many aspects of everyday life. In this work, we propose a clustering procedure to find a partition of ego-networks into homogeneous groups according to their features. We use data from the "Family and Social Subjects" (FSS) survey conducted by the Italian National Statistical Institute in 2009, on elderly couples with both partners aged 65 years and more. Preliminary results show the suitability of our proposal to analyze this kind of ego-network data.

Abstract *L'analisi delle caratteristiche delle reti ego-centrate (in particolare dimensione e composizione) è diventata cruciale nello studio di molti aspetti della vita quotidiana. In questo lavoro, proponiamo una procedura di clustering per determinare una partizione di reti ego-centrate in gruppi omogenei in base alle loro caratteristiche. La procedura viene illustrata sui dati dell'indagine Istat su "Famiglia e soggetti sociali" (FSS, 2009), con riferimento alle coppie di anziani con entrambi i partner di età pari o superiore a 65 anni. I primi risultati mostrano come il metodo proposto possa essere utile ad analizzare questo tipo di dati di rete.*

Key words: Ego-network, Italian couples, Hierarchical clustering

1 Introduction

In recent years the attention on the analysis of ego-networks in which individuals are embedded was extended to many areas of social sciences. Recent studies have

Elvira Pelle
Department of Communication and Economics, University of Modena and Reggio Emilia, Italy e-mail: elvira.pelle@unimore.it

Roberta Pappadà
Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Italy e-mail: rpappada@units.it

shown the importance of network characteristics, especially size and composition, and of their effects on many aspect of everyday life (such as, social support [2, 7], health and well-being [10, 13]).

In this work we present a clustering procedure to identify homogeneous groups of ego-networks with specific characteristics. This problem has been addressed by some authors in recent literature (see [4, 1]). We use data from "Family and Social Subjects" (FSS) survey carried out by the Italian National Statistical Institute (ISTAT) in 2009. Focusing on Italian elders living in couples (married or unmarried) with both partners aged 65 years and more, we propose a clustering procedure to detect groups of ego-networks that are similar according to their features.

The remainder of the article is organised as follow: in Section 2 we present the characteristics of elderly Italian people and ego-network characteristics. In Section 3 we describe the clustering procedure and main results. Section 4 ends the paper with some concluding remarks.

2 Data description and ego-network of contact construction

We focus on Italian elders living in couples (married or unmarried) with both partners aged 65 years and more without other members (we consider 1722 couples, that is $n = 3444$ individuals, 7.8% of the total of respondents). Looking at the place of residence, 44.2% of elder partners live in Northern Italy, 19.5% lives in the Center and the remaining 36.4% in the South or Islands. According to the age, 60% of elders has a partner in the same age group, while in 37% of couples man is older than woman; on average, women are 73 years old while men are 76.2 years old. The mean number of children is 2.1 and almost 70% of elder partners has at most 2 children. No striking differences appear between the distribution of the health conditions for women and men: in both cases about 80% declares to perceive its own health as good or fair, while about 20% of partners perceives a bad health status. More than 18% of men presents a medium/high level of education (high school diploma or university degree), while about 87% of women has a low level of education (compulsory or none). For almost the totality of men in a couple (about 94%), pension is the principal source of income, while this percentage decreases to 71% for women, for which the second source of income is represented by allowance by partners (about 25%).

As well-known an ego-network can be derived "looking at relations from the orientation of a particular person" [5, p. 509], usually referred to as "ego", while the persons or institutions connected to the ego by some relations of interest are referred to as "alters". We use relational information on non-cohabitant persons in FSS 2009 edition to construct the ego-network of contacts for each elderly partner, as proposed in [2]. In particular, we use data about relation with non-cohabitant siblings, children, grandchildren, as well as information on the type and the number (if any) of other relatives and friends, and on the presence of neighbours. Thus, aggregating alters by their role relation with ego, we identify a maximum of 6 different alters'

role in the resulting ego-networks. Note that for each alter role data supply information about the presence and the number, except for neighbours, for which only the presence is available. The considered alter roles can be naturally classified into 4 network typologies: the "kin network" with alters from immediate family (children) and extended family (siblings, grandchildren and relatives), the "non-kin network" composed only by neighbours and/or friends, the "comprehensive network" with alters from both kin and non-kin network, and "other network" collecting network typologies not in the aforesaid categories (see Table 1). A small percentage of persons is not related to any of these types of alter (6.5% and 5.6% for men and women, respectively). Note that in the following section we will focus on those egos that are related with at least one of the considered types of alters, i.e. we consider $n = 3235$ individuals, 50.23% female and 49.77% male partners.

Table 1 Distribution of network typologies by gender (M=mean size, SD= standard deviation)

Gender	Isolated	Kin	Non-kin	Comprehensive	Other
Men	0.065	0.435 (M:4.4; SD:2.7)	0.042 (M:3.2; SD:3.0)	0.362 (M:9.9; SD:5.8)	0.096 (M:6.0; SD:4.2)
Women	0.056	0.437 (M:4.5; SD:2.8)	0.041 (M:3.2; SD:3.3)	0.367 (M:9.8; SD:5.8)	0.099 (M:6.1; SD:4.4)

3 Clustering methodology

In this section we develop a clustering procedure in order to identify groups of similar ego-networks with respect to their features. To this aim, for each ego E_i ($i \in \{1,\dots,n\}$) we collect the total number of individuals in the types of alters *siblings*, *children*, *grandchildren*, *relatives* and *friends*. In addition, we consider the availability of neighbours as a binary variable taking value 1 in case of presence of *neighbours* an ego "can count on if necessary", and 0 otherwise. The resulting data matrix that we use for clustering purposes is a $n \times p$ matrix, where $p = 6$ and the i-th row $x_i = (x_{i1},\dots,x_{ip})^T$, corresponding to the feature vector for ego E_i, is composed by five numeric attributes (the counts on each alter category) and one binary attribute (neighbours).

We now consider the problem of computing a $n \times n$ dissimilarity matrix for each pair $(x_i, x_{i'})$ of ego-networks for the female and male partners, separately. In general, let p_n and p_b be the number of numerical and binary variables, respectively ($p = p_n + p_b$). As it is common in clustering mixed data types, the distance between two units can be computed using partial dissimilarities with suitable weights (see, e.g., [11, 12]). We therefore introduce the following measure

$$d(x_i, x_{i'}) = \sum_{r=1}^{p_n} d_1(x_{ir}, x_{i'r}) + \gamma \sum_{s=1}^{p_b} d_2(x_{is}, x_{i's}) \quad (1)$$

where $d_1(x_{ir}, x_{i'r}) = 0$ if $x_{ir} = x_{i'r} = 0$, and $d_1(x_{ir}, x_{i'r}) = |x_{ir} - x_{i'r}|/(x_{ir} + x_{i'r})$, otherwise. In the latter case d_1 corresponds to the well-known Canberra metric. The dissimilarity measure on binary attributes is the number of mismatches, i.e., we set $d_2(x_{is}, x_{i's}) = 0$ for $x_{is} = x_{i's}$, and $d_2(x_{is}, x_{i's}) = 1$ otherwise. In this case, $p_n = 5$ and $p_b = 1$. In addition, γ is a weight for the binary part, introduced to avoid favouring either type of attribute. The resulting dissimilarity matrix $D = (d_{ii'})$ can then be input to a standard clustering algorithm (see, e.g., [8]). We adopt agglomerative hierarchical clustering, although different methods could be applied in this context (see [1] and the references therein). To specify a value for γ, we exploit the information available on the network typology described in Section 2 (Table 1) that yields a clustering $\mathscr{C}$ of the units into four groups. A different partition $\mathscr{C}'$ into four groups is obtained via standard agglomerative hierarchical techniques using the distance defined in (1) with different values of γ. Then, γ is chosen so that the *normalized mutual information* (NMI) [9] between $\mathscr{C}$ and $\mathscr{C}'$ is maximized (maximum index value is 1). For alternative approaches to the computation of weighted distances in clustering mixed-type data see [6]. Following the described method, we obtain the dissimilarity matrix of ego-network data using the estimated γ ($\gamma = 1.25$ and $\gamma = 1.5$ for women and men, respectively).

To obtain the final clustering we use the *minimax-linkage* hierarchical clustering method, implemented in the `R` package `protoclust`, where the distance between clusters is measured by the minimax radius of the resulting merged cluster. This method allows to identify the associated cluster prototypes, i.e. central units, chosen from the original datasets (for more details see [3]). The resulting *dendrograms* for female and male partners are shown in Figure 1. In both cases, the dendrogram structure seems suggesting a partition into $k = 4$ groups may be appropriate. This is not surprising, given that the starting point for our proposal is represented by the prior knowledge of individual network type used to calibrate the distance in 1. The final clustering can then be used to gain a more exhaustive description of the grouping structure in the data.

Table 2 can be useful for the interpretation of the clustering results. For instance, looking at female partners, the larger group (Cluster 1) shows strong separation with respect to non-kin groups, being characterized by a higher number of kin contacts (36% of alters are grandchildren). More than 30% of female partners belonging to Cluster 2 are aged 70-74 years (see Figure 2), and are connected to people across different kind of alters, including non-kin. The persons in Cluster 3 are individuals connected to a larger number of alters (the network mean size is 10), where we observe that the number of friends is prevalent. Female partners in Cluster 4 are assimilated since none of the previously distinguished groups dominates these networks, and most alters are relatives and non-kin members. The limited number of children and grandchildren suggests that this cluster is characterized by women with none or one son/daughter. Similar considerations can be drawn from the clustering of ego-networks of male partners (see Table 2).

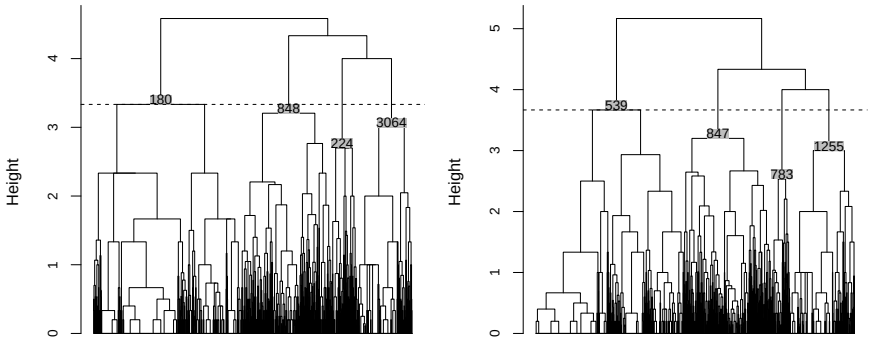


Fig. 1 The entire dendrogram of the ego-network dataset for females (*left*) and males (*right*) living in couple, with cut yielding the 4-cluster solution and prototypes displayed.

Table 2 Cluster size ($|C_i|$), mean size (MS) of ego-networks and relative frequency of alter types in the 4-cluster solution.

	$\|C_i\|$	MS	*siblings*(%)	*children*(%)	*grandchild.*(%)	*relatives*(%)	*friends*(%)	*neighb.*(%)
Female partner ($n = 1625$)								
1	753	4	14.65	34.69	35.79	14.87	0.00	0.00
2	276	5	1.28	22.20	19.65	19.36	17.94	19.57
3	465	10	7.19	17.04	19.08	20.80	26.36	9.53
4	131	8	10.53	3.23	1.02	33.15	39.97	12.10
Male partner ($n = 1610$)								
1	749	4	12.47	36.17	36.60	14.76	0.00	0.00
2	325	5	9.41	22.66	20.58	4.29	22.54	20.52
3	444	11	5.30	16.01	18.18	22.92	28.57	9.02
4	92	10	6.87	1.99	1.00	38.76	41.20	10.18

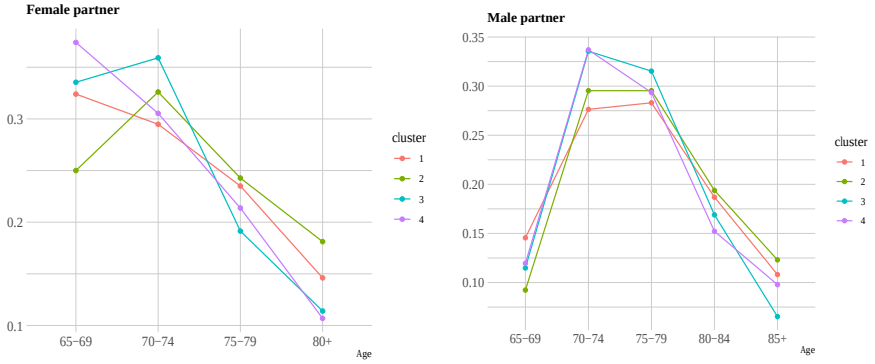


Fig. 2 Cluster profiles for attribute "Age" of ego-network data by gender. The values on the *y*-axis are proportions of variable' categories over each group in the clustering.

4 Conclusions and future work

In this paper the ego-network of contacts of Italian elders living in couple (FSS Survey, 2009) is analysed at individual level. Agglomerative hierarchical clustering is adopted, based on a suitable dissimilarity in role structures, in order to find subgroups of similar networks. The influence of the weights adopted in the computation of the dissimilarity will be subject of further analyses. Preliminary results on the analysis of male and female partners show how the proposed procedure can be used to investigate the existing patterns in ego-networks data, especially in cases where the data are defined over heterogeneous attributes. Further challenges arise from modelling specific features of network data, as in the case of zero-inflated counts. As an extension of the presented work, the latest FSS Survey data collected in 2016 will be analysed. Future work will also explore consensus functions in order to improve the robustness of clustering by combining the output of multiple algorithms.

References

1. Amati V., Meggiolaro, S., Rivellini G., Zaccarin S.: Relational Resources of Individuals Living in Couple: Evidence from an Italian Survey. Soc Indic Res, **134**, 547–590 (2017)
2. Amati V., Rivellini G., Zaccarin S.: Potential and effective support networks of young Italian adults. Soc Indic Res, **122**, 807–831 (2015)
3. Bien, J., Tibshirani, R.: Hierarchical Clustering With Prototypes via Minimax Linkage. J Amer Stat Assoc, **106**, 1075–1084 (2011)
4. Brandes, U., Lerner, J., Nagel, U.: Network ensemble clustering using latent roles. Adv Data Anal Classif, **5**, 81–94 (2011)
5. Breiger, R.: The analysis of social network. In: Hardy M., Bryman, A. (eds.) Handbook of data analysis, pp. 505–526. London: Sage (2004)
6. Budiaji, W., Leisch, F.: Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. Algorithms **12**, p. 177 (2019)
7. Dykstra, P. A., Bühler, C., Fokkema, T., Petrič, G., Platinovšek, R., Kogovšek, T., et al.: Social network indices in the Generations and Gender Survey: An appraisal. Demographic Research, **34**, 995–1036 (2016)
8. Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis, Oxford University Press (2001)
9. Fred, A.L.N., Jain, A. K.: Robust Data Clustering. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 128–136. CVPR (2003).
10. Ganster, D. C., Victor, B. The impact of social support on mental and physical health. British Journal of Medical Psychology, **61**, 17–36 (2011)
11. Gower, J.: A general coefficient of similarity and some of its properties. Biometrics **27**, 857–871 (1971)
12. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, pp. 21–34. Singapore (1997)
13. Taylor, S.E.: Social Support. In: Friedman, H.S., Cohen Silver, R. (eds.) Foundations of Health Psychology, pp. 145–171. Oxford University Press, Oxford (2007)

Analysing the mediating role of a network: a Bayesian latent space approach

Analisi di un network come mediatore: un approccio Bayesiano latent space

Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison

Abstract The use of network analysis for the investigation of social structures has recently seen a rise, due both to the high availability of data and to the numerous insights it can provide into different fields. Most analyses focus on the topological characteristics of networks and the estimation of relationships between the nodes. We adopt a different point of view, by considering the whole network as a random variable conveying the effect of an exposure on a response. This point of view represents a classical mediation setting, where the interest lies in the estimation of the indirect effect, that is, the effect propagated through the mediating variable. We introduce a latent space model mapping the network into a space of smaller dimension by considering the hidden positions of the units in the network. Furthermore, the mediation analysis is extended by using generalised linear models. A Bayesian approach allows to obtain the entire distribution of the indirect effect, generally unknown, and to compute highest density intervals, which give accurate and interpretable bounds for the mediated effect. Finally, an application to social interactions among a group of adolescents and their attitude toward smoking is presented.

Abstract *Di recente si è assistito a un aumento nell'uso della network analysis per analizzare le strutture sociali, grazie sia alla notevole disponibilità di dati sia ai numerosi contributi che essa può apportare in diversi campi. La maggior parte delle analisi si concentra sulle caratteristiche topologiche dei network e sulla stima delle relazioni tra i nodi. In questo lavoro gli autori adottano un punto di vista differente, considerando il network nella sua interezza come variabile casuale che ve-*

Chiara Di Maria
University of Palermo, Viale delle Scienze Building 13, Palermo (PA), Italy, e-mail: chiara.dimaria@unipa.it

Antonino Abbruzzo
University of Palermo, Viale delle Scienze Building 13, Palermo (PA), Italy, e-mail: antonino.abbruzzo@unipa.it

Gianfranco Lovison
University of Palermo, Viale delle Scienze Building 13, Palermo (PA), Italy, e-mail: gianfranco.lovison@unipa.it

icola l'effetto di un'esposizione su una variabile risposta. Questo è il classico caso di mediazione, in cui si è interessati alla stima dell'effetto indiretto, ovvero l'effetto che si propaga attraverso il mediatore. In particolare, si introduce un modello latent space che proietta il network in uno spazio di dimensione inferiore, considerando le posizioni latenti dei soggetti nel network. Inoltre, l'analisi di mediazione viene estesa a modelli lineari generalizzati. Un approccio Bayesiano consente di ottenere l'intera distribuzione dell'effetto indiretto, generalmente ignoto, e di calcolare gli intervalli di credibilità di massima densità, che forniscono valori soglia per l'effetto mediato accurati e interpretabili. Infine, viene presentata un'applicazione relativa alle interazioni sociali tra un gruppo di adolescenti e il loro atteggiamento nei confronti del fumo.

Key words: Network analysis, Bayesian methods, mediation analysis, longitudinal data, latent space model

1 Introduction

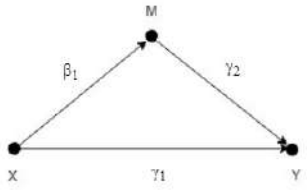
Network data consist of a set of n units and a relation tie m_{ij}, measured on each ordered pair of units $i, j = 1, \ldots, n$. Recently, there has been an explosion of network data in all corners of science [4]. Accordingly, several techniques have been proposed for the task of analysing network data. Graph theory is used to examine the network structure; fast and efficient algorithms are used to detect network communities; statistical models are adopted to understand the formation of connections between units. A review of statistical networks models, algorithms and software can be found in [8].

In this work, we view the network as a random variable M having a role in the mechanism through which an explanatory variable X affects a response Y. The aim is to decompose the total effect of X on Y into a direct and an indirect effect. Mediation analysis is a statistical technique widely used for this purpose [6]. The intermediate variable conveying the indirect effect is called *mediator*. For example, we consider an empirical analysis in which the social network of a sample of adolescents is regarded as a mediator in the relationship between sex and smoking status, and between the amount of pocket money each participant had per month and smoking status. The aim is to understand how social interactions can affect the propensity to smoke.

A major issue is related to the mismatch of dimensions between the network and the other variables. To deal with this evident mismatch, we do not use the network directly, instead we reduce its dimension through the latent space model proposed by [3]. This model projects the network in a space of smaller dimension, where each unit in the network is assumed to have an unknown position. These latent positions are estimated by modelling the presence of a link between two units in the network as dependent on their distance and possibly on additional covariates which may contribute to explaining the relationship. More details are provided in the next section.

2 Model specification: combining mediation analysis and latent space models

In the most straightforward setting, a mediation model includes three variables (see Fig. 1). If the mediator M and the outcome Y are assumed to be Normally distributed and to have linear relationships, the regression equations can be specified by:



$$\mathbb{E}[Y|X] = \tau_0 + \tau_1 X \tag{1}$$

$$\mathbb{E}[M|X] = \beta_0 + \beta_1 X \tag{2}$$

$$\mathbb{E}[Y|X,M] = \gamma_0 + \gamma_1 X + \gamma_2 M. \tag{3}$$

Fig. 1: Mediation model with three variables.

The equation (1) is the marginal model for the outcome; therefore, τ_1 represents the total effect of X on Y. The equation (2) is the mediator model, while in the last equation, the outcome model conditional on both the mediator and the exposure is specified. In the general associational framework, the indirect effect can be computed via the product method by multiplying β_1 and γ_2, that is, the coefficients corresponding to the arrows lying on the path connecting X to Y through the mediator M [1, 6]. In our analysis, the mediator is a network conveniently represented in a more parsimonious way through the latent space model, as described in [3]. This approach assumes the existence of a latent space where each unit in the network has a concealed position and relative distances predict the formation of a tie among units. Formally, a network with n nodes can be represented by an $n \times n$ matrix M, where each entry m_{ij} denotes a relationship between the units i and j. We focus on boolean relationships and, as a consequence, on adjacency matrices, but other kinds of relationships can be modelled as well. The probability that a link between i and j exists, denoted by p_{ij}, is assumed independent of all other ties in the network conditionally on the latent positions $\mathbf{z}$, and is modeled via a logistic regression

$$\text{logit}(p_{ij}|\mathbf{z}_i, \mathbf{z}_j, \alpha) = \alpha - \|\mathbf{z}_i - \mathbf{z}_j\|, \tag{4}$$

where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD})^\intercal$ and $\mathbf{z}_j = (z_{j1}, \ldots, z_{jD})^\intercal$ are the latent positions of units i and j and $\|\cdot\|$ is the Euclidean distance. The main advantage of this approach is that most networks can be represented in a space of dimension $D \ll n$. Moreover, the coordinates are orthogonal. Having established the setting, similar to [5], the idea is to use the components of the D-dimensional vector representing the position of each unit in the network as mediators in the relationship between an exposure and a response in order to estimate the indirect effect. If $D = 1$ we have a single mediator model, as in Figure 1. If $D > 1$ we have a multiple mediator model, as shown in Figure 2. The coordinates are mutually independent: joint mediation can then be addressed by performing separate mediation analyses considering one coordinate at a time and then combining the estimates of the indirect effects. Note that the

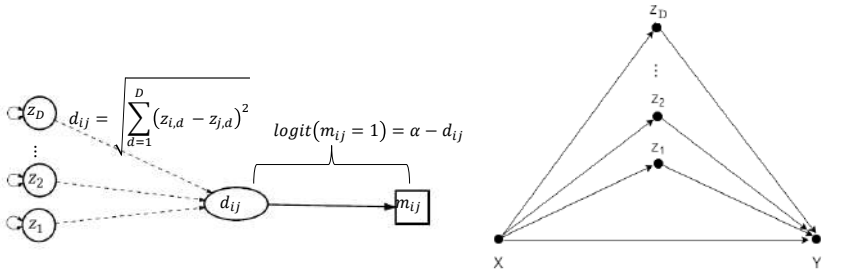


Fig. 2: Graphical representation of coordinates of a unit as mediators. The mediator for subject i is summarized by the latent position $z_{id}, d = 1, \ldots, D$ (see [5]).

assumption of linearity linking the mediator to the exposure, and the outcome to the exposure and the mediator is crucial since it ensures that the indirect effect can be obtained as a product of coefficients. For this reason, [5] used linear models for both the mediator and the outcome. We extend the model proposed in [5] by allowing the outcome to belong to the exponential family distribution, and to depend non linearly on its predictors. These extensions build on the works of [9] and [2]. [9] noticed that indirect effects could be seen as products of derivatives, the one with respect to the exposure in the mediator model and the one with respect to the mediator in the outcome model. For example, differentiating Equations (2) and (3) with respect to X and M, respectively, yields β_1 and γ_2, whose product is exactly the indirect effect in the classical Normal linear case. Formulas for the indirect effect when at least one between the mediator and the outcome model is nonlinear are more complex and depend on the values of X and (or) M: this is the reason why [2] suggest to call these effects *conditional*. The authors also point out that interpreting these effects as the increment in Y due to a unitary increment in X, mediated by M, is incorrect and that they should rather be commented in terms of increments of standard deviation, i.e. taking their magnitude into account.

Making inference on the indirect effect is not straightforward even in the linear case, since, even assuming that the two regression coefficients estimators are Normally distributed, generally their product is not. The distribution of the product may be highly skewed and, most importantly, unknown, and this complicates the estimation of confidence intervals for the mediated effect [7]. So, in this paper, inference is carried out within a Bayesian framework via Monte Carlo Markov Chain (MCMC) approach. The details of the inferential approach are omitted for the sake of brevity.

3 Data analysis

The data used in the analysis are a subsample of the 160 adolescents enrolled in the *Teenage Friends and Lifestyle Study*, a cohort study carried out in a secondary school of Glasgow between 1995 and 1997 intending to investigate how smoking behaviours change over time and the extent to which social interactions influence

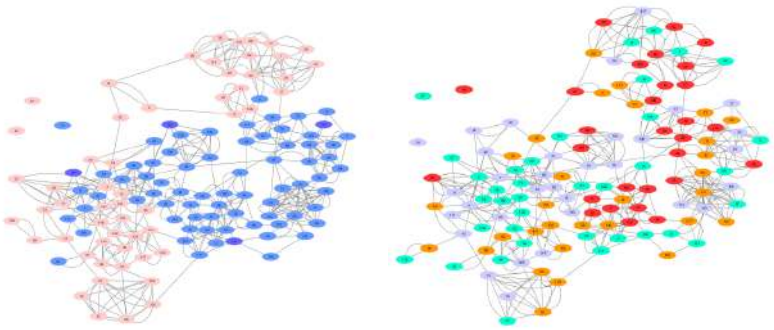


Fig. 3: Graphical representation of networks with nodes coloured according *Sex* and categories of *Money*: on the left, ■ = male and ■ = female, on the right ■ = 1 (0-7£), ■ = 2 (7-12£), ■ = 3 (12-20£), ■ = 4 (20-70£).

them. Except for participants' sex and age, which were recorded at baseline, other variables were collected at three different time occasions. Information on substance use (tobacco, alcohol, cannabis), leisure time activities, music taste, romantic relationships and several others are available. Moreover, social matrices representing friendship relationships among participants are included in the dataset. We conduct a cross-sectional analysis, choosing just variables relative to the last measurement occasion. Our interest lies on the way two exposures (*Sex*, participant's gender, and *Money*, the amount of pocket money each participant had per month) affect substance use through the relationship network. *Sex* is a binary variable, 0 denotes male and 1 female, *Money* is a continuous variable ranging from 0 to 70 pounds. We use the variable as it is and a categorical version obtained through its quarterlies so that we have four categories.

First, we fitted model in Equation (4) for different values of the latent space dimension D, and we compared the adjacency matrix corresponding to the actual network to the one obtained from the relative distances between the estimated latent positions in terms of F1-score. We obtained the highest F1-score for $D = 5$, and we noticed that increasing the dimension did not lead to a substantial improvement. Then, we used the estimated positions as mediators and estimated conditional indirect effects for each subject. We fit nine models, obtained combining the three exposure mentioned above with the three outcome variables *Smoke, Cannabis*, and *Alcohol*, binary variables where 0 indicates rare or no use of the correspondent substance and 1 moderate or high use.

We compute the direct effect of exposures on responses and individual conditional indirect effects and relative highest density intervals. The direct effect turns out to be significant in almost all analyses: in particular, *Sex* has a negative direct effect on *Smoke* and *Cannabis*, so the odds to make ample use of tobacco and cannabis is smaller for girls. In contrast, *Sex* has a positive direct effect on alcohol. Thus girls

seem to drink more than boys. The same pattern holds for *Money* in its categorical version. Indirect effects are non-significant for each subject and in each analysis.

4 Conclusions

We have addressed the issue of estimating the indirect effect in a mediational setting where the mediator is a network. We have used a latent space approach to reduce the dimensionality of the network, and the concept of conditional indirect effect to extend the estimation of the indirect effect on non-linear models. A Bayesian MCMC provides inferential results on the indirect effect. Finally, a dataset on adolescents shows that gender and the amount of pocket money available per month have a negative direct effect on smoking attitudes and cannabis use and a positive effect on alcohol consumption. Friendship relationships seem not to have any mediating role since indirect effects are non-significant in each analysis.

There are several directions this work can be developed on. It is necessary to carry out more research on how to extend latent space and mediation models to GLMs. Moreover, further extensions to the causal framework and longitudinal data to capture the dynamics of change would be promising.

References

[1] Baron, R. M. and Kenny, D. A.: The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations. J. Personal. Soc. Psychol. **51**(6), 1173–1182 (1986)

[2] Geldhof, G. J., Anthony, K. P., Selig, J. P., Mendez-Luck, C. A.: Accommodating binary and count variables in mediation: A case for conditional indirect effects. Int. J. Behav. Develop. **42**(2), 300–308 (2018)

[3] Hoff, P. D., Raftery, A. E., Handcock, M. S. : Latent Space Approaches to Social Network Analysis. J. Am. Stat. Assoc. **97**(460), 1090–1098 (2002)

[4] Kolaczyk, E. D. and Csárdi, G.: Statistical analysis of network data with R. Springer (2014)

[5] Liu, H., Jin, I.H., Zhang, Z., & Yuan, Y.: Social Network Mediation Analysis: a Latent Space Approach. arXiv:1810.03751 [stat.ME] (2018)

[6] MacKinnon, D. P.: Introduction to Statistical Mediation Analysis. New York: Taylor and Francis Group (2008)

[7] MacKinnon, D. P., Lockwood, C. M., Williams, J.:Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. Multivar. Behav. Res. **39**(1) (2004)

[8] Salter-Townshend, M. and White, A. and Gollini, I. and Murphy, T. B.: Review of statistical network analysis: Models, algorithms, and software. Statistical Analysis and Data Mining **5**, 243—264 (2012)

[9] Stolzenberg, R. M.: The measurement and decomposition of causal effects in nonlinear and nonadditive models. Sociol. Methodol. **11**, 459—488 (1980)

Network-time autoregressive models for valued network panel data

Modelli autoregressivi per dati panel di network pesati

Viviana Amati

Abstract We propose and discuss the application of network-time autoregressive models to analyse the evolution of valued networks over time. Compared with more traditional network models explicitly modelling the network dependence structures, those models incorporate tie dependence using covariates and correlated random terms. The models are an extension of the classical space-time autoregressive models. We illustrate the applicability of the models by using data on networks of international migration flows.

Abstract Proponiamo e discutiamo l'uso di modelli autoregressivi di network per analizzare i cambiamenti nel tempo di network pesati. A differenza di tradizionali modelli che descrivono esplicitamente la struttura di dipendenza tra i legami, i modelli che proponiamo incorporano questa struttura mediante l'uso di covariate. I modelli autoregressivi di network estendono l'applicazione dei classici modelli autoregressivi spazio-temporali alla network analysis. L'utilizzo e l'interpretazione dei modelli proposti è illustrato con l'analisi di reti di flussi migratori internazionali.

Key words: Space-time autoregressive models, Panel data, Valued networks

1 Introduction

A network is a set of entities and the relations among them. In some cases, binary information, indicating the presence or absence of a relationship among pairs of entities is collected. In other cases, information on the intensity of the relationships is also recorded. For instance, friendship can be ranked according to its strength, trade can be measured by the volume of materials that are exchanged, interactions can be timed, and migration flows can be quantified by the number of people moving. A

Viviana Amati
Chair of Social Networks, Department of Humanities, Social and Political Sciences, ETH Zurich, Weinbergstrasse 109, 8092, Zurich, Switzerland e-mail: viviana.amati@gess.ethz.ch

network in which the relationships are characterized by intensity is referred to as a valued network.

Due to the complex dependence structure between the ties, many of the models applied to make inference on an observed network have to date been limited to binary networks. Thus, the modelling of valued networks has often required a dichotomization of ties. Ties with a value greater than a specified threshold are assumed to be present and ties with a value lower than that threshold to be absent. The dichotomization leads to a loss of information and introduces biases.

While many of the descriptive statistics have been easily extended to valued networks, there is a paucity of models that have been developed to model valued ties. Examples of these models are multiple regression quadratic assignment procedures [3, 2], exponential random graph models [4] for count data and stochastic actor-oriented models [9] for ordinal data.

In this contribution, we extend space-time autoregressive models to valued networks. The spatial component accounts for tie dependence and the covariates incorporate the network structure. Hereafter, we refer to the proposed model as the network-time autoregressive model. We discuss and illustrate the applicability of the model by analysing the evolution of the network of international migration flows.

2 Network time autoregressive models

2.1 Notation

Let $\mathscr{N} = \{1, \ldots, n\}$ be a set of entities over which a relation $\mathscr{R} : \mathscr{N} \times \mathscr{N} \to \mathbb{R}$ is defined. This relation can be regarded as a digraph on the set $\mathscr{N}$, and will be called the network. We represent the network as an adjacency matrix $Y_{n\times n}$, whose cell Y_{ij} takes the value $\mathscr{R}(i,j) \in \mathbb{R}$. The relation is assumed to be non-reflexive, implying that $Y_{ii} = 0$ and directed, i.e. $y_{ij} = c$ does not imply that $y_{ji} = c$, with $c \in \mathbb{R}$.

We assume that the network Y is subject to change over time. The dependence on time t is denoted by Y_t. Networks are observed at $M > 1$ distinct time points $t_1, \ldots, t_M$ in a panel design. Observations are denoted by y_{t_m}, $m = 1, \ldots, M$.

Entity- and dyadic-level variables might also be collected besides network variables. Entity-level variables describe constant or time-dependent characteristics of the entities and are denoted by the array $v = (v_1, \ldots, v_M)$ where $v_m \in \mathbb{R}^{n\times p}$ is the standard data table at time t_m. Dyadic-level variables describe constant or time-dependent characteristics of dyads and therefore are recorded for each pair of actors. Dyadic variables are represented by the sequence of arrays $z = (z_1, \ldots, z_M)$ where $z_m \in \mathbb{R}^{n\times n\times q}$ is an $n \times n \times q$ array whose element z_{ijq} is the value of the $q-th$ dyadic covariate over the pair (i, j) at time t_m.

The time series of networks $Y = \{Y_{t_1}, \ldots, Y_{t_M}\}$ together with monadic and dyadic covariates v and w constitute the valued network panel data.

2.2 Model formulation

Space-time autoregressive (STAR) models [7, 5] apply to the study of the dynamics of a variable recorded at the same locations and distinct time points. Variations of the variable over time and space are explained by both spatial and temporal correlation, as well as characteristics of the locations. Since valued networks panel data can be thought of as the collection of relational intensities over ties and time, we can extend the framework of the STAR model by replacing the spatial component with the network component to capture tie dependence. Thus, the network-time autoregressive (NTAR) models belong to the class of models that incorporate the network structure through covariates [8]. Compared with previous models belonging to this class, the NTAR model relaxes the standard assumption of independent residuals.

The vectorial form of the NTAR model is defined by the equation:

$$Y_t = v'\beta + z \odot \gamma + \sum_k \delta_k W_k f_k(Y) + \varepsilon_t \quad , \tag{1}$$

with β, γ and δ statistical parameters that need to be estimated from the data, the symbols $'$ (prime) the transpose of a matrix and $\odot$ the Hadamard product.

The first two terms on the right side of Eq. (1) capture the dependence of the value of Y_t on an entity's characteristics v and dyadic attributes v, respectively. For instance, entities with certain characteristics might be more prone to send or receive ties with high values. The third term models time and network dependencies through the row-normalized matrices of weights W_k whose generic element w_k represents tie dependencies at different times and network lags as defined by the function $f_k : Y \to \mathbb{R}$. The subscript k indexes the diverse forms of dependence that are in the data as described by the following examples. The last term ε_t is the error term accounting for the network correlation between the ties at time t. We assume that $\varepsilon_t = \rho W \varepsilon_t$ with ρ the network correlation coefficient at time t and ε a random variable normally distributed with mean 0 and variance-covariance matrix ρW.

We describe a few NTAR models to clarify the role and the structure of the matrix of weights W_k and the function f_k. We consider that the time lag is equal to 1 without loss of generality.

When tie independence is assumed, i.e. $W_k = I_k$ is an identity matrix, the NTAR model reduces to a regular autoregressive model in which the tie value Y_{ijt} is explained only by its previous value $Y_{ij(t-1)}$ and by the entity and dyadic covariates.

When dyadic dependence is assumed, ties within the same dyad are considered to be dependent. This implies that the value of Y_{ijt} depends on the values of $Y_{ij(t-1)}$, $Y_{ji(t-1)}$ and Y_{jit} due to the time lag 1. The dependence is mathematically expressed by using two weighted matrices. The identity matrix $W_1 = I$ captures the dependence of Y_{ijt} on $Y_{ij(t-1)}$. The matrix W_2 is a block diagonal matrix describing the dependence of $Y_{ij(t-1)}$ on $Y_{ji(t-1)}$ when the matrix is used jointly with $f_k(Y_{ij(t-1)}, Y_{ji(t-1)})$. For instance, if $\mathscr{N} = \{1,2,3\}$, we obtain the matrix

$$W_2 = \begin{array}{c} \\ (1,2) \\ (2,1) \\ \\ (1,3) \\ (3,1) \\ \\ (2,3) \\ (3,2) \end{array} \begin{array}{c} \begin{array}{cccccc} (1,2) & (2,1) & | & (1,3) & (3,1) & \quad (2,3) \quad (3,2) \end{array} \\ \left[\begin{array}{cccccc} 0 & 1 & | & 0 & 0 & 0 \quad 0 \\ 1 & 0 & | & 0 & 0 & 0 \quad 0 \\ - & - & - & - & - & - \\ 0 & 0 & | & 0 & 1 & | \; 0 \quad 0 \\ 0 & 0 & | & 1 & 0 & | \; 0 \quad 0 \\ 0 & 0 & - & - & - & - \quad - \\ 0 & 0 & & 0 & 0 & | \; 0 \quad 1 \\ 0 & 0 & & 0 & 0 & | \; 1 \quad 0 \end{array}\right] \end{array}$$

The matrix W_2 can be used, for instance, to model the effect of past reciprocity at time $t-1$ on Y_{ijt}. One way to express reciprocity in valued networks is to use the difference between tie values, i.e. $f_k(Y_{ij(t-1)}, Y_{ji(t-1)}) = Y_{ij(t-1)} - Y_{ji(t-1)}$. The closer this difference to 0 is, the higher the level of reciprocity. Alternatives are averages or the maximum and the minimum between $Y_{ij(t-1)}$ and $Y_{ji(t-1)}$. The choice depends on the relation that defines the network.

According to this specification, the NTAR model takes the form

$$Y_t = v'\beta + z \odot \gamma + \delta_1 W_1 Y_{ij(t-1)} + \delta_2 W_2 (Y_{t-1} - Y'_{t-1}) + \rho W_2 \varepsilon_t \quad . \tag{2}$$

The inclusion of W_2 in the error term accounts for the dependence of Y_{ij} on Y_{ji} at time t.

Similarly, we can extend the model to account for the Markov dependence assumption. Markov-dependence implies that two ties are dependent if they share a vertex. A third matrix W_3 can be added to the terms on the right side of Eq. (3), e.g., to model the effect of clustering of ties with particular values. This effect is the valued counterpart of the transitivity effect that is often observed and included in standard network models. As for reciprocity, the challenge of the model specification is to define the mechanisms that we think foster the changes in the tie values and the corresponding functions $f_k(Y)$.

Once the model is specified, several methods can be implemented to estimate the parameters of the NTAR model, among them the general least squares estimation, as in regular STAR model [7], or generalised estimating equation [6] abbreviated as GEE.

The model defined in Eq. (1) can be extended in several ways. The introduction of random effects and time-dependent parameters allows relaxing the assumptions of constant parameters over time and including time lags with order higher than 1, respectively. This modification should be implemented with caution since the increased number of parameters might lead to over-parametrized models.

The NTAR model can also be generalized to the evolution of valued networks where the values are determined by count data as the STAR models have been defined also for discrete variables. Like in standard generalized linear models, the valued tie variables, are related to the linear predictor using a link function and a different distributional assumption is assumed for the random error ε. In the next paragraph, we illustrate the use of NTAR model for modelling network of migration flows, where the values of ties are defined as a count.

3 Application

We illustrate the applicability of the NTAR models by analysing the network of migration flows between world countries. The United Nations website provides data on the international migrant stocks by country of origin and destination. A migrant is a person who is not born in the respective country of residence. This data allows estimating the network of international migration flows according to the demographic accounting pseudo-Bayesian approach described in [1]. Data are available for the years 1990, 1995, 2000, 2005, 2010 and 2015. Due to the illustrative purpose of this section, we consider only the last two observations. Information on religion and human development index (HDI) for countries were also collected from UNDP and the CIA World Fact Book websites.

To model the evolution of the international migration network between 2010 and 2015, we consider the following log-linear NTAR model [6] based on time lag 1 and dyadic dependence:

$$\log(E[Y_t]) = \beta_1 v_i + \beta_2 v_j + z \odot \gamma + \delta_1 W_1 Y_{ij(t-1)} + \delta_2 W_2 (Y_{t-1} - Y'_{t-1}) + \rho W_2 \varepsilon_t \quad , \quad (3)$$

with v_i and v_j representing the value of the human development index of the sender and the receiver countries of the tie, z the matrix containing the Jaccard index measuring similarity in religion, W_1 and W_2 as described in Sec. 2.2 and ε assumed to follow a Negative Binomial distribution.

Table 1 reports the results obtained using the GEE method. The parameters of the HDI sender and receiver indicate that the migration flow from low HDI to high HDI countries is larger than the flow from high HDI to low HDI countries. The similarity in religion is not significant, indicating that similar religious beliefs are not determinants of the migration flows. The time lag 1 parameter δ_1 is significant and positive, suggesting that flows between countries tend to consolidate over time. The coefficient of reciprocity, measured as differences of tie values in opposite directions, is also significant and positive. Thus, flows from country i to country j increase as the difference in the flows from i to j and from j to i increases. This indicates that flows are not reciprocated in values at time $t-1$ and this asymmetry foster the flows towards the direction of the largest flow. The parameter ρ is negative underlying that there is a negative network dependence in reciprocity even at time t. The results suggest that the network flows are directed to countries with more favourable living conditions as already suggested by the application of simpler models (e.g. gravity model).

The example is only used to illustrate the model specification and interpretation. More complex specifications, including, for instance, clustering effects, provide more information concerning the mechanisms that explain the evolution of the network of international migration compared with traditional models. However, the definition and inclusion of additional effects require particular caution in NTAR models since different functions $f(Y)$ might represent multiple mechanisms and the number of explanatory variables and parameters can quickly increase.

Table 1 Network time autoregressive model for the migration flow networks.

	Est.	s.e.
HDI sender	-1.61*	0.45
HDI receiver	1.32*	0.35
Religion similarity	0.03	0.02
Time lag 1	0.89*	0.23
Reciprocity	0.63*	0.12
ρ	-0.45	

* Signif. < 0.05

4 Conclusion

In this paper, we proposed network time autoregressive models to analyse the evolution of valued networks over time. Compared with more traditional network models explicitly modelling the network dependence structures, those models incorporate tie dependence using covariates and correlated random terms. We illustrated the applicability and the parameter interpretation by using data on networks of migration flows.

The NTAR models we presented are quite simple in their structure. More complex time and network dependence assumptions can be implemented in the model, e.g., by using random effects or relaxing homogeneity assumptions. This increase in complexity results in a greater number of parameters which might make the estimation of the model problematic. How to address model complexity, by introducing parameter constraints or using dynamic structural equation models for multiple time-series data, is the object of future work.

References

1. Abel, G. J., Cohen, J. E.: Bilateral international migration flow estimates for 200 countries. Scientific data, 6(1), 1–13 (2019).
2. Dekker, D., Krackhardt, D., Snijders, T. A. B.: Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psych., **72**, 563–581 (2007).
3. Krackardt, D.: QAP partialling as a test of spuriousness. Soc. Net. **9**, 171–186 (1987)
4. Krivitsky, P. N.: Exponential-family random graph models for valued networks. Electronic journal of statistics, **6**, 1100 (2012).
5. Nainggolan, N., Titaley, J.: Development of generalized space time autoregressive (GSTAR) model. In AIP Conference Proceedings **1827**. AIP Publishing LLC (2017).
6. Melo, O. O., Mateu, J., Melo, C. E.: A generalised linear space–time autoregressive model with space–time autoregressive disturbances. J. Appl. Stat., **43**, 1198–1225. AIP Publishing LLC (2016).
7. Ruchjana, B. N., Borovkova, S. A., Lopuhaa, H. P.: Least squares estimation of Generalized Space Time AutoRegressive (GSTAR) model and its properties. In AIP Conference Proceedings, 61–64 (2012).
8. Snijders, T. A.: Statistical models for social networks. Ann. Revi. Soc., **37** (2011)
9. Snijders, T. A., Van de Bunt, G. G., Steglich, C. E.: Introduction to stochastic actor-based models for network dynamics. Soc. Netw., **32**, 44–60.

University student mobility flows and related network data structures

La mobilità studentesca universitaria e la definizione di strutture di dati di rete

Vitale Maria Prosperina, Giuseppe Giordano, Giancarlo Ragozini

Abstract The Italian student mobility in higher education represents a relevant issue for the academic system, especially for the regions located in Southern Italy. The aim of the contribution is to provide a first exploration of the structural patterns of student mobility flows among Italian geographical areas, focusing on South-North student migration routes and discovering the main destinations in the higher education migration phenomenon. A network perspective is considered by exploiting multimode network data structure. Data on cohorts of students enrolled at the Italian universities are considered.

Abstract *La mobilità studentesca rappresenta un tema di interesse per il sistema accademico soprattutto per le regioni del Sud Italia. L'obiettivo del presente contributo è fornire una prima esplorazione della struttura dei flussi degli spostamenti di studenti durante la carriera universitaria tra le regioni italiane, focalizzandosi sulla traiettoria Sud-Nord di tale catena migratoria. La prospettiva metodologica dell'Analisi delle Reti Sociali è utilizzata per individuare strutture di dati di reti multimodali. I dati su coorti di studenti iscritti ai vari corsi di laurea istituiti nelle università italiane sono oggetto di studio.*

Key words: Student mobility, Higher education, Multimode networks

Maria Prosperina Vitale
Dept. of Political and Social Studies, University of Salerno, Via Giovanni Paolo II, Fisciano (Salerno), Italy, e-mail: mvitale@unisa.it

Giuseppe Giordano
Dept. of Political and Social Studies, University of Salerno, Via Giovanni Paolo II, Fisciano (Salerno), Italy, e-mail: ggiordano@unisa.it

Giancarlo Ragozini
Dept. of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22, Naples, Italy, e-mail: giragoz@unina.it

1 Introduction

The Italian student mobility in higher education represents a relevant issue for the academic system [1, 14], especially for the regions located in Southern Italy. Such mobility generates massive student flows from the South to the North of the country [8, 13]. The choice to move from the place of residence is a complex decision that people often make to improve their quality of life, and economic drivers are only one of the motivation for people to migrate [13]. Several factors have been proposed in the literature as determinants of student mobility taking into account differences in socio-economic conditions of the origin and destination areas [10], cultural, institutional and organizational aspects. Students could move from their province of residence to find degrees in specific topics not available nearby, to enroll at Universities perceived as higher quality for research and teaching [6], and to join regions with better quality of life and/or with better job opportunities after graduation.

The university student mobility flows follow the traditional *South-to-North* migration chain. Indeed, considering the data provided by the National Student Archive (NSA) and related to the student enrollment at the first year of a Bachelor degree in the academic year (a.y.) 2018-2019, around 22% of students living in a region in Southern Italy and 24% of freshmen resident in the two main islands are enrolled in universities located in the Center and the North of Italy. Indeed, there are strong regional inequalities. For example, freshmen resident in the Basilicata region are enrolled in a university outside the region in 78% of the cases, among them 44% of students move to Center or North of Italy. On the contrary, students resident in the Campania region mainly stay in their own region and only around 11% of cases moves abroad the regional boundaries, showing an increasing power of the regional university system in retaining students. Considering one of the main Italian islands, Sicily shows the highest intellectual migration rate with 34% of students that enroll in universities outside the region. This phenomenon could yield some negative effects in terms of imbalance in government funding to universities, relevant brain drain on regional human capital accumulation [6], and demographic imbalance. Hence, the analysis of the factors explaining this phenomenon could help policy makers in promoting actions to counterbalance these negative effects of student migration in higher education.

Motivated by this framework, we aim at exploring the structural patterns of student mobility flows among Italian geographical areas, focusing on South-North student mobility route and discovering the destinations in the higher education migration phenomenon. A network perspective is considered in which regions as well as provinces or universities could be described as the set of nodes and student exchanges between units represent the set of links between them. The data on cohorts of students enrolled at the Italian universities are provided by the NSA archive. These data give rise to the definition of complex network data structures analyzed by considering community detection techniques in order to discover the presence of homogeneous territorial sub-areas in the student mobility phenomenon.

The contribution is organized as follows. Section 2 briefly describes the data on student mobility, while Section 3 introduces the methodological approach for ex-

ploring university student mobility data structure. Section 4 reports the first findings on the students' cohort under analysis.

2 Student Mobility Data

The NSA archive is considered to gather data on students enrolled in the Italian university system. Through this portal, the Ministry of University and Research offers the opportunity to consult, in an aggregate form, the information on student enrollments and careers of all registered universities. Here, data have been directly downloaded with information at individual level in anonymized form.[1]

Traditional data matrices could be defined by considering students' attributes (e.g., *sex*, *age*, *province of residence*, *type of secondary school*, *college grade*, *bachelor and master degree enrollment at university*, *type of bachelor and master degree*). These data can be analyzed as contingency tables in a classical statistical framework. Co-occurrence data might give rise to the definition of network data structures [4] in which frequencies weight the links connecting different sets of nodes, such as provinces of residence (province), universities of enrollment (university), and types of educational programs (educational program). Indeed, we can interpret frequencies as the number of student exchanges among geographical territories. The following network structures could be derived: *i) weighted one-mode networks* in which, for example, provinces are the nodes, and students' flows represent the edges' weights; *ii) weighted two-mode networks* in which the *province*×*university* or the *province*×*educational program* represent the two set of nodes, and students' flows represent the edges' weights; *iii) multimode* networks with different (three or more) modes (e.g., *province*×*university*×*educational program*; *iv) multiplex* networks with *province*×*university*, and the types of educational programs representing layers; *v) multilevel* networks with *province/university*×*educational program*, and regions representing nested level.

3 Multimode networks

In the following, the focus is on multimode networks to analyze complex network data structures derived from student mobility flows. Few papers are devoted to the study of multimode networks. A generalization of the conceptual basis and the matrix formalism of bipartite to tripartite graphs is described in the seminal paper of Fararo and Doreian [9], in which three types of nodes are defined and ties are present only between nodes of distinct types. The logic of this structure can be extended to

[1] The data at student level are available by the NSA archive for the universities involved in the project PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide". Principal Investigator Massimo Attanasio, n. 2017HBTK5P, CUP: B78D19000180001.

any number of modes in a network. An extension on the use of regular and structural equivalence to multimode data is proposed in Borgatti and Everett [3]. The combination of the logic of multimode networks with Newman's spectral partitioning of graphs into communities is discussed in Melamed et al. [12]. Recently, Everett and Borgatti [5] suggest to examine the collection of all two-mode networks in case of multimode data.

Formally, a multimode network $\mathscr{M}$ can be conceived as consisting of a pair $(\mathscr{V},\mathscr{E})$, being $\mathscr{V} = \{V_i\}_{i=1,\ldots,K}$ the collection of K set of nodes, one for each mode i, $V_i \cap V_j = \emptyset$, $\forall i \neq j$, and being $\mathscr{E} = \{E_{ij}\}_{(i,j=1,\ldots,K)}$, $E_{ij} \subseteq V_i \times V_j$, $\forall i \neq j$ $E_{ii} = \emptyset$, $\forall i$, the collection of edges existing among the nodes belonging to the i-th and j-th mode. As said, the multimode network can be seen as the collection of all possible two-mode networks $\mathscr{G}_{ij} = (V_i, V_j, E_{ij})$ among the K set of nodes. Following the original idea of tripartite graphs [9], given the multimode network $\mathscr{M}$, we define a unique adjacency matrix $\mathbb{A}$ given by the combination in a block matrix of the sociomatrix $\mathbf{A}_{ij}$ corresponding to the two-mode networks $\mathscr{G}_{ij}$.

In the present contribution, as first step, a three-mode network is considered in which V_1 is the set of province of residence of Italian students enrolled at the first year of any Italian Bachelor degree in a given a.y., V_2 the set of public and private Italian universities (excluding the distance learning universities), and V_3 the set of types of educational programs as coded by the Italian law. In this network data structure, the links are the number of students enrolled, and the corresponding two-mode networks are weighted. Thus, the related adjacency matrix is defined as follow:

$$\mathbb{A} = \begin{bmatrix} \mathbf{0} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{12}^T & \mathbf{0} & \mathbf{A}_{23} \\ \mathbf{A}_{13}^T & \mathbf{A}_{23}^T & \mathbf{0} \end{bmatrix}.$$

Note that the matrix $\mathbb{A}$ looks like the Burt matrix[11] (apart from the diagonal elements) used in the Multiple Correspondence Analysis. On the matrix A, we carry out the usual methods for weighted one-mode networks.

4 First results

The results on 223,908 students enrolled in a Bachelor degree of the Italian universities at the a.y. 2011-2012 are reported following their careers until the a.y. 2017-2018. 85.4% of the students were enrolled at the second year in the a.y. 2012-2013. 72.3% of the students confirmed their university choice, while the others moved in a different university and/or changed the educational program. About half of the students obtained the graduation in three years. This percentage increases to 60% considering the graduation after one year with respect to the regular duration of a study program.

To discover the presence of homogeneous territorial sub-areas in the student mobility phenomenon, we propose to use community detection algorithms [2] on two-

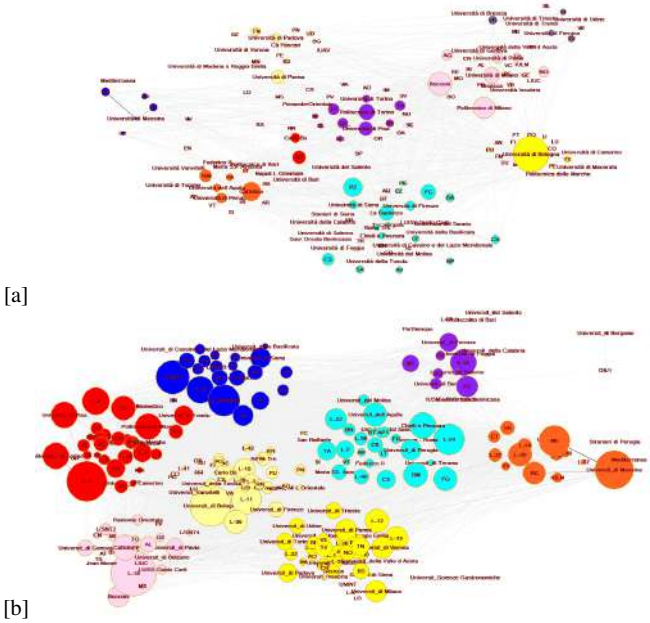
mode and three-mode networks: in the first step, we consider the two-mode network of *province*×*university*; in the second step, we analyze the full three-mode network ($\mathbb{A}$) *province*×*university*×*educational program*. In both cases, the edges with weight lower than 10 (i.e., pairs of entities sharing lower than a fixed threshold value) are removed. The interregional routes of student mobility, disregarding exchanges among the universities located within the same region, are taken into account.

Figures 1a and 1b report the visualization of two-mode and three-mode student mobility networks, respectively. In Figure 1a, on the top right, the group of the more attractive universities located in the North of Italy appears with the University of Bologna, the Polytechnic of Milan, and the Bocconi University emerging over the others; in the center of the network, the University of Turin and the University of Pisa attract students from Sicily and Sardinia; while the larger group, on the bottom, highlights the student mobility among Southern and Central provinces towards universities mainly located in the Center of Italy. On the left side, a small community emerges given by the link between Messina and Reggio Calabria exchanging students over the strait. In Figure 1b the information about the *type of educational program* is also considered. We note that the network structure of universities and provinces is rearranged in groups around specific topics. For example, in the bottom, business programs attract toward the Bocconi and the Luiss Guido Carli universities; on the top left, Engineering programs are linked to the Polytechnics and to the University of Pisa. Finally, in the center, Political Sciences and Foreign Languages programs are clustered with the University of Bologna and the University of Florence.

These first findings show that the network approach on multimode data structures provides interesting insights for the phenomenon under analysis. Besides the well-known South-to-North route, other interregional systems appear. However, we believe that specific tools for such complex network structures should be designed combining network analysis and multidimensional factorial techniques.

References

1. Attanasio, M., Enea, M. La mobilità degli studenti universitari nell'ultimo decennio in Italia. In: De Santis, G., Pirani, E., Porcu, M. (a cura di) Rapporto sulla Popolazione. L'istruzione in Italia, pp. 43–58. Il Mulino, Milano (2019)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. **10**, P10008 (2008)
3. Borgatti, S.P., Everett, M.G.: Regular blockmodels of multiway, multimode matrices. Soc. Networks **14**, 91–120 (1992)
4. Borgatti, S.P., Halgin, D.S.: Analyzing affiliation networks. In: Scott, J., Carrington, P.J. (eds.) The SAGE Handbook of Social Network Analysis, pp. 417–433. SAGE publications, London (2011)
5. Everett, M.G., Borgatti, S.P.: Partitioning multimode networks. In: Doreian, P., Batagelj, V., Ferligoj, A. (eds.) Advances in network clustering and blockmodeling, pp. 251-265. John Wiley & Sons, Chichester (2020)



[a]

[b]

Fig. 1 Visualization of: a) two-mode networks *province*×*university*; b) Representation of three-mode network *province*×*university*×*educational program*. Node color reflects the groups derived by community detection algorithm; node size is proportional to the betweenness centrality scores.

6. Ciriaci, D.: Does University Quality Influence the Interregional Mobility of Students and Graduates? The Case of Italy. Reg. Stud. **48**, 1592–1608 (2014)
7. Dotti, N.F., Fratesi, U., Lenzi, C., Percoco, M.: Local Labour Markets and the Interregional Mobility of Italian University Students. Spat. Econ. Anal. **8**, 443–68 (2013)
8. Enea, M.: From South to North? Mobility of Southern Italian Students at the Transition from the First to the Second Level University Degree. In: Perna C., Pratesi M., Ruiz-Gazen A. (eds) Studies in Theoretical and Applied Statistics. SIS 2016. Springer Proceedings in Mathematics & Statistics, vol 227. Springer, Cham (2018)
9. Fararo, T.J., Doreian, P.: Tripartite structural analysis: Generalizing the Breiger-Wilson formalism. Soc. Networks **6**, 141–175 (1984)
10. Giambona, F., Porcu, M., Sulis, I.: Students Mobility: Assessing the Determinants of Attractiveness Across Competing Territorial Areas. Soc. Indic. Res. **133**, 1105–1132 (2017)
11. Greenacre, M., Blasius, J. (Eds.): Multiple correspondence analysis and related methods. CRC press, Boca Raton (2006)
12. Melamed, D., Breiger, R.L., West, A.J.: Community structure in multi-mode networks: Applying an eigenspectrum approach. Connections. **33**, 25–30 (2013)
13. Santelli, F., Scolorato, C., Ragozini G.: On the determinants of student mobility in an interregional perspective: a focus on Campania region. Italian Journal of Applied Statistics , **31**, 119–142 (2019)
14. Sulis, I., Porcu M., Giambona F. (eds.): Student mobility, university and post-university choices. Electronic Journal of Applied Statistical Analysis **12**, 1–3 (2019)

Statistical Methods in Psychometrics

A simple probabilistic model to evaluate questionable interim analysis strategies

Un nuovo modello probabilistico per valutare strategie di analisi intermedie

Francesca Freuli and Luigi Lombardi

Abstract Similar to other *"Researcher degree of freedom"*, the running of questionable "*Interim Analysis*" strategies may lead to an increase in the likelihood of observing a false positive rejection of the null hypothesis. This work aims to present a new probabilistic model to characterize the evidence that a statistically significant result is actually caused by the application of a questionable interim analyses (which does not correct the p-value as a function of multiple analyses). In particular, it is described the context in which two one sample t tests are carried out on an incremental data set composed of two consecutive blocks of independent observations for which only the second analysis results statistically significant. The application of the model to the literature could lead to greater control regarding the reliability of reported (already published) results.

Abstract *Cosi come altri "gradi di libertà dei ricercatori" la conduzione di strategie di analisi intermedie portano a un aumento della probabilità (non più imputabile quindi al teorico 5%) che il risultato ottenuto rappresenti un falso positivo. Obiettivo del presente lavoro è quello di descrivere un nuovo modello di stima della probabilità che un risultato statisticamente significativo sia stato ottenuto mediante l'applicazione di analisi intermedie. In particolare, viene descritto il contesto inerente lo svolgimento di due t test a campione unico (definita n la numerosità campionaria finale e $n = n_1 + n_2$, la prima analisi viene svolta a n_1 mentre la seconda a n) e viene assunto che solo i risultati della seconda analisi siano significativi. Qualora venga applicato alla letteratura, il modello proposto può rappresentare un utile strumento di controllo dell'affidabilità dei risultati osservati.*

Key words: Interim analysis, Replication crisis, Researcher degree of freedom

Francesca Freuli
Department of Psychology and Cognitive Science, University of Trento, Corso Bettini, 84, 38068 Rovereto, e-mail: francesca.freuli@unitn.it

Luigi Lombardi
Department of Psychology and Cognitive Science, University of Trento, Corso Bettini, 84, 38068 Rovereto, e-mail: luigi.lombardi@unitn.it

1 Introduction

Uncontrolled interim analysis strategy represents one of the many questionable behaviors that researchers can adopt to force the publication of scientific results [3]. This strategy usually consists in carrying out multiple statistical analysis during the data collection phase, by adding, between the analysis sessions, a variable number of new observations. This recruiting and analysis "cycle" stops as soon as a significant result is observed.

Because a statistical control of the Type I error is lacking in this sort of practice, the observed results may easily lead to a 20% increment of false positives [7]. Several approaches have been proposed in the statistical literature to model rational interim analysis strategies to control for false positive rates [10]. For instance, the R package *GroupSeq* R [6] and another method [4] related to the Likelihood Ratio framework can be used to run interim analysis without incurring in statistical bias. Since, these approaches make it possible to plan a priori interim analysis while still preserving reliable results, they are frequently used in situations in which the recruitment process may be difficult, such as, for example, in clinical trials [2].

Another line of research regards the development of statistical procedures to estimate the reliability of the published statistical results. For instance, the method implemented in the R package *statcheck* makes it possible to replicate a paper's statistical analysis by comparing the model results against those reported in the published papers [5].

The present work aims to describe a new model to estimate the likelihood that a published statistical result was actually obtained via a questionable interim analysis. It is based on a transformed t-type density calculated by the convolution between a truncated t-density function linked to the first set of observations and the t-density function associated with the new set of observations in a two-step interim analysis. Importantly, this method is applied without having access to the original data (it only requires the final statistical test).Therefore, the proposed procedure can play a relevant role whenever a posteriori-analysis is required as a consequence of suspicious or alleged manipulations of the observed data or data analysis strategy. In the next sections, the procedure will be described together with a very simple application (one sample t test) in order to better highlight the underlying core logic of the novel proposal.

2 Model

2.1 Preliminary notation

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a random sample of n independent and identically distributed (i.i.d.) observations sampled from a normal distribution $N(\mu_X, \sigma_X^2)$ with unknown mean (μ_X) and variance (σ_X^2). Moreover, let t_o^n be the observed one-sample

t-test statistic computed on $\mathbf{x}$. In particular:

$$t_o^n = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

with $\bar{x}$, s, and μ_0 being the sample mean, sample standard deviation, and population mean under H_0, respectively. It is straightforward to verify that t_o^n is also equivalent to $\bar{z}/(1/\sqrt{n})$ where $\bar{z}$ is the mean of the standardized array $\mathbf{z}$ with elements $z_i = (x_i - \mu_0)/s$ (with $i = 1, \ldots, n$) such that z_i are distributed according to the standardized normal $N(0,1)$. In what follows, we will limit our representation (without loss of generality) on the derived standardized array $\mathbf{z}$, only.

Now, suppose that 2 one-sample t-test sequential analyses are performed on the concatenated random samples $(\mathbf{y}_1, \mathbf{y}_2)$ such that

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{z}_1 \\ \mathbf{y}_2 &= \mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2) \end{aligned}$$

where $(\mathbf{z}_1, \mathbf{z}_2)$ is a concatenation of 2 sequential sub-samples of $\mathbf{z}$ (here simply called *blocks*) with the 1-th block $\mathbf{z}_1$ having n_1 observations and the 2-nd block $\mathbf{z}_2$ with $n_2 = n - n_1$ observations. Note that performing the interim analyses according to the former scheme corresponds to set a sequential data analysis strategy S which specifies the positions in the sequence $1, 2, \ldots, n$ where to split the original sample $\mathbf{z}$ and therefore how to create the 2 sequential blocks. More formally, we define an interim analysis strategy as

$$S_2 = \langle (n_1, n_2), 2 \rangle \tag{1}$$

Finally, let $B = \{0,1\} \times \{0,1\}$ be the binary space of the final decisions for the 2 sequential analyses. An element $b \in B$ is a Boolean pair (b_1, b_2) where $b_h = 1$, if the result of the analysis on the concatenated sample $\mathbf{y}_h$ (with $h = 1, 2$) turns out to be statistically significant (on the basis of a fixed Type I error probability α). Here we assume that no correction for the p-values is applied on the interim analyses. In our context, we are interested in that very particular case, denoted as $b^* = (0,1)$. The latter corresponds to that situation where the first interim analysis boils down to a non-significant result, whereas the analysis on the complete sample $\mathbf{z}$ allows to reject the null hypothesis. This mimics the result of a cheating analyst who stops collecting new data as soon as the analysis turns out to be (apparently) statistically significant (without correcting the p-values). In our context, to simplify the description, we assume that this event occurs after one single additional data collection $\mathbf{z}_2$.

2.2 Model representation

Here we introduce the simplest model representation which set a fixed strategy for a two-step interim analyses. In the discussion section, we will illustrate how a more

general representation can be straightforwardly derived from this basic simple instance. We assume that the following elements are observable or known:

1. the output, t_o^n, of the final statistical analysis (one-sample t-test) applied on the complete sample $\mathbf{z}$ (which results statistically significant: $p < \alpha$),
2. the specific hypothetical interim analysis strategy S_2 is also known.

Note that, in this representation we do not have direct access to the observations of $\mathbf{z}$, we only have access to the observed statistic t_o^n. This mimics that particular situation where we get access to an already published statistically significant result without having access to the original data. It is known (e.g., [1]) that the adoption of an interim analysis strategy certainly increases the chance to incorrectly reject the null hypothesis. Therefore, we are interested in computing the likelihood that the observed statistic, t_o^n, is actually the result of an interim analysis strategy S_2 conditioned that the null hypothesis H_0 is true ($\mu_X = \mu_0$) and consequently the observed statistically significant result is a false positive. Of course, the computation of this likelihood only represents a first step for computing other relevant information such as, for example, the posterior probability of the strategy S_2 given t_o^n and H_0 or the ratio between the posterior probability of the strategy S_2 and the posterior probability of a standard analysis S_1 with no interim analyses.

Under the assumption that H_0 is true we have that the unknown means $\bar{y}_1$ is distributed according to $N(0,\delta_1)$ with δ_1 being the standard error, $1/\sqrt{n_1}$, for the mean of the first block. Moreover, b^* entails the following constraints:

$$\bar{y}_1 \in D_1 = [-|t_c^{n_1}|\sqrt{1/n_1}, |t_c^{n_1}|\sqrt{1/n_1}] = [d_1^-, d_1^+], \tag{2}$$

where $t_c^{n_1}$ denotes the critical value of the t-distribution with $n_1 - 1$ df. By contrast,

$$\bar{y}_2 \in D_2 =]-\infty, -|t_c^n|\sqrt{1/n}[\cup]|t_c^n|\sqrt{1/n}, \infty[\tag{3}$$

We recall the following simple identity:

$$\bar{y}_2 = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{z}_2 \tag{4}$$

where $\bar{z}_2 \sim N(0,\delta_2)$ with δ_2 being the standard error $1/\sqrt{n_2}$ for the mean of the second block. Now, let the two transformed random variables $\bar{y}_1^* = \frac{n_1}{n}\,\bar{y}_1$ and $\bar{z}_2^* = \frac{n_2}{n}\,\bar{z}_2$. Clearly,

$$\bar{y}_1^* \sim N\big(0, \delta_1^2\big(n_1/n\big)^2\big) \qquad \bar{z}_2^* \sim N\big(0, \delta_2^2\big(n_2/n\big)^2\big).$$

Since $\bar{y}_1 \in D_1$, the density distribution of $\bar{y}_2$ can be defined as the convolution of the truncated normal distribution of $\bar{y}_1^*$ and the distribution of $\bar{z}_2^*$ [9]

$$f(\bar{y}_2) = \gamma \exp\left(-\frac{(\bar{y}_2)^2}{2(s_2^2+s_1^2)}\right)\left[\Phi\left(\frac{\bar{y}_2 - d_1^- - \alpha}{\beta}\right) - \Phi\left(\frac{\bar{y}_2 - d_1^+ - \alpha}{\beta}\right)\right] \tag{5}$$

where $s_1^2 = \delta_1^2\big(\frac{n_1}{n}\big)^2$ and $s_2^2 = \delta_2^2\big(\frac{n_2}{n}\big)^2$ and

$$\alpha = \frac{s_2^2 \bar{y}_2}{s_2^2 + s_1^2}, \quad \beta^2 = \frac{s_2^2 s_1^2}{s_2^2 + s_1^2}, \quad \gamma = \frac{\sqrt{2\pi}\beta}{2\pi s_2 s_1 (\Phi(e) - \Phi(c))},$$

and where

$$c = \frac{0 - d_1^+}{s_1} \quad \text{and} \quad e = \frac{0 - d_1^-}{s_1}.$$

Figure 1 shows the behaviour of the density function derived in Eq. (5) and compares it against the density function of the strategy $S_1 = \langle n, 1 \rangle$ representing the H_0 distribution (t distribution with 389 df) for a basic t-test analysis. Density points were calculated considering the left tail of the curve ($\alpha < 0.025$). Due to the symmetry of the distributions, results can be replied also for the right tail. From a quick inspection of Figure 1, we may observe a higher densities for S_2 for t values closer to the critical t. This result supports the expectation that questionable interim analysis increases the chance to invalidly reject the null hypothesis for observed t values with very close proximity with the uncorrected critical value.

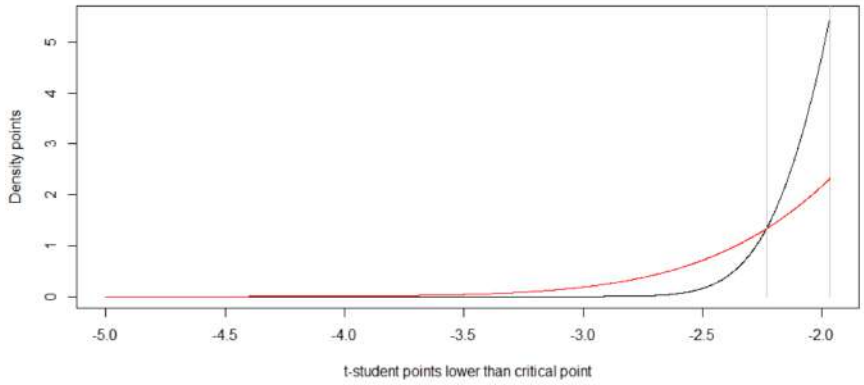


Fig. 1 Density function for the strategy S_2 (black line) and strategy S_1 (red line) model. In this representative example $n_1 = 360$ and $n_2 = 30$. The first grey line indicates the intersection point between two density function, the second one corresponds to the t critical point.

As stated earlier, the computation of the density function to represent the distribution of the test statistic under a questionable interim analysis strategy, constitutes only the first step to provide a full analysis about the evaluation of observed (and already published) statistical results. A subsequent step would require to combine or transform the derived densities to obtain useful information. For example, an analyst would be interested in computing the posterior probability of the strategy S_2 (resp. the strategy S_1) given the already published statistic t_o^n and H_0. Of course, the latter computation would require to set the prior probabilities for the two strategies, S_1 and S_2, on the basis of either speculative knowledge or previous empirical evidences which could naturally depend on the specific context of application.

3 Conclusions

In this short contribution, we presented a new model to estimate the density of a simple target statistic, the one-sample t-test, in case of a questionable interim analysis strategy. The ratio between these density values and the corresponding density values obtained under the "not questionable research practice" condition reflects the odds that a t value was obtained under the two conditions. For instance, for the condition described in figure 1, a t value of -2.1 is 1.65 times more likely to be associated with a questionable research practice than a not questionable one. Our approach can be useful to evaluate if an already published results may be consistent with questionable data analysis strategies, the so called researcher degree of freedom, which may drastically inflate the chance to erroneously reject a null hypothesis. By contrast, if an observed statistical result shows a very low evidence under the strategy, then this would indicate that rejection of the null is probably correct and not the result of a strategic manipulation adopted by the researcher to raise the chance to publish h(is/er) results.

References

1. Berry, D. A.: Interim analysis in clinical trials: The role of the likelihood principle. The American Statistician, 41.2, 117–122 (1987).
2. He, P., Lai, T. L., Su, Z. Design of clinical trials with failure-time end points and interim analyses: an update after fifteen years. Contemporary clinical trials, 45, 103–112 (2015).
3. Leslie, K.J., Loewenstein, G, Prelec, D.: "Measuring the prevalence of questionable research practices with incentives for truth telling." Psychological science 23.5, 524–532 (2012).
4. Murayama, K., Pekrun, R., Fiedler, K.: Research practices that can prevent an inflation of false-positive rates. Personality and Social Psychology Review 18.2, 107–118 (2014).
5. Nuijten, M.B., Van Assen, M., Hartgerink, C.HJ., Epskamp, S., Wicherts, J.: The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. PsyArXiv (2017).
6. Pahl, R., Ziegler, A., König, I.R.: GroupSeq: Designing clinical trials using group sequential designs. The Newsletter of the R Project Volume 6/2, May 2006. 21 (2006).
7. Simmons, J. P., Nelson, L.D., Simonsohn, U.: False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science 22.11, 1359–1366 (2011).
8. Team, R Core and others. R: A language and environment for statistical computing. Vienna, Austria (2013).
9. Turban, S. : Convolution of a truncated normal and a centered normal variable. Online in http://www. columbia. edu/ st2511/notes. html, Accessed, 4(17), (2010).
10. Wagenmakers, E.: A practical solution to the pervasive problems of p values. Psychonomic bulletin & review 14.5, 779–804 (2007).

Incorporating Expert Knowledge in Structural Equation Models: Applications in Psychological Research

Integrare il Parere degli Esperti con i Modelli di Equazioni Strutturali: Applicazioni nella Ricerca Psicologica

Gianmarco Altoè, Claudio Zandonella Callegher, Enrico Toffalini and Massimiliano Pastore

Abstract Structural Equation Modeling (SEM) is used in psychology to model complex structures of data. However, sample sizes often cannot be as large as ideal for SEM, leading to a problem of insufficient power. Bayesian estimation with informed priors can be beneficial in this context. Our simulation study examines this issue over a real case of a mediation model. Parameter recovery, power and coverage were considered. The advantage of a Bayesian approach was evident for the smallest effects. The correct formalization of the theoretical expectations is crucial, and it allows for increased collaboration among researchers in Psychology and Statistics.
Abstract *I Modelli di Equazioni Strutturali (SEM) sono spesso utilizzati in psicologia. Tuttavia, campioni limitati portano ad un problema di insufficiente potenza. Il nostro studio di simulazione esamina i vantaggi dell'approccio bayesiano con prior informative nel caso di un modello di mediazione. Sono state considerate la stima dei parametri, il coverage e la potenza. Il vantaggio dell'approccio Bayesiano è risultato evidente per gli effetti minori. La formalizzazione delle aspettative teoriche è cruciale e favorisce una fruttuosa collaborazione tra i ricercatori.*

Key words: Expert elicitation, Informative Priors, Structural Equation Models (SEM), Small sample sizes, Psychological research

---

Gianmarco Altoè
Department of Developmental Psychology and Socialisation, University of Padova
e-mail: gianmarco.altoe@unipd.it

Claudio Zandonella Callegher
Department of Developmental Psychology and Socialisation, University of Padova
e-mail: claudio.zandonellacallegher@phd.unipd.it

Enrico Toffalini
Department of General Psychology, University of Padova
e-mail: enrico.toffalini@yahoo.it

Massimiliano Pastore
Department of Developmental Psychology and Socialisation, University of Padova
e-mail: massimiliano.pastore@unipd.it

1 Introduction

Structural Equation Modeling (SEM) encompasses a range of multivariate statistical techniques. SEMs are composed of two parts: a *measurement model* and a *structural model* (see Fig. 1). The measurement model defines unobserved constructs (*latent variables*, circles in Fig. 1) according to a set of measured outcomes (*observed variables*, squares in Fig. 1), whereas the structural model describes the relationships between latent variables. SEMs are widely used in psychology to model complex relations between different latent psychological constructs. However, as the complexity of the model increases, more data are required to obtain accurate parameter estimates and model fit statistics [10]. Nevertheless, in many research settings, the number of participants may be limited and appropriate statistical techniques are required to enhance the reliability of the results.

Often in the literature, the Bayesian approach is suggested over frequentist estimation when limited data are available [1]. The inclusion of prior information can help in the parameter estimation, but researchers have to carefully consider priors choice. However, most of the studies rely on default software prior settings. A recent review, underlined that the use of diffuse default priors can result in severely biased estimates, and this bias can be decreased only by incorporating informative priors [8].

Informative priors allow researchers to include in the analysis relevant knowledge in the field. Researchers could also consider to include opinions of experts. On the base of their experience in the field, experts can evaluate relevant information and help researchers in the definition of a plausible range of values and priors choice. *Elicitation* is a structured procedure that allows experts to express their knowledge and uncertainty about quantities of interest in the form of probability distributions [4, 3]. Elicitation can be used to define priors according to experts' judgement.

The remainder of this article is structured as follows. In Sec. 2, we present a simulation study to evaluate the influence of different prior specifications in the case of SEM with small sample size. For the sake of simplicity (but without losing generalizability) we present a mediation model in which the measurement model is not considered. Following common procedures, all variables were standardized (i.e., mean = 0 and a standard deviation = 1) before fitting all models. In Sec. 3, we discuss the obtained results.

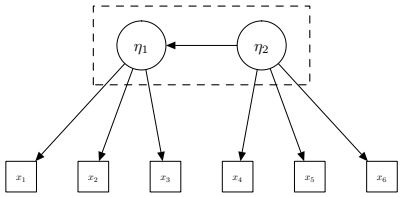


Fig. 1 A structural equation model. Within the dashed box is the structural model, outside is the measurement model. Circles for latent variables; rectangles for observed variables.

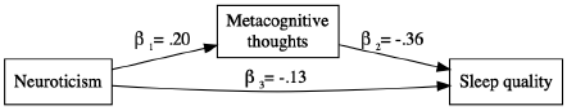


Fig. 2 Mediation model from [7]. The relation between Neuroticism and sleep quality is mediated by metacognitive thoughts.

2 Simulation

We considered a mediation model from [7] presented in Fig. 2. The study evaluated the relationship between participants' self-reported sleep quality (*Sleep quality*), participants' tendency to become anxious (*Neuroticism*), and negative beliefs about sleeping problems (*Metacognitive thoughts*). In particular, the association between *Neuroticism* and *Sleep quality* ($\beta_3 = -.13$) is mediated by *Metacognitive thoughts*. In other words, people with higher levels of distress and anxiety tend to have dysfunctional beliefs and attitudes about sleep ($\beta_1 = .20$) that, in turns, induce them to perceive and report a worse-quality sleep ($\beta_2 = -.36$).

2.1 Simulation details

The simulation was carried in `R` version 3.6.2 [5] using R-packages `lavaan` [6] and `blavaan` [2]. In the simulation, we considered as parameters of interest the regression coefficients ($\beta_1,\ \beta_2,\ \beta_3$). We compared the performance of Maximum Likelihood (ML) estimation and Bayesian estimation under four different sample size conditions (i.e., 20, 50, 100, 500).

Three different prior distribution specifications were used (see Fig. 3):

1. *Default prior* - $\beta_i \sim N(0, 10)$. These are intended to be non-informative.
2. *Reasonable prior* - $\beta_1 \sim N(.20,\ .50)$, and $\beta_{2;3} \sim N(-.20,\ .50)$. These are moderately informative to exclude excessively large values that are not reasonable within psychology research. Moreover, the mean of each prior is set to reflect the direction of the main results in the literature.
3. *Experts prior* - $\beta_1 \sim N(.20,\ .20)$, $\beta_2 \sim N(-.40,\ .20)$, and $\beta_3 \sim N(-.20,\ .20)$. These are intended to be highly informative representing experts' judgement.

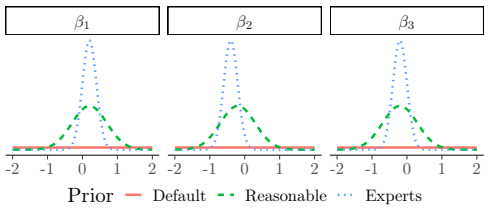


Fig. 3 Prior distribution in the three different settings. Default priors are intended to be non-informative. Reasonable priors are intended to exclude implausible values. Experts priors represent experts' judgement.

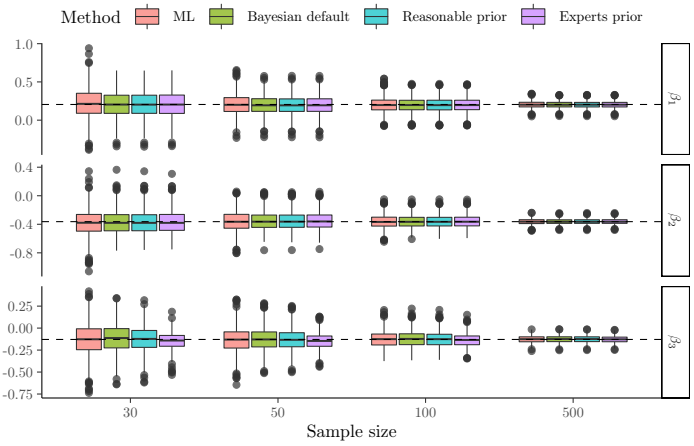


Fig. 4 Estimates distribution for each parameter across the different condition. Dashed lines represent the true population values.

Relative mean bias, relative median bias, mean square error (MSE), coverage and power were considered [9]. The relative mean bias (or median bias) evaluates the relative difference between mean estimate ($\bar{\theta}$; or median estimate $\widetilde{\theta}$) across replications and the population value (θ). Relative bias included between -.10 and .10 are considered acceptable [9]. MSE takes into account variability as well as bias of the estimates: $MSE = \sigma^2 + (\bar{\theta} - \theta)^2$, where σ is the standard deviation of the estimates across replications and $\bar{\theta}$ is the mean. Coverage is the proportion of replications in which the population value is included in the 95% confidence interval (CI; for the ML estimation) or 95% highest posterior density interval (HPD; for the Bayesian estimation). Instead, power is the proportion of replications in which the value zero is not included in the 95% CI or 95% HPD. Analyses were conducted considering the standardized parameters and for each condition 1000 replications were considered.

2.2 Results

The tables with detailed results for each parameter and condition are available at `https://osf.io/hwj8d/`. To interpret the results of relative mean and median bias, we considered the distribution of the estimated parameters (see Fig. 4). Only with very small sample sizes ($n = 30$) it is possible to observe some differences between estimation methods: Maximum likelihood approach produces the widest distributions, whereas Bayesian approach with experts prior has narrower distributions. However, differences between methods are noticeable for the parameter β_3 (i.e., the parameter with the smallest population value) but are less evident for the other parameters and, as the sample size increases, estimation methods perform similar to each other.

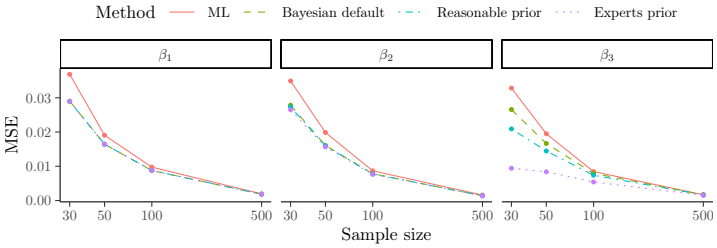


Fig. 5 Mean Squared Error (MSE) values for each parameter across the different conditions.

Considering the MSE (see Fig.5), we have the same results pattern. Differences between methods are bigger in the case of very small samples ($n = 30$), where Bayesian approach with experts prior performs better. However, differences between prior specification are noticeable only for the parameter β_3.

Finally, the result of coverage and power are presented in Fig. 6. Coverage reaches adequate levels in all conditions with sample size equal to or greater than 100. With smaller sample sizes, Bayesian approach with experts prior showed excessive coverage in the case of the parameter β_3. Power is extremely low when sample sizes are small. ML estimation performs slightly better in terms of power across all conditions, except for the parameter β_3 where is outperformed by Bayesian approach with experts prior. However, adequate levels of power are reached for all parameters only with large sample sizes ($n = 500$).

3 Discussion and conclusions

In the simulation, we evaluated the different estimation methods in the case of SEMs with small sample size. Overall, results indicate that informative priors are useful in

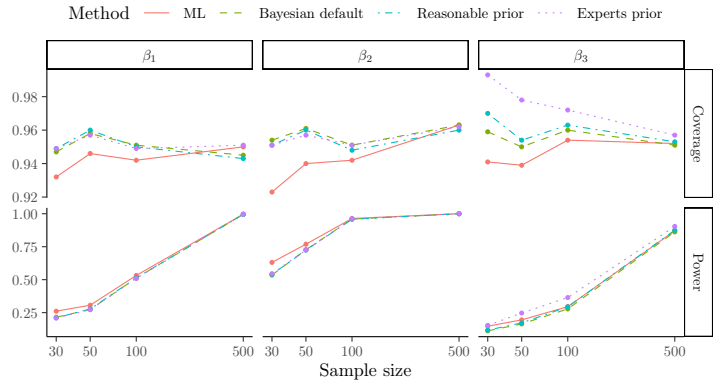


Fig. 6 Coverage and power values for each parameter across the different conditions.

the case of limited sample sizes and when the true population values are small. Parameter estimates were more stable across replications and extreme values were less likely. When the sample size increases the difference between estimation methods becomes less evident.

However, results are not consistent for all the parameters. In most conditions, Bayesian approach performs better than ML but results are very similar between the different prior specifications. Only in the case of small true population parameter values the Bayesian approach with expert priors performs much better than the other prior specifications. Future studies should focus on the role of prior definition in SEMs with different levels of complexity (e.g., also taking the measurement part into account) and in which the effect sizes vary on a larger range. Another important aspect that future studies should evaluate is the impact of prior knowledge misspecification, in particular in situations with small sample sizes.

Finally, we want to highlight that expert knowledge elicitation is not only useful to inform prior distributions but it can help also in other aspects of the analysis. Experts can help and inform researchers in the design of the experiments, definition of the models, interpretation of the results and make reasonable and informed choices along all the research process. Thus, the collaboration between different experts is a crucial point that should be encouraged in any applied research field.

References

1. McNeish, D.: On Using Bayesian Methods to Address Small Sample Problems. Structural Equation Modeling: A Multidisciplinary Journal **23**(5), 750–773 (2016). DOI 10.1080/10705511.2016.1186549
2. Merkle, E.C., Rosseel, Y.: *Blavaan* : Bayesian Structural Equation Models via Parameter Expansion. Journal of Statistical Software **85**(4) (2018). DOI 10.18637/jss.v085.i04
3. O'Hagan, A. (ed.): Uncertain Judgements: Eliciting Experts' Probabilities. Statistics in Practice. John Wiley & Sons, London ; Hoboken, NJ (2006)
4. O'Hagan, A.: Expert Knowledge Elicitation: Subjective but Scientific. The American Statistician **73**(sup1), 69–81 (2019). DOI 10.1080/00031305.2018.1518265
5. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2018)
6. Rosseel, Y.: *Lavaan* : An *R* Package for Structural Equation Modeling. Journal of Statistical Software **48**(2) (2012). DOI 10.18637/jss.v048.i02
7. Sella, E., Carbone, E., Toffalini, E., Borella, E.: Personality traits and sleep quality: The role of sleep-related beliefs. Personality and Individual Differences **156**, 109,770 (2020). DOI 10.1016/j.paid.2019.109770
8. Smid, S.C., McNeish, D., Miočević, M., van de Schoot, R.: Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review. Structural Equation Modeling: A Multidisciplinary Journal **27**(1), 131–161 (2020). DOI 10.1080/10705511.2019.1577140
9. Smid, S.C., Rosseel, Y.: Sem with Small Samples. In: R. van de Schoot, M. Miočević (eds.) Small Sample Size Solutions, first edn., pp. 239–254. Routledge (2020). DOI 10.4324/9780429273872-20
10. Wolf, E.J., Harrington, K.M., Clark, S.L., Miller, M.W.: Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. Educational and Psychological Measurement **73**(6), 913–934 (2013). DOI 10.1177/0013164413495237

Predicting social media addiction from Instagram profiles: A data mining approach

Uso del data mining per predire la dipendenza da Instagram in un campione italiano di studenti

Antonio Calcagnì, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale

Abstract In this short paper, we describe an application of data mining techniques to predict Instagram users' addiction from a set of features related to (i) Instagram captions extracted from photos, videos, comments, and stories, and Instagram indicators such as number of followers and following, blocked and closed friends, and frequency of use. We first applied text mining to explore and describe the main contents of Instagram captions. Next, we used a set of non parametric models and ensemble methods to predict Instagram addiction as measured by the Instagram addiction scale [1]. Models were compared via cross-validation using test and training (random) sets from the original dataset. Results showed that Instagram addiction is mainly predicted by the overall time spent on Instagram, writing stories and comments, and number of followers. Moreover, the results suggest that Instagram users made use of photos/videos and stories/comments differently, with the latter being mostly related to emoticons, experiences, and relationships with other users.

Abstract *Questo lavoro presenta alcuni risultati di una ricerca più ampia condotta su un campione di giovani donne italiane circa l'utilizzo di Instagram come facilitatore dell'oggetivizzazione sessuale. In questo lavoro si presentano i risultati relativi allo studio circa il modo con cui attività di Instagram come pubblicazioni di foto, video, commenti e realizzazione di storie possano predire la dipendenza dal social media, come misurato dall'apposita scala di Instagram addiction. Tale*

Antonio Calcagnì,
University of Padova, e-mail: antonio.calcagni@unipd.it

Veronica Cortellazzo,
University of Padova, e-mail: veronica.cortellazzo@studenti.unipd.it

Francesca Guizzo,
University of Padova, e-mail: francesca.guizzo@unipd.it

Paolo Girardi,
University of Padova, e-mail: paolo.girardi@unipd.it

Natale Canale,
University of Padova, e-mail: natale.canale@unipd.it

analisi è stata realizzata mediante data mining, utilizzando una serie di modelli non parametrici e metodi di ensemble. La validazione e la scelta del modello migliore è stata effettuata via validazione incrociata. I risultati mettono in evidenza il ruolo del tempo trascorso su Instagram, del numero di storie, commenti e followers come elementi di predizione della dipendenza dal social network. I risultati, inoltre, hanno altresì evidenziato come gli utenti tendano ad utilizzare video e foto in maniera diversa da commenti e storie: questi ultimi, infatti, sembrano maggiormente connessi a esperienze, emozioni e relazioni rispetto ai primi.

Key words: data mining, text mining, Instagram, social media addiction

1 Introduction

Instagram is a well-known social platform commonly used for personal reasons as well as business activities. Over the last years, it has gained wide popularity across the globe, becoming one of the most popular photo-sharing applications on the Facebook platform [2]. In particular, recent trends show that Instagram is the most important network being used among adolescents [3]. Recently, a number of research have shown the role of Instagram in several psychological processes, such as women objectification (e.g., see [4]). In this respect, appearance-related comments on women's bodies accompanying Instagram images seem to play a role in body dissatisfaction as well as women self-objectification [7]. This and other results suggest the importance of investigating the interplay between social media behaviors and social media addiction, especially in young users [5].

In this short paper we will focus on Instagram addiction in a sample of Italian students. In particular, we will investigate the predictive role of Instagram contents such as text (comments, hashtags, photo captions/descriptions), indicators (number of followers/followings, blocked users, closed friends), and activity frequency on social media addiction, as measured by the *Instagram Addiction Scale* [1]. Data mining techniques have been used to analyse the data. Particularly, text mining was applied to textual components of the dataset (comments, emoticons, captions) whereas a set of non parametric models and ensemble methods has been used to choose the best model and predictors for the response variable *Instagram addiction* [6]. The results suggest that the latter is mainly predicted by activities like interactions with other users of the social networking system via messages, comments, and likes.

2 Data and Methods

2.1 Data

The *Instagram Addiction Scale* [1] was administered to $N = 97$ female participants, all of them using Instagram on a daily basis. Data were collected by using an online survey.[1] Subjects were between 18 and 31 years old with an average age of 23.64 years (standard deviation 2.23). They were asked to answer fifteen items grouped into two sub-scales (i.e., Social effect, Compulsion) using an ordinal scale with six anchors. As scales were standardized, final scores were computed by summing the items corresponding to each sub-scale. Thus, the aggregated variable *addiction* was then defined, with higher scores being indicative of higher levels of Instagram addiction. For each participant, all Instagram data were also available in compressed format (up to six months before). They consisted in *comments* to other posts, *connections* with other users (followers, following, friends, blocked users), *likes* given to media (photos and videos) as well as *comments*, *media* (photo, stories, videos), *searchers* (texts, hashtags), and *stories*. For all these data, temporal information regarding the use of the social network (time, day) were also available. The final dataset comprises 94 subjects and 95 variables, including the response variable regarding Instagram addiction. Before running the analyses, 3 subjects were excluded as they included missing observations (row-wise exclusion) whereas 36 variables were left out from the analyses because of multicollinearity (as indicated by the threshold $r \geq 0.9$).

2.2 Methods

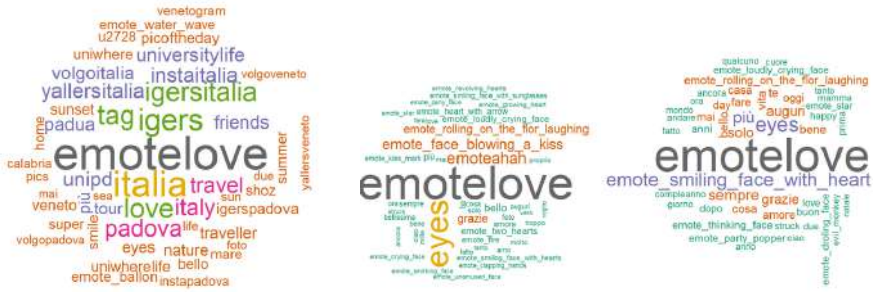
Data were analysed by means of data mining methods. With regards to the textual part of the data (photo captions, comments, hashtags), text mining descriptive techniques (i.e., most frequent terms, bigrams, graphical analyses, sentiment analysis) were used as implemented by the `R` library `TextWiller` [8] and `tidytext` [9]. All texts from Instagram were pre-processed according to standard text-mining pre-processing procedure (i.e., text-normalization, stopwords elimination, tokenization) [10]. In order to predict Instagram addiction with respect to 95 features of Instagram use, the following non-parametric and ensemble methods were instead adopted: (i) Principal component regression (tuning: number of variables), (ii) Partial least squares regression (tuning: number of variables), (iii) Lasso regression (tuning: λ penalty parameter), (iv) Regression trees (tuning: α complexity parameter), (v) Random forest (tuning: number of variables), (vi) Boosting (tuning: number

[1] The institutional review board at the University of Padova gave ethical approval for the study (protocol number: 2956)

and depth of trees), (vii) Bagging. They were compared via 10-fold cross validation, with prediction errors being computed via Out Of Bag (OOB) approach.[2]

3 Results and Discussion

Figure 1 shows wordclouds of the most frequent terms as extracted from captions of Instagram photos, videos, comments, and stories. Results indicate how distributions of photos/videos' comments were different from those of stories: indeed, words used in the first category were essentially descriptive (i.e., they just described photos/videos in terms of reference to places or experiences like traveling) whereas words used in stories/comments contained mainly emoticons and tag to users. Therefore, Instagram users made use of stories/comments and photos/videos differently. In general, sentiment analyses of captions revealed the presence of neutral sentiments more frequently than negative and positive sentiments. In particular, comments and stories showed a positive sentiment when compared to photos and videos, which suggest that users tended to use photos/videos and comments/stories with a different strategy.



(a) Photos and videos (b) Comments (c) Stories

Fig. 1: Wordcloud for captions of Photos, videos, comments, and stories.

In order to select the best subset of predictors of Instagram addiction from the large set containing 95 features, we proceeded by running a set of non-parametric models and ensemble methods as describe in section 2.2. Table 1 show the results of model comparison on training and test sets, respectively. Among others, Random

[2] More in details, ten train/test sets were created by bootstrapping the original dataset row-wise. Then, models were estimated for each training data using a 10-fold cross validation; cv-errors were computed and used to select the optimal tuned model. The best model was then chosen among those having minimum RMSE, which was computed on the test sets using the out-of-bag observations. Finally, the prediction error of the model was obtained as average of the ten RMSEs. Note that cross-validation was used as implemented in the `caret` library [11].

forest achieved lower MAE and RMSE w.r.t. parameters estimation (training set) and, in a similar way, it showed lower RMSE in the set as well. Thus, Random forest was selected as the best predictive model of Instagram addiction.

	RMSE	Test set error
PCA regression	0.7769	0.8715
PLS regression	0.663	1.0411
LASSO regression	0.7189	1.0016
Regression Tree	0.7387	0.8615
Bagging	0.6017	0.7631
Random Forest	0.6403	0.7324
Boosting	0.6749	0.7909

Table 1: Model comparison on test set. Note that Test set error is computed via bootstrap prediction error using 10 samples whereas the best model is represented in gray tones.

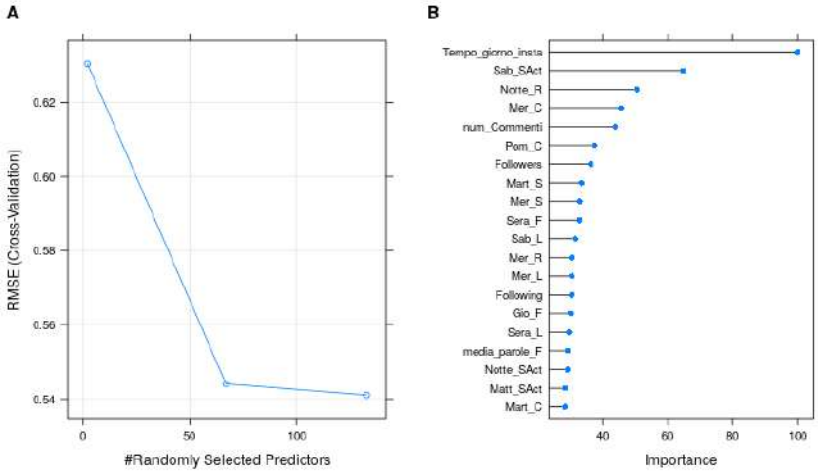


Fig. 2: Random forest to predict Instagram addiction: (A) RMSE as a function of the best number of predictors, (B) Variable importance plot for the final model. Note that variable importance measures are computed via the `importance()` function as implemented in the `randomforests` library [12].

The Random forest model was built on 500 trees which selected a final set of 67 predictors. Figure 2 shows the results of the final model. Overall, they suggested the following main predictors for Instagram addiction (in order of importance): Overall time spent on Instagram (`Tempo_giorno_insta`), replies to users' stories on Saturday (`Sab_SAct`), content searches during the night (`Notte_R`), writing com-

ments overall (`num_Commenti`) or in a specific temporal window, namely on afternoon (`Pom_cC`) and on Wednesday (`Mer_C`), number of followers (`Followers`).

4 Conclusion

We investigated the role of Instagram user's data (captions, photos, comments, likes, stories, hashtags, following and followers) in predicting social media addiction as measured by *Instagram Addiction Scale* [1] in a sample of Italian students. We used data mining techniques in order to (i) describe the use of Instagram commonly made by users and (ii) find predictors of Instagram addiction from a large database containing 95 features. The results suggested the importance of some variables, such as time spent on Instagram, writing stories, comments, and number of followers, as predictors of Instagram addiction. Moreover, the results also showed that users tend to use photos/videos and stories/comments differently: the first were mainly adopted for describing the objects they are related to. By contrast, the latter are more than simple descriptions, as they include emoticons, reference to places and experiences, tags to other users. Further studies need to be conducted to establish insights into the various mechanisms of Instagram addiction. For instance, text mining results (e.g., emoticons, sentiments) may be used to define new predictors of social media addiction, which would shed light on the role played by *typed emotions* on this emerging phenomena.

References

1. Kircaburun, K., Griffiths, M.D., Journal of behavioral addictions **7**(1), 158–170 (2018)
2. Li, Z., Agarwal, A., Management Science **63**(10), 3438–3458 (2017)
3. Brown, R.C., Fischer, T., Goldwich, A.D., Keller, F., Young, R., Plener, P.L., Psychological medicine **48**(2), 337–346 (2018)
4. Meier, E.P., Gray, J., Cyberpsychology, Behavior, and Social Networking **17**(4), 199–206 (2014)
5. Wallace, P., EMBO reports **15**(1), 12–16 (2014)
6. Azzalini, A., Scarpa, B., Springer Science & Business Media (2009)
7. Tiggemann, M., Barbato, I., Body image **27**, 61–66 (2018)
8. Solari, D., Sciandra, A., Finos, L.: Textwiller, Journal of Open Source Software **4**(41), 1256 (2019). DOI 10.21105/joss.01256. URL https://doi.org/10.21105/joss.01256
9. Silge, J., Robinson, D.: tidytext, Journal of Open Source Software **1**(3), 37 (2016)
10. Silge, J., Robinson, D.: Text mining with R: A tidy approach. " O'Reilly Media, Inc." (2017)
11. Kuhn, M., et al., Journal of statistical software **28**(5), 1–26 (2008)
12. Liaw, A., Wiener, M., R News **2**(3), 18–22 (2002). URL https://CRAN.R-project.org/doc/Rnews/

Structural entropy based modeling for psychological measurement

Modeling Strutturale basato sull'entropia per le misure in Psicologia

Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo

Abstract The contribution introduces the Entropy Based Structural Models and their dynamical evolution: the *Streaming entropy*. The new variant is described and tested on a typical clinical psychology scenario, that is, the psychotherapy process, as no previous work made us of behavioral data to study the psychotherapeutic relationship. Textual data consist of a psychotherapy transcript, whose word blocks have been processed and classified according to their valence: positive, negative and abstract. At first, the Stre-GCE algorithm computes one model parameter for each interaction components, then parameters get updated in an alternate way (therapist-patient, patient-therapist and so forth). Results show that Stre-GCE accounts for the fluctuating nature of the psychotherapy interaction.

Abstract *Il contributo introduce i Modelli Strutturali basati sull'Entropia ed una loro evoluzione dinamica: la Streaming Entropy. La nuova variante viene descritta e testata in uno scenario tipico della psicologia clinica, il processo psicoterapeutico, in quanto non esiste ancora un lavoro che utilizzi dati comportamentali per esaminare la relazione psicoterapeutica. I dati testuali consistono nel trascritto di un'intera psicoterapia, i cui blocchi di parole sono stati classificati in categorie: positive, negative, astratte. Inizialmente, l'algoritmo Stre-GCE stima i parametri di un modello per ciascuno dei componenti dell'interazione terapeutica, per poi aggiornarli in modo alterno con riferimento all'interlocutore precedente (terapeuta-paziente, paziente-terapeuta, ecc..). I risultati riflettono le fluttuazioni tipiche dell'interazione psicoterapeutica.*

Enrico Ciavolino
University of Salento, e-mail: enrico.ciavolino@unisalento.it

Mario Angelelli
University of Salento and INFN Lecce e-mail: mario.angelelli@unisalento.it

Paola Pasca
University of Salento e-mail: paola.pasca@unisalento.it

Omar Carlo Gioacchino Gelo
University of Salento e-mail: omar.gelo@unisalento.it

Key words: Entropy Measures, Streaming Data, Structural Models, psychotherapy process

1 Introduction

The maximum entropy principle can be employed to quantify, keep track or estimate the information content in a system. Among the various applications of this principle, the Generalized Cross Entropy (GCE) proved to be useful in the estimation of model parameters subject to different forms of uncertainty (randomness, fuzziness). In this work, we examine the information content from a dynamical perspective through the Streaming Generalized Cross Entropy (Stre-GCE) method. As the Stre-GCE performed well in several simulations (see [2]), this work considers a real-life scenario in the clinical psychology context. In particular, we examine how the patient-therapist relationship unfolds over the psychotherapeutic process. Starting from an initial prior, we estimate the parameters evolution of an underlying statistical model through the Stre-GCE algorithm. This allows one to quantify the response to the stimuli and to evaluate the effects of the therapy in terms of the mutual actions in the patient /therapist pair.

2 Entropy models

In the following sections GCE theoretical framework and the Streaming version are introduced.

2.1 Generalized Cross Entropy (GCE)

Golan et al. [8] proposed by GCE as a generalization of the Maximum Entropy principle (MEP) developed by E.T. Jaynes [11, 12]. The GCE is based on the reformulation of regression parameters as expected values. Let us consider the i^{th} unit of the following regression model (n observations and m variables) by omitting the intercept:

$$y_i = \sum_j^m x_{ij}\beta_j + \varepsilon_i \tag{1}$$

the parameters (β_j, ε_i) are re-formulated as follows:

$$\beta_j = \sum_k^K z_{jk}^{\beta} p_{jk}^{\beta} \qquad \forall j, \qquad \varepsilon_i = \sum_h^H z_{ih}^{\varepsilon} p_{ih}^{\varepsilon} \qquad \forall i$$

where z^{β}_{jk} and z^{ε}_{ih} are the elements, called fixed points, of the support vectors $\mathbf{z}^{\beta}_{j}$ and $\mathbf{z}^{\varepsilon}_{i}$ equally spaced, uniformly and symmetrically chosen around zero. For the definition of these vectors see ([4, 8]). The associated vectors $\mathbf{p}^{\beta}_{j}$ and $\mathbf{p}^{\varepsilon}_{i}$ are the probability distributions of the unknown parameters β_j, ε_i, estimated by the minimization of the following cross-entropy function:

$$\begin{cases} \text{Minimize :} & \sum_j \sum_k p^{\beta}_{jk} \log p^{\beta}_{jk}/q^{\beta}_{jk} + \sum_i \sum_h p^{\varepsilon}_{ih} \log p^{\varepsilon}_{ih}/q^{\varepsilon}_{ih} \\ \text{Subjectto :} & \sum_k p^{\beta}_{jk} = 1, \quad \forall j \\ & \sum_h p^{\varepsilon}_{ih} = 1, \quad \forall i \\ & \sum_j x_{ij} \sum_k z^{\beta}_{jk} p^{\beta}_{jk} + \sum_h z^{\varepsilon}_{ih} p^{\varepsilon}_{ih} = y_i, \quad \forall i \end{cases} . \tag{2}$$

The key features and the main advantages of the GCE are the flexibility to introduce prior information in the model, defined by the two vectors $\mathbf{q}^{\beta}_{j}$ and $\mathbf{q}^{\varepsilon}_{i}$.

Entropy-based models do not require any distributional assumption, for instance they may be extended to cases where errors are correlated. In addition, contrary to other methods such as OLS, GCE can be applied without special algebraic requirements, e.g. when multicollinearity among covariates occurs. Finally, it is worth remarking that GCE works well in the case of small sample size too [8, 4, 5, 6, 9].

2.2 Streaming GCE

The choice of the criterion to update probability distributions starting from data may depend on the specific situation and on data characteristics. In several contexts, there is a temporal dimension associated with data. In particular, modern ICT systems involve *streams* of data, which are captured in different moments.

In order to explore the GCE approach in a dynamic context and extend it with adaptive aspects, we introduced the Streaming Generalized Cross Entropy algorithm in [2]. The algorithm involves two phases: the *batch* phase and the *streaming* phase.

2.2.1 The Batch phase

The *batch* is represented by m data samples to be used to assess the initial estimates. The Stre-GCE process continues until n data samples are collected. Formally, these n samples provide consistency constraints and $m < n$ distinguished samples represent the batch. They are processed using (2) with uniform prior:

$$\begin{cases} \text{Maximize:} & \left(-\sum_{j=1}^{J} \sum_{k=1}^{K} p^{\beta}_{j,k} \cdot \ln p^{\beta}_{j,k} - \sum_{h=1}^{H} \sum_{i=1}^{m} p^{\varepsilon}_{i,h} \cdot \ln p^{\varepsilon}_{i,h}\right) \\ \text{Subject to:} & \sum_{k=1}^{K} p^{\beta}_{j,k} = 1, \quad \forall j \\ & \sum_{h=1}^{H} p^{\varepsilon}_{i,h} = 1, \quad \forall i \\ & y_i = \sum_{j=1}^{J} \sum_{k=1}^{K} p^{\beta}_{j,k} \cdot z^{\beta}_{j,k} \cdot x_{i,j} + \sum_{h=1}^{H} p^{\varepsilon}_{i,h} z^{\varepsilon}_{i,h}, \quad \forall i \end{cases} . \tag{3}$$

The output $(\hat{\mathbf{P}}^{\beta(m)}, \hat{\mathbf{p}}_1^{\varepsilon}, \ldots, \hat{\mathbf{p}}_m^{\varepsilon})$ constitutes initial estimates:

$$\begin{aligned} \hat{\beta}_j^{(m)} &:= \sum_{k=1}^{K} \hat{p}_{j,k}^{\beta(m)} \cdot z_{j,k}^{\beta}, \quad j \in \{1,\ldots,J\}, \\ \hat{\varepsilon}_i &:= \sum_{h=1}^{H} \hat{p}_{i,h}^{\varepsilon} \cdot z_{i,h}^{\varepsilon}, \quad i \in \{1,\ldots,m\} \end{aligned} \tag{4}$$

2.2.2 The Streaming phase

In the update process, the upcoming information follows to the already present one (that is, the actual probability distribution used to compute the model estimates). Formally, the new estimates will be produced through the minimization of the cross entropy

$$\mathrm{D_{KL}}(\mathbf{P}^{\beta} || \hat{\mathbf{P}}_i^{\beta}) - H(\mathbf{p}^{\varepsilon}). \tag{5}$$

So we find the new distribution as the solution to the following constrained optimization problem:

$$\begin{cases} \text{Minimize :} & \left(\sum_{j=1}^{J}\sum_{k=1}^{K} p_{j,k}^{\beta} \ln p_{j,k}^{\beta} - \sum_{j=1}^{J}\sum_{k=1}^{K} p_{j,k}^{\beta} \ln \hat{p}_{j,k}^{\beta(i)} + \sum_{h=1}^{H} p_{i+1,h}^{\varepsilon} \ln p_{i+1,h}^{\varepsilon}\right) \\ \text{Subjectto :} & \sum_{k=1}^{K} p_{j,k}^{\beta} = 1, \qquad \forall j \\ & \sum_{h=1}^{H} p_{i+1,h}^{\varepsilon} = 1, \qquad \forall i \\ & y_{i+1} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{j,k}^{\beta} \cdot z_{j,k}^{\beta} \cdot x_{j,i+1} + \sum_{h=1}^{H} p_{i+1,h}^{\varepsilon} z_{i+1,h}^{\varepsilon}, \qquad \forall i \end{cases} \tag{6}$$

It is noted that the difference between signal and error is described in terms of effects of the temporal dimension, (i.e. note the different priors in 5: at the $(i+1)$th step, the probability distributions selected are chosen as prior $\hat{\mathbf{P}}_i^{\beta}$ for the signal, while the uniform prior is chosen for errors.

3 Theoretical model for synchronization in Psychotherapy

The flow of the psychotherapeutic interaction has been traditionally studied and located in the theoretical framework of the self-organization theory [7]: according to this view, a transition from a behavioral pattern to another occurs as a consequence of a temporary destabilization of a relatively stable system. So far, self-report measures were the primary source of information about instability/discontinuity moments [10]. Thus, the use of behavioral data to investigate the therapeutic interaction is still an unexplored territory.

Our textual data comprise a whole psychotherapy transcript from the York I Depression Study. It consists of a number N_S of sessions, and a number w_t of word

blocks for the t-th session. Words in each block are partitioned into three classes, i.e. positive words, negative words, and abstract words. This partition provides each block i with two probability distributions, i.e. the relative frequencies of each class compared to the total number of words in the i-th block. Therefore, we also get two entropies $h_i^{(T)}$ and $h_i^{(P)}$ associated with the therapist's and the patient's frequencies in the i-th block, respectively.

Two simultaneous models are provided to describe the effects of the patient's (respectively, the therapist's) word block: for the i-th word block , we consider the following pair of relations describing the mutual influence of the two actors:

$$h_i^{(T)} = \beta^{P\to T} \cdot h_i^{(P)} + \varepsilon_i^{P\to T}, \tag{7}$$

$$h_i^{(P)} = \beta^{T\to P} \cdot h_i^{(T)} + \varepsilon_i^{T\to P}. \tag{8}$$

The number of session is used as an index to switch between the updates of the patient's and the therapist's estimates, while the word-blocks within a given session represent the block stream that feeds the block Stre-GCE algorithm.

In principle, the two parameters $\beta_s^{P\to T}$ and $\beta_s^{P\to T}$ are distinct. In the models (7-8) $\beta_s^{P\to T}$ and $\beta_s^{P\to T}$ are uni-dimensional (i.e., scalars), but they can be extended to a multi-dimensional case when more synchronization "measures" are considered. We also stress that the present model envisages two different entropies: the pair of entropies $\left(h_i^{(P)}, h_i^{(T)}\right)$ associated with the individual i-th word block, and the cross entropy representing the objective function in (6) to get the estimates for parameters. It is worth remarking that the relation among entropies at different scales has been explored in a different framework in [1].

We will assume that the models alternately update: (7) is updated when s is odd, while (8) is updated when s is even. For each session s, we focus on the corresponding word blocks indexed by $i \in \{1,\dots,w_s\}$. Each such block contains the data for a single update. Therefore, the Stre-GCE algorithm discussed in this work is an adaptation of the Stre-GCE algorithm already discussed in [2]. A different approach is to run the Stre-GCE algorithm on the individual data within a given block. The output of this process is the update of the therapist's (respectively, the patient's) response model estimate. In particular, the updated distribution represents the prior distribution in the subsequent update. As for the Stre-GCE algorithm, only the distributions $\mathbf{p}_s^{\beta,P\to T}$ and $\mathbf{p}_s^{\beta,P\to T}$ associated with parameters $\beta_{P\to T}$ and $\beta T \to P$ are updated, while the distributions associated with errors are not.

4 Stre-GCE in the psycho-therapeutic context: Results and Discussion

We consider each session as a domain in which the actors affect one another in a specific way. At the considered domain of the interaction, we assume that only one of the actors influences the other one. Specifically, this means that sessions labelled

by odd integers preserve one of the models, say (8) while updating the other one (7). Conversely, sessions labelled by even integers preserve (7) and update (8).

The Stre-GCE algorithm described in previous sections gives the estimates shown in Table 1: they represent a quantitative response of the patient to the psychotherapist and vice-versa.

	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}_{P\to T}$	0,803		1,030		0,741		0,602		0,844	
$\hat{\beta}_{T\to P}$		0,643		0,719		0,952		0,991		0,746
	11	12	13	14	15	16	17	18	19	20
$\hat{\beta}_{P\to T}$	0,867		0,850		0,772		0,977		0,921	
$\hat{\beta}_{T\to P}$		0,979		1,066		0,661		0,884		0,722

Table 1 Stre-GCE estimates of model parameters β_{PT} and β_{TP} in (7)-(8).

Results show fluctuations between the two interactans. As the psychotherapeutic relationship comprises moments of engagement and disengagement, these findings are not surprising [7, 10]. In summary, the Stre-GCE algorithms proves to be a viable means to further explore the psychotherapy interaction via behavioral data.

Acknowledgements We are grateful to Les Greenberg for providing the transcript of this case.

References

1. Angelelli, M. (2017). "Tropical limit and a micro-macro correspondence in statistical physics". *J. Phys. A-Math. Theor.* **50**(41), 415202.
2. Angelelli, M., Ciavolino, E., and Pasca, P. (2019). "Streaming generalized cross entropy". *Soft Comput.*, 1-15.
3. Caticha, A., and Giffin, A. (2006). "Updating Probabilities". *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, A. M.-Djafari (ed.), AIP Conf. Proc. **872**, 31.
4. Ciavolino, E., Al-Nasser, A.D. (2009). "Comparing generalised maximum entropy and partial least squares methods for structural equation models". *J. Nonparametr. Stat.* **21**(8), 1017-1036.
5. Ciavolino, E., Calcagnì, A. (2015) "Generalized cross entropy method for analysing the SERVQUAL model". *J. Appl. Stat.* **42**(3), 520–534.
6. Ciavolino, E., Carpita, M. (2015) "The GME estimator for the regression model with a composite indicator as explanatory variable". *Qual. Quant.* **49**(3), 955–965.
7. Gelo, O. C. G., and Salvatore, S. (2016). "A dynamic systems approach to psychotherapy: A meta-theoretical framework for explaining psychotherapy change processes". *J. Couns. Psychol.* **63**(4), 379.
8. Golan, A., Judge, G.G., Miller, D.: *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, New York (1996).
9. Golan, A.: *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*. New York, NY: Oxford University Press, 2018.
10. Gumz, A., Kästner, D., Geyer, M., Wutzler, U., Villmann, T., and Brähler, E. (2010). "Instability and discontinuous change in the experience of therapeutic interaction: An extended single-case study of psychodynamic therapy processes". *Psychother. Res.* **20**(4), 398-412.
11. Jaynes, E.T. (1957). "Information theory and statistical mechanics". *Phys. Rev.*, **106**(4), 620.
12. Jaynes, E.T. (1968). "Prior probabilities". *IEEE T. Syst. Sci. Cyb.*, **4**(3), 227–241

Statistical modelling in environmental epidemiology

A Time Varying Coefficient Model to Estimate the Short-Term Effects of Air Pollution on Human Health

Un Modello con Coefficienti Variabili nel Tempo per la Stima degli Effetti di Breve Periodo dell'Inquinamento Atmosferico sulla Salute Umana

Valentini Pasquale, Ippoliti Luigi and Grazian Clara

Abstract Misspecification in regression models can lead to biased coefficients and, in turn, to incorrect inference. In this paper we consider the problem of omitted variable bias in a regression framework where we estimate the relationship between air pollution and health data. In this context, we show that the problem of the omitted variable bias of OLS estimators can be overcome by using a time-varying coefficient model. By relying on a state-space representation, some different model specifications are presented.

Abstract *Alcuni casi di errata specificazione dei modelli di regressione possono portare a coefficienti distorti e, a sua volta, a un'inferenza errata. In questo lavoro consideriamo il problema della distorsione derivante dall'omissione di una o più variabili in un contesto di regressione in cui si vuole stimare la relazione tra inquinamento atmosferico e salute umana. In particolare, si dimostra che il problema può essere affrontato usando un modello state-space a coefficienti variabili nel tempo. Per il modello proposto vengono presentate alcune parametrizzazioni corrispondenti a specifici interventi correttivi della distorsione.*

Key words: Hierarchical model, spatio-temporal model, time varying coefficients

1 Introduction

A rich literature on health care research has provided substantial statistical evidence of the adverse health effects associated with air pollution. Most of the studies are

Valentini Pasquale
DeC, University "G.d'Annunzio" of Chieti-Pescara e-mail: pasquale.valentini@unich.it

Ippoliti Luigi
DeC, University "G.d'Annunzio" of Chieti-Pescara e-mail: luigi.ippoliti@unich.it

Grazian Clara
DeC, University "G.d'Annunzio" of Chieti-Pescara, e-mail: clara.grazian@unich.it

usually based on time series models, developed both in single and multi-sites frameworks (see, for example, [1]).

Time series studies allow to estimate associations between day-to-day variations in air pollution concentrations and day-to-day variations in adverse health outcomes. In this context, the nature of the data make risk estimation challenging, requiring complex statistical methods sufficiently sensitive to detect effects that can be small relative to the combined effect of other time-varying covariates. It is also common for researchers to be confronted with situations where unobservability of variables, or unavailability of data, force them to omit relevant variables correlated with both pollutants and the outcome (e.g. measured confounders). Confounding by unmeasured variables ([1]) also remains a large concern in this framework. Influenza and respiratory infections, for example, may produce seasonal and long-term trends in health data and not accounting for these potential confounders can lead biased conclusions on the effect of exposure on health.

Though the omitted variable bias (OVB) is well known and has been discussed for decades, the mechanics of OVB are not yet fully understood. In the next Section, we discuss the use of time-varying coefficient (TVC) models which results useful to improve inference in the presence of omitted variables. The theoretical framework considers the possibility of using a set of *driving* variables as well as a state-space representation of the model. Some model specifications are presented.

2 Model Specification

Assume that Y is a spatio-temporal process observed at temporal instants $t = 1,2,\ldots,T$ and generic sites, $\mathbf{s}$, located within a fixed region $\mathscr{D}_y$. Usually, health data (Y) are collected over time in a fixed study region with N locations typically in the form of mortality and morbidity counts or hospital admissions, coded according to the type of disease (e.g. cardiovascular, acute respiratory, etc).

The model is based on Poisson conditional distribution

$$Y(\mathbf{s},t)|Y^*(\mathbf{s},t) \overset{ind}{\sim} Poi[\exp(Y^*(\mathbf{s},t)]$$

with the logarithm of the conditional mean given by

$$Y^*(\mathbf{s},t) = \beta_0 + \beta_1 X(\mathbf{s},t) + \sum_{p=1}^{P} \alpha_p Z_p(\mathbf{s},t) + \varepsilon(\mathbf{s},t) \tag{1}$$

where $X(\mathbf{s},t)$ is the pollutant, $Z_p(\mathbf{s},t)$ are time-varying covariates to control for measured and unmesaured confounders, and $\varepsilon(\mathbf{s},t)$ is assumed to be Gaussian with zero-mean and variance σ_y^2 constant over space and time. Of course, it is also possible to introduce a spatial dependence in $\varepsilon(\mathbf{s},t)$.

Without loss of generality, assume $P = 1$ so that

$$Y^*(\mathbf{s},t) = \beta_0 + \beta_1 X(\mathbf{s},t) + \alpha_1 Z(\mathbf{s},t) + \varepsilon(\mathbf{s},t). \tag{2}$$

Our interest lies in the estimation of the linear exposure-outcome relationship β_1, conditional on Z.

By using only predictor X it is known that the OLS estimator b_1 of β_1 is biased:

$$E(b_1) = \beta_1 + \alpha_1 \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X(\mathbf{s}_i,t)-\bar{X})(Z(\mathbf{s}_i,t)-\bar{Z})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X(\mathbf{s}_i,t)-\bar{X})^2} \tag{3}$$

that is, it is equal to β_1 in expectation only if $\sum_{i=1}^{N}\sum_{t=1}^{T}(X(\mathbf{s}_i,t)-\bar{X})(Z(\mathbf{s}_i,t)-\bar{Z}) = 0$ (the omitted variable is uncorrelated with X), or if $\alpha_1 = 0$. In any other case we have to face with the OVB problem. In epidemiolologic-based time-series studies, various covariates, such as temperature, humidity, or ozone, are likely to be confounders because they do co-vary with pollution. Therefore, if the model is specified incorrectly (either because the confounding variables are unknown or because the data do not exist) we can lead to wrong conclusions on the effect of exposure on hospital admissions.

Swamy and Metha ([2]) and, subsequently, Granger ([3]) state that any misspecification problem can be solved by relying on a model which is linear in variables, but with time varying coefficients. The implication of this result is that, even if we do not know the correct specification form of a relationship, we can always represent this relationship as the following TVC model

$$Y^*(\mathbf{s},t) = \beta_0(t) + X(\mathbf{s},t)\beta_1(t) + \varepsilon(\mathbf{s},t). \tag{4}$$

To make TVC estimation fully operational, we specify the following state equation for the time varying coefficients:

$$\beta(t) = \Phi\beta(t-1) + \Gamma\mathbf{C}(t) + \omega(t) \tag{5}$$

where $\beta(t) = [\beta_0(t)\ \ \beta_1(t)]'$, Φ denotes a state-transition matrix for the unobserved components, $\mathbf{C}$ is a vector $(r \times 1)$ containing a set of *driving variables* which are correlated with the omitted variable $Z(\mathbf{s},t)$, Γ is a $(2 \times r)$ coefficient matrix, and $\omega(t)$ represents the (2×1) vector of temporally uncorrelated error terms with distribution $N(\mathbf{0},\mathbf{Q})$.

Equations (4) and (5) may be rewritten as

$$Y^*(\mathbf{s},t) = \beta_0 + \delta_0(t) + X(\mathbf{s},t)\beta_1 + X(\mathbf{s},t)\delta_1(t) + \varepsilon(\mathbf{s},t) \tag{6}$$

$$\delta(t) = \Phi\delta(t-1) + \Gamma\mathbf{C}(t) + \omega(t) \tag{7}$$

where $\delta(t) = [\delta_0(t) \;\; \delta_1(t)]'$, $\delta_0(t) = \beta_0(t) - \beta_0$ and $\delta_1(t) = \beta_1(t) - \beta_1$.

Considering equations (6) and (7), we note that imposing some restrictions on the parameters it is possible to obtain three nested models. In particular, imposing $\mathbf{Q} = \mathbf{0}$, $\Phi = \mathbf{I}$, $\Gamma = \mathbf{0}$ and the initial state $\delta(0) \sim N(\mathbf{0}, \mathbf{0})$, the regression coefficients remain constant for all periods, and accordingly we have the traditional OLS solution.

The second specification, known as the Random Coefficient (RC) models, is obtained by setting $\Phi = \mathbf{0}$ and $\Gamma = \mathbf{0}$.

Finally, fixing $\mathbf{Q} = diag(0, q_{2,2})$, $q_{2,2} > 0$, , $\Phi = \mathbf{I}$, $\Gamma = \mathbf{0}$ and the initial state $\delta(0) \sim N(\mathbf{0}, \mathbf{0})$, the Time Varying Parameters (TVP) model is obtained.

The general model given in equations (6) and (7) is developed specifying the prior distributions of all hyperparameters within a state-space framework. The full probabilistic inference for the parameters is facilitated by a Markov chain Monte Carlo (MCMC) scheme for multivariate dynamic systems.

The performance in terms of bias reduction for the general and the nested models will be tested trough both an extensive simulation study and real data in an extended version of this paper.

References

1. Peng, R.D. and Dominici, F. and Louis, T.A.: Model choice in time series studies of air pollution and mortality. J. R. Stat. Soc. Ser. A. Stat. Soc., **169**, 179–203 (2006)
2. Swamy, P. A. V. B. and Mehta J. S.: Bayesian and Non-Bayesian Analysis of Switching Regressions and of Random Coefficient Regression Models. Journal of the American Statistical Association, **70**:351a, 593–602 (1975)
3. Granger, C. W. J.: Nonlinear models: Where do we go nextTime-varying parameter models?. Stud. Nonlinear. Dyn. E., **12**(3), 1–9 (2008)

Joint Analysis of Short and Long-Term Effects of Air Pollution

Analisi congiunta degli effetti a breve ed a lungo termine dell'inquinamento dell'aria

Annibale Biggeri[1], Dolores Catelan[1], Giorgia Stoppa[1], Corrado Lagazio[2]

Abstract We present a bivariate Bayesian space-time geostatistical model for exposure assessment and disease risk estimation. We use data from a panel study of 113 children on respiratory health in the high risk area of Valle del Mela (Sicily, IT). Data and gaseous pollutants were collected on 12 weeks in 21 locations 2007–2008. The model consists in an exposure model to predict pollutant concentrations at children's residential addresses; a bivariate disease model. The original features are the joint specification of a spatial long-term effect and a spatiotemporal short-term effect of the pollutant concentrations and uncertainty propagation.

Abstract *Presentiamo un modello geostatistico bayesiano spazio-temporale bivariato per la predizione dell'esposizione e la stima del rischio di malattia, applicato ai dati di uno studio longitudinale sulla salute respiratoria di 113 bambini dell'area a rischio Milazzo-Valle del Mela. Dati sanitari e inquinanti gassosi furono misurati su 12 settimane in 21 postazioni tra 2007-2008. Il modello include un modello predittivo dell'esposizione e un modello bivariato per lo stato di salute. L'originalità sta nella specificazione congiunta di un effetto a breve ed un effetto a lungo termine dell'inquinante e nella corretta propagazione dell'incertezza.*

Key words: Bayesian model-based geostatistics, environmental epidemiology

[1] Annibale Biggeri, Department of Statistics, Computer Science, Applications "G. Parenti" University of Florence; email: annibale.biggeri@unifi.it

Dolores Catelan, Department of Statistics, Computer Science, Applications "G. Parenti"University of Florence; email: dolores.catelan@unifi.it

Giorgia Stoppa, Department of Statistics, Computer Science, Applications "G. Parenti"University of Florence; email: giorgia.stoppa@unifi.it

Corrado Lagazio, Department of Economics, University of Genoa; email: corrado.lagazio@unige.it

1 Introduction

In Environmental Epidemiology exposure to pollutants may affect disease risk on a short or very short time lag. The exposure is acting as a precipitating event and some cases are simply anticipated of few days. If the exposure dose is high enough we can observe acute toxicological effects. In the case of chronic degenerative diseases like cardiovascular diseases or cancer, we cannot observe an acute effect because the pathogenesis requires several steps to be completed before the disease become clinically detectable. In this case, to measure a change in disease risk a potentially life-long observation period is needed.

Short-term effects are measured by longitudinal designs with average daily concentration of pollutants as exposure metrics. Long-term effects are measured by population cohort studies in which individuals at different baseline levels of exposure are followed-up over decades. The exposure metrics are annual averages.

Few studies attempted to joint model short and long-term effects of air pollution on the occurrence of health outcomes (Kloog (2013)). The methodological approach was to specify a generalized linear mixed model including in the linear predictor the two exposure metrics – daily and annual averages – and enrolling several population cohorts over a wide geographical region. Currently, most efforts are devoted to create huge datasets covering millions of people over long calendar periods. Exposures is obtained as daily averages predicted at a fine geographical resolution by integrating remote sensing data via machine learning and ensemble modelling (Shtein (2019)). We expect that joint modelling be soon a data science hot topic.

In this paper we present a joint Bayesian space-time geostatistical model for exposure assessment and disease risk estimation. The originality of this modelling is the joint specification of a spatial long-term effect and a spatiotemporal short-term effect of the pollutant concentrations and an appropriate uncertainty propagation.

2 Motivating example

The data came from an epidemiological project on the National Remediation Site of Milazzo-Valle del Mela funded by WHO and Regione Siciliana (Biggeri (2014)). Main sources of pollution were oil refinery, petrochemical plant, oil-powered energy plant, foundries, seaport and traffic. In this paper we used a panel study on 113 susceptible children (compliance rate 73%) selected from all 2506 children attending the primary schools of the area. We analysed for the present study inflammatory symptoms (persistent cough). In the study period 21 passive monitors for gaseous pollutants were located in the backyard of each primary school. We will concentrate on Nitrogen Dioxide weekly averages ($\mu g/m^3$) monitored on 12 weeks between November 2007 and April 2008. Baseline information on demographics, respiratory health, lifestyles and risk factors were collected by parents' questionnaires. Weekly diaries were collected during the study period as well as lung function examinations.

3 Methods

We develop a joint bivariate disease and exposure model, within model-based spatio-temporal geostatistics (Diggle and Giorgi (2019)). Below, we introduce the disease, the exposure, the bivariate model and the Cronbach model.

3.1 Disease model

Let assume that $s \in R^d$ be a location – residential address - in *d*-dimensional space and $Y_t(s)$ be a data response variable observed at location *s* on time *t* for subject *i*

$$\{\mathbf{Y}_{it}(s): s \in D\},\ D \subset R^d$$

the disease model can be expressed as

$$\boldsymbol{Y}_{it}(s) = g\left(\beta_0 + \beta_1 \tilde{X}_{it}(s) + \gamma W_{it}(s) + S(s,t) + Z_i(s) + Z_{it}(s)\right)$$

where $g(\cdot)$ is the link function, $\tilde{X}_{it}(s)$; $W_{it}(s)$ are predicted exposure and observed fixed confounder(s), β; γ regression coefficients, S(s,t) a continuous spatio-temporal process and $Z_i(s)$, $Z_{it}(s)$ random error terms.

3.2 Exposure model

Let assume now that $s' \in R^d$ be a location - ambient monitoring site - in *d*-dimensional space and $X_t(s')$ be a data exposure value observed at s' on time *t*

$$\{X_t(s'): s' \in D\},\ D \subset R^d$$

the **exposure model** can be expressed as

$$X_t(s') = g\left(\beta_0' + \gamma W_t'(s') + S'(s',t) + Z_t'(s')\right)$$

where $X_t(s')$; $W'_t(s')$ are exposure and confounder(s), γ' regression coefficients, S'(s',t) a continuous spatio-temporal process and $Z'_t(s')$ a random error term.

The disease and the exposure models are linked by the predictive distribution. To obtain $\tilde{X}_{it}(s)$ - the exposure value at location *s* on time *t* – we use the known covariate vector $W'_{it}(s)$ and the predictive distribution from a Bayesian Gaussian Spatio-temporal Exponential model. MCMC methods are used taking advantage of the posterior samples and the conditional normal distributions arising from the joint multivariate distribution of $X_{it}(s')$ and $X_{it}(s)$ (Banerjee, Carlin and Gelfand (2014)).

3.3 Short and Long-term effects modelling

The regression coefficient of interest are respectively short-term – associated to the exposure at time *t* and location *s* - and long-term – associated to the average exposure at location *s*.

The disease model could be expressed in two alternative forms, the Firebaugh model (not shown) and the Cronbach model:

$$Y_{it}(s) = g\left(\beta_0 + \beta_1\left(\tilde{X}_{it}(s) - \tilde{X}_i(s)\right) + \beta'_2\tilde{X}_i(s) + \gamma W_{it}(s) + S(s,t) + Z_i(s) + Z_{it}(s)\right)$$

In the Cronbach model under specific assumptions on confounders, the regression coefficient β'_2 is equal to the coefficient of the model on the aggregated data:

$$Y_i(s) = g\left(\beta''_0 + \beta'_2\bar{X}_i(s) + \gamma'' W_i(s) + S''(s) + Z_i(s)\right)$$

This equality can be used to specify a computationally more robust bivariate model which takes the form - $\eta_{2i}(s)$ denoting the linear predictor of $Y_{2i}(s)$:

$$\begin{Bmatrix} Y_{1it}(s) \\ Y_{2i}(s) \end{Bmatrix} = \begin{Bmatrix} g\left(\beta_{0,i} + \beta_1\left(\tilde{X}_{it}(s) - \bar{X}_i(s)\right) + \gamma W_{it}(s) + S(s,t) + Z_{it}(s)\right) \\ g\left(\beta''_0 + \beta'_2\bar{X}_i(s) + \gamma'' W_i(s) + S''(s) + Z_i(s)\right) \end{Bmatrix}$$

$$\beta_{0,i} = \left(\eta_{2i}(s) + Z_i(s)\right)$$

where we can recognize the Cronbach model specification and the specification of a model on the aggregate data – i.e. averaged over time.

4 Results

We fitted a hierarchical Bayesian model. The exposure model was:

$$X_t(s') = exp\left(\beta'_0 + \gamma W'_t(s') + S'(s',t) + Z'_t(s')\right)$$

$$S'(s',t) = \pi(s') + \pi(t) + \psi(s',t)$$

$$\pi(s') \sim GP\left(\mu(s'),\ {}^{s}\sigma^2, \rho(d)\right)$$

$$\pi(t) \sim G\left(\bar{\pi},\ {}^{t}\lambda\ \ n_t\right)$$

$$\psi(s',t) \sim GP\left(\mu(s',t),\ {}^{st}\sigma^2, \rho(d)\right)$$

where $\pi(s')$ is a Gaussian process (GP) with covariance matrix $\Sigma_{ij} = {}^{s}\sigma^2 \cdot \rho(d)$ induced by the correlation function $\rho(d) = exp(-\varphi \cdot d)$ that depends on the distance d between pairs of points and by the spatial variance parameter ${}^{s}\sigma^2$; φ is chosen to get almost zero correlation at the maximum distance (9.32 Km) and almost one at the minimum distance (0.103 Km); $\pi(t)$ is a conditional autoregressive term of order 1; $\bar{\pi}$ is the mean of the (t-1)-th and (t+1)-th terms, ${}^{t}\lambda$ is the precision and n_t the number of adjacencies; $\psi(s',t)$ is a Gaussian process with time-dependent mean $\mu(s',t) \sim G\left(\bar{\mu},\ {}^{\mu.st}\lambda\ \ n_t\right)$ and covariance matrix according to the specification of the spatial main effect $\pi(s')$. A GIS for the study area was developed using as data-layers: digital elevation, slope and aspect of the study area (spatial resolution = 40 m); land use (Corinne Land Cover), population density (LandScan 2010) and road map (TeleAtlas). The selection of the covariates to be included in the model was done outside the Bayesian model through both an automatic stepwise regression and information from the literature. The selected covariates $W'_t(s')$ were: altitude above the sea level, distance to the main roads, population density and land use covariates calculated as the proportion of surface area with the characteristic over a buffer of 500 meters radius (industrial, urban and

suburban land cover). Information of all these covariates was obtained for each residential address and pollution monitor location (Vicedo-Cabrera (2013)).
The multivariate disease model takes the form:

$$\begin{Bmatrix} Y_{1,it}(s) \\ Y_{2,i}(s) \end{Bmatrix} = \begin{Bmatrix} logit\left(\beta_{0,i} + \beta_1\left(X_{it}(s) - \bar{X}_i(s)\right) + \gamma W_{it}(s)\right) \\ logit\left(\beta''_0 + \beta'_2 \bar{X}_i(s) + \gamma'' W_i(s)\right) \end{Bmatrix}$$

$$\beta_{0,i} = \left(\eta_{2,i}(s) + Z_i(s)\right)$$

We consider inflammatory symptoms as bivariate 0-1 responses ($Y_{1,it}(s)$cough or phlegma outside common cold in the *t*-th measurement week; $Y_{2,i}(s)$ cough or phlegma outside common cold for more than three months in the last 12 months - as defined in the ISAAC study). As time-dependent confounder $W_{it}(s)$ we include use of β_2 agonist in the *t*-th measurement week, and as subject specific confounders $W_i(s)$ we include age, sex, parents education, mother smoking, mould or dampness in the children bedroom, familiarity for asthma. The hierarchical Bayesian model specification includes two joint Bernoulli likelihoods for the two responses, weakly informative priors for all regression coefficients, $Z_i(s)$ as a Gaussian noise. The joint exposure and disease model provides uncertainty propagation and potentially adjust for the measurement error bias in the regression coefficients. In figure 1 we report as example the average predicted Nitrogen Dioxide concentration surface. In figure 2 we show the posterior density for the long-term coefficient (log Odds Ratio) of inflammatory symptoms for a 10 μg/m^3 increase in the annual average of Nitrogen Dioxide. In light yellow we show for comparison the posterior density if we do not consider uncertainty in the predicted exposures.

Figure 1: Average Nitrogen Dioxide concentration surface. Results of the exposure model (see text). Milazzo – Valle del Mela study (Biggeri et al. 2014).

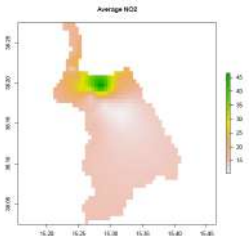
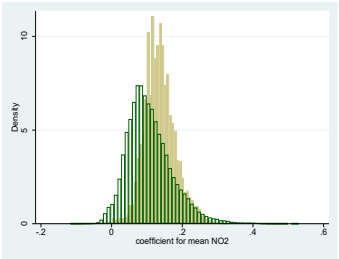


Figure 2: Posterior density of mean NO2 coefficient (green). In light yellow the posterior density of a model which does not consider the uncertainty in the exposure model (see text).



Discussion and Conclusions

Alternatively to our modelling choice, a Firebaugh multilevel model could have been specified. Fitting it on our data gave very unstable results on estimating variance components, multimodality of posterior distributions and correlation among model parameters which affect proper mixing in MCMC algorithm. A multinomial model, reparametrizing the bivariate response in a trinomial response could also have been considered. Fitting this model results in effect estimates which may partially explain the bimodality in the posterior distributions of the multilevel model. Different space-time interaction specification can be considered in the exposure model, but it was not the case in our data, which shows a very stable time pattern (see figure 1). The effect estimates are consistent with the literature. In conclusion we stress that accounting for prediction uncertainty provided a more accurate effect estimate. Short-term and long-term effects are difficult to jointly model and we present a robust approach which take advantage of aggregate data. This modelling can be connected with the literature on hybrid designs in environmental epidemiology (e.g. Smoot (2015)).

Acknowledgments

This study was funded by the MUR INSIDE project (PRIN 20152T74ZL).

References

1. Banerjee S et al. (2014). Hierarchical modeling and analysis for spatial data. Crc Press.
2. Biggeri A et al. (2014). Epidemiological investigations of air pollution and asthma symptoms in children living in the Milazzo–Valle del Mela high-risk area. Chp 11 in Mudu, Terracini, Martuzzi (eds), WHO Regional Office for Europe, ISBN: 978 9 289 05005 0.
3. Diggle PJ and Giorgi E (2019). Model-Based Geostatistics for Global Public Health, Crc Press.
4. Kloog I et al. (2013). Long- and short-term exposure to PM2.5 and mortality: using novel exposure models. Epidemiology 24(4):555-61.
5. Shtein A et al. (2019). Estimating Daily PM(2.5) and PM(10) over Italy Using an Ensemble Model. Environ Sci Technol. 2019 Dec 10. [Epub ahead of print]
6. Smoot E and S. Haneuse S (2015). On the Analysis of Hybrid Designs that Combine Group- and Individual-Level Data. Biometrics 71: 227–236
7. Vicedo-Cabrera AM et al. (2013). A Bayesian kriging model for estimating residential exposure to air pollution of children living in a high-risk area in Italy. Geospatial health, 8(1): 87-95.

Statistical Modelling of Scientific Evidence for Forensic Investigation and Interpretation

DNA mixtures with related contributors

Misture di DNA con contributori correlati parentalmente

Peter J. Green and Julia Mortera

Abstract In both criminal cases and civil cases there is an increasing demand for the analysis of DNA mixtures where relatives are involved. The goal might be to identify the contributors to a mixture where the donors may or may not be related, or to determine the relationship between individuals based on a DNA mixture. Here we present examples that might occur in forensic casework, where we wish to determine which among a group of relatives involved in criminal activity left a mixed DNA trace at the crime scene.

Abstract *Sia nei processi penali che civili vi è una crescente domanda per l'analisi di misture di DNA in cui vi siano tra i potenziali contributori dei parenti. L'obiettivo potrebbe essere quello di identificare se i contributori alla mistura possono o meno essere correlati parentalmente, o di determinare la relazione di parentela tra individui sulla base di una mistura di DNA. In questo lavoro presentiamo due esempi che potrebbero verificarsi in casistica forense reale, dove si vuole determinare quali tra un gruppo di parenti coinvolti in attività criminali abbiano lasciato una traccia mista di DNA sulla scena del crimine.*

Key words: coancestry, deconvolution, disputed relationship, identity by descent, kinship, DNA mixtures, likelihood ratio

1 Introduction

DNA is now routinely used in criminal and civil investigations. DNA samples are of varying quality and therefore present complex problems for their interpretation.

Peter J. Green
UTS, Sydney, Australia and University of Bristol, UK. e-mail: P.J.Green@bristol.ac.uk

Julia Mortera
Università Roma Tre, Via Silvio D'amico 77, 00145 Roma. e-mail: julia.mortera@uniroma3.it

The identification of the DNA composition of mixed samples gives rise to a wide range of challenging statistical questions, some associated with uncertainties and artefacts in the measurement processes and some associated with population genetic variations. In both criminal cases and civil cases based on relationship inference there is an increasing demand for the analysis of DNA mixtures where relatives are involved. The goal might be to identify the contributors to a mixture where the donors may or may not be related, or to determine the relationship between individuals based on a mixture. Here we use probabilistic genotyping methods for DNA mixtures, under hypotheses about the relationships among contributors to the mixture and to other individuals whose genotype is available.

The basis for any model-based DNA mixture analysis is a joint model for the peak heights $\mathbf{z}$ in the electropherogram and genotypes represented as allele counts $\mathbf{n}$ $p(\mathbf{n},\mathbf{z}|\psi) = p(\mathbf{n}) \times p(\mathbf{z}|\mathbf{n},\psi)$, having parameters $\psi = (\phi,\rho,\xi,\eta)$. Given a hypothesis on the DNA mixture contributors, the database allele frequencies, the parameters ψ, the DNA mixture model consists of two components: (a) the joint distribution $p(\mathbf{n})$ of the contributors' genotypes; (b) the conditional distribution $p(\mathbf{z}|\mathbf{n},\psi)$ of the peak heights as observed in the electropherogram, given the genotypes. We base the analysis of the DNA mixture on the model described in [1]. This model takes fully into account the peak heights and the possible artefacts, like stutter and dropout, that might occur in the DNA amplification process. We refer to the review on DNA mixtures by [8] for further details.

In the standard case, unknown contributors to the mixture are assumed drawn i.i.d. from the gene pool – multinomial sampling. When contributors are related, there is positive association between their genotypes. [6] extend the peak height model above to allow for inference about the relationships between contributors to a DNA mixture with unknown genotype and other individuals of known genotype: a simple example would be testing whether a contributor to a mixture is the father of a child of known genotype. The evidence for these relationships is presented as the likelihood ratio LR for a hypothesis H_p about the specified relationship versus the null hypothesis H_0, that there is no relationship, so the person of interest is taken to be a random member of the population.

Here we extend the analysis to different scenarios where:

(i) we have the genotypes of some of the actors, the others being contributors to a mixture;
(ii) we do not have the genotypes of the actors but they are contributors to one or more mixtures;
(iii) a combination of (i) and (ii).

In all cases were one unknown contributor to the mixture is higher up the relationship pedigree than the other contributor we need to test whether the major unknown contributor U_1 is a descendant or an ancestor of U_2 – the minor unknown contributor. Examples 1 and 2 in Section 2 illustrate scenarios of type (ii) and (iii), where we make inference about the two-way relationships between two mixture contributors with and without information on relatives' genotypes.

We also present an example where a group of relatives are involved in criminal activity. This case can occur when a family is engaged in a joint criminal activity and a DNA mixture might be found on, *e.g.* a getaway car, a balaclava, banknotes, a crowbar, a gun *etc.* that might have been handled by some members of the family.

Software used to analyse the examples is the new `KinMix R` package [4] that generalises the `DNAmixtures R` package [2] to allow for modelling DNA mixtures with related contributors.

2 Relationships among contributors

In both civil and criminal cases, we sometimes need to allow for dependence between genotypes of contributors. This possibility will obviously affect the evidential value of the mixture. We distinguish two cases: populations with high relatedness, due to inbreeding, and specific close relationships, *e.g.* father and son, or brothers engaged in a joint criminal activity. In both cases, the genotypes of two or more actors will be positively associated through identity by descent (IBD), the phenomenon that two genes may be identical because they are copies of the same ancestor gene, rather than being independent draws from the 'gene pool'. The phenomenon is the same in both cases, but they require different modelling approaches.

(a): Populations with high relatedness – ambient IBD

As observed in [5] the same probabilistic model for the joint distribution of multiple genes arises in a simple model for uncertainty in allele frequencies, in which the true allele frequencies are treated as unknowns with a Dirichlet distribution and the database used for calculation regarded as a multinomial sample from these true frequencies.

In discussion of [1], [3] observes that when genotypes are represented by allele counts arrays $(n_{ia})_{a=1}^{A}$, the number of alleles a of individual i, the conditional distributions over allele counts have Beta–Binomial conditional distributions:

$$n_{ia}|(n_{j.})_{j=1}^{i-1}, \{n_{ib}, b<a\} \sim \mathtt{BB}((2-n_{i,<a}),(q_a+n_{<i,a}),(q_{>a}+n_{<i,>a})),$$

where q_a are the database allele frequencies. The new `KinMix` software contains code to implement this model.

(b): Specific close relationships

Two non-inbred actors, can either share none, one or both of their genes IBD with probabilities $\kappa = (\kappa_0, \kappa_1, \kappa_2)$, $\sum_i \kappa_i = 1 \;\; i = 0, \cdots, 2$ determined by the pedigree [9]. Table 1 gives the values of κ for various pairwise relationships between non-inbred

individuals. For example, half-sibs have $(\kappa_0, \kappa_1, \kappa_2) = (0.5, 0.5, 0)$ and quadruple half-first cousins have $(\kappa_0, \kappa_1, \kappa_2) = (7/32, 7/16, 1/32)$. The IBD probabilities are used in the new `KinMix` to implement inference about DNA mixtures with related contributors. `KinMix` also handles situations where more than two contributors are related, and allows them to be inbred.

Example 1: Testing relationships among DNA mixture contributors

In this example, we test whether two contributors to a mixture U_1 and U_2 have a specific relationship, with respect to the hypothesis that all contributors are unrelated, *i.e.* H_p*: U_1 has relationship R to U_2 versus* H_0*: U_1 is unrelated to U_2*. Table 2 shows the results for a group of relationships $R = \{$parent-child; half-sibs; cousins; quadruple-half-cousins$\}$.

We first simulate several DNA mixtures using `pcrsim` [7], which simulates the DNA amplification process. Each DNA mixture is simulated from the pedigrees in R where the two actors have a specified "true" relationship. Table 2 gives the $\log_{10}$LR for testing H_p *versus* H_0.

Table 1 Values of κ for various relationships between two non-inbred individuals.

Pairwise relationships	κ_0	κ_1	κ_2
unrelated	1	0	0
parent-child	0	1	0
sibs	0.25	0.5	0.25
quadruple-half-cousins	0.5312	0.4375	0.0312
half-sibs	0.5	0.5	0
cousins	0.75	0.25	0

Table 2 Median $\log_{10}$LR over 4 replicates for H_p: U_1 and U_2 have relationship R versus H_0: U_1 and U_2 are unrelated.

	"True" relationship				
Relationship of U_1 to U_2 under H_p	parent-child	sibs	half-sibs	cousins	quadruple half-cousins
parent-child	3.76	2.01	-Inf	-60.56	-79.93
sibs	2.16	6.50	3.05	-2.33	0.37
half-sibs	2.54	3.65	1.91	-0.54	0.44
cousins	1.60	2.69	1.28	-0.12	0.44
quadruple-half-cousins	2.31	4.38	2.12	-0.57	0.62

The highest median $\log_{10}$LR's are attained along the diagonal for close relationships, like parent-child, $\log_{10}\text{LR} = 3.76$ and sibs $\log_{10}\text{LR} = 6.5$, showing that these are correctly identified. However, more distant relationships among contributors to a mixture are harder to distinguish. When there is additional information, like the

genotypes of some individuals potentially related to mixture contributors the evidence becomes much stronger, as we will see in Example 2.

Example 2: Family involved in a joint criminal activity

This case concerns a family with 6 male children (some having a different mother) hypothesised as being engaged in a joint criminal activity. A two-person DNA mixture was found on the door handle of a getaway car abandoned after a crime. DNA samples were taken from two brothers B1 and B2, the other 4 children having escaped the country. A first analysis excluded that B1 or B2 were contributors to the mixture. The question then was to test which if any of the escaped brothers or half brothers were contributors to the incriminating mixture. Table 3 gives the $\log_{10}$LR for various different hypotheses, such as, H_p: the major contributor U_1 is the brother (half-brother) of B1 and B2 (or of B1, or B2 alone) against a defence hypothesis H_0: B1 and/or B2 are unrelated to U_1. Similarly for the second contributor U_2.

The results in Table 3 show that it is likely that the major contributor U_1 to the mixture was the brother of B1 and the half-brother of B2. The $\log_{10}$LR for testing this was 9.86, whereas it seems plausible that the second contributor to the mixture is the brother of B2.

Table 3 The $\log_{10}$LR for testing different hypotheses H_p on the potential relationship between the contributors to a two-person DNA mixture and the typed individuals B1 and B2 versus H_0 the null hypotheses of no relationship.

	$\log_{10}$LR for H_p vs. H_0	
Relationship in H_p	U_1	U_2
brother of both B1 and B2	7.65	10.64
brother of B1	6.86	2.73
brother of B2	1.71	8.77
half-brother both B1 and B2	6.02	7.33
half-brother of B1	1.10	2.22
half-brother of B2	5.37	5.38
brother of B1 and half brother of B2	9.86	–

3 Conclusions

We have shown a wide range of relationship inference problems, where one or more actors appear as contributors to a DNA mixture, can be handled coherently. We handle relationships among contributors, and between contributors and typed individuals. This approach provides a toolkit for new problems of this kind that can arise in forensic casework.

References

1. Cowell, R.G., Graversen, T., Lauritzen, S.L., Mortera, J.: Analysis of DNA mixtures with artefacts (with discussion). Journal of the Royal Statistical Society: Series C **64**, 1–48 (2015)
2. Graversen, T.: DNAmixtures: Statistical Inference for Mixed Traces of DNA (2013). URL dnamixtures.r-forge.r-project.org/. R package version 0.1-4. http://dnamixtures.r-forge.r-project.org
3. Green, P.J.: Contribution to discussion of analysis of DNA mixtures with artefacts (with discussion) by cowell et al. Journal of the Royal Statistical Society: Series C **64**, 1–48 (2015)
4. Green, P.J.: KinMix: DNA mixture analysis with related contributors (2020). URL https://petergreenweb.wordpress.com/kinmix2-0/. R package 2.0, `https://petergreenweb.wordpress.com/kinmix2-0`
5. Green, P.J., Mortera, J.: Sensitivity of inferences in forensic genetics to assumptions about founder genes. Annals of Applied Statistics **3**, 731–763 (2009). DOI 10.1214/09-AOAS235
6. Green, P.J., Mortera, J.: Paternity testing and other inference about relationships from dna mixtures. Forensic Science International: Genetics **28**, 128–137 (2017). `http://dx.doi.org/10.1016/j.fsigen.2017.02.001`
7. Hansson, O.: `pcrsim`: Simulation of the Forensic DNA Process (2017). URL https://github.com/OskarHansson/pcrsim. R package version 1.0.2, `https://github.com/OskarHansson/pcrsim`
8. Mortera, J.: DNA mixtures in forensic investigations: The statistical state of the art. Annual Review of Statistics and Its Application **7**, 1–34 (2020). `https://doi.org/10.1146/annurev-statistics-031219-041306`
9. Thompson, E.A.: Statistical inference from genetic data on pedigrees. In: NSF-CBMS regional conference series in probability and statistics, pp. i–169 (2000)

Forensic Statistics: How to estimate life expectancy after injury

Statistiche forensi: come stimare l'aspettativa di vita dopo un infortunio

Jane L Hutton

Abstract People who are injured in a traffic, industrial or medical accident, or contract a disease or cancer through negligence of an employer, doctor or local authority, can bring a civil case for compensation. Approaches to establishing causation differ between medical and legal professionals. If responsibility is assigned, the value of compensation will often depend on how long the person is expected to live. In the UK, for common injuries, there are life tables known as the Ogden Tables. For some particular injuries, such as spinal cord or brain injuries, an expert witness might be instructed to provide individual report. Such a report might also take into account other factors, such as alcohol consumption or medical history of diabetes or depression. I will compare the approaches used by actuaries, statisticians and medical doctors to estimating an individual person's life expectancy. The underlying assumptions of independence of different factors, or additive or multiplicative modifiers of mortality rates differ substantially.

Abstract *Le persone che sono vittime di incidenti stradali, industriali o di errori medici, o che contraggono una malattia o un cancro per negligenza del datore di lavoro, di un medico o di un'autorità locale, possono intentare una causa civile al fine di ottenere un risarcimento. Gli approcci utilizzati per stabilire il possibile nesso causale differiscono tra professionisti operanti in campo medico o giuridico. Nel caso di attribuzione di responsabilità, l'ammontare del risarcimento è spesso legato all'aspettativa di vita della persona coinvolta. Nel Regno Unito, per gli infortuni più comuni, si fa ricorso a delle tavole di mortalità note con il nome di 'Ogden Tables'. Per lesioni di una certa gravità, ad esempio lesioni del midollo spinale o lesioni cerebrali, un esperto potrebbe essere incaricato di redigere una perizia individuale. Quest'ultima potrebbe tenere conto di diversi fattori, incluso il consumo di alcol, o di altri elementi caratterizzanti la storia clinica della persona, ad esempio patologie quali il diabete o la depressione. In questo lavoro, verranno messi a confronto i diversi approcci utilizzati da attuari, statistici e medici per stimare l'aspettativa di vita di una persona. Ne emergerà una differenza sostanziale, in merito sia alle*

Department of Statistics
University of Warwick, Coventry, UK e-mail: J.L.Hutton@warwick.ac.uk

ipotesi di indipendenza tra i diversi fattori, che di effetti additivi o moltiplicativi in termini di tassi di mortalità.

Key words: Compensation, injury, life expectancy, methods of estimation

1 Introduction

Accidents are part of life. A road traffic accident might leave a young woman tetraplegic, unable to use her limbs, but still able to learn and pass examinations. An industrial accident might result in a man losing several fingers. A mismanaged birth might leave a baby with cerebral palsy. A factory might expose workers to dangerous chemicals, which increase the risks of cancer. In such cases, if some person or company can be shown to be responsible, the injured person might be given financial compensation. The differences between professions in methods for establishing causation is not considered here (Hutton, 2018). In some countries and situations, compensation is almost automatic. In others, a civil case for damages will be lodged with a court. If adequate care is estimated to cost £100,000 in 2020, and the claimant is expected to live for 10 years, a total sum of £1 million is implied. Of course, this should be adjusted to take into account assumptions about the investment of the lump sum and inflation for the next 10 years.

This article considers how life expectancy is estimated. The two main challenges are finding reliable data or estimates of mortality risks (Hutton, 2018; Pharoah and Hutton, 2006), and deciding on a sensible method of estimation. General population estimates are considered, then smaller datasets which focus on particular populations. Medical doctors methods of estimation are considered before those used by statisticians.

2 Population life tables

In countries with adequate vital statistics registration, population life tables are estimated from age-specific death rates calculated from records of deaths and estimates of the numbers alive derived from censuses. There are some difficulties to address, such as the accuracy of population numbers between censuses and delays in death registration. However, population life tables are standard demographic resources.

National Statistics Offices publish period life tables and cohort life tables. The estimates which determine the tables are the observed central mortality rates at age x, $\hat{m}_x$, given by y_x/N_x, where y_x is the number of deaths of people aged x last birthday, and N_x is the number aged $[x, x+1)$ at the midpoint $x+0.5$. Period life tables use the current age-specific death rates with no changes for the future, so the observed mortality rate of 4.7/1000 per year for a 59 year old woman in 2013-15 would be used for a life expectancy estimate for woman aged 59 in 2020. Cohort life tables

use predictions of death rates. Using a simple 0.5% decrease per year, the mortality rate of 4.7/1000 in 2014 for a woman age 59 decreases by $1.005^6 = 1.0303$ to $4.0/1000$ in 2020. The older a person is, the more observed mortality rates rather than projected rates are used in the calculation of life expectancy.

In the UK, the size of lump sum compensation awarded in personal injury and fatal accident cases is usually informed by particular life tables, the Ogden tables (Government Actuary's Department, 2011). A prescribed discount rate is set by the government. The Ogden tables give multipliers for annual costs by adjusting 2008 cohort life expectancies for different rates of return. The UK has recently revised the methods used to project mortality rates (Dodd et al, 2018).

3 Medical doctor's life expectancy estimates

Lawyers for a claimant who lost several fingers might be content with general population tables, but those acting for parents of a child with severe disabilities due to cerebral palsy will usually seek an expert opinion on life expectancy. In the past, such opinions were sought from medical doctors. However, medical doctors generally lack the relevant statistical and epidemiological knowledge, and do not have resources to follow-up patients to death (Pharoah and Hutton, 2006). A pediatrician is asked to give a life expectancy estimate for a child will not know about adult life. I was called by lawyers already in court in Dublin because a pediatrician's report assumed that a ten year old boy was more likely than not to live to age 15. She then used a research article (Brooks et al, 2014, Table III) which gave estimated additional year life expectancy at age 15 of 18 to argue that the boy would live to age 30 'on balance of probability'. She therefore assumed he would be alive at age 30, again used life expectancy of 19, and gave her final estimate of expected age at death life expectancy of 59, 14 years from age 45 years. The judge had no difficulty in understanding that being likely to live to age 30 'on balance of probability' is not the same as certainty, and that the probability of death before age 30 must be taken into account in the estimate.

A rehabilitation specialist provided many expert witness reports on life expectancy for people who had had serious brain or spinal cord injuries. His approach was to review some of the research literature, and then decide how many years to substract from general population life expectancy for various factors. For an older person, such as a 70 year old man in 2000 with a life expectancy of 13.9 years, there was a possibility that subtracting some years for several factors would mean zero or negative life expectancy. This approach was described as a "top-down" approach. After a judge expressed his disapproval, the specialist changed his approach to deciding on a percentage of remaining life expectancy based on one or two research articles.

Another physician uses a method referred to as the "Rating of Substandard Lives", a widely accepted methodology for the assessment of life expectancy amongst medical insurance companies (Kita, 2006). He also selects factors which

might affect life expectancy, references some research articles and then chooses a mortality risk, such as 200% for smoking, 150% for obesity or 400% for brain injury. These risks are *added* together, and a look-up table used (Bowen-Jones et al, 2014). The natural method of combining percentages is multiplication, not addition. The incoherence of adding or subtracting mortality risks as a method can be seen by considering the effect of subtracting from a baseline mortality 100%, factors for being a non-smoker, say 40%, and for being fit with no family history of disease, say 60%. This gives 100% - 40% - 60% = 0%, so the person is immortal.

No indication is given of how interdependence of risk factors is assessed, or how such dependence should be used to reduce mortality rates. If rates for, say, obesity and diabetes and cholesterol are all included, then there will be double-counting, as increased mortality associated with obesity includes deaths from diabetes and cholesterol related causes.

In a 2018 joint report from medical doctors, estimates are given as feelings: "Dr X feels that there is a reasonable range of 7-12 years of further life expectaiton. Professor Y feels that Mrs A is likely to live around another 19 years (say a range of 18-20 years)."

4 Statistical approaches to individual life expectancy estimates

I use data from a geographically-based cohort of about people with cerebral palsy for estimates (Hutton and Pharoah, 2002). For the most severely affected children, it is possible to estimate the median survival using Kaplan-Meier estimates conditional on the severity of the person's impairments, and his current age. Lawyers often like large data sets, so assume that data on 12,709 clients of the California Department of Developmental Services labelled as having cerebral palsy (Strauss et al, 1998) must be more relevant than smaller data sets. However, the definition of cerebral palsy used to include people was incoherent. The distribution of impairments differed from geographical cohorts, as more severely impaired individuals are more likely to be clients of the state (Hutton et al, 2000). Even if there is a large, high quality patient cohort, it is unlikely that all deaths have been observed. Such cohorts will include people born, or injured, in past decades, we cannot assume there are no secular improvements in mortality rates. In general, there is limited access to good quality data on specific patient groups. Further, life expectancy before injury might need to allow for factors such as obesity or smoking.

An approach which does not modify the age-specific mortality rates, $\hat{m}_x$, is to "rate up," which means assuming that the effect of a medical condition is to increase a person's effective age by, say, 10 years, so that the population life expectancy at 70 years might be used to estimate the life expectancy of a 60 year old heavy smoker. Estimated death rates or relative risks inform the up-rated age.

Common statistical approaches begin with estimates of absolute mortality rates, or excess mortality rates or relative risks, either directly from available datasets, or from published research. These estimates are used to adjust the mortality rates

given in National Statistics Office life tables, and the adjusted mortality rates used to create a customised individual life tables. The choice between period and cohort life tables as baseline is debated. A Californian group claims that there have been no improvements in survival of people with spinal cord injury (Shavelle et al, 2015), therefore period life tables should be used to estimate compensation. The UK courts usually require the use of cohort life tables.

Let $\hat{d}$ be an estimated death rate, and $\hat{r}$ an estimated relative risk of death. The difference between two death rates is an excess death rate. If $\hat{d}$ is large, say $> 1\%$, the excess death rate will be close to the estimated death rate. Age-specific mortality rates, $\hat{m}_x$, can be modified by adding an excess death rate, and using $\hat{m}_x + \hat{d}$ to derived a new life table. Assuming a constant excess death rate, the risk of death *relative* to the general population rate will decrease towards one as age increases, as the general population death rate increases. Alternatively, the mortality rates can be multiplied by a relative risk, so that $\hat{m}_x \times \hat{r}$ determines the customised life table. There is debate as to whether either of these assumptions is appropriate, with Shavelle et al (2015) asserting that a constant relative risk is correct. However, observed relative risks tend to decrease with age (Middleton et al, 2012). Other proposals include log-linear declining relative risk, which requires a choice of age at which to apply the relative risk, and the age at which the log relative risk should reach zero (Strauss et al, 2005). A smooth function of relative risk which tends to an asymptote at unity might be more sensible. A further suggestion is to have proportional life expectancy constant: men might be assumed to have life expectancy 95% of women at all ages.

Medical statisticians have developed guidelines to assist in reporting research which aims to provide reliable predictions for individuals (Collins et al, 2015; Moons et al, 2015). Moons et al (2015) document different types of prediction model studies, and performance measures. An important distinction is made, between apparent performance of a model which is only validated on the data used to define the model, and the performance of that model in future use.

Within forensic practice, there are no common statistical approaches to deciding between different methods of predicting life expectancy for individual people. Shavelle et al (2015) used absolute errors, in years and percentages, for predictions at age 30 only, averaged across general and four patient populations to conclude that log-linear declining relative risk is the best approach.

5 Conclusion

The open issues for estimation of life expectancy to inform legal decisions are sources of reliable estimates of absolute and relative mortality rates, exploring methods for projecting life expectancy and determining which criteria to use to decide which methods are optimal.

References

Bowen-Jones D, Nwaneri C, Morris J (2014) Prediction of life expectancy in individuals in the United Kingdom using current cohort tables. J Insur Med 44:164–169

Brooks J, Strauss D, Shavelle RM, Tran L, Rosenbloom L, Wu Y (2014) Recent trends in cerebral palsy survival. part ii: individual survival prognosis. Dev Med Child Neurol 56(11):1065–1071, DOI 10.1111/dmcn.12519

Collins G, Reitsma J, Altman D, Moons K (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350, DOI 10.1136/bmj.g7594

Dodd E, Forster J, Bijak J, Smith P (2018) Smoothing mortality data: the english life tables, 20102012. Journal of the Royal Statistical Society: Series A (Statistics in Society) 181(3):717–735, DOI 10.1111/rssa.12309

Government Actuary's Department (2011) Actuarial Tables With explanatory notes for use in Personal Injury and Fatal Accident Cases, 7th edn. The Stationery Office, London, The Ogden Tables

Hutton J (2018) Expert evidence: civil law, epidemiology and data quality. Law, Probability And Risk 17:101–110, DOI doi.org/10.1093/lpr/mgy004

Hutton JL, Pharoah POD (2002) Effects of cognitive, sensory and motor impairment on the survival of people with cerebral palsy. Arch Dis Ch 86:84–89

Hutton JL, Colver AF, Mackie PC (2000) Effect of severity of disability on survival in north east England cerebral palsy cohort. Arch Dis Ch 83:468–474

Kita M (2006) The Rating of Substandard Lives, vol Brackenridge's Medical Selection of Life Risks, Palgrave Macmillan, London, chap 5, pp 71–98

Middleton J, Dayton A, Walsh J, Rutkowski S, Leong G, Duong S (2012) Life expectancy after spinal cord injury: a 50-year study. Spinal Cord 50:803–11, DOI 10.1038/sc.2012.55, epub 2012 May 15

Moons K, Altman D, Reitsma J, Ioannidis J, Macaskill P, Steyerberg E, Vickers A, Ransohoff D, Collins G (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine 162(1):W1–W73, DOI 10.7326/M14-0698

Pharoah POD, Hutton JL (2006) Life expectancy in severe cerebral palsy. Arch Dis Ch 91:254–258, doi:0.1136/adc.2005.075002

Shavelle R, DeVivo M, Brooks J, Strauss D, Paculdo D (2015) Improvements in long-term survival after spinal cord injury? Arch PMR 96:645–651

Strauss D, Shavelle RM, Anderson TW (1998) Life expectancy of children with cerebral palsy. Ped Neurol 18:143–149

Strauss D, Vachon P, Shavelle (2005) Estimation of future mortality rates and life expectancy in chronic medical conditions. J, Ins Med 37:20–34

The additional contribution of combining genetic evidence from multiple samples in a complex case[1]

Il contributo supplementare fornito dalla ridondanza di analisi genetiche da più campioni in un caso complesso

Giampietro Lago[2]

Abstract In an Italian well known case, indicated by the media as "Yara murder", supplementary information was obtained from mixed DNA profiles thanks to the application of statistical models. The analysis is based on an interpretation model for DNA mixtures that takes into account the heights of the peaks and possible artifacts, such as stutters or dropouts that could occur in the DNA amplification process. The proposed interpretation model illustrates how to compose tests of multiple samples and different marker systems. The combined evidence is used for deconvolution, where the focus is to find likely profiles for the donors to the sample. Furthermore, a mixture can be used to establish familial relationships between a reference profile and a donor to the mixed DNA sample. Results based on a single mixed DNA profile can be compared with combination of replicates, combinations of different samples and combinations of different kits.
The purpose of this contribution is to highlight, also in the context of judicial police investigations and that of a criminal trial such as the combination of different redundant data sets (different samples, different kits) works in itself an added value that provides to produce novel information unknowingly overlooked by investigators, investigators, prosecutors, lawyers and judges.
Abstract *In un caso italiano molto noto, indicato dai media come "omicidio di Yara", dai profili di DNA misti si sono potute ottenere informazioni suppletive grazie all'applicazione di modelli statistici. L'analisi si basa su un modello di interpretazione per misture di DNA che tiene conto delle altezze dei picchi e dei possibili artefatti, come gli stutter o il dropout che potrebbero verificarsi nel processo di amplificazione del DNA. Il modello di interpretazione proposto illustra come comporre prove di più campioni e di diversi sistemi di marcatori. La*

[1] This contribution is largely extracted from the paper The Yara Gambirasio case: Combining evidence in a complex DNA mixture case (2019) by T. Graversen, J. Mortera, G. Lago, Forensic Sci. Int. Genetics *40* 53-63).

[2] Giampietro Lago is forensic biologist, colonel of the Carabinieri Armed Force and commander of RIS Parma, the forensic department located in Parma for northern Italiy of the Raggruppamento Carabinieri Investigazioni Scientifiche.

combinazione dell'evidenza viene utilizzata per il miglioramento del processo di deconvoluzione in cui l'obiettivo è quello di trovare probabili profili per i contributori della traccia mista. Inoltre, una mistura può essere utilizzata per stabilire relazioni familiari tra un profilo di riferimento e un contributore al campione di DNA misto. La comparazione degli esiti basati sullo studio di un singolo profilo di DNA misto e su quello della concorrenza di replicati, di diversi campioni e diversi kit mostra l'efficacia del modello.
Lo scopo di questo contributo è quello di evidenziare, anche nell'ambito delle investigazioni giudiziarie e nei processi come la combinazione stessa di diversi set di dati ridondanti (diversi campioni, diversi kit) costituisce in sé un valore aggiunto che consente di produrre informazioni nuove altrimenti inconsapevolmente tralasciate dagli investigatori, pubblici ministeri, avvocati e giudici.

Key words: Combining evidence, DNA mixtures, forensic statistics, kinship, likelihood ratio, mixture deconvolution, probabilistic genotyping.

1 Introduction

The genetic data production on casework can support different purposes. These goals are generally obvious but in some cases, with the support of advanced statistical tools, the analysis of the data produce innovative information generally not considered by forensic geneticists and in criminal trials.
How statistical models can help extract information from mixed DNA profiles? (i) One common use of statistical models for mixtures focuses on computing a likelihood ratio that quantifies the evidence for the presence of DNA from a specific person of interest. (ii) A second use is for deconvolution. (iii) A third aspect is the possibility of studying a mixture to establish familial relationships between a reference profile and a donor to the sample.
All DNA analysis for forensic purposes can be based on different combinations of evidence. The simplest analyses are based on a single mixed DNA profile or possibly on a set of replicates thereof.
In this paper we focus on an unusual, more complex approach. We may wish to use a combination of mixed DNA profiles taken from different samples, or profiles that are typed using different kits.

1.1 The case

On Friday 26 November 2010 at 17:15 13-year-old Yara Gambirasio left home in Brembate di Sopra, a small town near Milan, Italy, to go to the gym. An hour and a half later she left the gym never to return home. Three months later her body was found in an abandoned field in an industrial area about 10 km away. She had suffered multiple injuries from a sharp weapon, which had pierced her clothing at various points. It seemed that she had been attacked and abandoned. She had died slowly from hemorrhage and hypothermia.
The DNA from the genetic material that was taken from the victim's clothes was analysed. The DNA extracted from the front and the waistband of her underpants showed the presence of male DNA. The analysis of these DNA profiles that we will show here, lead to the profile of an unknown contributor, referred to by the media as Ignoto 1 (U1[3]). It was assumed that this profile was from the murderer, who had left his DNA on the girl's underpants.

[3] U1: Unknown man #1

The profile of 'Ignoto 1' was compared to many thousands of individuals who were either local or known to have been in the area around the time of Yara's disapperance. Comparisons were previously also made with different criminal databases but had not given any leads.
Familial search showed that two brothers, who were visitors to a nearby nightclub and unrelated to the crime, shared many alleles with 'Ignoto 1' and could therefore potentially be related to the murderer. A DNA sample from their mother revealed that she shared no alleles with Ignoto 1. The brothers' father, GG, was a bus driver who had died in 1999, eleven years before the crime. A DNA profile was at first retrieved from a stamp he had licked, and in March 2013 DNA was extracted from his exhumed body. The resulting profile identified him as overwhelmingly likely to be the father of U1.
However, it was apparently totally unknown to anyone that GG had any other children, so it was hypothesized that GG had an illegitimate child. The investigators then decided to screen women who potentially could have borne him a child decades earlier. By combing through the population registers of the time they found a woman, EA. Before moving in other town she had, in fact, lived in the same village as GG and an analysis of EA's DNA showed that she was very likely to be the mother of Ignoto 1. Thus EA's son, MGB, became the chief suspect.
MGB was sentenced to life imprisonment on 1 July 2016 for the murder. On 18 July 2017 the appeal court upheld the life sentence. On 12 October 2018 the "Corte di Cassazione", the Italian Supreme Court, confirmed the sentence to life imprisonment.

1.2 Crime scene profiles

A thorough analysis of Yara's clothes revealed the presence of male biological material in an area on her underpants (exhibit #31). This area was further inspected through 24 virtual grid cells (G1 to G24). Each grid cell was split into two parts.
Serological analyses were made on one part of the grid cell, in order to determine the biological nature of the male contribution (e.g. blood, saliva, or sperm). The other part of each grid cell was used for DNA profiling with at least three different kits among NGM, Identifiler, PowerPlex ESI16, and PowerPlex ESX16. Some grid cells, e.g. G20, were also analysed with Argus X, Y-Filer, and PowerPlex 16 in order to obtain information about the X- and Y-chromosomes and the two Penta loci.
The forensic lab analyzed the following pieces of evidence pertaining to Yara's underpants, but emphasize that many more pieces of evidence played a part in resolving the case: (i) an NGM profile from each of the grid cells G13-G16 and G20; (ii) an ESX16 profile from grid cell G14; (iii) replicates (R1 and R2) of an Identifiler profile obtained from grid cell G20.
Genetic outputs show all the observed peaks above 150RFU. Some loci, e.g. D18 for sample G13, exhibit complete dropout at this level of detection (an excerpt of the profiles can be seen in Graversen et al. 2019).

2. Combining evidence from multiple samples or different marker systems

As detailed above there are multiple crime scene profiles available in the Yara case. These are either replicates or originate from different samples. Furthermore, the samples were also typed with different kits.

There are many reasons for combining evidence, one important being that it strengthens the information about the profiles of any shared contributors. The model presented and discussed by Graversen et al. (2019) readily allows a combined analysis of multiple crime scene profiles.
Combining the information in multiple profiles requires a slightly more complex analysis than that of single profiles, since it is necessary to make assumptions about which – if any – contributors may be in common. When combining replicates it is natural to make an assumption that contributors are the same, however when combining profiles from different samples one needs to carefully consider whether there is perhaps only a partial overlap. However, once a hypothesis describing the contributors is formulated, the mathematical details in extending a peak height model from one to multiple crime scene profiles are completely straightforward.
Imagine a pool of k persons, who are all (proposed) contributors to one or more of the profiles under analysis. The model then describes the DNA profiles of the set of k persons and the associated peak heights across the set of profiles. Each profile (EPG[4]) has an individual set of model parameters that determines the variability of the peaks that we may observe just for this profile.
We assume that, conditionally on the DNA profiles of the entire pool of contributors (and the model parameters), the peak heights in one EPG are independent of the peak heights in the other EPGs. The variability of the peak heights for each profile is then described by the gamma distribution as described in Cowell et al. (2015).
A pictorial representation of the combination of peak height information from four crime scene profiles can be seen in Fig.1. Ignoring the node R, the figure depicts how the peak-height information is conditionally independent across the four EPGs given the full DNA profiles of all the contributors to the mixture. Here the set of contributors is assumed to consist of the victim and two unknown contributors, U1 and U2.

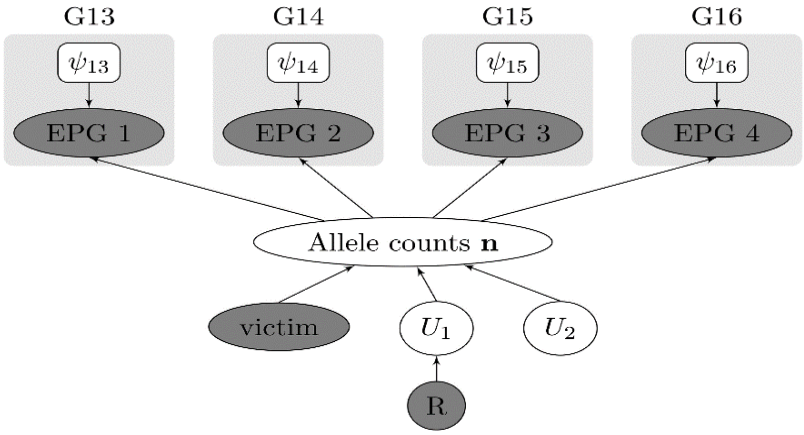


Fig. 1. Pictorial DAG representation of the combination of different types of evidence: four crime scene profiles (EPG 1–4) for samples G13-16 as well as the reference profiles of the victim and a putative parent of U1. The hypothesis under investigation here assumes a total of at most three contributors to the four samples. Each EPG has its own set of model parameters.

[4] EPG: ElectoPheroGram

3 Discussion

Model checking methods suggest that the model captures well the pattern of peaks for each profile, as described by the set of alleles observed in the profile. The set of model checking methods applied here to the statistical model in DNAmixtures (Graversen, 2015). Two model checking examples for samples G13-G16 of the Yara case are shown in Figs 2 and 3. The two figures each address different aspects of the modelling and jointly enable a thorough assessment of the joint analysis. From the probability plot in Fig.2, we see that the model adequately captures the overall variability of peak heights above the detection threshold of 150RFU. When the model fits well, the points should resemble the diagonal line as they do here. Fig.3 shows a prequential monitor plot (Seillier-Moiseiwitsch & Dawid, 1993), which assesses the ability of the model to predict whether or not a peak is observed at the next allelic position in the EPG. When the model fits well, the monitor should stay below the two upper prediction limits. We see that the model captures these aspects of the data very well.

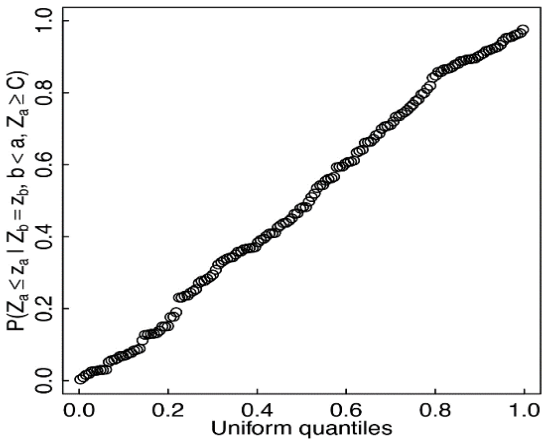


Fig. 2 Probability plot assessing the variability of peak heights for all peaks above the detection threshold of 150RFU. When the model fits well, the points should follow the diagonal as they do here.

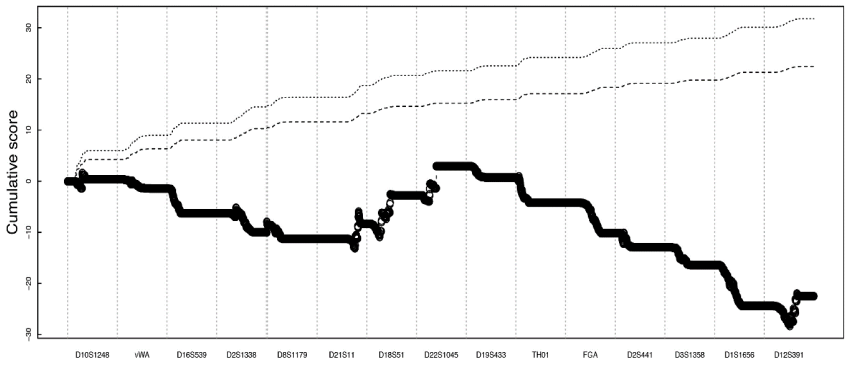


Fig.3 Prequential monitor plot confirming that the model is able to predict at which allelic positions a peak is observed. When the model fits well, the monitor should stay below the two upper prediction limits (dashed 95%, dotted 99%)

4 Final remarks

The author is not a statistician and the description and discussion of the interpretation model of the data presented is not an objective of this paper. The aim of this contribution is to highlight, also from the point of view of casework applications, how it is possible and of great interest to extract coherent and different information from a dataset of EPG profiles. The replication of analyses on a given area is a typical approach, used by forensic biologists point of view, on exhibits of particular complexity such as that described in the present paper. The reasons for this choice are substantially technical and of two types: (i) the search for the best possible outcome and (ii) the confirmation of the robustness of the outcome itself.
A novel lesson from this approach is useful for a forensic biologist. The redundancy of complex and heterogeneous results is not only a generic confirmation of the technical robustness of the same but is, in itself, a source of additional coherent information.
The redundancy of certain outcomes can therefore be translated into further scientific proof at the criminal trial. This is the challenge for the near future in which forensic biologists and statisticians will face in finding a common language capable of respecting the rigor of scientific methodology and the need for its entry into court.

References

1. T. Graversen, J. Mortera, G. Lago The Yara Gambirasio case: Combining evidence in a complex DNA mixture case Forensic Sci. Int. Genetics 40 (2019) 52-63
2. C.C. Benschop, S.Y. Yoo, T. Sijen, Split DNA over replicates or perform one amplification? Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e532–e533.
3. R.G. Cowell, T. Graversen, S.L. Lauritzen, J. Mortera, Analysis of DNA mixtures with artefacts (with discussion), J. R. Stat. Soc. Ser. C 64 (2015) 1–48.
4. [R.G. Cowell, S.L. Lauritzen, J. Mortera, A gamma model for DNA mixture analyses, Bayesian Anal. 2 (2) (2007) 333–348.
5. R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic expert systems for handling artefacts in complex DNA mixtures, Forensic Sci. Int. Genet. 5 (2011) 202–209.
6. A.P. Dawid, Present position and potential developments: some personal views. Statistical theory. The prequential approach (with discussion), J. R. Stat. Soc. Ser. A 147 (1984) 278–292.
7. G. Dørum, N. Kaur, M. Gysi, Pedigree-based relationship inference from complex DNA mixtures, Int. J. Legal Med. 131 (2017) 629–641, https://doi.org/10.1007/s00414-016-1526-x.
8. T. Graversen, Statistical and Computational Methodology for the Analysis of Forensic DNA Mixtures with Artefacts, DPhil, University of Oxford, 2014, https://ora.ox.ac.uk/objects/ora:9362.
9. T. Graversen, DNAmixtures: Statistical Inference for Mixed Traces of DNA. R package version 0.1-4, (2015) http://dnamixtures.r-forge.r-project.org.
10. T. Graversen, S. Lauritzen, Computational aspects of DNA mixture analysis, Stat. Comput. 25 (2015) 527–541.
11. N. Kaur, M. Bouzga, G. Dørum, T. Egeland, Relationship inference based on DNA mixtures, Int. J. Legal Med. 130 (2016) 323–329.
12. M.W. Perlin, Combining dna evidence for greater match information, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e510–e511 Progress in Forensic Genetics 14.
13. K. Ryan, D.G. Williams, D.J. Balding, Encoding of low-quality DNA profiles as genotype probability matrices for improved profile comparisons, relatedness evaluation and database searches, Forensic Sci. Int. Genet. 25 (2016) 227–239.
14. F. Seillier-Moiseiwitsch, A.P. Dawid, On testing the validity of sequential probability forecasts, J. Am. Stat. Assoc. 88 (1993) 355–359.
15. K. Slooten, Identifying common donors in DNA mixtures, with applications to database searches, Forensic Sci. Int. Genet. 26 (2017) 40–47.
16. C. Steele, M. Greenhalgh, D. Balding, Evaluation of low-template DNA profiles using peak heights, Stat. Appl. Genet. Mol. Biol. 15 (2016) 431–445.

The history of forensic inference and statistics: a thematic perspective

La storia dell'inferenza e della statistica in campo forense: una prospettiva tematica

Franco Taroni and Colin Aitken

Abstract In the last two decades, publications in judicial and scientific literature reveal that the trend among researchers in forensic science is to adopt a Bayesian approach to the evaluation of trace evidence, such as glass, fibres, fingerprints and biological evidence. In many areas of forensic science, however, such as those involving tool marks, shoe marks, gunshot residues and document examination, the Bayesian approach remains ignored or untrusted.
The presentation argues that it is time for Bayesian methods of evaluating evidence to be generalized. Such a broad use of the Bayesian perspective not only follows from the recent achievement of statistical arguments in forensic science, but also from the history of its earlier and effective use, at the turn of the nineteenth and twentieth centuries, in a great variety of cases and contexts. The historical development of forensic inference and statistics is presented through fifteen themes that are identified as important in the development of the ideas for reasoning in forensic science.

Abstract
Negli ultimi due decenni, pubblicazioni in riviste giuridiche e scientifiche rivelano che la tendenza tra i ricercatori in scienze forensi è quella di adottare un approccio bayesiano per la valutazione di 'prove' scientifiche, quali frammenti di vetro, fibre tessili, tracce digitali e materiali biologici. In molte aree della scienza forense, tuttavia, come quelle relative a residui di polvere da sparo ed esami di documenti, l'approccio bayesiano rimane ignorato o considerato come non applicabile.

Franco Taroni, School of Criminal Justice, The University of Lausanne, Switzerland; Franco.Taroni@unil.ch

Colin Aitken, School of Mathematics and the Maxwell Institute, The University of Edinburgh, United Kingdom; C.Aitken@ed.ac.uk

Lo scopo della presentazione è di sottolineare la necessità di estendere l'applicazione dell'approccio bayesiano alla valutazione di tutte le prove scientifiche.
Un più vasto ricorso alla prospettiva bayesiana è suffragato non solo dal recente sviluppo di argomenti statistici nella scienza forense, ma anche dalla storia che mette in evidenza il suo uso efficace, a cavallo tra il XIX e il XX secolo, in una grande varietà di casi e processi. Lo sviluppo storico dell'inferenza e della statistica forense è presentato attraverso quindici temi chiave nello sviluppo delle idee per il ragionamento nelle scienze forensi.

Key words: forensic science, forensic statistics, evidence evaluation, Bayesian framework

1 The choice of relevant themes describing the historical development of the area

The historical development of forensic inference and statistics is presented through fifteen important themes. The themes are identified as important in the development of the ideas for probabilistic and statistical reasoning in forensic science and have been chosen as the ones that created, and in some cases are still creating, important debates. The choice of themes is a personal choice of the authors and some readers may not agree. This form of presentation aims to clarify thinking around past and current problems in forensic statistics and suggests ways in which the subject may develop further.
It is only the role of statistics in the evaluation of evidence in criminal cases that is discussed.

The first theme is the recognition in the early 20th century of the need for the evaluation of scientific findings in the administration of criminal justice through the use of probability as a measure of uncertainty. Early examples are those of Bertillon (1893, 1898) and Locard (1920).

The next two themes (2 and 3) concern ideas for the integration of scientific information with other relevant information from a particular criminal case and the increasing support of judicial disciplines for the scientific presentation of evidence. From a theoretical point of view, we can refer back to 1897 when Mr Justice Holmes, then of the Supreme Judicial Court of Massachusetts and latterly of the Supreme Court of the United States (Holmes, 1897) wrote:

> For the rational study of the law the black-letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics. (at p. 469)

From a practical point of view, with reference to the Dreyfus' military trial held in 1908, Henri Poincaré and others invoked the use of Bayes' Theorem as the only way in which the court ought to revise its opinion about the issue of forgery (Darboux et al., 1908).

These ideas led to recognition of the importance of the separation of the evidence from the propositions of prosecution and defence and the correct conditioning of one on the other depending on the role of the person making the judgement (theme 4). Such separation should help avoid confusion (as presented by Thompson and Schumann, 1987) between, on one side, a small probability of finding the evidence on a person who is innocent of a crime and, on the other side, a small probability that a person on whom the evidence is found is innocent of the crime.

Various attempts to quantify the value of the evidence before the general acceptance of the likelihood ratio as the best way to do this are described in theme 5, followed by a description of the discrediting of the idea of a match (theme 6). Procedures for the evaluation of evidence in forensic science were changed dramatically by a paper by Dennis Lindley in 1977 (Lindley, 1977) emphasizing the change from a so-called 'two-stage approach' to a continuous approach.

The advent of DNA profiling in the mid-1980's led to consideration of many new factors in the evaluation of evidence which are outlined in theme 7, for example, how to assess the value (a) for a result after a database selection, (b) by taking errors into account, and (c) by conditioning on a suspect's profile in a particular population.
One factor in particular, that of the possibility of extremely small probabilities for a DNA profile and correspondingly large values of the likelihood ratio, merits a theme on its own (8). The factor relates to the acceptance as accurate and reliable experts' testimony that quote astronomical figures such as 1 in ten billion or trillion (Kaye, 1993).

The concept of propositions was developed further in the late 1990's with the introduction of differing levels of propositions (theme 9), a so-called 'hierarchy of propositions' in a project called 'Case Assessment and Interpretation' (Cook et al., 1998).
The general use of the likelihood ratio and the difficulty jurists had with its interpretation led to attempts to summarise its numerical value verbally (theme 10). The translation of a numerical value into a verbal equivalent dated back to Jeffreys (1983) (first edition in 1939) and is still under discussion (Berger and Stoel, 2018).

Though the role of the likelihood ratio was generally accepted by forensic scientists in part of Europe, there could be several different values in a particular case arising

from the use of different assumptions and statistical models. Recognition of these differences led to consideration of methods for the assessment of the performance of different models (theme 11). Measures of performance are generic. They are applied to the general performance of the model with measures obtained from datasets where the correct answer is known. Earlier works dated back to the use of scoring rule used for evaluation of evidence (Good, 1950).

The role for the forensic scientist in the investigation of a crime (the investigative role) before they are asked to evaluate evidence in a trial (the evaluative role) was recognised in a separate development in the 1990's, a role that is described as theme 12.

Statistical research in the late 20th century led to probabilistic graphical networks for complicated problems of inference. These networks had an intuitively satisfying application in forensic science, in particular for the management of many different pieces of evidence, and this application is described in theme 13. The use of Bayesian networks to manage information dates back to Aitken and Gammerman (1989) with the charting method developed by Wigmore (1913) as a predecessor of this modern network approach to inference and decision analyses.
Ultimately a decision has to be reached by the jurist (jury or judge) concerning the outcome of a criminal trial. Scientists also have decisions to make, earlier in the process, for example concerning sample size or choice of analysis. The role of decision theory for the scientific process is described in theme 14; a description that is heavily inspired by the seminal chapter written by Kaye (1988).

Finally, the early years of the 21st century have seen the questioning of the presentation of a single value for evidential value with the likelihood ratio. The alternative suggestion is that the single value of the likelihood ratio should be replaced by an interval for, or a lower bound on, its value. A comment on this debate is given in the final theme 15. The debate is still continuing; some quarters are thinking about the existence, or otherwise, of a true value of the evidence.

Acknowledgment

The authors thank the *Swiss National Science Foundation* for their support (grant n. IZSEZ0_191147).

References

Aitken, C.G.G., Gammerman, A.: Probabilistic reasoning in evidential assessment. Journal of the Forensic Science Society, **29**, 303-316 (1989)

Aitken, C.G.G., Taroni, F.: The history of forensic inference and statistics: a thematic perspective. In Handbook of Forensic Statistics, ed by D. Banks, K. Kafadar, D. Kaye, M.Tackett, Boca Raton, Florida, CRC Press (2020)

Berger, C.E.H., Stoel, R.D: Letter to the Editor - Response to Arscott et al ; S&J, 2017, 57, 221-227 'A study of the perception of verbal expression of the strength of evidence'. Science & Justice, **58**, 76-77 (2018)

Bertillon, A.: Instructions signalétiques. Imprimerie Administrative, Melun (1893)

Bertillon, A.: La comparaison des écritures et l'dentification graphique. In: Revue Scientifique, Dec.18,1897 - Jan. 1, 1898. Typographie Chamerot et Renouard, Paris (1898)

Cook, R., Evett, I.W., Jackson, G., Jones, P.J., Lambert, J.A.: A hierarchy of propositions: deciding which level to address in casework. Science & Justice, **38**, 231-239 (1998)

Darboux, J.G., Appell, P.E., Poincaré, J.H.: Examen critique des divers systèmes ou études graphologiques auxquels a donné lieu le bordereau. In L'affaire Dreyfus - La révision du procès de Rennes - Enquête de la chambre criminelle de la Cour de Cassation. Ligue française des droits de l'homme et du citoyen, Paris (1908)

Good, I.J.: Probability and the Weighing of Evidence. Griffin, London (1950)

Holmes, O.W.: Path of the law. Harvard Law Review, **10**, 457-478 (1897)

Jeffreys, H.: Theory of Probability. Clarendon Press, Oxford, 3rd edition (1983)

Kaye, D.H.: What is Bayesianism? In P. Tillers and E.D. Green, editors, Probability and Inference in the Law of Evidence, The Uses and Limits of Bayesianism (Boston Studies in the Philosophy of Science), pages 1-19. Springer, Dordrecht (1988)

Kaye, D.H.: DNA evidence: probability, population genetics, and the courts. Harvard Journal of Law & Technology, **7**,101-172 (1993)

Lindley, D.V.: A problem in forensic science. Biometrika, **64**, 207-213 (1977)

Locard, E.: L'enquête criminelle et les méthodes scientifiques. Flammarion, Paris (1920)

Thompson, W.C., Schumann, E.L. : Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. Law and Human Behaviour, **11**, 167-187 (1987)

Wigmore, J.H.: The problem of proof. Illinois Law Review, **8**, 77-103 (1913)

Topological learning: interpretable representations of complex data

Comparing Neural Networks via Generalized Persistence

Persistenza generalizzata per la comparazione di reti neurali

Mattia G. Bergomi, Pietro Vertechi

Abstract Artificial neural networks are often used as black boxes to solve supervised tasks. At each layer, the network updates its representation of the dataset in order to minimize a given error function, which depends on the correct assignment of predetermined labels to each observed data point. On the other end of the spectrum, topological persistence is commonly used to compare hand-crafted low-dimensional data representations. Here, we provide an application of *rank-based persistence*, a generalized persistence framework that allows us to characterize the data representation generated by each layer of an artificial neural network, and compare different neural architectures.

Abstract *Le reti neurali artificiali sono spesso usate come scatole nere per risolvere task supervisionati. Durante il training la rete aggiorna la rappresentazione del dataset ad ogni layer, in modo da minimizzare una certa funzione errore che dipende dall'assegnamento di etichette predeterminate ad ogni sample analizzato. In maniera antipodale, la persistenza omologica è usata per comparare rapidamente dati, in base a caratteristiche specificate dall'utente. Proponiamo di utilizzare* rank-based persistence*, una generalizzazione della teoria della persistenza topologica, per caratterizzare la rappresentazione dei dati ottenuta ad ogni livello di una rete neurale e comparare architetture neurali differenti.*

Key words: Generalized persistence, rank-based persistence, interpretable neural networks.

Mattia G. Bergomi
Veos Digital, Via Gustavo Fara 20, 20124 Milano, e-mail: mattia.bergomi@veos.digital

Pietro Vertechi
Champalimaud Research, Av. Braslia, 1400-038 Lisboa, e-mail: pietro.vertechi@neuro.fchampalimaud.org

1 Introduction

Topological persistence allows for swift quantitative comparison of topological spaces [1]. However, often data are not organized as, or easily mappable to such spaces. In [2], we generalize persistence to work with a broader set of categories and functors than topological spaces and homology.

Here, we exemplify one of the use cases discussed in [2], where persistent homology is used to characterize labeled point clouds, by working on the category of metric spaces and the poset induced by the inclusion on non-empty subsets of the label set. We focus on point clouds generated by each layer of an artificial neural network when solving a supervised classification task. We use the generalized persistence framework for both the evaluation of the layer-wise representation of different subsets of labels and to compare neural architectures.

In section 2, we give an intuition about the classical and generalized persistence frameworks. Section 3 shows how data represented at each layer of a feed-forward neural network can be summarized via *multicolored persistence diagrams*. Afterwards, two neural architectures are compared by computing the *multicolored bottleneck distance* between their diagrams.

2 From classical to generalized persistence

Classical persistent homology is based on three main ingredients: 1. A filtered topological space; 2. the homology functor H_k mapping topological spaces to finite vector spaces; 3. a notion of *rank*, such as the dimension of the vector space, or the cardinality in the category of sets [3]. See fig. 1 for an example and an intuitive introduction to persistent homology, and [4] for details.

In [2], we introduce a framework that extends classical persistence to new categories, functors and rank-like functions. We will not discuss the technical details of rank-based persistence here. See [2, Table 1] for an intuitive summary.

The main theoretical tool used in the following applications is what we called *multicolored persistence*. Multicolored persistence allows one to use persistence in semisimple categories (which can have more than one non-isomorphic indecomposable object), by preserving the fundamental properties of persistent homology: flexibility (dependence on the filtering function), stability [5], and resistance to occlusions [6]. In [2, Section 4] we discuss the stability conditions of such construction; we define multicolored persistence diagrams (MPD), and adapt the bottleneck distance to the semisimple case. These results allow us to use the classical Vietoris-Rips construction to study the interactions at the homological level of cycles generated by labeled points in a metric space.

Vietoris-Rips filtration with labeled data. Let us consider a metric space X, a finite set of points $d \subset X$ and a labeling function $l : X \to L = \{l_1, \dots, l_n\}$ associating each point with a label in the finite set L. Let $\{X_i = l^{-1}(l_i)\}_{i \in \{1,\dots,n\}}$ be the family of subdatasets corresponding to each label.

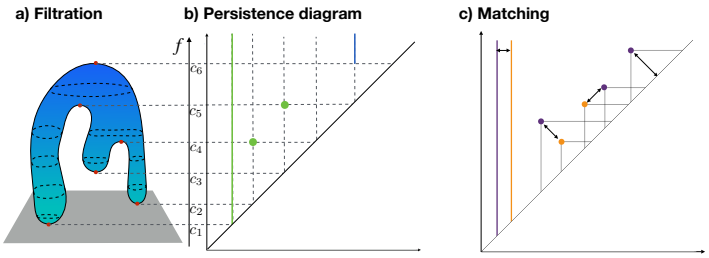


Fig. 1 Let (X, f) be a a topological space X and a continuous function $f : X \to \mathbb{R}$, respectively. The (homological) critical values $\{c_1, \ldots, c_6\}$ of f induce a sub-level set filtration of X, depicted in panel a. The change in number of generators of the kth homology groups is represented as a persistence diagram (panel b). Birth and death of homological classes are represented as points and half-lines, named cornerpoints and cornerlines, color coded depending on their associated degree: connected components in green, and the void obtained at the last sublevel set as the blue half-line. Two persistence diagrams can be compared by computing the optimal cornerpoints matching (panel c). Unmatched points are associated with their projection on the diagonal.

In the classical framework, it would be straightforward to build a filtration of X through the Vietoris-Rips construction and compute its persistent homology. However, this procedure cannot address the problem of quantifying how points belonging to different subdatasets contribute to the evolution of homological classes throughout the filtration. To do this we consider the poset $(P_n, \subseteq)$ of non-empty subsets of $\{1, \ldots, n\}$ ordered by inclusion, and the functor $F : P_n \to \mathbf{Met}$ mapping $\{1, \ldots, k\} \mapsto \sqcup_{i \in \{1,\ldots,k\}} X_i$, where **Met** is the category of metric spaces and $k \leq n$. The Vietoris-Rips construction allows us to build a $(P_n, \subseteq)$-indexed diagram (see [2, Remark 3.8]) and consequently a multicolored persistent diagram, i.e. a persistence diagram in which the information concerning the subset of labels contributing to a homological class is retained and color coded.

Implementation. Let $\mathscr{C} = \{c_1, \ldots, c_k\}$ be a neural network classifier composed of k layers, and (d, l_d) and (t, l_t) the training and test labeled datasets. We compute the Vietoris-Rips multicolored persistence as follows: 1. We train $\mathscr{C}$ on (d, l_d), by using the cross-entropy loss function and stochastic gradient descent. 2. We obtain the space $X = \sqcup_i X_i$ by considering the activation of the j-th layer on the samples belonging to the test dataset (250 randomly chosen samples per label). 3. We filter X with the classical Vietoris-Rips construction, labeling each simplex according to the labels associated to its vertices. 4. We sort the simplices of the filtration according first to their sublevel set, and then by considering the mode of their associated labels. 5. Finally, we compute the persistent homology of the sorted filtration and retrieve the labeling information associated to each homology class.

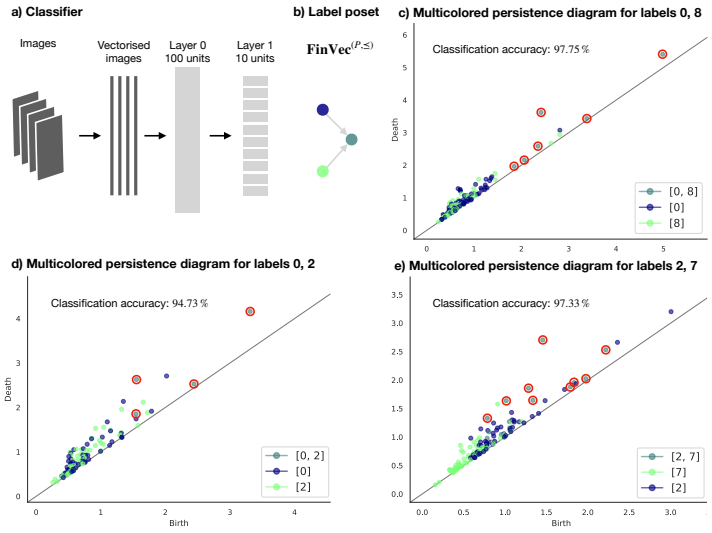


Fig. 2 Multicolored persistence. We encode the data representation obtained considering the activity pf a layer of an artificial neural network (panel *a*), by considering the multicolored persistence diagrams induced by the labeling defined on the dataset (panel *b*). In panels *c*,*d* and *e*, we show the multicolored persistence diagrams computed by considering the activation of the last layer of the trained network on the corresponding MNIST test samples. Cornerpoints originated by the interaction of simplices associated with multiple labels as highlighted (red circles).

3 Applications

In the following applications we use a shallow feed-forward neural network composed only by fully-connected layers, to classify the images of handwritten digits of MNIST [7]. The neural network is composed of two layers of 100 and 10 units, where the first layer is followed by a ReLU [8] nonlinear function.

Layer-wise embedding evaluation. As a first application we want to show how homological classes arising from the interaction of different labels carry information concerning the separability of those labels at a given layer. We consider samples belonging to pairs of labels in the MNIST dataset, train the neural network described above for 20 epochs, and evaluate it on the test set. We then consider the label poset of fig. 2, panel (b) and compute the multicolored persistence diagram by following the algorithm described in section 2.

The 1-multicolored persistence diagrams corresponding to the pairs of labels $(0,8),(0,2)$ and $(2,7)$ obtained considering the activity of the last layer of the network on the samples belonging to the test dataset are shown in panels (c, d, e) of fig. 2. An artificial neural network applies nonlinear transformations in order to separate samples belonging to different classes. Indeed, the cornerpoints associated with cycles generated by samples belonging to a single label appear overlapping and with low persistence. Observe how, although present in all the examples we

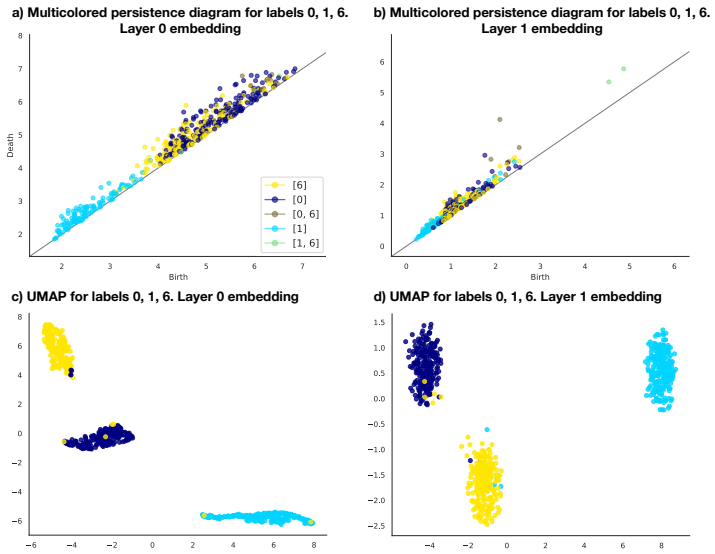


Fig. 3 Multicolored persistence diagrams obtained by considering the point clouds generated by the activation of the first (panel a) and last (panel b) layers of the neural network described in fig. 2, after training to distinguish the MNIST samples labeled with $0, 1$ and 6. Panels c and d are obtained by reducing the dimensionality of the same point clouds.

showcase, it modulates depending on the considered pair of labels, reflecting the validation accuracy of the classifier, it being $97.75\%, 94.73\%, 97.33\%$ for each of the considered pair of labels. However, cornerpoints associated with multiple classes are born later along the filtration (they are *larger holes*) and have larger persistence, again correlating with the classifier's score.

The same analysis for samples with labels $(0, 1, 6)$ are reported in fig. 3. There, however, we considered the point cloud generated by the first and second (last) layer of the network. The corresponding multicolored persistence diagrams are shown in panels (a) and (b) of fig. 3. As a comparison, we reduced the dimensionality of the point clouds by using UMAP [9].

A distance for neural architectures. One of the main advantages in using stable data representations such as multicolored persistence diagrams is that they can be compared through the multicolored bottleneck distance. This distance is essentially a color-wise version of the classical bottleneck distance used in persistent homology [10], i.e. it only admits matching between cornerpoints that respect their coloring (labeling).

As a proof of concept we compared the neural architecture and a second one, identical in its structure, but with 10 units in the first layer. The classification accuracy of the two models after 20 training epochs are 98.6% and 99%, respectively. Although the difference in accuracy is low, the multicolored bottleneck distance between the 1-MDPs computed by considering the activity of the last layer of the two trained architectures can be used to discriminate between them.

The values of the multicolored bottleneck distance (label-subset $\to$ distance) are: $\{0\} \to 0.6$, $\{1\} \to 0.2$, $\{6\} \to 0.3$, $\{1,6\} \to 0.5$, $\{0,6\} \to 1$. Coherently with our previous applications, cycles produced by multiple labels or associated with labels that are easily misclassified (0 and 6) have higher distances.

4 Discussion

After providing a short introduction to persistence homology and its rank-based generalization, we showed how this latter technique can be used to represent and compare in a robust and stable fashion the transformations learned by artificial neural networks in supervised classification tasks.

This work is intended as an exemplification of the theoretical framework described in [2], and in particular of the generalization of persistence to semisimple categories. In a forthcoming work, we plan to apply this technique to the evaluation and selection of more complex architectures (e.g. convolutional neural networks) and biological neural networks.

References

1. M. Ferri, Persistent topology for natural data analysis - A survey, arXiv:1706.00411 [math]ArXiv: 1706.00411.
   URL http://arxiv.org/abs/1706.00411
2. M. G. Bergomi, P. Vertechi, Rank-based persistence, Theory and applications of categories 35 (2020) 34.
3. M. G. Bergomi, M. Ferri, P. Vertechi, L. Zuffi, Beyond topological persistence: Starting from networks, arXiv preprint arXiv:1901.08051.
4. H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, in: Proceedings 41st Annual Symposium on Foundations of Computer Science, 2000, pp. 454–463. doi:10.1109/SFCS.2000.892133.
5. D. Cohen-Steiner, H. Edelsbrunner, J. Harer, Stability of Persistence Diagrams, Discrete & Computational Geometry 37 (1) (2007) 103–120. doi:10.1007/s00454-006-1276-5.
   URL https://doi.org/10.1007/s00454-006-1276-5
6. B. Di Fabio, C. Landi, A mayer–vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions, Foundations of Computational Mathematics 11 (5) (2011) 499.
7. L. Deng, The mnist database of handwritten digit images for machine learning research [best of the web], IEEE Signal Processing Magazine 29 (6) (2012) 141–142.
8. R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, H. S. Seung, Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, Nature 405 (6789) (2000) 947–951.
9. L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426.
10. M. d'Amico, P. Frosini, C. Landi, Using matching distance in size theory: A survey, International Journal of Imaging Systems and Technology 16 (5) (2006) 154–161.

On the topological complexity of decision boundaries

Sulla complessita' topologica delle superfici di separazione

António Leitão and Giovanni Petri

Abstract Complex data requires complex models, or so the saying goes. However, the reason why classifying two concentric circles is challenging is not because they are circles, but because they are concentric. Our contribution aims at shifting the focus from data-driven to task-oriented models. To this aim, we study the topology of the decision boundary of a classification problem as a measure of its intrinsic complexity. We employ a topological approach to understand how the architecture of a model can limit its topological expressiveness and how complex data might not equate with difficult classification problems.

Abstract *I dati complessi richiedono modelli complessi, o almeno così si dice. Tuttavia, la ragione per la quale classificare due cerchi concentrici è difficile non è perché sono cerchi, ma perché sono concentrici. Il nostro contributo mira a spostare l'attenzione da modelli basati sui dati a quelli orientati ai compiti. A questo scopo, studiamo la topologia della superficie di separazione di un problema di classificazione come misura della sua complessità intrinseca. Utilizziamo un approccio topologico per comprendere come l'architettura di un modello possa limitare la sua espressività topologica e in che modo dati complessi possano non corrispondere a problemi di classificazione difficili.*

Key words: Neural Networks, persistent homology, topological complexity

1 Introduction

Neural Networks have taken over as one of the most popular methods in machine learning, due to both their versatility and power. However, in many domains, one of the major difficulties is how to select the optimal neural network architecture for a given task. While a large effort has been poured into developing effective

António Leitão
NOVA IMS, Campolide Lisboa, e-mail: aleitao@novaims.unl.pt

Giovanni Petri
ISI Foundation, Turin, Italy e-mail: giovanni.petri@isi.it

architecture for computer vision tasks [Simonyan and Zisserman, 2014, Szegedy et al., 2014, He et al., 2015], it is still unclear what defines an optimal architecture for an arbitrary task. Although recent work [Bianchini and Scarselli, 2014, Raghu et al., 2016, Daniely et al., 2016] has shed some light on how the parameters of a neural network influence its effectiveness, the adequate complexity of the model must be known beforehand in order to find the optimal set of parameters for a problem.

We propose to bridge these two concepts by adopting a novel perspective on the problem. In particular, we propose a method to retrieve from data the minimal model complexity required to accurately generalize to an arbitrary classification problem. This is done by assessing the topological complexity of the decision boundary of the classification problem. Our contribution hinges on the fact that harder problems display more topologically complex decision boundaries and that neural networks learn by disentangling the representations of different classes from one another. We also provide some insight into how the architecture of a neural network is directly involved in how effective it is in disentangling classes.

If we consider neural networks as a set of transformations between spaces, topology becomes a natural lens through which to study them. For example, Guss and Salakhutdinov [2018] showed that persistent homology can be used to characterize the capacity of neural architectures in direct relation to their ability to generalize on data. Rieck et al. [2019] observed that the 0-dimensional homology class of the weights and their connections is an adequate indicator of a given Neural Network's learning performance. Finally, Bianchini and Scarselli [2014] adopt the total sum of Betti numbers as a measure of how the depth of feedforward networks affects their ability to implement high complexity functions. Other works [Olah, 2014, Brahma et al., 2015, Guss and Salakhutdinov, 2018] instead assume the complexity of the classification problem to be solely expressed by the complexity of the data. Ramamurthy et al. [2018] address this issue yet do not offer a solution that is scalable to multi-classification problems.

Our work differs from these because we base our analysis in the topology of the decision boundary and show evidence that in many cases the topological characterization of the data does not reflect the topological characterization of a given classification problem (Fig. 1a and b). Our contribution is threefold: (i) we introduce a topological characterization of neural networks in simple cases; (ii) we present a method to characterize classification problems by the topological complexity of their decision boundary; and, (iii) we provide empirical results pointing to how a neural network's data representation becomes less and less disentangled during training.

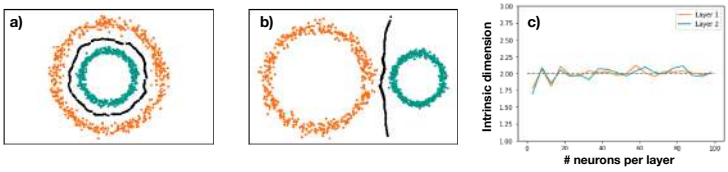


Fig. 1: Difference in the complexity of decision boundary for datasets with the same topological characteristics but different task structure. a) The boundary has non-trivial topology, while b) on the right is homologically trivial. c) The intrinsic dimension of the layers' representation of a dataset of 2 interlinked Tori for trained networks (in the conditions of theorem 1) with increasing number of neurons.

2 Neural Networks from a topological perpective

Definition 1 (Neural Network). A fully connected dense Neural Network of n-layers can be seen as a chain:

$$X_0 \xrightarrow{L_0} X_1 \xrightarrow{L_1} \dots X_{n-1} \xrightarrow{L_{n-1}} X_n \tag{1}$$

Where X_0 is called the *input* and X_n the *output*. Each L_i is in general defined to be the composition of an affine transformation with a non-linear continuous monotonous function, called *activation* function. $L_i(x) = a(W_i x + B_i)$.

The expressiveness of a neural network is then fully characterized by the nature of the chain of transformations $L_0 \dots L_{n-1}$. Interestingly, it is possible to show that there are situations in which these transformations do not alter the topological properties of the data: this is possible when a given L_i is a homeomorphism.

Theorem 1. *Let W_i be a $m \times n$ matrix such that $m \geq n$ and $rank(W_i) = n$, and let a be a continuous bijection with continuous inverse. If we consider: $L_i(X_i) = a(W_i x + B_i) \quad \forall x \in X_i$, then we have that X_i and $X_{i+1} = L_i(X_i)$ are homeomorphic.*

Proof. Let us first consider the case when $m = n$. In these conditions W_i is a linear function with a linear inverse. Since linear functions are continuous W_i is a homeomorphism. If a is a continuous bijection with a continuous inverse then L_i is a composition of homeomorphisms which is also a homeomorphism.

Let us take now $m > n$. Without loss of generality we only need to find a homeomorphism between X_i and $W_i X_i$. Since $rank(W_i) = n$, $dim(W_i x) = n$. Since $W_i x \subset \mathbb{R}^m$ there exist $m - n$ linearly independent vectors $\{e_1, \dots e_{m-n}\}$ such that the matrix $W_i' = \left[W_i |\ e_1\ \dots\ e_{m-n} \right]$ has $rank(W_i') = m$. We have then that:

$$W_i x = W_i' x' = \begin{bmatrix} & e_1^1 & \dots & e_{m-n}^1 \\ W_i & \vdots & & \\ & e_1^m & \dots & e_{m-n}^m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \hline 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \Big\} \text{n} \\ \\ \\ \\ \Big\} \text{m-n} \end{matrix} \tag{2}$$

By construction W_i' is a linear function with a linear inverse such that

$$L_i(X_i) = a(W_i x + B_i) = a(W_i' x' + B_i) \quad \forall x \in X \tag{3}$$

Along with the first case we conclude that X_i and $X_{i+1} = L_i(X)$ are homeomorphic. □

We have then that if the number of neurons does not decrease from one layer to the next, neural networks will never be able to alter the topology of the data. These conditions are somewhat restrictive, but this result is important because it challenges a very common misunderstanding that a larger number of neurons directly translates into a more powerful model.

We show this in a simple example: consider a dataset of 2 interlinked Tori (a 2-dimensional manifold) given by a uniform sample of vectors in $\mathbb{R}^3$. Consider also

a neural network with two hidden layers and sigmoid activation functions, aimed at classifying in which Torus a given point lies on. Since, by construction, the layers of the neural network are homeomorphisms, no matter how much we increase the number of neurons in each layer, the intrinsic dimension (a topological property) of the representation of the dataset through the layers is unchanged (Figure 1c). The intrinsic dimension can be accurately estimated with a method presented by Facco et al. [2017].

This simple observation allows us to contextualize previous results. For example, Ansuini et al. [2019] showed that the first layers of a convolutional neural network dramatically increase the intrinsic dimension of the data representation, while the last ones steadily decrease it. In the light of Theorem 1, this is not surprising. In fact, the neural networks studied in [Ansuini et al., 2019] do not satisfy Theorem 1, and thus different results may appear. Interestingly, they also find that in the last few layers, which are fully connected layers with constant dimension, the intrinsic dimension does not change anymore, as predicted by Theorem 1.

3 Decision Boundaries

We now try to characterize the complexity of a task from a topological perspective.

We propose to shift the focus from the datasets to the tasks themselves. The complexity of the model should not be determined by the topological features of the data, but by that of the decision boundary. Here, we consider the boundary that maximizes the distance between each point of different classes, which corresponds to the edges of adjacent Voronoi cells corresponding to points of different classes.

Definition 2 (Voronoi Cell). For each point in a given set $S = \{s_1, ..., s_k\}$ we define a Voronoi cell as the set:

$$x \in V_{s_i} \quad \Leftrightarrow \quad d(x, s_i) \leq d(x, s_j) \quad \forall x \in \mathbb{R}^n \wedge i \neq j \tag{4}$$

Intuitively, the Voronoi cell of a point s_i is the set of all the points that are the closest to s_i than to any other datapoint. Given two points of different classes a_i, b_i, if they have adjacent Voronoi cells, their edge is the set:

$$DB_{a_i,b_i} = \{x \in \mathbb{R}^n \mid d(x, a_i) = d(x, b_i) \leq d(x, s_j)\} \tag{5}$$

We call *decision boundary* the collection of all these edges. Therefore, for a classification problem in $\mathbb{R}^n$ with c classes, the decision boundary is a $n-1$-manifold that divides the space in c sections.

Definition 3 (Decision Boundary). We call *Decision Boundary* to the union of the edges of adjacent Voronoi cells corresponding to points of different classes.

3.1 Finding the decision boundary

While they are fundamental structures, computing a Voronoi diagram on n points in $\mathbb{R}^d$ requires $O(nlogn + n^{\lceil d/2 \rceil})$[Aurenhammer and Klein, 1996] making it prohibitive in high dimensions. We introduce an algorithm to sample the decision

boundary that is computationally feasible and theoretically exact for high dimensions. The central idea is to sample randomly a point and then "push" it into the decision boundary, by iteratively projecting it to the hyperspace orthogonal to its closest points belonging to different classes.

Algorithm 1: Sample the decision boundary

```
Q ← Uniform cover of n points.;
for each epoch do
    for each point q in Q do
        A ← closest point of a class to q ;
        B ← closest point of a different class to q;
        O ← hyperplane orthogonal to A − B;
        do an affine projection of q onto O;
        q ← P_O(q) ;
    end
end
```

Theorem 2. *The algorithm converges to the edges of adjacent Voronoi cells corresponding to points of different classes.*

Proof. By definition the Voronoi cell associated with point $s_i \in S$ is the set $\{x \in \mathbb{R}^n \mid d(x,s_i) \leq d(x,s_j) \forall i \neq j\}$. Given a point a_i belonging to a class, and b_i belonging to another class, we have that the set of points in the common edge of their Voronoi cells is given by: $DB_{ab} = \{x \in \mathbb{R}^n \mid d(x,a_i) = d(x,b_i) \leq d(x,s_j)\}$.
Therefore, at a given iteration of the algorithm, if point P does not belong to the set DB_{ab} then, by definition of Voronoi cell, there has to exist a point a_j (or b_j) such that $d(P,a_j) < d(P,a_i) = d(P,b_i)$. And therefore this point is considered the new closest neighbor in the next iteration. It follows that the algorithm only stops when all points reach the decision boundary. □

3.2 Topological Complexity

It now becomes possible to characterize the complexity of a classification model by looking at the topology of the decision boundary. We consider a complex decision boundary to be the opposite direction of linear separability, meaning that we consider richer topological structures to be more complex. In this perspective, we would like to measure how far a decision boundary is from being a hyperplane. In practice, we do this by computing the persistent homology [Silva and Carlsson, 2004, Edelsbrunner and Harer, 2010] of the sampled decision boundary. We define *topological complexity* as the sum of the persistences of the generators of the persistent homology groups in each dimension.

We tracked this complexity through training of a neural network and found that –even in very simple cases– one can clearly see how the topological complexity decreases (the classes become less and less entangled) the deeper we go into the network, and the further we are in the training. In Figure 2 we show the results for a neural network with 2 hidden layers with ReLU activation functions and with variable number of neurons (10/20/100) for the MNIST, Fashion-MNIST and CIFAR-10 datasets respectively.

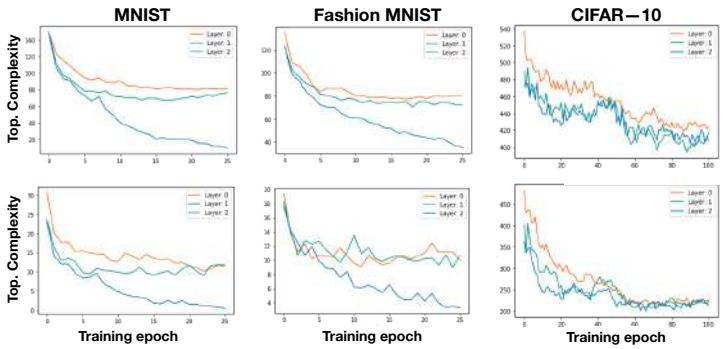


Fig. 2: Topological complexity of decision boundary. Results aree shown for H_0 (top row) and H_1 (bottom row). The studied datasets are MNIST, Fashion-MNIST and CIFAR-10.

Acknowledgements GP acknowledges support from Compagnia San Paolo and from Intesa Sanpaolo Innovation Center. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

References

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks, 2019.

Franz Aurenhammer and Rolf Klein. *Voronoi Diagrams*. Informatik-Berichte. Karl-Franzens-Univ. Graz & Techn. Univ. Graz, 1996. URL https://books.google.pt/books?id=27vkSgAACAAJ.

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, 27(10):1997–2008, 2015.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity, 2016.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied Mathematics. American Mathematical Society, 2010. ISBN 9780821849255. URL https://books.google.pt/books?id=MDXa6gFRZuIC.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

William H Guss and Ruslan Salakhutdinov. On characterizing the capacity of neural networks using algebraic topology. *arXiv preprint arXiv:1802.04443*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Christopher Olah. Neural networks, manifolds, and topology, 2014. URL https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks, 2016.

Karthikeyan Natesan Ramamurthy, Kush R Varshney, and Krishnan Mody. Topological data analysis of decision boundaries with application to model selection. *arXiv preprint arXiv:1805.09949*, 2018.

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxkijC5FQ.

Vin Silva and Gunnar Carlsson. Topological estimation using witness complexes. *Proc. Sympos. Point-Based Graphics*, 06 2004. doi: 10.2312/SPBG/SPBG04/157-166.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

Persistence-based Kernels for Data Classification

Kernel basati sulla Persistenza per la Classificazione di Dati

Ulderico Fugacci

Abstract In the last decades, topological data analysis and, specifically, persistent homology have widely proven their capabilities in extracting from large and unorganized datasets stable and discriminative information. Despite this, the retrieved topological descriptors need to further pass a "translation" process before being suitable for statistics and machine learning. In this paper, we will show how such a process can be achieved thanks to the introduction of persistence-based kernels aiming at endowing the space of persistent diagrams with an inner product. Moreover, we present a brief overview of the various definitions of a kernel for persistent and multi-parameter persistent homology given in the literature.

Abstract *Negli ultimi decenni, l'analisi topologica di dati e, in particolare, l'omologia persistente hanno ampiamente dimostrato la loro capacità nell'estrarre informazioni stabili e discriminanti da insiemi di dati non organizzati e di grandi dimensioni. Nonostante ciò, i descrittori topologici ottenuti necessitano di un ulteriore processo di "traduzione" prima di essere utilizzati in contesti statistici e di apprendimento automatico. In questo lavoro, mostreremo come tale passaggio possa essere realizzato grazie all'introduzione di kernel basati sulla persistenza che permettano di dotare lo spazio dei diagrammi di persistenza di un prodotto interno. Inoltre, presenteremo una breve panoramica delle varie definizioni di kernel per l'omologia persistente e l'omologia persistente multi-parametrica fornite in letteratura.*

Key words: Topological Data Analysis, Persistent Homology, Kernels.

1 Introduction

Nowadays, one of the most challenging problems with which researchers from all the scientific disciplines have to deal consists in extracting the core information

Ulderico Fugacci
Polytechnic University of Torino, Torino, Italy; e-mail: ulderico.fugacci@polito.it

from huge, high-dimensional, and noisy datasets. *Topological data analysis (TDA)* provides an effective collection of tools satisfying this need and notably succeeding in several applications in science and engineering. More specifically, TDA and, in particular, *persistent homology* enable to endow (almost) any kind of datasets (e.g., complex networks, point clouds, and multi-variate functions) with a notion of shape (more properly, with a topological structure) and, based on that, to study the considered data in terms of its topological features. Thanks to their proven stability, the topological descriptors obtained by such a strategy have revealed to be effective as criteria for discriminating and classifying data. Unfortunately, the retrieved information, typically represented by multi-sets of points in $\mathbb{R}^2$ and lacking a definition of inner product, cannot be straightly applied in statistics and machine learning. Recently, in order to overcome this limitation and in view of fully integrating topology among the tools that any statistician and data scientist can fruitfully adopt, kernels for persistent homology have been proposed in the literature. The main aim of this paper is, after having introduced all the required preliminaries (Section 2), to overview the currently available kernels introduced in TDA (Section 3).

2 Topological Notions

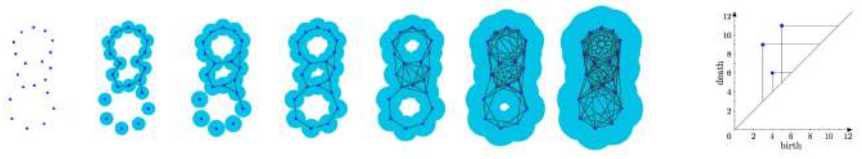
In this section, we introduce some preliminary notions at the basis of our work, namely simplicial complexes and persistent homology.

Simplicial Complexes. Simplicial complexes are collections of well-glued bricks, called simplices, typically adopted to represent discretized objects. Formally, a *k-simplex* σ is the convex hull of $k+1$ affinely independent points. A 0-simplex is a single point, a 1-simplex an edge, a 2-simplex a triangle, a 3-simplex a tetrahedron, and so on. Given a k-simplex σ, the dimension of σ is defined to be k and denoted as $dim(\sigma)$. Any simplex which is the convex hull of a non-empty subset of the points generating σ is called a *face* of σ. A *simplicial complex* Σ is a finite set of simplices which satisfies the gluing conditions requiring that each face of a simplex in Σ belongs to Σ, and each non-empty intersection of any two simplices in Σ is a face of both. We define the *dimension* of a simplicial complex Σ, denoted as $dim(\Sigma)$, as the largest dimension of its simplices.

Homology. Homology is a powerful topological tool which enables to describe a shape in terms of its "holes". More formally, given a simplicial complex Σ, the chain complex $C_*(\Sigma) := (C_k(\Sigma), \partial_k)_{k\in\mathbb{Z}}$ associated with Σ consists of: vector spaces $C_k(\Sigma)$ (over $\mathbb{Z}_2$) generated by the k-simplices of Σ; homomorphisms $\partial_k : C_k(\Sigma) \to C_{k-1}(\Sigma)$, called boundary maps, which encode the boundary relations between the k-simplices and the $(k-1)$-simplices of Σ and such that $\partial_k \circ \partial_{k+1} = 0$. We denote as $Z_k(\Sigma) := \ker(\partial_k)$ the group of the k-cycles of Σ and as $B_k(\Sigma) := Im(\partial_{k+1})$ the group of the k-boundaries of Σ. Then, we define the *k^{th} homology group of Σ* as

$$H_k(\Sigma) := H_k(C_*(\Sigma)) = \frac{Z_k(\Sigma)}{B_k(\Sigma)}.$$

A non-null element of $H_k(\Sigma)$ is an equivalence class of cycles that are not the boundary of any collection of $k+1$-simplices of Σ. Such classes represent, in dimension 0, the connected components of complex Σ, in dimension 1, its tunnels and its loops, in dimension 2, the shells surrounding voids or cavities, and so on.



(a) (b)

Fig. 1 (a) A filtration of a simplicial complex Σ obtained starting from a set of points and increasing the radius of the balls centered in each point. A k-simplex generated by vertices $v_0,\ldots,v_k$ is created when every two balls centered in $v_0,\ldots,v_k$ have a non-null intersection. (b) The persistence diagram collecting the persistence pairs describing the evolution of the 1^{st} homology groups of the filtration depicted in (a). Images courtesy of [2].

Persistent Homology. In applications, quite rarely a dataset can be described by a unique simplicial complex. Conversely, there are common constructions (such as the one achieved by the Vietoris-Rips complexes [12]) enabling for associating with a data a growing collection of simplicial complexes. Persistent homology [8] represents a development of standard homology aiming at studying the changes in homology that occur during the evolution of such a collection. Given a simplicial complex Σ, a filtration of Σ is a finite sequence of subcomplexes $\Sigma^f := \{\Sigma^p \,|\, 0 \leq p \leq m\}$ of Σ such that $\emptyset = \Sigma^0 \subseteq \Sigma^1 \subseteq \ldots \subseteq \Sigma^m = \Sigma$ (see Figure $1(a)$). For $p,q \in \{0,\ldots m\}$ such that $p \leq q$, the (p,q)-*persistent* k^{th} *homology group* $H_k^{p,q}(\Sigma)$ of Σ consists of the k-cycles included from $C_k(\Sigma^p)$ into $C_k(\Sigma^q)$ modulo boundaries. Formally, it can be defined as $H_k^{p,q}(\Sigma^f) := Im(i_k^{p,q})$, where $i_k^{p,q}$ denotes the linear map between $H_k(\Sigma^p)$ and $H_k(\Sigma^q)$ induced by the inclusion of complexes between Σ^p and Σ^q. While homology captures cycles in a shape by factoring out the boundary cycles, persistent homology allows for the retrieval of cycles that are non-boundary elements from a certain step p of the filtration and that will turn into boundaries in some subsequent step q. Such a contribution can be represented as a pair $(p,q) \in \{0,\ldots,m\} \times (\{0,\ldots,m\} \cup \infty)$. These pairs, called persistence pairs, can be visualized as a collection, called *persistence diagram*, of points in $\mathbb{R}^2$ (see Figure $1(b)$). Specifically, persistence pairs lay in the portion of the first quadrant of $\mathbb{R}^2$ above the main diagonal. Points far from the diagonal are typically considered as associated with relevant topological features while points close to the diagonal are usually interpreted as noise. The relevance of persistence diagrams as effective topological descriptors is confirmed by several stability results (see, for instance,

[5]) ensuring that slight perturbations in the original dataset will produce similar persistence diagrams.

3 Persistence-based Kernels

Thanks to their stability, the topological information and, specifically, persistence diagrams are effective data discriminants. On the other hand, interfacing persistence diagrams directly with statistics and machine learning poses technical difficulties, because the space of persistence diagrams is not endowed with a structure of an inner product or, more properly, of a Hilbert space structure. This lack prevents that lengths, angles, and means of persistence diagrams can be defined as well as that kernel-based learning methods such as kernel SVMs or PCAs [11] can be adopted.

Kernel Trick. Given an input space X, a *kernel* for X is a map $k : X \times X \to \mathbb{R}$ such that there exist a Hilbert space H and a map $\phi : X \to H$, called *feature map* for which, for any $x, y \in X$, $k(x,y) = \langle \phi(x), \phi(y) \rangle$. Equivalently, k is a kernel if the following diagram commutes:

$$\begin{array}{ccc} X \times X & \xrightarrow{\;k\;} & \mathbb{R} \\ {\scriptstyle \phi\times\phi}\downarrow & \nearrow_{\langle\cdot,\cdot\rangle} & \\ H \times H & & \end{array}$$

Recall that a Hilbert space H is a complete metric space with respect to the distance induced by the inner product $\langle \cdot, \cdot \rangle$. Given a (pseudo) distance $d : X \times X \to \mathbb{R}$ of a space X, a kernel k for X is called *stable* with respect to d if there exists a constant $C > 0$ such that, for any $x, y \in X$,

$$\|\phi(x) - \phi(y)\|_H \leq C \cdot d(x,y),$$

where $\| \cdot \|_H$ is the norm induced by the inner product of H.

Kernels for Persistent Homology. The idea of defining a kernel for the space X of persistence diagrams has been introduced in the late nineties in [9, 7] but it has been widely adopted just in the last few years. Usually, in the definition of a persistence-based kernel, the feature map is explicitly provided and the chosen Hilbert space is typically the $L^2(\mathbb{R}^2)$ space of the square-integrable functions. In the following, we will list and briefly discuss the main features of all (to the best of our knowledge) persistence-based kernel introduced in the literature.
Roughly, we can classify them into four groups:

- Gaussian kernels [11, 10, 1];
- persistence landscapes [2];
- sliced Wasserstein kernel [3];

- kernels for multi-parameter persistent homology [6].

The first class of persistence-based kernels is defined thanks to the explicit introduction of a feature map ϕ. Focusing on [11], the map ϕ sends a persistence diagram D into a square-integrable function $\phi(D)$ on $\mathbb{R}^2$ defined as the solution of a certain heat diffusion problem satisfying suitable boundary conditions and setting as an initial condition a sum of Dirac deltas centered at each point (p,q) of the persistence diagram D. In a nutshell, this is equivalent to returning an L^2 function having a Gaussian peak centered at each point of the considered persistence diagram. Moreover, in order to obtain a stable kernel, the height of each peak is set as proportional to the distance of the corresponding point (p,q) from the diagonal of the first quadrant of $\mathbb{R}^2$.

The strategy adopted in [2] is called persistence landscapes and, as in the previous case, it is based on the definition of a proper feature map from the space of the persistence diagrams to $L^2(\mathbb{R}^2)$. In short words, the L^2 function associated to a persistence diagram D is defined as the map sending the point $(x,y) \in \mathbb{R}^2$ into the number of persistence pairs $(p,q) \in D$ such that $p < x < y < q$.

A completely different strategy is adopted in [3]. In this case, in fact, a feature map is not explicitly defined. Given a real value $\sigma > 0$, a kernel $k : X \times X \to \mathbb{R}$ for X can be defined as

$$k(x,y) := \exp\left(-\frac{f(x,y)}{\sigma^2}\right)$$

as long as f satisfies the suitable condition of being a conditionally negative definite function. Based on this fact, the introduction in [3] of a proper conditionally negative definite function, called sliced Wasserstein distance, on pairs of persistence diagrams is a sufficient condition for defining a new persistence-based kernel.

Finally, in [6], the authors have introduced a strategy for extending a kernel for persistent homology to a kernel for multi-parameter persistent homology. This latter is a generalization of standard persistent homology but it considers filtrations of simplicial complexes evolving in accordance with more than just one parameter. Inspired by the definition of a distance between multi-parameter persistence modules [4], the strategy presented in [6] allows for returning a kernel for multi-parameter persistent homology given a kernel for persistent homology. Moreover, if the input kernel is stable and can be computed in polynomial time, the same properties will be inherited by the newly retrieved kernel for multi-parameter persistent homology.

4 Conclusions

In this short paper, we have given a quick and compact excursus on topological data analysis and on the adoption of kernels to make the topological information suitable

for statistical analyses. This research field is still far from being completely investigated and we are sure that a deeper integration between topologists and experts in statistics and machine learning can seriously improve the effectiveness of such strategies.

Acknowledgements This work has been partially supported by the Italian MIUR Award "Dipartimento di Eccellenza 2018-2022"- CUP: E11G18000350001, and by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

References

1. H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.
2. P. Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.
3. M. Carriére, M. Cuturi, and S. Oudot. Sliced Wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 664–673, 2017.
4. A. Cerri, B. Di Fabio, M. Ferri, P. Frosini, and C. Landi. Betti numbers in multidimensional persistent homology are stable functions. *Mathematical Methods in the Applied Sciences*, 36(12):1543–1557, 2013.
5. D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.
6. R. Corbet, U. Fugacci, M. Kerber, C. Landi, and B. Wang. A kernel for multi-parameter persistent homology. *Computers & Graphics: X*, 2:100005, 2019.
7. P. Donatini, P. Frosini, and A. Lovato. Size functions for signature recognition. In *Vision Geometry VII*, volume 3454, pages 178–183. International Society for Optics and Photonics, 1998.
8. H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Society, Providence, RI, USA, 2010.
9. M. Ferri, P. Frosini, A. Lovato, and C. Zambelli. Point selection: a new comparison scheme for size functions (with an application to monogram recognition). In *Asian Conference on Computer Vision*, pages 329–337. Springer, 1998.
10. G. Kusano, K. Fukumizu, and Y. Hiraoka. Persistence weighted Gaussian kernel for topological data analysis. *Proceedings of the 33rd International Conference on Machine Learning*, 48:2004–2013, 2016.
11. J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748, 2015.
12. A. Zomorodian. Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271, 2010.

Topological and Mixed-type learning of Brain Activity

Determinanti Topologiche e di tipo misto per pattern di attivazione in dati di neuroimmagini

Tullia Padellini, Pierpaolo Brutti, Riccardo Giubilei

Abstract Topological characterization of brain imaging data is gaining momentum in the statistical literature, however, the resulting representation of brain imaging data are often defined in complex spaces, not amenable for standard statistical methods. Ad hoc procedures must thus be adopted in order to perform standard statistical tasks such as regression or classification with topological summaries, drastically limiting their use. Exploiting distance-covariance based tests of independence, which can assess the presence of association between object defined in different metric spaces, we build a new class of conditional inference trees, which we call *energy trees*, that allows to use topological summaries together with more standard covariates, such as categorical variables or graphs, for the analysis of fMRI data.
Abstract *Nonostante sintesi topologiche stiano acquisendo popolarità nell'analisi dei dati di brain-imaging, spesso la complicata struttura degli spazi in cui sono definite preclude il loro utilizzo in metodi statistici classici. Problemi standard di classificazione e regressione con covariate topologiche possono comunque essere risolti, ma solo tramite procedure ad hoc, cosa che pone un limite alla diffusione di questo tipo di strumenti. In questo lavoro introduciamo una nuova classe di alberi di regressione che, sfruttando test di indipendenza basati su distance-covariance, permettono di usare variabili "complesse" (come quelle topologiche) e "semplici" (come quelle numeriche e categoriche) simultaneamente.*

Key words: Topological Data Analysis, Conditional Inference Trees, Energy Statistics, fMRI data

Tullia Padellini
Imperial College London, e-mail: t.padellini@imperial.ac.uk

Pierpaolo Brutti, Riccardo Giubilei
Sapienza University of Rome e-mail: {pierpaolo.brutti},{riccardo.giubilei}@uniroma1.it

1 Topological Summaries of Brain Activity

The ever-growing complexity of modern data has been fostering the development of new tools for statistical analysis, whose aim is to provide easily interpretable and low dimensional representations. This is especially true for brain imaging data such as EEG and fMRI, which represent a challenge for statisticians as well as clinicians due to their pervasive yet still unknown dependency structure. The emerging field of Topological Data Analysis (TDA) has shown to be able to provide new insights on brain activity [8, 3]; as it characterizes data through connected structures of any dimension (i.e. connected components in dimension 0, loops in dimension 1 and so on), in fact, TDA provides information not only which on areas of the brain are connected (i.e. work together) but also what kind of connection relates them.

In short, Topological Data Analysis is a growing branch of statistics, devoted to the study of the connectivity structure of the data, typically summarised in terms of connected components (or dimension 0 topological features), loops (or dimension 1 features) or higher dimensional voids.

While some of these feature have been object of interest in network analysis for some time [7], what makes TDA a new approach is the mathematical formulation of these topological invariants. As opposed to network analysis, which is typically more concerned with the presence of connection rather than with its nature, the main goal of TDA is to estimate Homology Groups of the data, which, depending on their dimension, inform us on different types of connectivity. More specifically, for a generic space X, the Homology Groups of dimension 0 ($H_0(X)$) represent its "plain" connected components, those of dimension 1 ($H_1(X)$) represent its cyclical structures or loops and so on [11].

Since the resolution of the data, which may vary depending on the sampling density, noise level and other factors that are in no relation with the shape we are trying to uncover, affects their topological structures, TDA computes Homology Groups at different scales, i.e. at different levels of aggregation. The topological features found at each scale are then ranked according to how persistent they are with respect to the aggregation level; the longer in the aggregation scheme a feature "survives", the more important it is. If a feature is very dependent on the connectivity structure we impose by selecting a reference scale for the analysis, it may be noise, and may be neglected. The persistence of every feature in the data is collected into a multiset

$$D = \{(b_i, d_i), \quad i = 1, \ldots, n_D\}$$

called the *Persistence Diagram*, which records all the different scales at which feature i can be found, from when it first appears (b_i, the "birth-time"), to the level of aggregation at which it disappears (d_i, the "death-time").

Persistence Diagrams can be compared through the Wasserstein distance, which, for

a pair of Persistence Diagrams D, D' can be defined as follows

$$W_p(D,D') = \left[\inf_{\gamma} \sum_{x \in D} \|x - \gamma(x)\|^p \right]^{\frac{1}{p}},$$

where the infimum is taken over all bijections $\gamma : D \mapsto D'$.

Brain imaging data are the ideal playground to explore these new set of tools, due to the complexity of their dependency structure. In order to produce understandable representations, brain activity is typically studied by building brain networks, which consist in relating areas of the brain which are activated together. In a functional brain network, areas of the brain are represented by nodes and edges are drawn by thresholding some measure of similarity [4]. Brain activity is then investigated through the topology of the network, at the resolution given by the threshold used for building the graph [2].

One of the advantages of using TDA tools is that they are multi–scale by definition, hence it is not necessary to choose such a threshold. The network can be investigated at every resolution, hence no information is lost. More importantly, TDA allows to investigate different types of dependence beside connected components, such as loops and voids, which may represent more complex activation patterns.

2 Energy trees

While some work has been done on investigating the brain network topology, even if at a fixed scale, yet the use of TDA beside descriptive statistics is still largely unexplored [1]. The space of Persistence Diagrams has in fact the noticeable shortcoming of only having a metric structure, and does not allow for the vector-space representation needed for most statistical algorithm [10]. In practice this means that it is either necessary to convert it into a vector by means of feature extractions, which inevitably result in a loss of information, or to design procedures especially tailored for its metric space, which makes it impossible to mix it with other types of covariates.

As both solutions seem rather limiting, we develop a new procedure, which extends Regression trees to the case of mixed-type data. We base our proposal on Conditional Inference Trees (ctree)[5], a class of methods that exploits independence tests for the choice of the splitting variable when building the tree.

Let Y be a response variable to be predicted by a set of covariates $X_1, \ldots, X_k$, and let $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$ with $x_i = (x_{1i}, \ldots, x_{ki})$ be the corresponding observed sample. The ctree algorithm can be summarized in the following steps:

Step 1 Test the global null hypothesis of independence between $X_1, \dots, X_k$ and Y. If this hypothesis is rejected, select the covariate X_{j^*} with the strongest association to Y, otherwise stop.

Step 2 Partition the support $\mathscr{X}_{j^*}$ of the selected covariate X_{j^*} onto two sets A and $\bar{A}$ and define

$$x_A = \{x_{ji}, \text{ with } j = 1, \dots, k \text{ and } i \text{ s.t. } x_{j^*i} \in A\} \qquad x_{\bar{A}} = x \setminus x_A$$

Step 3 Recursively repeat **Step 1** and **Step 2** for x_A and $x_{\bar{A}}$.

Our extension of this algorithm, the Energy tree (etree), allows for mixed-type covariates, by adopting an energy-based test in **Step 1** and a clustering procedure to define the sets A and $\bar{A}$ in **Step 2**. The key feature of energy-based test is that it allows to test for the presence of association between random objects defined in different metric spaces [9]. This is especially relevant for the case of Persistence Diagrams, whose space, when endowed with the Wasserstein distance, is metric, however is a more general procedure and can be exploited to combine complex objects such as functional data or graphs together with more standard scalar variables in regression problems.

3 Nathan Kline Institute-Rockland Sample

We conclude by testing etree on real brain imaging data taken from the Nathan Kline Institute-Rockland Sample (NKI-RS), an ongoing study aimed at uncovering links between physiological and psychological assessments, genetic information, and neuroimaging data [6]. We consider a sample of $n = 158$ subjects, and we try to investigate the relationship between brain connectivity and IQ score, which is taken to be the response variable in our regression model. Our goal is to understand whether brain activity is best understood by means of *structural* or *functional* brain connectivity, and how to combine them.

In order to account for both kinds of connectivity, we take as covariates Diffusion Tensor Imaging (DTI), from which we infer the *anatomical* connectivity between brain areas, and functional MRI (fMRI), which embeds the *functional* connectivity instead. Since raw images of both DTI and fMRI are preprocessed and aggregated on the Craddock 200 Atlas, for each tool and each subject we have measurements on 200 areas of the brain (usually denoted by Region of Interest or ROI). DTI and fMRI data in their vector form may not be very informative, we thus adopt a different representation for them to isolate connectivity information. We represent the DTI as a graph, where each ROI is a node and an edge is drawn if the two ROI are physically connected, and we build Persistence Diagrams from the fMRI, where the Homology Groups are computed for the ε-correlation graph that has the ROI as node and the edges representing a correlation between them that is higher than ε, while the "resolution" level across which "features persists" is taken to be the

threshold ε. The difference in the choice of summary is motivated by the fact that we are trying to represent two different kinds of connectivities. While a graph may be a simpler yet reasonable representation for the structural connectivity in fact, it may be limiting in the case of the functional one, as there may be relevant information at different correlation levels.

Figure 1 shows the etree for the NKI-RS data, fitted using the `etree` R-package. Nodes denoted with "1" correspond to splits with respect to the structural graph, while nodes denoted by "2" correspond to splits with respect to the functional Persistence Diagram. Functional connectivity seems to be more discriminative than Structural one, as most of the splits are determined by the Persistence Diagram. This is somehow expected, as it is intuitive to think that brain activity is best understood in terms of assessing which areas work together and how. At the same time, it is interesting to see how the structural component does retain some predictive power, meaning that the anatomic component cannot be completely neglected. Finally, we wish to stress that at this preliminary stage of the analysis we did not include other covariates, such as age or sex, as their impact on their connectivity structure is unclear and they could be confounders, however the method is flexible enough to easily account for them.

References

1. P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer. Persistent homology analysis of brain artery trees. *The annals of applied statistics*, 10(1):198, 2016.
2. E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.
3. C. Geniesse, O. Sporns, G. Petri, and M. Saggar. Generating dynamical neuroimaging spatiotemporal representations (dyneusr) using topological data analysis. *Network Neuroscience*, 3(3):763–778, 2019.
4. C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.
5. T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.
6. K. B. Nooner, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li, et al. The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6:152, 2012.
7. A. Ponce-Alvarez, G. Deco, P. Hagmann, G. L. Romani, D. Mantini, and M. Corbetta. Resting-state temporal synchronization networks emerge from connectivity topology and heterogeneity. *PLoS computational biology*, 11(2), 2015.
8. M. Saggar, O. Sporns, J. Gonzalez-Castillo, P. A. Bandettini, G. Carlsson, G. Glover, and A. L. Reiss. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nature communications*, 9(1):1–14, 2018.
9. G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.
10. K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.
11. L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.

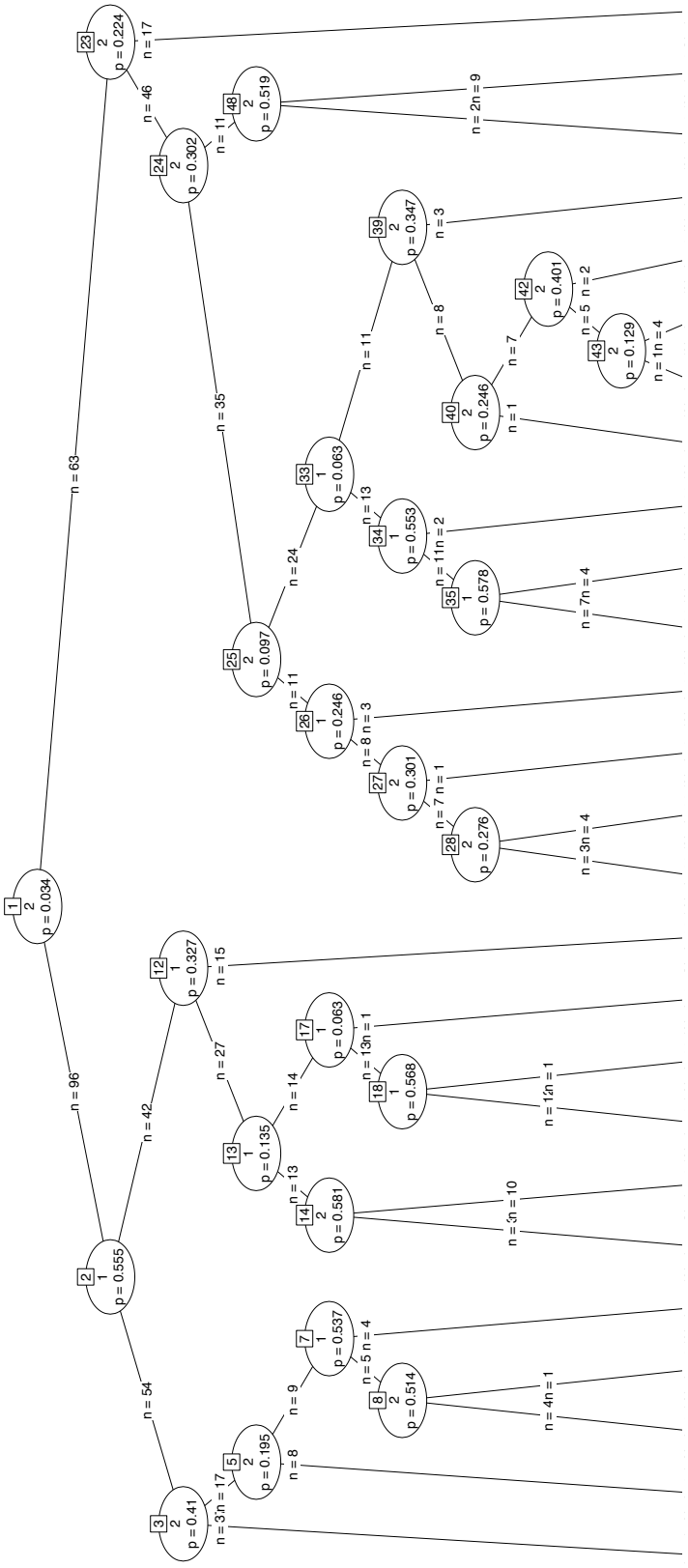


Fig. 1 Regression etree for the NKI-RS data

Contributed papers and Posters

Bayesian Statistics

A Bayesian approach for modelling dependence among mixture densities

Un modello bayesiano per indurre dipendenza tra misture di densità

Mario Beraha, Matteo Pegoraro, Riccardo Peli and Alessandra Guglielmi

Abstract We propose a Bayesian model for areal data, when multiple observations are available for each area. This situation is often encountered in practical applications, for example when geo-referenced data are collected in an aggregated form, reporting only the region or state instead of the exact GPS location. We propose to model the density of each area through a finite mixture of Gaussian distributions, for accurately approximating smooth densities. The prior for the weights of the mixtures encourages densities referring to areas that are geographically close to be similar. We derive efficient Bayesian inference based on Markov Chain Monte Carlo (MCMC) simulation and illustrate the proposed approach through simulation studies.

Abstract *Proponiamo un modello bayesiano per dati areali, qualora più osservazioni siano disponibili in ogni area. Questa situazione si incontra spesso nelle applicazioni, ad esempio quando dati georeferenziati vengono raccolti in maniera aggregata, riportando solamente la regione o lo stato invece dell'esatta posizione GPS. Proponiamo di modellare la densità in ogni area attraverso misture finite di distribuzioni gaussiane, al fine di approssimare accuratamente densità regolari. La distribuzione a priori dei pesi delle misture induce similarità tra densità di aree che sono geograficamente vicine. Basandoci su un algoritmo Markov Chain Monte Carlo (MCMC), deriviamo un approccio efficiente per l'inferenza bayesiana e illustriamo il metodo proposto in casi simulati.*

Key words: spatially dependent mixtures, logistic MCAR distribution, density estimation, conditional autoregressive prior

Mario Beraha[1,2], Matteo Pegoraro[3], Riccardo Peli[3] and Alessandra Guglielmi[1]
[1] Department of Mathematics, Politecnico di Milano, Milano, Italy
[2] Università degli Studi di Bologna, Bologna, Italy
[3] MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: {mario.beraha, matteo.pegoraro, riccardo.peli, alessandra.guglielmi}@polimi.it

1 Introduction

Finite mixture models provide an extremely rich and flexible framework, that can easily be employed to approximate densities that do not belong to standard parametric families as well as to group together observations for clustering purposes. Starting from early works [6], a huge literature has been developed around the topic; see for example [3, 4] for a detailed discussion. Although mixture models are often used in problems with exchangeable samples from a single unknown distribution, recent attention has been given to extension for multiple dependent distributions.

In spatial applications, it is often the case that data arising nearby exhibit a similar behaviour. For geo-referenced data, this similarity can be expressed assuming that data geographically close should also be close in distribution such as in the case of the generalized spatial Dirichlet Process [2], for example. On the other hand, when dealing with areal data, the notion of geographical distance is replaced by a definition of neighborhood and the similarity can be introduced using the family of Conditionally Autoregressive (CAR) distributions.

In this work, we consider the problem of modelling areal data, when, for each area $i = 1,\dots,I$, a set of observations $y_{ij},\ j = 1,\dots,N_i$ is available. This is a common scenario in real world applications, for example when geo-referenced data are presented in an aggregated form, e.g. reporting only the region or state instead of the exact GPS location, possibly to protect users' privacy.

In this setting, the usual CAR models would imply that data in each area are suitably modelled by a single parametric family, for example the Gaussian distribution. This assumption can be over-restrictive, as in the case when data in each area present a large heterogeneity because of the aggregation procedure. With this aim in mind, we propose to model the I distributions, each associated to an area, through a finite mixture of Gaussian distributions, keeping in mind the flexibility of Gaussian mixture models to accurately approximate smooth densities. We let all the mixtures share the same set of atoms, while introducing similarity between different mixtures through a novel CAR model for vectors on the simplex, that we employ as a prior for the weights of the mixtures.

2 Spatially dependent mixture models

Consider data $\boldsymbol{y}_1,\dots,\boldsymbol{y}_I$, where $\boldsymbol{y}_i = (y_{i1},\dots,y_{iN_i})^T$. Assume that a neighboring structure G between the I different areal units is known. We can assume G as a $I \times I$ matrix, its entries g_{ij} indicating the strength of the spatial association between two different areas. It might be useful to think $g_{ij} = 1$ if i and j are neighbors and 0 otherwise, but we will not be limited to this particular case.

The conditional distribution of our data is specified as follows:

$$y_{ij} \mid \mathbf{w}_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \overset{\text{iid}}{\sim} \sum_{h=1}^{H} w_{ih} \mathcal{N}(\mu_h, \sigma_h^2) \quad j = 1,\dots,N_i \tag{1}$$

where $\mathbf{w}_i = (w_{i1}, \cdots, w_{iH})^T$ is a H–dimensional vector in the simplex S^H, i.e $w_{ij} \geq 0$, $\sum_j w_{ij} = 1$ for each $i = 1, \cdots, I$ and $\mathscr{N}(\mu, \sigma^2)$ denotes the the Gaussian density with mean μ and variance σ^2.

Observe how in (1) the atoms (μ_h, σ_h^2) are shared across all the spatial locations. This is a key aspect of our model as it allows to introduce dependency between mixtures associated to different areas only through the prior for the weights.

2.1 The Logistic MCAR prior

To define a joint distribution for vectors on the simplex S^H, we start from the logistic-normal distribution in [1]. Formally, we say that $\boldsymbol{w} \in S^H$ follows a logistic-normal distribution of parameters $\boldsymbol{\mu} \in \mathbb{R}^{H-1}$, $\Sigma \in \mathbb{R}^{(H-1)\times(H-1)}$ if its additive log-ratio transformation $\widetilde{\boldsymbol{w}} := \text{alr}(\boldsymbol{w}) \sim \mathscr{N}_{H-1}(\mu, \Sigma)$, where $\text{alr} : \boldsymbol{w} \mapsto \widetilde{\boldsymbol{w}}$ is defined as:

$$\widetilde{w}_i = \log \frac{w_i}{w_H}, \quad i = 1, \ldots, H-1 \tag{2}$$

We consider instead the following model on the transformed variables $\widetilde{\boldsymbol{w}}_i$, $i = 1, \cdots, I$: the alr transformation of the mixture weights $\boldsymbol{w}_i$ at area i is given by a Multivariate Conditionally Autoregressive (MCAR) model

$$\widetilde{\boldsymbol{w}}_i \mid \widetilde{\boldsymbol{w}}_{-i}, \Sigma, \rho \sim \mathscr{N}_{H-1}\left(\rho \sum_j g_{ij} \widetilde{\boldsymbol{w}}_j, \Sigma\right) \quad i = 1, \cdots, I \tag{3}$$

It is well known that this model defines a unique joint distribution when $\rho \in (-1, 1)$. See [5] for a comprehensive overview of MCAR distribution.

Definition 1. We say that a sequence of vectors $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I)$ follows a Logistic Multivariate CAR distribution of parameters (ρ, Σ) on the adjacency graph G, if the transformed variables $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, where $\widetilde{\boldsymbol{w}}_i = \text{alr}(\boldsymbol{w}_i)$ for each $i = 1, \cdots, I$, follow the MCAR model with the same parameters, and we write $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim \text{LogisticMCAR}(\rho, \Sigma; G)$

2.2 The Bayesian model

We assume that observations $\{y_{ij}\}$ are conditionally independent between areas and iid within each area according to likelihood (1), with the following prior for the parameters:

$$(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \mid \rho, \Sigma \sim \text{LogisticMCAR}(\rho, \Sigma; G) \quad \mu_h, \sigma_h^2 \overset{\text{iid}}{\sim} P_0 \quad h = 1, \cdots, H \tag{4}$$

where P_0 is the *base measure*; we assume P_0 as an absolutely continuous distribution on $\mathbb{R} \times \mathbb{R}^+$. In our examples, we have assumed P_0 to be the well known Normal-inverse-Gamma distributions with parameters μ_0, a, b, λ, i.e. $\mu_h \mid \sigma_h^2 \sim \mathscr{N}\left(\mu_0, \sigma_h^2 \lambda^{-1}\right)$, $(\sigma_h^2)^{-1} \sim \text{Gamma}(a, b)$.

The model can be extended to include hyper-priors on the parameters of the LogisticMCAR prior. Specifically, we choose $\Sigma \sim \text{Inv-Wishart}(\nu, V)$ and $\rho \sim \mathscr{U}(0, 1)$.

Note that, though (3) gives a unique multivariate distribution for every $\rho \in (-1,1)$, we assume $\rho > 0$ to ensure the similarity among spatial neighbours, as in [5].

3 Simulation studies

We present two simulation studies to illustrate the flexibility of our model in capturing the spatial dependence in the densities among close areas. In the first simulation example, we make a comparison with the Hierarchical Dirichlet Process (HDP)-mixture model, whereas in the second one we apply our model to data generated from spatial dependent densities. We have built a MCMC scheme for simulating from the posterior, and we have coded it in C++. We do not give details of the algorithm here. We fix the hyperparameters of the HDP such that the expected number of components matches the parameter H of our model. We run the MCMC algorithm for 20,000 iterations, discarding the first 10,000 as burn-in, with thinning equal to 5, for a final sample size of 2,000.

As a metric to compare the density estimates, we use the Monte Carlo estimate of the Hellinger distance between the estimated densities and the true ones.

3.1 Non-Gaussian simulated data

We consider three scenarios. In each scenario we generate, for $I = 6$ different spatial areas, an iid sample from a density that is not the Gaussian one. The three scenarios vary with the number of data in each area and with the data generating densities employed, as reported in Table 1. They cover extremely different cases: in Scenario I a huge number of data is available in each area, so that borrowing strength from nearby areas would be superfluous; we actually expect our model to perform worse than the HDP, being the latter fully nonparametric. On the other hand, in Scenario II there are three areas (2, 4 and 6) with few data points (only 10). In this situation we expect our model to express its strength and give a better density estimate than the HDP, especially in those areas where few data are present. Finally, Scenario III is an in-between situation, where not so many observations as in Scenario I are available in each area.

	Scenario I				Scenario II				Scenario III			
Area	Density	N_i	d_H sp	d_H HDP	Density	N_i	d_H sp	d_H HDP	Density	N_i	d_H sp	d_H HDP
1	$t(6,-4,1)$	1000	0.074	0.029	$t(6,-4,1)$	1000	0.087	0.054	$t(6,-4,1)$	100	0.045	0.366
2	$t(6,-4,1)$	1000	0.074	0.026	$t(6,-4,1)$	10	0.089	0.162	$t(6,-4,1)$	100	0.045	0.364
3	$SN(4,4,1)$	1000	0.088	0.069	$SN(4,4,1)$	1000	0.098	0.074	$SN(4,4,1)$	100	0.084	0.452
4	$SN(4,4,1)$	1000	0.088	0.069	$SN(4,4,1)$	10	0.0101	0.202	$SN(4,4,1)$	100	0.084	0.452
5	$\chi^2(3,0,1)$	1000	0.091	0.096	$\chi^2(3,0,1)$	1000	0.103	0.098	$cauchy(0,1)$	100	0.044	0.291
6	$\chi^2(3,0,1)$	1000	0.091	0.096	$\chi^2(3,0,1)$	10	0.103	0.228	$cauchy(0,1)$	100	0.046	0.283

Table 1: Data generating densities in each area; t denotes the t-distribution, while SN the skew normal.

Table 1 includes also Monte Carlo estimates of the Hellinger distance between the true density and the estimate under our model (sp) and the HDP-mixture model

(HDP). From Table 1, we see that our model performs slightly worse than the HDP-mixture in Scenario I, but overall the density estimates are close enough to the true density; moreover, by visual inspection of the plots, not reported here, they are similar to the HDP-mixture density estimates. As mentioned before, under Scenario II, our model gives a better density estimate (than the HDP-mixture) in areas 2, 4 and 6, where only 10 data points are available; see Figure 1. Interestingly, our model performs much better, in all the areas, under Scenario III. We believe this is due to the presence of extreme data in areas 5 and 6, where we generate data from a Cauchy distribution. This behavior is evident from Figure 1, being the 95% pointwise credible interval much wider in HDP than in our approach.

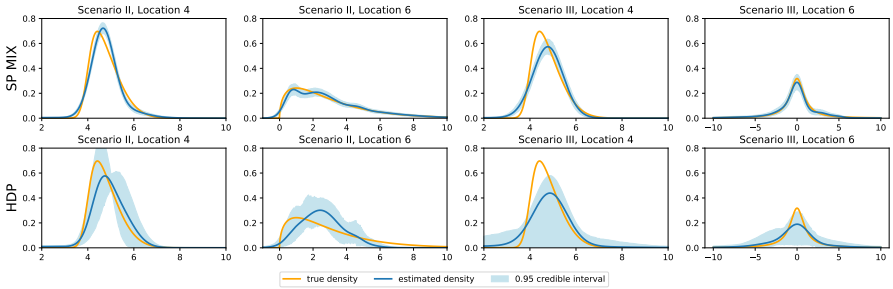


Fig. 1: Scenario II and III, non-gaussian data: estimated and true densities for areas 4 and 6

3.2 Simulation from spatially dependent weights

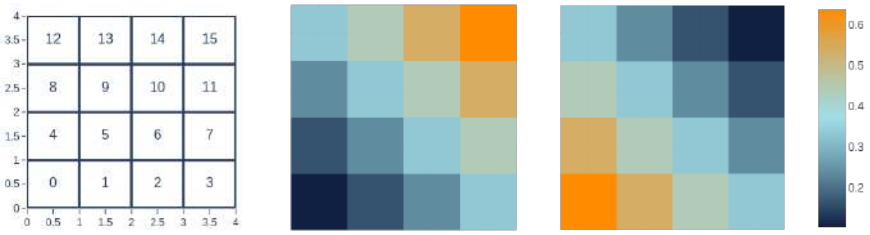
In the second simulation case, we consider 16 areas displayed in Figure 2a. Two areas are neighbors if they share an edge. At the i-th area, we draw a sample according to this distribution

$$y_{ij} \overset{\text{iid}}{\sim} w_{i1}\mathcal{N}(-5,1) + w_{i2}\mathcal{N}(0,1) + w_{i3}\mathcal{N}(5,1) \quad j = 1,\dots,100 \tag{5}$$

where the weights are chosen as $alr^{-1}(\tilde{\boldsymbol{w}}_i)$ and the transformed weights $\tilde{\boldsymbol{w}}_i$ have the following form

$$\tilde{w}_{i1} = 0.3(x_i - 2) + 0.3(y_i - 2) \quad \tilde{w}_{i2} = -0.3(x_i - 2) - 0.3(y_i - 2) \tag{6}$$

where (x_i, y_i) are the coordinates of the area center. In this way, we have introduced a strong spatial dependence, induced by (6), among the weights of different areas. The simulated weights $\{w_{i1}\}$ and $\{w_{i2}\}$ are shown in Figures 2b and 2c; of course $w_{i3} = 1 - w_{i1} - w_{i2}$. After running the MCMC sampler, we show the estimated densities in four areas in Figure 3. As expected, our approach produces accurate estimates in areas surrounded by four neighbors or whenever the true density is close to the average of the nearby densities. This is not the case in areas 0 and 15, where, indeed, we lack accuracy in density estimation.



(a) Areas (b) w_{i1} (c) w_{i2} (d)

Fig. 2: Areas and simulated weights for data in Section 3.2

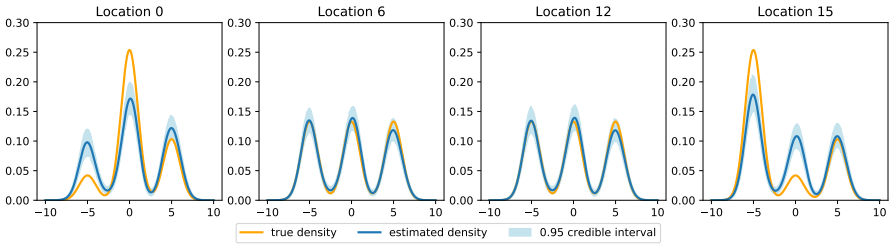


Fig. 3: Simulated data from spatially dependent weights: estimated and true densities for four locations.

4 Discussion and Conclusions

In this paper we present a new model for spatial dependent densities. In particular, each density is modeled as a mixture of Gaussian kernels with a Logistic MCAR prior on the weights that induces similarity between nearby areas. Via MCMC, we estimate the densities of different areas in simulation examples, showing the goodness of fit of our model.

References

1. J. Atchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.
2. J. A. Duan, M. Guindani, and A. E. Gelfand. Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825, 2007.
3. S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.
4. S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert. *Handbook of mixture analysis*. Chapman and Hall/CRC, 2019.
5. A. E. Gelfand and P. Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15, 2003.
6. K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

A change of glasses strategy to solve the rare type match problem

Un cambio di prospettiva per risolvere il problema del rare type match

Giulia Cereda and Fabio Corradi

Abstract We propose a solution to a forensic statistics problem known as the "rare type match case". It happens when the characteristics of the crime and the suspect's traces match but they have not been observed yet in previously collected databases. The proposed solution relies on a "change-of-glasses" strategy and consists of ignoring the specific evidence characteristics thus only modeling equalities and inequalities among different types. For the rare type match case this reduces to consider the event of seeing twice a never observed type, along with a database, now coded in form of a partition, losing reference to the specific characteristics observed. We propose to use a Bayesian nonparametric approach and derive the likelihood ratio required for forensic assessment. MCMC inference is carried on and compared to MLE through a toy example.

Abstract *Si propone una soluzione al problema forense "match di tipo raro", che si verifica quando la traccia trovata sulla scena del crimine e quella di un sospetto corrispondono ma non sono mai state osservate in precedenza. La strategia proposta è un "cambio di occhiali" in cui si ignora la specifica caratteristica delle tracce osservate, modellizzando solamente uguaglianze e disuguaglianze tra le diverse caratteristiche. Per il match di tipo raro, questo significa considerare l'evento di osservare due volte una nuova caratteristica, senza riferimento a quale essa sia, insieme a un database codificato sotto forma di partizione. Proponiamo per questo problema un approccio Bayesiano non parametrico, derivando il rapporto di verosimiglianza richiesto dal protocollo forense insieme a un'inferenza MCMC e stimatori di massima verosimiglianza per i parametri.*

Key words: Rare type match problem, Forensic Statistics, Bayesian nonparametrics, two-parameter Poisson Dirichlet, MCMC methods.

Introduction

On the crime scene, a trace has been retrieved showing characteristics that match the suspect's characteristics. The forensic statistician, who is given this piece of evidence along with a database of reference containing n traces, is asked to assess the likelihood ratio (LR), in order to weight the data D under the prosecution's (identification) and the de-

Giulia Cereda
Mathematical Institute, Leiden University, e-mail: giulia.cereda7@gmail.com

Fabio Corradi
Dipartimento di Statistica, Informatica, Applicazioni Firenze, e-mail: fabio.corradi@unifi.it

fence's (no-identification) hypotheses, h_p and h_d:

$$\mathrm{LR} = \frac{\Pr(D \mid H = h_p)}{\Pr(D \mid H = h_d)}. \tag{1}$$

In the rare type match case no other traces with the same characteristics are among those contained in the database of reference and the "rarity principle" is not operational since there are no frequencies to evaluate the rarity.

We propose a "change-of-glasses strategy", which considers the event that a never observed characteristic has been observed twice regardless of its specific type. Overall, data D are made of $n+2$ observations (n from a database and two traces from the suspect and crime scene). We are thus reducing D to a partition of the set $[n+2] = \{1,...,n+2\}$, by assigning to each class the indexes of the observations with equal characteristics.

Focusing on partitions allows us to use a nonparametric Bayesian approach. More specifically, as proposed in [2], we make use of the two-parameter Poisson Dirichlet prior to model the relative ranked sizes of the classes of the partition. This choice is motivated by the power-law shape often encountered using forensic data, such as Y-STR profiles.

1 Random partitions: an example

Consider a sequence of integer-valued random variables $I_1,...,I_n$, representing units with some characteristics expressed by the value of the Is and the equivalence relation $i \sim j$ if and only if $I_i = I_j$. The equivalence classes formed by subsets of indices with identical I form a random partition of $[n]$, which will be denoted as $\Pi_{[n]}(I_1, I_2, ..., I_n)$. For instance, the following random partition

$$\pi_{[10]} = \{\{1,3\},\{2,4,10\},\{5,6\},\{7\},\{8\},\{9\}\} \tag{2}$$

corresponds to $I_1 = I_3$, $I_2 = I_4 = I_{10}, I_5 = I_6$, while I_7, I_8, and I_9 are singletons. What we have retained is the composition of the classes themselves, but we have lost the information about the characteristics of each I_i. Assume partition $\pi_{[10]}$ as representing a database of 10 individuals. The rare type match case partition is obtained by augmenting $\pi_{[10]}$ by:

$$\pi_{[12]} = \{\{1,3\},\{2,4,10\},\{5,6\},\{7\},\{8\},\{9\},\{11,12\}\}.$$

The 11-th and the 12-th observed traces constitute a new class by themselves since they are equal one another but different from those previously observed. The strategy is thus to focalise on the classes of the partition, taking only account of similarities and dissimilarities among traces.

2 The two-parameter Poisson-Dirichlet distribution

The two-parameter Poisson Dirichlet distribution [8], is a distribution over the infinite simplex of the form $\nabla_\infty = \{(p_1, p_2, ...) \mid p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$. It has two parameters, $\alpha \in [0,1)$, and $\theta > -\alpha$ and operatively can be constructed in two steps by sorting the well-known GEM(α, θ) [5] also known as 'stick breaking prior'. One of the interesting feature of PD is its ability to represent different power-law distributions. By assuming that there is an infinite number of different characteristics, that their ordered frequencies follow a two-parameter PD distribution, and that the database is an i.i.d. sample from $\mathbf{p}$,

$$\mathbf{P} \mid \alpha, \theta \sim \mathrm{PD}(\alpha, \theta), \quad I_1, \ldots, I_n \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p} \tag{3}$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(I_1, ..., I_n)$ has the following distribution:

$$\Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{4}$$

where $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$, $[x]_{a;b} := \begin{cases} \prod_{i=0}^{a-1}(x + ib) & \text{if } a \in \mathbb{N}\backslash\{0\} \\ 1 & \text{if } a = 0 \end{cases}$, and n_i is the size of the ith class of $\pi_{[n]}$.

Relation (4), known as the *Pitman sampling formula* [6], will be used as likelihood for deriving inference for α and θ by using an MCMC scheme.

There is an alternative characterization of this model, called "Chinese restaurant process", studied in detail in [7]. It is defined as follows: consider a restaurant with infinitely many tables, each one infinitely large. Let $S_1, S_2, ...$ be integer-valued random variables representing the seating plan of the restaurant: tables are ranked in order of first occupancy: $S_i = j$ means that the ith customer seats at the jth table. The process is described by the following conditional probabilities:

$$S_1 = 1, \qquad \Pr(S_{n+1} = j | S_1, ..., S_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } j = k + 1 \\ \dfrac{n_j - \alpha}{n + \theta} & \text{if } 1 \leq j \leq k \end{cases} \tag{5}$$

where k is the number of tables occupied by the first n customers, and n_j is the number of customers already occupying table j. The process depends on two parameters α and θ constrained as the PD parameters. Clearly $S_1, ..., S_n$ are not i.i.d., nor exchangeable but in [7] it is shown that $\Pi_{[n]}(S_1, ..., S_n)$ is distributed as $\Pi_{[n]}(I_1, ..., I_n)$, with $I_1, ..., I_n$ defined by (3) and they are distributed according to the Pitman sampling formula (4).

3 Likelihood ratio

Reducing the data to partitions and assuming that the two-parameter Poisson Dirichlet distribution models the ordered frequencies of the (infinite) characteristics, the LR in (1) can be defined and derived as follows:

$$\begin{aligned}
\mathrm{LR} &= \frac{p(\pi_{[n+2]} \mid h_p)}{p(\pi_{[n+2]} \mid h_d)} = \frac{p(\pi_{[n+2]}, \pi_{[n+1]} \mid h_p)}{p(\pi_{[n+2]}, \pi_{[n+1]} \mid h_d)} && \text{since } \pi_{[n+1]} \subset \pi_{[n+2]} \\
&= \frac{p(\pi_{[n+2]} | \pi_{[n+1]}, h_p) p(\pi_{[n+1]} \mid h_p)}{p(\pi_{[n+2]} | \pi_{[n+1]}, h_d) p(\pi_{[n+1]} \mid h_d)} \\
&= \frac{1}{p(\pi_{[n+2]} \mid \pi_{[n+1]}, h_d)} && \text{since } \Pi_{[n+1]} \perp\!\!\!\perp H \text{ and } p(\pi_{[n+2]} | \pi_{[n+1]}, h_p) = 1 \\
&= \frac{1}{\int p(\pi_{[n+2]} | \pi_{[n+1]}, \alpha, \theta, h_d) p(\alpha, \theta | \pi_{[n+1]}, h_d) d\theta d\alpha} \\
&= \frac{1}{\int \frac{1-\alpha}{n+1+\theta}\, p(\alpha, \theta \mid \pi_{[n+1]}, h_d) d\alpha d\theta} && \text{by (5)}
\end{aligned} \tag{6}$$

Note that under h_p the $n+2$-th and the $n+1$th characteristics are equal with probability 1. Result (6) is formally reminiscent of the likelihood ratio employed in usual forensic identification, whenever the same characteristic is observed from the crime's and the suspect's trace. There, the crucial quantity – to be integrated with respect to the posterior distribution of unknown parameters – is the probability of observing the suspect's evidence given the database enlarged with the crime trace, under the defense hypothesis [3]. Changing glasses, we now integrate, with respect to the posterior of the Poisson Dirichlet parameters, the event of observing the $n+2$-th profile, identical to the $n+1$th, both never observed before. Using a prior PD, the probability of this event, conditionally to the model parameters is provided by (5) (bottom line with $n_i = 1$) and is equal to $\frac{1-\alpha}{n+1+\theta}$. Also, from (6), it is apparent the crucial role played by the posterior of α and θ, obtained by conditioning on $\pi_{[n+1]}$. This motivates our interest in studying inference on the PD parameters.

4 MCMC inference

For a budget of T simulations, Algorithm 1 summarizes the implementation of the Metropolis-Hastings inference for the parameters (α, θ) of the Poisson Dirichlet distribution, conditionally to an observed random partition $\pi_{[n]}$.

Algorithm 1 MH

Initialize $\theta_0 \sim p(\theta)$, $\alpha_0 \sim p(\alpha)$
for $t = 1, ..., T$ **do**
  Propose a θ_{t+1}, α_{t+1} from $p(\theta_{t+1}|\theta_{t^*})p(\alpha_{t+1}|\alpha_{t*})$
  Evaluate the ratio $R = \frac{\ell(\theta_{t+1},\alpha_{t+1};\pi_{[n]})}{\ell(\theta_{t^*},\alpha_{t^*};\pi_{[n]})} \frac{p(\theta_{t+1})}{p(\theta_{t^*})} \frac{p(\alpha_{t+1})}{p(\alpha_{t^*})} \frac{p(\theta_{t^*}|\theta_{t+1})}{p(\theta_{t+1}|\theta_{t^*})} \frac{p(\alpha_{t^*}|\alpha_{t+1})}{p(\alpha_{t+1}|\alpha_{t^*})}$
  Accept θ_{t+1} with probability R
end for

In particular, we propose as prior for θ: $\theta \sim U(0, \theta_{max})$ and $p(\theta_{max}|\theta_0, \tau) = \tau\theta_0^{\tau}(\theta^{-\tau-1})$, a Pareto$(\theta_0, \tau)$. This requires to express a prior opinion on θ_0, the smallest value that θ_{max} can assume. As prior for α: $\alpha \sim \text{Unif}(0,1)$.

The proposal distribution for θ is $\theta_{t+1} \mid \theta_{t^*} \sim \text{Exp}(\frac{1}{\theta_{t^*}})$, so the mean of the proposal distribution is equal to the last accepted θ_{t^*}. The proposal for α is a reflecting random walk to take into account that $\alpha \in [0,1]$.

5 A simulated example

To derive inference for the two parameters we explore two methods:

1. MLE estimators $\widehat{\alpha}_{\text{MLE}}$ and $\widehat{\theta}_{\text{MLE}}$ obtained by using the Pitman's sampling formula (4).
2. the Metropolis Hasting method described in Section 4 that provides a sample from the posterior of α and θ given an observed partition of size $\pi_{[n]}$.

The likelihood ratio of formula (6) can be obtained by plugging-in the MLE estimates for method 1, or by Montecarlo approximation using the MCMC sample, for method 2.

In order to compare the two approaches, we apply them to observed partitions obtained from two distinct populations that are distributed according to the two-parameter Poisson Dirichlet with known parameters. This allows us to concentrate on the quality of the

inference and intentionally avoid the influence of possible model mi-specification. More specifically, we create two distinct populations of size $N = 10^6$ using the Chinese Restaurant process (5), then we draw a sample of size $n = 1000$ for each population, and obtain the corresponding partition $\pi_{[1000]}$. The true parameters of each population and the MLE estimators are shown in Table 1.

	True values		MLE		MCMC specifications				Effective sample size	
	α	θ	α_{MLE}	θ_{MLE}	n. iterations	thinning	burn-in	accepted	α	θ
Example 1	0.5	20	0.48	19.37	10^6	50	75'000	18'500	44'893.88	54'257.52
Example 2	0.2	2	7×10^{-7}	3.57	10^6	80	100'000	11'250	32'075.81	56'373.49

Table 1 Values of α and θ used to simulate two populations of size $N = 10^6$ along with the MLE for α and θ, obtained from a i.i.d sample of size $n = 1000$ reduced in form of partition. The specifications of the Metropolis Hasting algorithm are also displayed, along with the effective sample size.

To stress the inferential procedures, the partition sample of the first example (from Population 1) is very close to the average partition obtained by repeatedly applying the simulation process. On the opposite, the partition sample of the second example (from Population 2) is more "pathological", since it is quite different from the average partition. In particular, it is outstanding with regard to the number of singletons observations (only two) and of duplets (four). This heterogeneity allows us to have an insight into the robustness of the two methods over "extreme" observations from the population. As expected, in the second example the MLE provides a weak inference. The Metropolis Hasting algorithm is used with $S = 10^6$ iterations: burn-in and thinning interval are assessed by using the diagnostic tests of the R package `coda` (see Table 1).

Figure 1 shows the MLE estimates and true values along with the posterior distributions of $\alpha, \theta \mid \pi_{[n+1]}$ obtained with MH with 95% credible intervals. In the first example, both methods provide good inference for α and θ, while in the second case, with the pathological partition, MLE is practically useless for α and bad for θ. The MCMC approach, even though not optimal, represents an improvement, since at least the true values for α and θ are reached by the credible intervals. The same considerations can be made regarding LR values (last column of Figure 1).

6 Conclusions

At first glance, the rare type match problem appears as an odd issue. Actually the "not yet observed" condition is very common also with the most widespread forensic identification evidence, the autosomal DNA-STR profiles. Indeed, if considered as whole profiles, they are often unique and only resorting to some independence assumptions it is possible, to consider each locus separately. In other circumstances, if such forms of independence do not hold, as it happens for the Y-STR profiles or for non-genetic characteristics, the rare type match problem remains a common and challenging issue. Our proposal is very general and only relies on a few conditions, such as the existence of a high number of different modalities for the considered characteristic and their power-law behavior. For the Y-STR rare type match problem, [1] proposed a different solution that does not reduce data to a partition but makes use of some genetic assumptions and the knowledge of some population parameters (such as mutation rates, and IBD parameters). In the future, it will be interesting to compare their approach with ours. Other areas of application concern important qualitative evidence used in forensic science for identification, such as glass

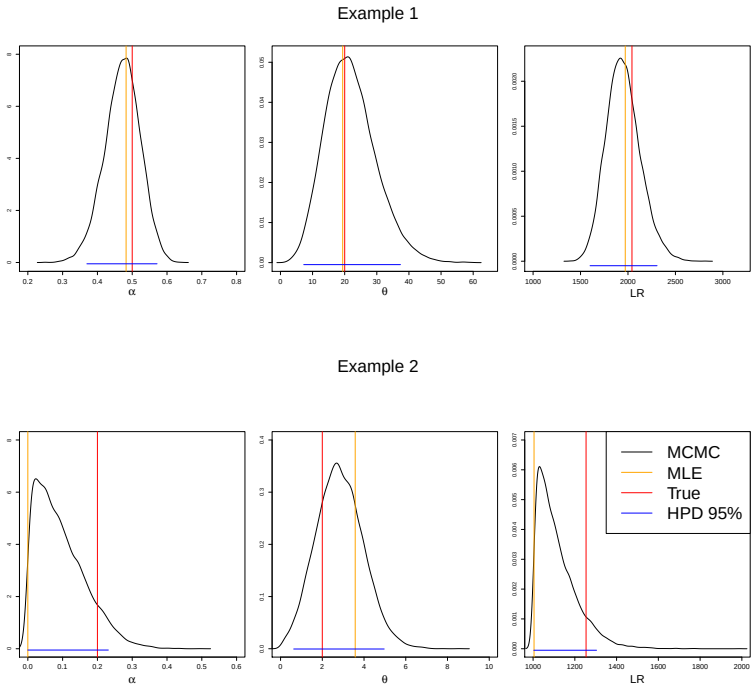


Fig. 1 Comparison of the inference provided by MCMC simulation and MLE. The first column corresponds to inference for α, the second column for θ, and the third column to the LR values obtained by plugging in MLE estimate or by Montecarlo approximation of the integral. The vertical lines represent the true value (red) and the MLE estimates (orange). The credible intervals with probability 95% of the distribution obtained through MH are also displayed through horizontal blue segments.

fragments and tire marks. Many efforts have been devoted to solve the rare type match problem in these areas, see [4] and we hope our contribution would be helpful.

References

1. M. M. Andersen and D. J. Balding. How convincing is a matching y-chromosome profile? *PLOS Genetics*, 13:1–16, 2017.
2. G. Cereda and R. D. Gill. A nonparametric Bayesian approach to the rare type match problem. *Entropy* 22(4): 439, 2020.
3. G. Cereda. Bayesian approach to LR in case of rare type match. *Statistica Neerlandica*, 71:141–164, 2017.
4. J. M. Curran, T.N. Hicks, and J.S. Buckleton, *Forensic Interpretation of Glass Evidence*. CRC Press, Boca Raton, Florida, 2000.
5. R. C. Griffiths. Exact sampling distributions from the infinite neutralalleles model. *Advances in Applied Probability*, 11:326–354, 1969.
6. J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.
7. J. Pitman. *Combinatorial Stochastic Processes*. École D'Été de Probabilités de Saint-Flour XXXII - 2002. Springer, Berlin, 2006.
8. J. Pitman. The two-parameter generalization of Ewens' random partition structure. Technical report, Department of Statistics U.C. Berkeley CA, 2003.

A new prior distribution on the simplex: the extended flexible Dirichlet

Una nuova distribuzione a prior sul simplesso: l'extended flexible Dirichlet

Roberto Ascari, Sonia Migliorati and Andrea Ongaro

Abstract The Dirichlet is a very popular prior distribution for the multinomial's probability vector parameter, due to its simplicity. Nonetheless, the Dirichlet density function cannot model many reasonable shapes (i.e. multi-modalities and/or positive covariances). This work aims to perform a preliminary study of the extended flexible Dirichlet (EFD) as a possible prior distribution. In particular, we show that the EFD prior is conjugate to the multinomial scheme and explain how the hyper-parameters change once a sample is observed.

Abstract *La distribuzione Dirichlet è spesso utilizzata come distribuzione a priori per il vettore di probabilità di una distribuzione multinomiale. Nonostante la sua semplicità, la funzione di densità della Dirichlet non riesce a modellare molti fenomeni che possono prensentarsi (ad esempio, multi-modalità e/o covarianze positive). L'obiettivo di questo lavoro è quello di condurre uno studio preliminare della distribuzione extended flexible Dirichlet (EFD) come possibile alternativa. In particolar modo, andremo a dimostrare che la EFD è una distribuzione a priori coniugata alla multinomiale e spiegheremo come gli iperparametri vengono aggiornati tramite le osservazioni campionarie.*

Key words: bayesian inference, prior, simplex, dirichlet distribution, mixture model

Roberto Ascari
University of Milano-Bicocca, e-mail: roberto.ascari@unimib.it

Sonia Migliorati
University of Milano-Bicocca, e-mail: sonia.migliorati@unimib.it

Andrea Ongaro
University of Milano-Bicocca, e-mail: andrea.ongaro@unimib.it

1 Introduction

The Dirichlet distribution is the most widespread prior distribution for parameters defined on the D-part simplex $\mathscr{S}^D$. It provides a convenient conjugate prior for Bayesian analyses involving multinomial proportions (e.g. in population genetics studies [3] or in modeling the proportion of people suffering from a disease [6]). In addition, it represents the standard prior for the mixing weights within mixture models [4]. More recently it is often encountered in the context of topic modeling in natural language processing, where it's commonly used as part of a latent Dirichlet allocation (LDA) model [1, 2]. Nonetheless, the Dirichlet has several drawbacks mainly due to its poor parameterization. In particular, variances are forced to be proportional to the corresponding means. Moreover, it cannot model multi-modalities and/or positive covariances.

Aim of this work is to explore the performance of the extended flexible Dirichlet (EFD) [5] as a prior distribution for parameters defined on the simplex. The EFD is an identifiable finite mixture with Dirichlet components, i.e.:

$$\text{EFD}(\boldsymbol{\theta};\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p}) = \sum_{r=1}^{D} p_r \text{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}+\tau_r\mathbf{e}_r), \tag{1}$$

where $\text{Dir}(\cdot;\boldsymbol{\alpha})$ denotes the Dirichlet distribution with parameter $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and $\mathbf{p}$ lie in $\mathscr{S}^D$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_D)^\intercal$, $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_D)^\intercal$, $\alpha_r > 0$, $\tau_r > 0$, $r = 1,\ldots,D$, and $\mathbf{e}_r$ is a vector of zeros except for the r-th element which is equal to one. Its probability density function (p.d.f.) can be written as:

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\alpha},\boldsymbol{\tau},\mathbf{p}) = \left(\prod_{r=1}^{D}\frac{\theta_r^{\alpha_r-1}}{\Gamma(\alpha_r)}\right)\sum_{i=1}^{D} p_i \frac{\Gamma(\alpha^+ + \tau_i)\Gamma(\alpha_i)}{\Gamma(\alpha_i+\tau_i)}\theta_h^{\tau_i} \tag{2}$$

where $\alpha^+ = \sum_{r=1}^{D}\alpha_r$.

The Dirichlet is obtained as special case of the EFD distribution setting:

$$\begin{cases} \tau_r = 1 \\ p_r = \bar{\alpha}_r = \frac{\alpha_r}{\alpha^+} \end{cases} \qquad r = 1,\ldots,D. \tag{3}$$

Adding the equality $\alpha_r = 1$, $r = 1,\ldots,D$, to conditions (3), it is possible to obtain a uniform distribution on $\mathscr{S}^D$, meaning that the EFD can be used as weakly informative prior.

The mixture component means can be written as:

$$\boldsymbol{\lambda}_r = \frac{\boldsymbol{\alpha}+\tau_r\mathbf{e}_r}{\alpha^+ + \tau_r} = \frac{\alpha^+}{\alpha^+ + \tau_r}\bar{\boldsymbol{\alpha}} + \frac{\tau_r}{\alpha^+ + \tau_r}\mathbf{e}_r, \qquad r = 1,\ldots,D. \tag{4}$$

From Equation (4) it is possible to observe that each component mean can be expressed as a weighted mean of two vectors: a common barycenter $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+$ and one vertex of the simplex. Each τ_r affects only the r-th cluster barycenter: the higher

τ_r the closer $\boldsymbol{\lambda}_r$ to $\mathbf{e}_r$. An example of component mean configuration with $D = 3$ is given in Figure 1 (dashed lines): given $\bar{\boldsymbol{\alpha}}$ (red square), when τ_r varies each $\boldsymbol{\lambda}_r$ can lie on the dashed segment joining the corresponding vertex $\mathbf{e}_r$ and $\bar{\boldsymbol{\alpha}}$.

The first two moments of the EFD have the form:

$$\mathbb{E}[X_i] = \alpha_i k_1 + \tau_i \frac{p_i}{\alpha^+ + \tau_i}, \tag{5}$$

$$\mathrm{Var}(X_i) = \alpha_i^2(k_2 - k_1^2) + \frac{p_i\tau_i(2\alpha_i + \tau_i + 1)}{(\alpha^+ + \tau_i)(\alpha^+ + \tau_i + 1)} + \alpha_i k_2 - \frac{p_i^2\tau_i^2}{(\alpha^+ + \tau_i)^2} - k_1\frac{2\alpha_i p_i \tau_i}{\alpha^+ + \tau_i}, \tag{6}$$

$$\begin{aligned}\mathrm{Cov}(X_i, X_l) = \alpha_i\alpha_l(k_2 - k_1^2) - \frac{p_i p_l \tau_i \tau_l}{(\alpha^+ + \tau_i)(\alpha^+ + \tau_l)} + \\ + \frac{\alpha_i p_l \tau_l}{\alpha^+ + \tau_l}\left(\frac{1}{\alpha^+ + \tau_l + 1} - k_1\right) + \frac{\alpha_l p_i \tau_i}{\alpha^+ + \tau_i}\left(\frac{1}{\alpha^+ + \tau_i + 1} - k_1\right),\end{aligned} \tag{7}$$

$(i, l = 1, \ldots, D, i \neq l)$ where: $k_1 = \sum_{r=1}^{D} \frac{p_r}{\alpha^+ + \tau_r}$ and $k_2 = \sum_{r=1}^{D} \frac{p_r}{(\alpha^+ + \tau_r)(\alpha^+ + \tau_r + 1)}$.

An analysis of these expressions shows two very interesting properties of the EFD, i.e. variances are not necessarily proportional to the corresponding means, and covariances may take positive values. This overcomes some of the major drawbacks of the Dirichlet distribution. In particular, note that positive correlations on the simplex can be useful in many scenarios, e.g. modelling topic correlation in LDA [1, 2].

2 EFD as prior for the multinomial model

Let us now focus on the basic, but very widespread, multinomial scheme. Suppose we observe a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\intercal$ from $\mathbf{X}_j|\boldsymbol{\theta} \sim \mathrm{Multinomial}(n_j, \boldsymbol{\theta})$, $j = 1, \ldots, N$. Then, the likelihood function is defined as:

$$f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}) = \left(\prod_{j=1}^{N} \frac{n_j!}{x_{1j}! \ldots x_{Dj}!}\right)\left(\prod_{r=1}^{D} \theta_r^{\sum_{j=1}^{N} x_{rj}}\right). \tag{8}$$

Let $x_r^+ = \sum_{j=1}^{N} x_{rj}$, $\mathbf{x}^+ = (x_1^+, \ldots, x_D^+)^\intercal$ and $n^+ = \sum_{j=1}^{N} n_j$. We can select an $\mathrm{EFD}(\boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p})$ prior distribution for $\boldsymbol{\theta}$. Then, it is possible to write the posterior p.d.f. $f_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta})$ as:

$$f_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta}) = \left(\prod_{r=1}^{D} \frac{\theta_r^{\alpha_r + x_r^+ - 1}}{\Gamma(\alpha_r + x_r^+)}\right) \sum_{i=1}^{D} \frac{p_i^*}{p_+^*} \frac{\Gamma(\alpha^+ + n^+ + \tau_i)\,\Gamma\left(\alpha_i + x_i^+\right)}{\Gamma\left(\alpha_i + x_i^+ + \tau_i\right)} \theta_i^{\tau_i} \tag{9}$$

where $p_i^* = p_i \dfrac{(\alpha_i+\tau_i)^{[x_i^+]}}{(\alpha^+ + \tau_i)^{[n^+]}\alpha_i^{[x_i^+]}}$, $p_+^* = \sum_{r=1}^D p_r^*$, and $\alpha^{[x]} = \Gamma(\alpha+x)/\Gamma(\alpha)$. Equation (9) shows that the posterior is still an EFD distribution, so that the EFD is a conjugate prior for the multinomial likelihood. Specifically, we have:

$$\boldsymbol{\theta}|\mathbf{x} \sim \mathrm{EFD}(\boldsymbol{\alpha}+\mathbf{x}^+, \boldsymbol{\tau}, \mathbf{p}^*/p_+^*). \tag{10}$$

From (10) it is possible to make some considerations regarding the upgrade of the hyper-parameters. The sample updates the parameters $\boldsymbol{\alpha}$ and $\mathbf{p}$, whereas $\boldsymbol{\tau}$ is left unchanged. The parameter $\boldsymbol{\alpha}$ is updated exactly in the same way as the parameter of the Dirichlet. To understand prior to posterior modification induced by this parameter, it is convenient to derive the updated mean of the mixture components, i.e.:

$$\begin{aligned}
\boldsymbol{\lambda}_r^* &= \frac{\boldsymbol{\alpha}+\mathbf{x}^+ + \tau_r \mathbf{e}_r}{\alpha^+ + n^+ + \tau_r} \\
&= \frac{\alpha^+}{\alpha^+ + n^+ + \tau_r}\bar{\boldsymbol{\alpha}} + \frac{\tau_r}{\alpha^+ + n^+ + \tau_r}\mathbf{e}_r + \frac{n^+}{\alpha^+ + n^+ + \tau_r}\bar{\mathbf{x}}^+ \\
&= \frac{\alpha^+ + \tau_r}{\alpha^+ + \tau_r + n^+}\boldsymbol{\lambda}_r + \frac{n^+}{\alpha^+ + \tau_r + n^+}\bar{\mathbf{x}}^+
\end{aligned} \tag{11}$$

where $\bar{\mathbf{x}}^+ = \mathbf{x}^+/n^+$. The posterior mixture component mean vector can be expressed as a weighted average of the prior $\boldsymbol{\lambda}_r$ and the sample barycenter $\bar{\mathbf{x}}^+$. Thus, the $\boldsymbol{\lambda}_r^*$'s lie on the solid lines (as n^+ varies) displayed in Figure 1 . Here and in the following figures we set $D = 3$ to facilitate graphical representations.

The behavior of the posterior mixing weights $\mathbf{p}^*/p_+^*$ is more intricate and it can be best understood through a graphical analysis. Figures 2, 3 and 4 illustrate how the posterior mixing weights $\mathbf{p}^*/p_+^*$ depend on sample quantities.

Figure 2 is obtained increasing n^+ and computing $\mathbf{x}^+ = n^+ \cdot \bar{\mathbf{x}}^+$. From the right panel it emerges that increasing n^+ makes the component mean closest to $\bar{\mathbf{x}}^+$ to augment its posterior weight.

Figures 3 and 4 study the posterior weights behavior for various x_r^+. In general each weight p_r^*/p_+^* depends increasingly on the corresponding x_r^+, given the other x_i^+, $i \neq r$. This is shown in Figure 3, where x_3^+ has been fixed to 50. Furthermore, in Figure 4 the weights p_r^*/p_+^* $(r = 1,2)$ are plotted as functions of x_1^+ and x_2^+/x_1^+, still for given x_3^+. This choice serves to evidentiate the further feature that for large values of x_1^+ the weights depend essentially only on the ratio x_2^+/x_1^+.

The obtained results seem to indicate that the EFD displays an interesting behavior in terms of interpretability and dependence, and suggest that a deeper analysis of this prior is worth to be undertaken.

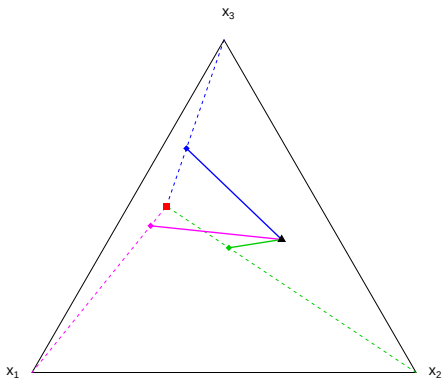


Fig. 1 Dashed lines represent the set of points where $\boldsymbol{\lambda}_r$'s can lie, solid lines represent the set of points where the $\boldsymbol{\lambda}_r^*$'s can lie ($r = \{1,2,3\}$). Colored diamonds represent $\boldsymbol{\lambda}_1$ (violet), $\boldsymbol{\lambda}_2$ (green) and $\boldsymbol{\lambda}_3$ (blue) with $\boldsymbol{\alpha} = (12,3,15)^\intercal$ (red square), $\boldsymbol{\tau} = (4,10,16)^\intercal$ and $\mathbf{x} = (0.15,0.45,0.4)^\intercal$ (black triangle).

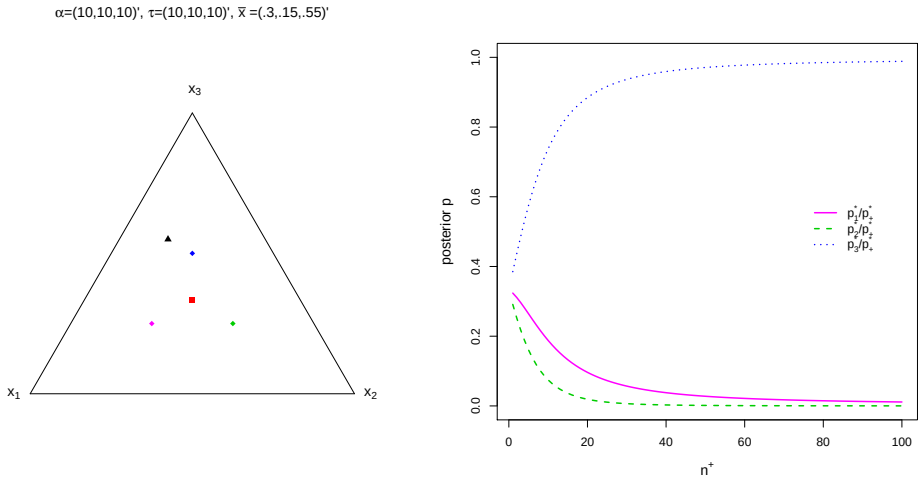


Fig. 2 Posterior mixing weights p_r^*/p_+^* as function of n^+. $\mathbf{p}$ is fixed equal to $(1/3,1/3,1/3)^\intercal$.

References

1. Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

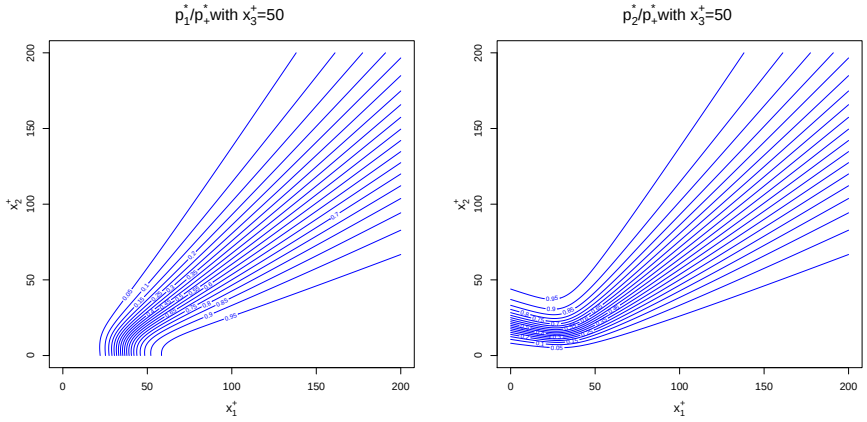


Fig. 3 Posterior mixing weights p_r^*/p_+^* as function of $\mathbf{x}^+$. $\mathbf{p}$ is fixed equal to $(1/3,1/3,1/3)^\intercal$.

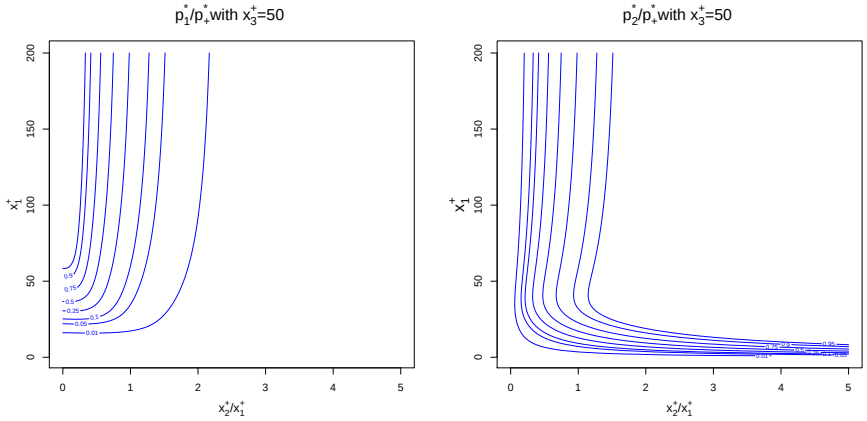


Fig. 4 Posterior mixing weights p_r^*/p_+^* as function of $\mathbf{x}^+$ (x_2^+/x_1^+ on the x-axis). $\mathbf{p}$ is fixed equal to $(1/3,1/3,1/3)^\intercal$.

2. Blei, D. M., Lafferty, J. D. (2007). A correlated topic model of Science. The Annals of Applied Statistics, 1(1), 17–35. https://doi.org/10.1214/07-aoas114
3. Etheridge, A. (2011). Some Mathematical Models from Population Genetics, Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16632-7
4. Frühwirth-Schnatter, S. (2006). Finite mixture and markov switching models. https://doi.org/10.1007/978-1-4614-6022-0
5. Ongaro, A., Migliorati, S., Ascari, R. (2020). A new mixture model on the simplex. Statistics and Computing. https://doi.org/10.1007/s11222-019-09920-x
6. Pullen, G. A., Kumaran, M. (2010). Application of Multinomial-Dirichlet conjugate in MCMC estimation: a breast cancer study, Int. Journal of Math. Analysis, 4, 41, 2043–2049

ABC model choice via mixture weight estimation

ABC model choice mediante stima del peso di mistura

Gianmarco Caruso, Luca Tardella, Christian P. Robert

Abstract Approximate Bayesian Computation (ABC) methods are widely employed to obtain approximations of posterior distributions without having to calculate likelihood functions. Nevertheless, the general impossibility to find statistics which are sufficient across models leads to unreliability of the classical tools for ABC model choice. To overcome this issue, a different kind of modelling is here proposed by replacing the traditional comparison between posterior probabilities of candidate models with posterior estimates of the weights of a mixture of these models. A simulation study highlights several strengths of this alternative approach, presenting it as a robust and flexible extension of the classical one.

Abstract *I metodi di ABC (*Approximate Bayesian Computation*) sono largamente utilizzati per ottenere approssimazioni di distribuzioni a posteriori senza dover calcolare funzioni di verosimiglianza. Tuttavia, in generale, non è possibile trovare statistiche che siano sufficienti tra modelli e ciò rende poco attendibili gli strumenti classici su cui si basa l'*ABC model choice*. Al fine di superare tale problema, si propone di rimpiazzare il tradizionale confronto tra probabilità a posteriori dei modelli candidati con la stima a posteriori dei pesi di una mistura di tali modelli. Uno studio di simulazione mette in luce diversi punti di forza di questo approccio alternativo, presentandolo come una robusta e flessibile estensione di quello classico.*

Key words: Approximate Bayesian Computation, Bayesian statistics, model choice, Bayes Factor, mixture models, intractable likelihoods, model misspecification

---

Gianmarco Caruso
Dipartimento di Scienze Statistiche, La Sapienza Università di Roma, Italy
e-mail: gianmarco.caruso@uniroma1.it

Luca Tardella
Dipartimento di Scienze Statistiche, La Sapienza Università di Roma, Italy
e-mail: luca.tardella@uniroma1.it

Christian P. Robert
CEREMADE, Université Paris Dauphine, PSL Research University, France; University of Warwick, UK; Università Ca' Foscari di Venezia, Italy
e-mail: xian@ceremade.dauphine.fr

1 Introduction

In the last decades, *Approximate Bayesian Computation* (ABC, henceforth) methods have become popular as a class of likelihood-free algorithms which aim to draw samples from an approximate posterior distribution, in the cases where the likelihood is unavailable or intractable, but it is still possible to generate data from the corresponding distribution.
Suppose to observe a sample $\mathbf{y} = (y_1, \ldots, y_n)$ of realizations from iid random variables Y_i, $i = 1, \ldots, n$, with a (*complex*) density $p(\cdot|\theta)$, where θ is an unknown parameter of interest with prior distribution $\pi(\theta)$. In order to sample from an approximate posterior distribution of θ, one can use a basic version of ABC rejection sampling algorithm (Pritchard et al. 1999): the idea is to draw iid parameter values $\theta_1^*, \ldots, \theta_N^*$ from $\pi(\theta)$, and to use each of these values to generate a synthetic sample of iid pseudo-observations, $\mathbf{z}$, from the sampling distribution $p(\cdot|\theta)$. If $\mathbf{z}$ is *similar* to the observed data $\mathbf{y}$, the corresponding θ^* is accepted as a value generated from the posterior distribution $\pi(\theta|\mathbf{y})$.
The concept of *similarity* between two datasets is expressed by three tools: a vector of statistics (or *summaries*), a distance and a tolerance threshold. The vector of summaries $\eta(\cdot) = \big(\eta_1(\cdot), \ldots, \eta_k(\cdot)\big)$ is used to summarise the information contained in a dataset, so that a pair of vectors of k statistics is compared instead of a pair of vectors of n observations (being $k \ll n$). This vector of statistics is most often not sufficient, but the consequent loss of information is tolerated with the idea to avoid the *curse of dimensionality* and to reduce the running time of the ABC algorithms. The distance $\rho(\cdot)$ quantifies how much $\eta(\mathbf{z})$ is close to $\eta(\mathbf{y})$. The threshold ε allows to accept all the θ^*'s which generate datasets whose associated vector of summaries is *close enough* to the observed vector of summaries. Therefore, the posterior sample is generated from $\pi\big(\theta\big|\rho(\eta(\mathbf{y}), \eta(\mathbf{z})) < \varepsilon\big)$ as surrogate of $\pi(\theta|\mathbf{y})$: the more informative the vector of statistics $\eta(\mathbf{y})$ and the smaller ε, the better the approximation.

1.1 Classical ABC model choice and drawbacks

When M models are compared, one considers the model index $\mathscr{M}$ as an additional unknown parameter, with prior distribution $\pi(\mathscr{M} = m)$, $m = 1, \ldots, M$ (Grelaud et al. 2009). The classical ABC model choice algorithm is summarised in **Algorithm 1**.

Algorithm 1 classical ABC model choice (ABC-mc)

for $i = 1$ to N **do**
  repeat
    Generate m^* from the prior $\pi(\mathscr{M} = m)$
    Generate $\theta_{m^*}^*$ from the prior $\pi_{m^*}(\cdot)$
    Generate $\mathbf{z}$ from the sampling distribution $p_{m^*}\Big(\cdot\Big|\theta_{m^*}^*\Big)$
  until $\rho\big(\eta(\mathbf{z}), \eta(\mathbf{y})\big) < \varepsilon$
  Set $m^{(i)} = m^*$ and $\theta_i = \theta^*$
end for

The posterior probability of model m can be estimated with the frequency of acceptances from model m, namely $\hat{\pi}_\varepsilon(\mathscr{M}=m|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(m^{(i)}=m)$, while $\hat{B}_{12,\varepsilon}(\mathbf{y}) = \frac{\pi(\mathscr{M}=2)}{\pi(\mathscr{M}=1)}\frac{\sum_{i=1}^{N}\mathbb{I}(m^{(i)}=1)}{\sum_{i=1}^{N}\mathbb{I}(m^{(i)}=2)}$ can be used to approximate the Bayes Factor (BF),

$$B_{12}(\mathbf{y}) = \int_{\Theta_1} p_1(\mathbf{y}|\theta_1)\pi_1(\theta_1)\,d\theta_1 \Bigg/ \int_{\Theta_2} p_2(\mathbf{y}|\theta_2)\pi_2(\theta_2)\,d\theta_2\,, \tag{1}$$

where $p_m(\mathbf{y}|\theta)$ is the likelihood function associated with the m-th model, $m = 1,2$. These approximations are valid as long as identical summaries, distance and tolerance threshold are used over both models.
Robert et al. (2011) thoroughly investigate the drawbacks of model choice in ABC. First of all, they show that the approximated BF, $\hat{B}_{12}(\mathbf{y})$, converges to the BF based on the vector of observed statistics, $\eta(\mathbf{y})$,

$$B^{\eta}_{12}(\mathbf{y}) = \int \pi_1(\theta_1)p_1^{\eta}\big(\eta(\mathbf{y})|\theta_1\big)\,d\theta_1 \Bigg/ \int \pi_2(\theta_2)p_2^{\eta}\big(\eta(\mathbf{y})|\theta_2\big)\,d\theta_2\,, \tag{2}$$

as ε goes to zero. This quantity is only based on the observed vector of statistics and, therefore, insufficient statistics yield a BF which converges to a quantity different from (1). Even in the favourable case where $\eta(\mathbf{y})$ is sufficient for both models, sufficiency *across* models can be hardly obtained and this leads to a discrepancy between (2) and (1) which cannot be computed. Insufficiency of the statistics for models or across models is thus the main source of unreliability of the ABC model choice based on the estimated BF.

2 ABC model choice via mixture weight estimation

The lack of confidence on the classical ABC model choice may be solved by introducing a different kind of modelling: the idea is to replace the inference on the posterior probabilities of the models with the posterior estimate of the weights of a mixture of the candidate models. This approach is an extension to the ABC realm of the inferential procedure proposed by Kamary et al. (2014). Here one considers the case of two candidate models (i.e. $M = 2$) sharing an unknown parameter of interest θ which has a common meaning for both models.
One considers the data $\mathbf{y}$ as produced by a mixture of $\mathscr{M}_1$ and $\mathscr{M}_2$, namely

$$\mathscr{M}_w : \mathbf{y} \sim w\,p_1(\mathbf{y}|\theta) + (1-w)\,p_2(\mathbf{y}|\theta), \quad 0 \le w \le 1\,, \tag{3}$$

with θ following a prior distribution $\pi(\theta)$ and associated to the sampling distribution $p_w(\cdot|\theta)$. The parameters w and θ will be considered independent *a priori*. Notice that this mixture model is an *encompassing* model since it contains both models as special cases: for $w = 1$ it is equivalent to $\mathscr{M}_1$, while for $w = 0$ it is equivalent to $\mathscr{M}_2$. The weight w represents the probability that an observation is sampled from p_1, so that it may be interpreted as the proportion of data which support $\mathscr{M}_1$.

A posterior inference on w may then offer interesting information about which one of the two models is the most suitable according to the observed data as well as the degree of support of one model against the other.
Algorithm 2 shows the way the ABC rejection sampling algorithm is applied on this mixture model in order to estimate θ and w.

Algorithm 2 ABC model choice via mixture weight estimation (ABC-mix)

for $i = 1$ to N **do**
  repeat
    Generate θ^* and w^* from their respective prior distributions
    Generate $\mathbf{z}$ from the sampling distribution $p_{w^*}(\cdot|\theta^*)$ of the model $\mathscr{M}_{w^*}$
  until $\rho\left(\eta(\mathbf{z}),\eta(\mathbf{y})\right) < \varepsilon$
  set $w_i = w^*$ and $\theta_i = \theta^*$
end for

This kind of modelling has several advantages. First of all, a whole posterior distribution on w allows, *inter alia*, for posterior point estimates, region of credibility, quantification of the uncertainty on the results and sensitivity analysis. On the other hand, the BF offers poor information since it is merely a scalar which suggests which model is more adequate and the relative degree of evidence of models. In addition, in the case where a same scalar value provided by the BF may seek either a misspecification of both models or a general acceptance of both - perhaps suggesting a more cautious approach via model averaging -, this new model choice approach may allow to conclude that both models or none could be acceptable, in the sense that a mixture of them may be a better choice. In particular, Kamary et al. (2014) show that this approach leads to a consistent testing procedure, not only when one of the two models is the true one, but also when neither are correct, since the posterior on w tends to concentrate around the value which minimizes the Kullback-Leibler divergence from the true distribution. A further attractive feature of this approach is the fact that it allows for improper priors on θ, where the BF totally prohibits this kind of assumption.
A standard prior for w is a symmetric $Beta(a_0,a_0)$, with $a_0 \in \mathbb{R}^+$ representing the degree of uncertainty *a priori* about the fact that one of the two candidate models is indeed the true one. A small a_0 may offer a regularization tool, being most of the density placed around the boundary values 0 and 1. For $a_0 \downarrow 0$, the proposed values from the prior distribution of w tends to be only 0's and 1's, so that the ABC-mix behaves like the ABC-mc with symmetric prior probabilities on the models. In this sense, the ABC-mc may be substantially seen as a limiting case of the ABC-mix.

3 Simulation study

One considers the case of comparison between two models, that is, the α-stable distribution (Borak, Härdle, and Weron 2005) and the skew-Normal distribution (Azzalini 2013), where the common unknown parameter is the location one. A reparametrisation of the location parameter of the skew-Normal distribution is con-

sidered, so that θ corresponds to the first moment of the two families of distributions. One considers the case where both models are wrong, since the data are iid simulations from a skew-*t* distribution (Azzalini 2013) with expected value equal to 0. The likelihood of the α-stable model is not available in closed form but it is possible to simulate from this model: the inference on the unknown parameter θ and the model choice are thus carried out through ABC-mc and ABC-mix. *Fig. 1* shows that the approximated posterior density of θ provided by ABC-mc is vague and tends to recover the corresponding prior, while ABC-mix produces posterior density estimates which put their masses around 0, in most of the cases.

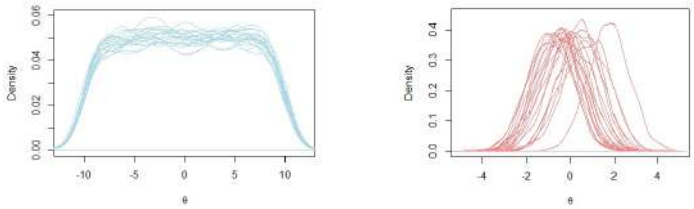


Fig. 1: Density estimate of the posterior distributions of θ *computed via ABC-mc* (left panel) *and ABC-mix* (right panel) *over* 20 *different datasets of* $n = 2000$ *iid samples from a skew-t with mean* $\theta = 0$*. Both algorithms are based on* 10^5 *simulations from* $\theta \sim Unif[-10,10]$*. The prior probabilities of the models are* 0.5 *(for ABC-mc) and* $w \sim Beta(0.5,0.5)$ *(for ABC-mix). The tolerance corresponds to an acceptance rate of* 0.01*. The distance used to measure the discrepancy between the observed and the j-th simulated vector of summaries (*$j = 1,\ldots,N$*) is* $\rho(\eta(\mathbf{y}),\eta(\mathbf{z}^j)) = \sum_{i=1}^{k} \frac{|\eta_i(\mathbf{z}^j)-\eta_i(\mathbf{y})|}{\text{mad}_{j=1}^{N}\eta_i(\mathbf{z}^j)}$*, where the denominator is the median absolute deviation (*mad*) of the i-th summary statistic over all the N simulations: this avoids a distance dominated by the variable with the greatest magnitude. The two summaries (i.e.* $k = 2$*) are the* mad*, which is a good option for ABC model choice (Marin et al. 2014), and the median.*

The difficulty for ABC-mc to provide reasonable posterior density estimates for θ can be understood by analysing the asymptotic behaviour of the posterior probability of the first model (i.e. the α-stable). In fact, *Fig. 2* shows that the posterior probability of the first model seems to converge to 0, by always supporting the second model. On the other hand, the weight of the first component of the mixture of the models, w, does not concentrate near one of the boundary values as the sample size increases. In particular, as n increases, the posterior medians of w tends to concentrate around 0.2, regardless of the prior specifications.

4 Conclusions

ABC-mix has shown better performances than ABC-mc when both models are misspecified, which is the most likely situation in real applications. ABC-mix behaves as a flexible extension of ABC-mc, since a tuning of the hyperparameter a_0 of the prior of the weight w offers a further dimension to the prior specification of the model. This allows to specify the degree of prior uncertainty about the true model and this may be useful as regularization tool or for carrying out sensitivity analysis.

A whole posterior on a mixture weight offers richer information than the one provided by the posterior probability of a model, it allows for measure of uncertainty on the estimates and it may lead to propose a mixture of the candidate models as better choice. Finally, it may be a valid solution to overcome the problem of insufficiency of the summary statistics *for* models and - above all - *across* models which dramatically affects the classical ABC model choice based on the BF.

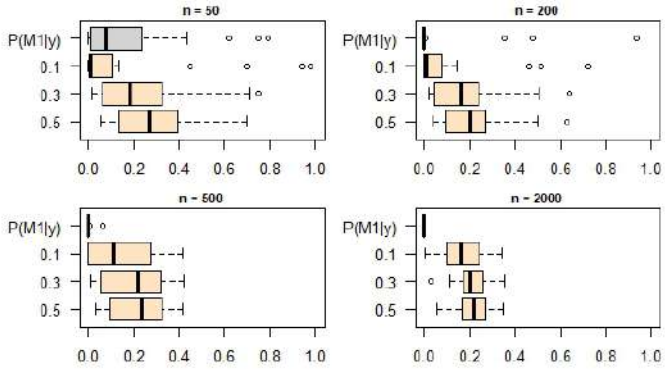


Fig. 2: Boxplot of the posterior medians of w for different values of a_0 (wheat) - *where* a_0 *is the hyperparameter of the prior* $Beta(a_0, a_0)$ *of w - and of the posterior probabilities* (grey) *of the* α*-stable model computed over* 20 *iid datasets from a skew-t with mean* $\theta = 0$ *for different sample sizes (*$n = 50, 100, 500, 2000$*). The posterior of w and the posterior probability of the first model have been estimated via ABC-mix and ABC-mc (respectively), both based on* 10^5 *simulations and an acceptance rate set to* 0.01.

References

Azzalini, Adelchi (2013). *The skew-normal and related families*. Vol. 3. Cambridge University Press.

Borak, Szymon, Wolfgang Härdle, and Rafał Weron (2005). "Stable distributions". In: *Statistical tools for finance and insurance*. Springer, pp. 21–44.

Grelaud, Aude et al. (2009). "ABC likelihood-free methods for model choice in Gibbs random fields". In: *Bayesian Analysis* 4.2, pp. 317–335.

Kamary, Kaniav et al. (2014). "Testing hypotheses via a mixture estimation model". In: *arXiv preprint arXiv:1412.2044*.

Marin, Jean-Michel et al. (2014). "Relevant statistics for Bayesian model choice". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.5, pp. 833–859.

Pritchard, Jonathan K et al. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." In: *Molecular biology and evolution* 16.12, pp. 1791–1798.

Robert, Christian P et al. (2011). "Lack of confidence in approximate Bayesian computation model choice". In: *Proceedings of the National Academy of Sciences* 108.37, pp. 15112–15117.

An ABC algorithm for random partitions arising from the Dirichlet process

Un algoritmo ABC per partizioni aleatorie da un processo di Dirichlet

Mario Beraha and Riccardo Corradin

Abstract Dealing with latent random partitions might be a tedious task, due to mathematically and computationally tractability of the problem. We propose an approximate Bayesian computation (ABC) approach to deal with the estimation of random partitions latent in sets of exchangeable data. Furthermore, we present some preliminary simulation results of the novel proposal, investigating both the quality of the sample produced and the computational time required.

Abstract *Lavorare con partizioni aleatorie latenti può rivelarsi un compito impegnativo, a casua della intrattabilità, matematica e numerica, del problema. In questo lavoro, proponiamo un approccio ABC (Approximate Bayesian Computation) per stimare la partizione latente in una collezione di dati scambiabili. Presentiamo inoltre dei risultati preliminari su un esempio sintetico, dove investighiamo sia la qualità del'approssimazione prodotta e il costo computazionale.*

Key words: ABC, MCMC, random partition, Bayesian statistics

1 Introduction

Approximate Bayesian Computation (ABC) is a vibrant area of research that has been witnessing an exponential growth in the last decade. ABC provides a coherent methodology to deal with problems whose estimation can be intractable. We distinguish two main cases of intractability: when the problem is mathematically

Mario Beraha[1, 2]
[1] Department of Mathematics, Politecnico di Milano, Milano, Italy
[2] Università degli Studi di Bologna, Bologna, Italy
e-mail: mario.beraha@polimi.it

Riccardo Corradin[3]
[3] Department of Economics, Management and Statistics, University of Milano Bicocca
e-mail: riccardo.corradin@unimib.it

intractable, e.g. when the likelihood is not available in analytical form, and when the problem is computationally intractable, i.e. it is not possible to provide an estimation in a reasonable time.

Although the ABC family encompasses a large variety of methods, in this work we are mainly concerns the class of ABC-MCMC algorithms, in the spirit of [7]. The goal of this methods is to perform Markov Chain Monte Carlo sampling from an invariant distribution sufficiently close to the posterior distribution of interest. The basic idea of ABC-MCMC is to replace the evaluation or sampling from intractable distributions with the calculation of a distance d between the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ and a synthetic dataset $\boldsymbol{s} = (s_1, \ldots, s_n)$, generated from a surrogate model. Then, when the true and synthetic data are close, according to a metric d and with respect to a threshold ε, a Metropolis-Hastings step is performed.

The choice of the metric d and threshold ε is then crucial. Choosing the wrong distance or threshold may produce a chain that never moves from the initial state, while choosing a threshold too large may lead to sampling from a law that is very far from the posterior distribution of interest.

In this work, we derive an ABC-MCMC scheme to sample from the approximate law of the latent random partition $\boldsymbol{\rho}$ of a set of exchangeable observations $\boldsymbol{y} = (y_1, \ldots, y_n)$. We will show that, for a particular class of priors on the partitions, it is possible to obtain a proposal that leads to accepting all the values, as long as the distance of the synthetic data from the observed data is smaller than ε. Moreover, by choosing the Wasserstein metric as distance d, the random partition of $\boldsymbol{y}$ can be easily inferred from the random partition of the simulated data.

2 Exchangeable random partitions

The exchangeability assumption on the observed data, in the sense that the law of $\boldsymbol{y}$ is invariant with respect to permutations of the elements, plays a crucial role in our proposal. Furthermore, by assuming exchangeability for the data, the latent random partition $\boldsymbol{\rho}$ is exchangeable too. Let $[n] = \{1, 2, \ldots, n\}$ denotes a set of n integers. A partition of $[n]$ is a collection of non empty and mutually disjoints sets $\{C_1, \ldots, C_k\}$, i.e. $C_i \cap C_j = \emptyset$ for $i \neq j$ and such that $\cup_i C_i = [n]$.

By restricting our attention to the *exchangeable random partitions* cases, the probability distribution of $\boldsymbol{\rho}$ becomes a tractable object. From [8], a random partition $\boldsymbol{\rho}$ of the integers set $[n]$ is exchangeable if and only if $P(\boldsymbol{\rho} = \{C_i, \ldots, C_k\})$ depends on $\{C_i, \ldots, C_k\}$ only through the cardinalities of each block $n_i = |C_i|$, and there exists a symmetric function $\pi_k^{(n)}(n_1, \ldots, n_k)$ such that

$$P(\boldsymbol{\rho} = \{C_i, \ldots, C_k\}) = \pi_k^{(n)}(n_1, \ldots, n_k). \tag{1}$$

The function $\pi_k^{(n)}$ is termed *exchangeable partition probability function* (EPPF) and plays a key role in modern Bayesian nonparametric statistics.

We focus on random partitions structure arising from the Dirichlet process (DP), early introduced in [6], and we denote by ϑ its mass parameter. In this case, the EPPF has the following expression

$$\pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\vartheta^k}{(\vartheta)_{(n)}}\prod_{i=1}^{k}(n_i-1)!, \tag{2}$$

where the $(\vartheta)_{(n)} = \vartheta(\vartheta+1)\ldots(\vartheta+n-1)$ denotes the Pochhammer symbol. Although an EPPF as in (2) can be extended in several directions, mainly to more general structures or to the partially exchangeable case. We furthermore denote by $\boldsymbol{\theta} \sim DP(\vartheta, G_0)$ a sample distributed as a DP with mass parameter ϑ and base measure G_0, such that the distribution of the latent random partition in $\boldsymbol{\theta}$ is fully described by (2).

3 The ABC-MCMC algorithm

Let $\boldsymbol{y}$ be a set of exchangeable observations, and consequentially the latent random partitions $\boldsymbol{\rho}$ object of interest is exchangeable as well. Common strategies to deal with exchangeable latent random partitions in a DP context mostly resort to MCMC methods, see for example [3] and the discussion within. We propose a simple algorithm, also from the implementation's point of view, with the drawback that it produces samples from an approximation of the posterior distribution. Nevertheless, we are able to control the degree of the approximation by varying the threshold parameter ε. The proposed algorithm is essentially an extension of the ABC-MCMC introduced in [7] to the random partition case, where we employ as proposal distribution $q(\boldsymbol{\theta} \to \boldsymbol{\theta}')$ the predictive law of $\boldsymbol{\theta}' \mid \boldsymbol{\theta}$, i.e.

$$P(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = P(\theta_1' \mid \boldsymbol{\theta})P(\theta_2' \mid \theta_1', \boldsymbol{\theta})\ldots P(\theta_n' \mid \theta_1', \ldots, \theta_{n-1}', \boldsymbol{\theta}), \tag{3}$$

where $P(\theta_i' \mid \theta_1', \ldots, \theta_{i-1}', \boldsymbol{\theta})$ denotes the generic predictive distribution for the i-th element, and by assuming a $DP(\vartheta, G_0)$ governing sequence of elements, it becomes

$$P(\theta_i' \in dt \mid \theta_1', \ldots, \theta_{i-1}', \boldsymbol{\theta}) = \frac{\vartheta}{\vartheta+n+i-1}G_0(dt) + \sum_{j=1}^{k}\frac{n_j}{\vartheta+n+i-1}\delta_{\theta_j^*}(dt), \tag{4}$$

where θ_j^* denotes the j-th unique element in $\{\boldsymbol{\theta}, \theta_1', \ldots, \theta_{i-1}'\}$ and n_j its frequency.

Observe that, similarly to Gibbs sampling step, this choice produces an acceptance rate α of the Metropolis-Hastings step equal to 1, indeed

$$\alpha = \min\left(1, \frac{P(\boldsymbol{\theta}')q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}{P(\boldsymbol{\theta})q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}\right) = \min\left(1, \frac{P(\boldsymbol{\theta}')P(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}{P(\boldsymbol{\theta})P(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}\right) = 1.$$

By exploiting a proposal distribution as in (4), we propose a partition that reflects a trade-off between the prior guess, expressed in terms of new values sampled from $G_0(\cdot)$, and the groups already observed in the previous state $\boldsymbol{\theta}$. With this particular choice, we encompass a problem commonly arising in the random partition framework, where a random sampling on the partitions' space leads most of the cases to get an unrepresentative element of the partition of the data.

As mentioned in the introduction section, a sensible choice to be made is the metric for the comparison of the observed and the synthetic data. Despite the multitude of possible distance known in literature, we assume the Wasserstein metric as distance d, in the spirit of [1, 2]. We refer to [10] for technical details and historical background on the Wasserstein distance. In our framework, we aim at comparing the two empirical distributions of observed and synthetic data, $\boldsymbol{y}$ and $\boldsymbol{s}$ respectively, with the same number of elements. Denoting by $W_q(\boldsymbol{y},\boldsymbol{s})$ the Wasserstein distance between the empirical distributions of $\boldsymbol{y}$ and $\boldsymbol{s}$, we have

$$W_q(\boldsymbol{y},\boldsymbol{s}) = \min_{P \in M_{n\times n}} \left(\sum_{i,j} |y_i - s_j|^q P_{ij} \right)^{1/q}, \tag{5}$$

where $M_{n\times n}$ denotes the set of matrices with size $n \times n$, non-negative entries, and with each row and column summing up to n^{-1}. For the framework considered, since $\boldsymbol{y}$ and $\boldsymbol{s}$ have the same number of elements, there exist a solution P^* to the optimization problem (5), which corresponds to a matrix with only one non-zero element for each row and for each column, see [2, 9] for further details. The matrix P^* describes also a permutation $\lambda : \mathbb{N}_n \to \mathbb{N}_n$ such that the distance between $\boldsymbol{y}$ and $\boldsymbol{s}_\lambda$ is minimized, where $\boldsymbol{s}_\lambda$ denotes the set of elements $\boldsymbol{s}$ reordered with respect to λ. Thus we use $\boldsymbol{\rho}_\lambda$, the partition of $\boldsymbol{\theta}'_\lambda$ latent parameters of the reordered vector $\boldsymbol{s}_\lambda$, as proposed elements in the ABC-MCMC algorithm.

Algorithm 1: ABC-MCMC

```
[1] Inizialize θ^(0) and ρ^(0);
[2] for each iteration r = 1,...,R do
[3]     propose a move from θ^(r−1) to θ' with proposal distribution q(θ^(r−1) → θ')
          according to (3) and (4);
[4]     sample s vector of synthetic data according to a kernel, S_i ~ k(s, θ'_i);
[5]     if W_q(y, s) ≤ ε then
[6]         accept ρ'_λ as new state and go to the next iteration;
[7]     else
[8]         go back to [3];
[9] end
```

In Algorithm 1 we report the pseudo-code for the implementation of a generic ABC-MCMC sampling scheme, proposing the synthetic data from a kernel function $k(s,\theta)$ and with proposal distribution according to (3) and (4). Due to the generality

of algorithm 1, the same procedure can be applied to either univariate and multivariate observed data $\boldsymbol{y}$ by setting an opportune kernel function $k(s,\theta)$.

4 Simulation study

In this paper, we report the results of a preliminary simulation study. We consider sets of simulated data $\boldsymbol{y} = y_1,\ldots,y_n$ coming from a mixture of equally weighted Gaussian components

$$y_i \overset{\text{iid}}{\sim} \frac{1}{3}\mathcal{N}(-5,1) + \frac{1}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(5,1) \quad i = 1,\ldots,n,$$

and we compare the proposed ABC-MCMC algorithm with the marginal sampler for the Dirichlet mixture model (DPM) implemented in the R package `BNPMix` [4]. For both models, the point estimates of the partition were obtained the least square procedure described in [5]. In order to get a fair comparison, we set both the likelihood in the DPM $f(y \mid \theta)$ and the synthetic data generating kernel k in 1 to be the Gaussian density with parameters $\theta = (\mu,\sigma^2)$. We set as G_0 in (4) and the base measure in the DPM are the Normal-inverse-gamma prior with matching parameters. We consider four different sample sizes $n \in \{40,80,160,320\}$, and the study was replicated 100 times. Convergence of the chains was assessed by visual investigations of randomly selected replicates, without finding any indication against it.

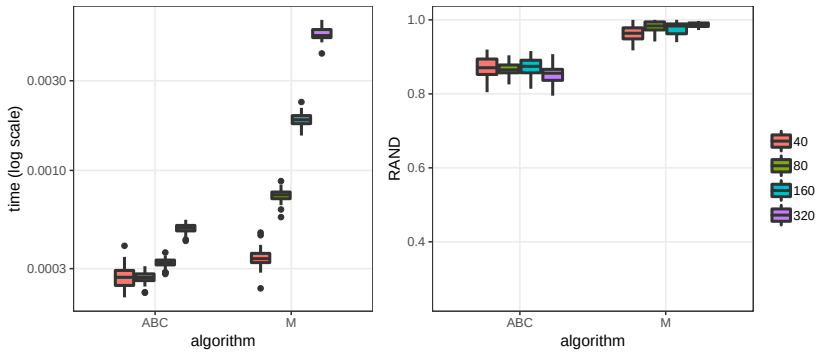


Fig. 1: Left plot: time for a single iteration (log-scale). Right plot: RAND index of the partition estimate from the posterior distribution and the true partition. Different sample sizes. In each plot a comparison of ABC-MCMC algorithm (ABC) and the marginal sampler (M). Results aggregated over 100 of replications.

Figure 1 reports the result of both algorithms. It is clear that the our model outperforms the marginal sampler in terms of time per iteration, while retaining a comparable ability to estimate the latent partition of the data.

5 Discussion and future work

In this work we presented an ABC-MCMC algorithm for the estimation of random partitions. The two main novelties in our approach were using the predictive law $P(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$ as proposal distribution in the ABC algorithm and the use of Wasserstein metric to evaluate the distance of observed and synthetic data.

There are numerous criticalities to our approach that we want to address in the future. First of all, it remains unclear how to properly choose the parameter ε, the threshold value used to accept or discard a synthetic dataset. This threshold has an impact on the efficiency of our algorithm and we would like to derive a strategy to avoid its specification.

The second main question to be addressed is the extension to the multivariate setting. Although the theoretical development presented remains valid, computational issues may arise in the evaluation of the Wasserstein distance, and we aim to explore possible numerical approximate solutions to solve this step.

Out proposal is a first step in the direction to define a general ABC framework to random partitions, a key quantity in the Bayesian nonparametric setup. Indeed, although not explored here, we could use our ABC-MCMC algorithm to expand the range of application of the Bayesian nonparametric methodology to those models whose likelihood is intractable.

References

1. E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society: Series B*, 81(2):235269, Feb 2019.
2. E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 10 2019.
3. A. Canale, R. Corradin, and B. Nipoti. Importance conditional sampling for pitman-yor mixtures, 2019.
4. R. Corradin, A. Canale, and B. Nipoti. *BNPmix: Bayesian Nonparametric Mixture Models*, 2019.
5. D. B. Dahl. Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, 4:201–218, 2006.
6. T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
7. P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.
8. J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.
9. C. Villani. Topics in optimal transportation theory. 58, 01 2003.
10. C. Villani. *Optimal transport – Old and new*, volume 338. 2008.

Bayesian Inference of Undirected Graphical Models from Count Data

Inferenza bayesiana di modelli grafici adirezionali da dati di conteggio

Pier Giovanni Bissiri, Monica Chiogna, Nguyen Thi Kim Hue

Abstract Bissiri et al. (2016) present a general Bayesian approach where the likelihood is replaced more generally by a loss function, which is used to derive the posterior distribution from the prior. In this paper, we exploit this idea for learning the structure of undirected graphical models over discrete variables.

Abstract *Bissiri et al. (2016) presentano un approccio bayesiano generalizzato in cui la verosimiglianza è sostituita da una funzione di perdita più generale, che viene utilizzata per derivare la distribuzione a posteriori dalla distribuzione a priori. In questo lavoro, l'approccio viene contestualizzato al problema di apprendimento della struttura di modelli grafici non direzionati per variabili discrete.*

Key words: loss functions, general Bayesian approach, graphical models, undirected graphs, structure learning

1 General Bayes

Bissiri et al. (2016) propose a framework for general Bayesian inference which is based on the updating of a prior belief distribution to a posterior when the parameter of interest is connected to observations via a loss function rather than the traditional likelihood function, which is recovered as a special case. The parameter of interest θ_0 is the value minimizing the expected loss with respect to the unknown population probability distribution F_0 that generates the data. In other words, θ_0 minimizes the loss

Pier Giovanni Bissiri
Department of Statistical Sciences, University of Bologna, e-mail: piergiovanni.bissiri@unibo.it

Monica Chiogna
Department of Statistical Sciences, University of Bologna, e-mail: monica.chiogna2@unibo.it

Nguyen Thi Kim Hue
Department of Statistical Sciences, University of Padova e-mail: nguyen@stat.unipd.it

$$\theta \in \Theta \to \int l(\theta, \mathbf{x}) \mathrm{d}F_0(\mathbf{x}), \tag{1}$$

where Θ denotes the parameter space and $l(\cdot, \mathbf{x})$ is a loss function on Θ for every piece of data $\mathbf{x}$.

From a classical perspective, a typical procedure for estimating the parameter θ is based on the minimization of the cumulative loss,

$$\sum_{i=1}^{n} l(\theta, \mathbf{x}_i) \tag{2}$$

where θ varies in Θ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the data. The loss (2) is clearly the finite sample version of $\int l(\theta, x) \mathrm{d}F_0(x)$.

In what follows, let the data be $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. Typically, in statistics, a parametric family of densities (or probability mass functions) $\mathscr{F} = \{f(\mathbf{x}, \theta) : \theta \in \Theta, \mathbf{x} \in \mathbb{R}^p\}$ is considered. In this setting, a loss function on the parameter space Θ arises naturally, which is called the self–information loss and is equal to the negative log–likelihood, namely:

$$\theta \to \log f(\mathbf{x}, \theta),$$

for every $\mathbf{x} \in \mathbb{R}^p$. In this case, let the unknown value θ_0 be defined as the minimum point of the expected loss (1) where $l(\theta, \mathbf{x})$ is the self–information loss. If the density of the unknown population probability distribution F belongs to $\mathscr{F}$ then θ_0 turns out to be the value such that the population density coincides with $f(\mathbf{x}, \theta_0)$. If instead the density of the unknown population probability distribution F does not belong to $\mathscr{F}$, then θ_0 is the value in Θ minimizing the Kullback–Leibler divergence between the probability distributions defined by the densities in $\mathscr{F}$ and the unknown population probability distribution F_0. Moreover, the estimate obtained minimizing (2) is the maximum likelihood one.

Due to robustness reasons, one may wish to replace the self–information loss with a more general loss $l(\cdot, \mathbf{x})$ which does not rely on particular assumptions about the unknown population distribution (see Hüber 2009). Estimators defined through minimization of losses of the form (2) are called *M–estimators*. They turn out to be particularly useful when a probability model for the data is too difficult to formulate or is unavailable. Another common circumstance in statistics is when the likelihood is intractable and is therefore replaced by a pseudo–likelihood. In such scenario, it is reasonable to let loss $l(\cdot, \mathbf{x})$ be the negative log-pseudo-likelihood and to obtain an estimate of the parameter θ_0 through minimization of the loss (2). Bissiri et al. (2016) present a general Bayesian extension of this idea of making inference while replacing the traditional likelihood with a general loss function. They aim at updating a prior distribution π of θ on the basis of the loss function l and they claim that the appropriate posterior distribution should be absolutely continuous with respect to the prior π with density proportional to $\exp\{-\lambda l(\cdot, \mathbf{x})\}$, where λ is a positive constant. This general form for a posterior distribution is also considered by Zhang (2006*a*,*b*), Jiang & Tanner (2008), Bissiri & Walker (2010) and Bissiri & Walker (2012).

Bissiri et al. (2016) and Bissiri & Walker (2019) characterize the general Bayesian approach through a set of assumptions, or axioms. The result given by Bissiri et al. (2016) is restricted to the case of Θ being a finite probability space, but Bissiri & Walker (2019) have improved this result including the case of Θ being a subset of the real line, besides reorganizing the set of axioms which the result is based on.

Let π be a (prior) probability measure on the parametric space Θ, let Θ be a (Borel) subset of the real numbers, and for every $\mathbf{x} \in \mathbb{R}^p$ let $l(\cdot,\mathbf{x})$ be a loss function defined on Θ that is a measurable function valued into $[0,\infty]$ and not identically infinite. Bissiri et al. (2016), Bissiri & Walker (2019) argue that a valid and coherent update of π on the basis of $\mathbf{x}$ is the probability measure $\pi_{\mathbf{x}}$ that is absolutely continuous with respect to π and

$$\mathrm{d}\pi_{\mathbf{x}}/\mathrm{d}\pi(\theta) = \frac{\exp\{-\lambda l(\theta,\mathbf{x})\}}{\int_{\Theta}\exp\{-\lambda l(\tau,\mathbf{x})\}\mathrm{d}\pi(\tau)} \tag{3}$$

for some $\lambda > 0$. Their argument is based on a set of axioms. In particular, the updating rule (3) ensures that a very natural coherence property is satisfied, since we end up with $\pi_{\mathbf{x}_1,\mathbf{x}_2}$ as the same object whether we update with $(\mathbf{x}_1,\mathbf{x}_2)$ together on the basis of the additive loss function $l(\cdot,\mathbf{x}_1)+l(\cdot,\mathbf{x}_2)$ or we update with $\{\mathbf{x}_1,\mathbf{x}_2\}$ one after the other, using first the loss $l(\cdot,\mathbf{x}_2)$ and then the loss $l(\cdot,\mathbf{x}_2)$. See Bissiri & Walker (2019) for all details.

In presence of n pieces of data $\mathbf{x}_1,\ldots,\mathbf{x}_n$, in virtue of this coherence property, the general posterior distribution takes the following form:

$$\mathrm{d}\pi_{\mathbf{x}_1,\ldots,\mathbf{x}_n}/\mathrm{d}\pi(\theta) = \frac{\exp\{-\lambda\sum_{i=1}^{n} l(\theta,\mathbf{x}_i)\}}{\int_{\Theta}\exp\{-\lambda\sum_{i=1}^{n} l(\tau,\mathbf{x}_i)\}\mathrm{d}\pi(\tau)}$$

The assessment of the calibrating parameter λ is a delicate issue. See, for instance, Bissiri et al. (2016), Holmes & Walker (2017), Syring & Martin (2018). In this paper, we consider the loss $l(\theta,\mathbf{x})$ being equal to $-\log L(\theta,\mathbf{x})$, where $L(\theta,\mathbf{x})$ is a pseudo–likelihood. Being such pseudo–likelihood intended as an approximation of the true likelihood, it is reasonable to assess $\lambda = 1$.

2 The proposal

Consider a p dimensional vector $\mathbf{X} = (X_1,\ldots,X_p)$ such that each random variable X_s corresponds to a node of a graph $G = (V,E)$ with index set $V = \{1,2,\ldots,p\}$. An edge between two nodes s and t will be denoted by (s,t). The neighbourhood of a node $s \in V$ is defined to be the set $N(s) = \{t \in V : (s,t) \in E\}$ consisting of all nodes connected to s. The random vector $\mathbf{X}$ satisfies the Markov property, namely X_s and $\mathbf{X}_{V\setminus\{N(s)\cup\{s\}\}}$ are conditionally independent given $\mathbf{X}_{N(s)}$.

When the structure of a graph is not known, its recovery is not an easy task. On assuming that the graph is perfectly Markovian, that is, it satisfies the global Markov property and its reverse implication, known as faithfulness, neighborhood

selection strategies learn the graph G by determining the neighborhood $N(s)$ of each node s. Estimating $N(s)$ often corresponds to variable selection in a (penalized) regression problem (Allen & Liu 2013) . Hue Nguyen & Chiogna (2018) propose a new algorithm for learning the structure of undirected graphs for (possibly right-truncated) count data by substituting regression strategies with hypothesis testing, following the lines of the PC algorithm (Spirtes et al. 2000). To infer a graphical model from data, PC-like algorithms use sequences of conditional independence tests. Hue Nguyen & Chiogna (2018) prove that their proposal has high-dimensional consistency properties when conditional independences are tested using Wald-type tests of conditional independence.

Although consistent in the limit of infinite observations, in finite settings performances of the algorithm may depend on the choice of the statistic used to perform the tests and on the chosen significance level. In this work, we try to overcome these limitations by adopting a Bayesian perspective (see also Banerjee & Ghosal (2015) for a different approach in the Bayesian setting).

To this aim, we assume that each conditional distribution of node X_s given other variables $\mathbf{X}_{V\setminus\{s\}}$ follows a Poisson distribution truncated at R, $R > 0$

$$\mathbb{P}_{\boldsymbol{\theta}}(x_s|\mathbf{x}_{V\setminus\{s\}}) = \exp\left\{\theta_s x_s + x_s\langle\boldsymbol{\theta}_s, \mathbf{x}_{V\setminus\{s\}}\rangle - \log x_s! - D(\langle\boldsymbol{\theta}_s, \mathbf{x}_{V\setminus\{s\}}\rangle)\right\}, \quad (4)$$

where $\boldsymbol{\theta} = \{\theta_{st} : s,t \in V, s \neq t\}$, $\boldsymbol{\theta}_s = \{\theta_{st},\ t \in V, t \neq s\}$ denotes the set of conditional dependence parameters, $\langle .,. \rangle$ denotes the inner product, and $D(\langle\boldsymbol{\theta}_s, \mathbf{x}_{V\setminus\{s\}}\rangle) = \log\left(\sum_{k=0}^{R} \exp\left\{\theta_s k + k\langle\boldsymbol{\theta}_s, \mathbf{x}_{V\setminus\{s\}}\rangle - \log k!\right\}\right)$. A missing edge between node s and node t corresponds to the condition $\theta_{st} = \theta_{ts} = 0$. On the other side, one edge between node s and node t implies $\theta_{st} \equiv \theta_{ts} \neq 0$.

The neighborhood selection strategy is related to the concept of pseudo-likelihood. We borrow the idea of resorting to a pseudolikelihood and consider the function:

$$L(\boldsymbol{\theta}) = \prod_{s\in V} \mathbb{P}_{\boldsymbol{\theta}}(x_s \mid \mathbf{x}_{V\setminus\{s\}}).$$

Let $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ be the collection of n samples with $\mathbf{x}^{(i)} = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. In such setting, the generalized posterior is absolutely continuous with respect to the prior with density which is equal to:

$$\frac{\prod_{i=1}^{n}\prod_{s\in V}\mathbb{P}_{\boldsymbol{\theta}}\left(x_{is} \mid \mathbf{x}^{(i)}_{V\setminus\{s\}}\right)}{\mathbb{E}_{\pi}\left(\prod_{i=1}^{n}\prod_{s\in V}\mathbb{P}_{\tilde{\boldsymbol{\theta}}}\left(x_{is} \mid \mathbf{x}^{(i)}_{V\setminus\{s\}}\right)\right)}, \quad (5)$$

where the expectation in the denominator is intended to be taken with respect to a random element $\tilde{\boldsymbol{\theta}}$ with (prior) distribution π. We consider a prior π of the parameter vector $\boldsymbol{\theta}$ such that the elements θ_{st} of $\boldsymbol{\theta}$ are i.i.d. for every pair (s,t) such that $s,t \in V$ and $s < t$. Such prior is constructed in the following way: let $\{W_{st} : t,s \in V, s \neq t\}$ be a collection of i.i.d. random variables such that, for each s,t, W_{st} is equal to 1 with probability p, where $p \in (0,1)$, and is equal to zero with probability $1-p$ and let the θ_{st} be equal to zero if W_{st} is zero and let the conditional distribution of θ_{st}

given $W_{st} = 1$ be Normal with mean zero and variance σ^2, where σ^2 is a large value. In other words, $\theta_{st} = X_{st} W_{st}$ where X_{st} and W_{st} are independent and X_{st} is Gaussian with zero mean and variance σ^2. This is also equivalent to say that the distribution function of θ_{st} is equal to:

$$P(\theta_{st} \leq t) = (1-p)\,\delta_0((-\infty, t]) + p\,\Phi(t/\sigma),$$

for every $t \in \mathbb{R}$, where δ_0 is the probability distribution of a random variable degenerate at zero and Φ denotes the standard Normal distribution function.

The generalized posterior given by (5) can be simulated via the Metropolis–Hastings algorithm. Conditional independence tests can be based on the posterior probability of each individual coefficient, this being larger or smaller that 1/2.

Through extensive simulation studies, validity of the proposal can be illustrated numerically, along with its scalability to large graphs.

3 Conclusions

We presented a general Bayesian nonparametric approach for learning the structure of a graphical model when count data are at hand. The approach leverages the use of a pseudo likelihood as a loss function. Compared with the more common frequentist approaches, this strategy allows to deal with situations in which the existence of the joint distribution of the variables is not guaranteed. We developed appropriate procedures for posterior computation.

References

Allen, G. & Liu, Z. (2013), 'A local Poisson graphical model for inferring networks from sequencing data', *NanoBioscience, IEEE Transactions on* **12**(3), 189–198.

Banerjee, S. & Ghosal, S. (2015), 'Bayesian structure learning in graphical models.', *J. Multivariate Anal.* **136**, 147–162.

Bissiri, P. G., Holmes, C. C. & Walker, S. G. (2016), 'A general framework for updating belief distributions', *J. Roy. Statist. Soc. Ser. B* **78**(5), 1103–1130.

Bissiri, P. G. & Walker, S. G. (2010), 'On Bayesian learning from Bernoulli observations', *J. Statist. Plann. Inference* **140**(11), 3520–3530.

Bissiri, P. G. & Walker, S. G. (2012), 'Converting information into probability measures with the Kullback–Leibler divergence', *Ann Inst Stat Math* **64**(6), 1139–1160.

Bissiri, P. G. & Walker, S. G. (2019), 'On general Bayesian inference using loss functions.', *Statist. Probab. Lett.* **152**, 89–91.

Holmes, C. C. & Walker, S. G. (2017), 'Assigning a value to a power likelihood in a general Bayesian model', *Biometrika* **104**(2), 497–503.

Hüber, P. (2009), *Robust Statistics*, 2nd edn, John Wiley & Sons Inc, Hoboken, NJ, USA.

Hue Nguyen, T. K. & Chiogna, M. (2018), 'Structure learning of undirected graphical models for count data', *ArXiv e-prints* . 1810.10854.

Jiang, W. & Tanner, M. (2008), 'Gibbs posterior for variable selection in high-dimensional classification and data mining', *Ann. Statist.* **36**, 2207–2231.

Spirtes, P., Glymour, C. N. & Scheines, R. (2000), *Causation, prediction, and search*, MIT press.

Syring, N. & Martin, R. (2018), 'Calibrating general posterior credible regions', *Biometrika* **106**(2), 479–486.

Zhang, T. (2006*a*), 'From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation', *Ann. Statist.* **34**, 2180–2210.

Zhang, T. (2006*b*), 'Information theoretical upper and lower bounds for statistical estimation', *IEEE Trans. Inform. Theory* **52**, 1307–1321.

Bayesian IRT models in NIMBLE

Modelli IRT bayesiani in NIMBLE

Sally Paganin, Chris Paciorek, Perry de Valpine

Abstract IRT models relate observed data to some latent traits typically encoding item characteristics, as well as individual abilities. Often these lasts are assumed to follow a standard normal distribution, but there are situations in which such assumption may be unrealistic. A possible extension for such models uses a Dirichlet process mixture of normal distributions, which is seldom employed in real data analysis due to the lack of guidelines and software tools. We contribute to fill this gap by reviewing both parametric and semiparametric versions of such models. Using 2PL model as an example we also illustrate how these models can be easily implemented via the novel NIMBLE software.

Abstract *I modelli IRT caratterizzano la relazione tra dati osservati e variabili latenti. Quest'ultime solitamente descrivono sia caratteristiche degli item che l'abilità degli individui. Spesso si assume che tali abilità seguano la distribuzione di una normale standard, ma ci sono situazioni in cui tale assunzione non risulta appropiata. In questi casi, una possibile estensione si può ottenere utilizzando una mistura di distribuzioni normali per le abilità latenti basata sul processo di Dirichlet. Tuttavia questi modelli risultano essere poco diffusi in pratica a causa della mancanza di linee guide e software che li implentino. In questo lavoro presentiamo le versioni parametriche e non dei modelli IRT. Usando il modello 2PL come esempio, illustriamo anche come implementarli facilmente attraverso l'uso del software NIMBLE.*

Key words: 2PL, Bayesian nonparametrics, IRT, Dirichlet Process, NIMBLE.

1 Motivation

Item response theory (IRT) refers to a family of models that investigate the relationship between responses to a set of items and some latent traits, typically encoding

Department of Statistics, Department of Enviromental Science, Policy & Management
UC Berkeley e-mail: sally.paganin@berkeley.edu

individual or item characteristics. Such models are employed in different application domains, with educational measurement and psychometrics being the most popular. Models for binary responses are among the most common among IRT models, comprising the one, two or three parameter logistic models (1PL, 2PL and 3PL). These models assume that the probability of a correct answer is related to the individual's latent ability, as well as items difficulty and potentially other item characteristics.

Standard approaches rely on the assumption that latent abilities follow a standard normal distribution. This assumption is sometimes considered for computational convenience, but it may be unrealistic in many situations [9]. For example, [7] gives a comprehensive review of many psychometric datasets where the latent traits distribution does not respect the normality assumption and presents instead asymmetries, heavy-tails or multimodality.

Different proposals have been made in literature relaxing the normality assumption for the latent abilities. Arguably, the most general approach in a Bayesian framework uses a Dirichlet process [4] mixture of normal distributions as a nonparametric distributions for latent abilities. Such models are semiparametric because they retain other, parametric, assumptions of binomial mixed models. Within this approach, the semiparametric 1PL model has been the focus of more effort as well as software tool [6, **DPpackage**]. [10] investigate semi-parametric generalization of Rasch-type models from a theoretical perspective, while [5] provides results from simulation studies considering the 1PL model. An example using the 2PL model is given in [3], but there is a lack of comprehensive studies of such models as well as general tools for model estimation in real data analysis. In this work we review the semiparametric 1PL and 2PL models, and illustrate how to easily implement them in the NIMBLE software.

2 NIMBLE

NIMBLE [2] is a flexible R-based system for hierarchical modeling, which extends BUGS language used in WinBUGS, OpenBUGS, NAGS [8], providing efficient execution of algorithms via custom-generated C++ code. Besides offering new degrees of customization of MCMC algorithms, one of the latest NIMBLE features added support for MCMC inference for Bayesian nonparametric mixture models. In particular, NIMBLE provides functionality for fitting models using a Dirichlet process prior, either via the Chinese Restaurant Process (CRP) [1] or a truncated stick-breaking (SB) [11] representation of the Dirichlet process prior. These features allows Dirichlet process priors to be embedded in very general hierarchical models, supporting extensions of the approaches we illustrate here.

3 IRT models background

In this context, observed data are typically answers to exam questions or items from a set of individuals. Let y_{ij} denote the answer of an individual j to item i for $j = 1,\dots,N$ and $i = 1,\dots,I$, with $y_{ij} = 1$ when the answer is correct and 0 otherwise. Typically, different individuals are assumed to work independently, while responses from the same individuals are assumed independent conditional to the latent trait (*local independence assumption*). Hence each answer y_{ij}, conditionally to the latent parameters, is assumed to be a realization of a Bernoulli distribution, and the probability of a correct response is typically modeled via logistic regression.

In the two-parameter logistic (2PL) model, the conditional probability of a correct response is modeled as

$$\Pr(y_{ij} = 1|\eta_j,\lambda_i,\beta_i) = \frac{\exp\{\lambda_i(\eta_j-\beta_i)\}}{1+\exp\{\lambda_i(\eta_j-\beta_i)\}} \quad i = 1,\dots,I, j = 1,\dots,N, \tag{1}$$

where η_j represents the latent ability of the j-th individual for $j = 1,\dots,N$, while β_i and λ_i encode the item characteristics for $i = 1,\dots,I$. The parameter λ_i is often referred as *discrimination*, since items with a large λ_i are better at discriminating between subjects with different abilities, while β_i is called *difficulty* because the probability of a correct response is equal to 0.5 when $\eta_j = \beta_i$. Discrimination parameters λ_i are typically assumed positive. When $\lambda_i = 1$ for $i = 1,\dots,I$ model in (1) reduces to the one-parameter logistic (1PL) model. Often, conditional log-odds in (1) are reparametrized as $\lambda_i\eta_j + \gamma_i$, with $\gamma_i = -\lambda_i \times \beta_i$. Sometimes this is reffered to as *slope-intercept* (SI)parameterization as opposed to the *IRT* parameterization in (1) traditional cosidered for interpretation.

Traditional literature assumes that $\eta_j \sim \mathcal{N}(0,1)$ for $j = 1,\dots,N$, but there are situations in which such assumption can be too restrictive. To add more flexibility, we can extend the model in (1) via a DP prior as

$$\eta_j|G \sim G \quad G \sim DP(\alpha,G_0) \tag{2}$$

where α is the concentration parameter and G_0 the base measure. The DP process is often represented via the Chinese Restaurant Process representation, introducing a set of indicator variables z_j for $j = 1,\dots,N$ indicating the cluster assignment for the ability η_j. The prior in (2) becomes

$$(\eta_j|z_j = h) = \eta_h \quad \eta_h \sim \mathcal{N}(\mu_h,\sigma_h^2) \tag{3}$$

with typically hyperpriors on μ_h and σ_h^2.

4 Model comparison

We compare estimation of the parametric and nonparametric estimation of the parametric 2PL models via simulation. Typically parameters of the 2PL model are not identifiable, so constraints are either included in the model or one can post-process posterior samples to meet the constraints. We consider this last option and use sum-to-zero constraints on the item parameters, i.e. $\sum_{i=i}^{I}\beta_i = 0, \sum_{i=i}^{I}\log(\lambda_i) = 0$ and estimate the 2PL under IRT and SI parameterizations.

We simulate data from two different scenarios changing the distribution generating the latent abilities. We simulate responses from $N = 2,000$ individuals to $I = 20$ binary items. Values for the discrimination parameters $\{\lambda_i^0\}_{i=1}^{20}$ are sampled from a $Unif(0.5, 1.5)$, while values for difficulty parameters $\{\beta_i^0\}_{i=1}^{20}$ are taken as equally spaced between $(-3, 3)$. In particular we considered for the latent abilities η_j^0 for $j = 1, \ldots, 2000$:

1. **Unimodal scenario**. Latent abilities comes from a normal distribution with mean 0 and standard deviation 1.25.
2. **Bimodal scenario**. Latent abilities comes from a mixture of two normal distribution with means $\{-2, 2\}$ and common standard deviation 1.25.

We implement all the strategies in NIMBLE, choosing moderately vague priors for the item parameters, $\beta_i \sim \mathcal{N}(0,3), \gamma_i \sim \mathcal{N}(0,3), \log(\lambda_i) \sim \mathcal{N}(0.5, 0.5)$ for $i = 1, \ldots, I$. We assume normal latent abilities $\eta_j \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2)$ for $j = 1, \ldots, N$, and placed a $\mathcal{N}(0,3)$ on μ_η and $Unif(0, 10)$ on the standard deviation σ_η in the parametric case, while in the nonparametric setting we choose $G_0 \equiv \mathcal{N}(0,3) \times Inv-Gamma(1.01, 2.01)$. We run the MCMC for $50,000$ iterations using a 10% burn-in of $5,000$ iterations, and check traceplots for convergence.

Table 1 reports the minimum effective samples size (ESS) per second relative to the strategies, computed by dividing the ESS for the computation time. As expected there is a loss in efficiency when moving from the parametric to the semiparametric specification, given that sampling from the Dirichlet Process requires more computational effort. Flexibility comes with a price, but also with a benefit for inference when abilities are not normal. While in the unimodal scenario results match, in the bimodal there are substantial differences.

	unimodal	simulation	bimodal	simulation
Model	parametric	bnp	parametric	bnp
IRT unconstrained	0.78	0.33	2.43	1.27
SI unconstrained	1.06	0.57	0.23	0.06

Table 1 Minimum ESS/seconds for different estimations strategies of the 2PL model parameters under the two simulated scenarios.

For example, Figure 1 compares the density estimates of the posterior mean latent abilities from the parametric and semiparametric models, computed taking the posterior means of the $\{\eta_j\}_{j=1}^{N}$. It can be notice that the parametric model detect just

one mode because of the underlying normal assumption, while the semiparametric specification recover the true density structure. Better estimation of the latent abilities helps to avoid bias in inference, for example when estimating item parameters or item characteristics curves (ICC).

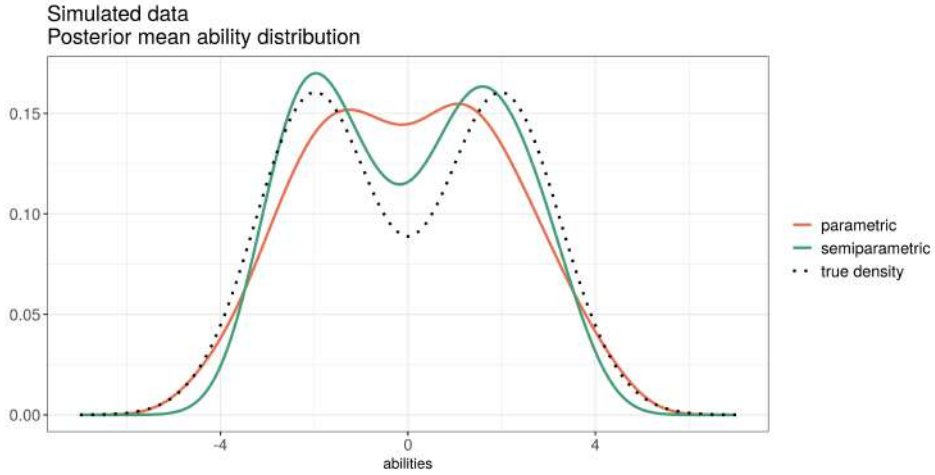


Fig. 1 Density estimates of the posterior mean latent abilities under the parametric and semiparametric 2PL models under the bimodal simulated scenario. Both models are estimated under the unconstrained IRT parameterizations, the most efficient from Table 1.

References

[1] D. Blackwell, J. B. MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[2] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming with models: Writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

[3] K. A. Duncan and S. N. MacEachern. Nonparametric bayesian modelling for item response. *Statistical Modelling*, 8(1):41–66, 2008.

[4] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 03 1997.

[5] H. Finch and J. M. Edwards. Rasch model parameter estimation in the presence of a nonnormal latent trait using a nonparametric bayesian approach. *Educational and Psychological Measurement*, 76(4):662–684, 2016.

[6] A. Jara, T. E. Hanson, F. A. Quintana, P. Müller, and G. L. Rosner. Dppackage: Bayesian semi-and nonparametric modeling in r. *Journal of statistical software*, 40(5):1, 2011.

[7] T. Micceri. The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, 105(1):156, 1989.

[8] M. Plummer. Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 124(125):10, 2003.

[9] F. Samejima. Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, 62(4): 471–493, 1997.

[10] E. SanMartín, A. Jara, J.-M. Rolin, and M. Mouchart. On the bayesian nonparametric generalization of irt-type models. *Psychometrika*, 76(3):385–409, Jul 2011. ISSN 1860-0980.

[11] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

Bayesian modelling of Facebook communities via latent factor models

Modellazione Bayesiana di comunità Facebook tramite modelli a fattori latenti

Emanuele Aliverti

Abstract Network data are routinely collected in a variety of applications, ranging from marketing to social sciences. In this article I focus on the analysis of a large Facebook network involving thousands of public pages. The dependence structure is characterized trough a Bayesian latent factor model for network data, which allows to model the architecture of nodes' connectivity as a function of their position in a latent space. Inference proceeds via Variational Bayes leveraging a Coordinate Ascent routine. Results indicate good fit for the algorithm, while inference on the latent coordinates suggests the presence of different communities.

Abstract *Dati di rete vengono regolarmente raccolti in diversi campi applicativi. In questo articolo mi concentro su una grande rete Facebook composta da alcune migliaia di pagine pubbliche. La struttura di dipendenza è caratterizzata attraverso un modello Bayesiano a fattori latenti per dati di rete, che consente di modellare la connettività dei nodi in funzione della loro posizione in uno spazio latente. La stima procede tramite un approccio variazionale. I risultati indicano un buon adattamento del modello, mentre l'inferenza sulle coordinate latenti suggerisce la presenza di diverse strutture di comunità.*

Key words: Bayesian Inference, Factor Model, Network Data, Variational Bayes

1 Introduction

Network data are ubiquitous in many areas of science and industry. Some notable examples are social science [10], biology [9] and neuroscience [5, 2], where network analysis can provide valuable insight into the functionality of an entire system. In practice, it is interesting to study various aspects of network data, ranging from sim-

Emanuele Aliverti
Department of Statistical Sciences, University of Padova e-mail: aliverti@stat.unipd.it

ple descriptive statistics to the full specification of the network generating process and its dependence structure. In general terms, a network can be defined as a collection of n interconnected units (*nodes*), which can be conveniently expressed as an $(n \times n)$ *adjacency* matrix $\mathbf{Y}$ with elements y_{ij} characterising the connection (*edges*) from node i to node j. In this article I focus on binary undirected networks, leading to symmetric adjacency matrices. Therefore, it is sufficient to characterise the lower-triangular part of $\mathbf{Y}$, thereby letting $y_{ij} = y_{ji} = 1$ if there is and edge between the pair (i, j) with $i = 2, \dots, n$ and $j = 1, \dots, i-1$ and 0 otherwise. Specifically, I will focus on a large network involving $n = 22470$ public Facebook pages collected through the Facebook API in November 2017 [11]. A link between page i and page j is present if they share a mutual like relationship. My interest is on characterising the dependence structure of such network, to provide insights on the community structure underlying such Facebook community. Such an aim is accomplished via a Bayesian Latent Factor Model (LFM) for networks, which characterizes the probability to observe and edge between two Facebook pages as a function of their position in an H-dimensional latent space. The more similar two nodes are in the latent space, the more likely it is to observe a connection among them. Such a structure facilitates interpretation on the underlying community structure, reducing the dimensionality from $n(n-1)/2$ to nH free parameters. In addition, posterior inference is performed efficiently leveraging a scalable Variational Bayes algorithm.

2 Latent factor model

Let $\text{pr}(y_{ij} = 1 \mid \pi_{ij}) = \pi_{ij} \in (0,1)$, denote the population probability of an edge between node i and j, for each $i = 2, \dots, n$, $j = 1, \dots, i-1$. The LFM for networks [6] is specified as follows.

$$\begin{aligned} (y_{ij} \mid \pi_{ij}) &\sim \text{Bern}(\pi_{ij}) \\ \pi_{ij} &= \Phi(\eta_{ij}) \\ \eta_{ij} &= \beta_0 + w_i^{\intercal} w_j \end{aligned} \tag{1}$$

for each pair $i = 2, \dots, n$ and $j = 1, \dots, i-1$, with Φ denoting the Gaussian cumulative distribution function. The vector $(w_i) \in \mathbb{R}^H$ corresponds to the position of the node i in the H-dimensional latent space. Therefore, the quantity $\eta_{ij} \in \mathbb{R}$ can be interpreted as a weighted similarity among region i and j in the latent space, with more similar regions having larger similarities resulting in greater probability of being connected. See [7] for a recent overview and a comparison among the LFM and other latent variable methods for networks.

I seek simple specification of the prior distributions, leading to efficient computational algorithms. With this motivation in mind, the prior distributions for the coefficients and latent structures are specified as follows.

$$\beta_0 \sim \text{N}(0,1), \quad w_i \sim \text{N}_H(0, I_H) \quad i = 1, \dots, n \tag{2}$$

Conditional conjugacy is obtained leveraging the data-augmentation strategy of [1], introducing a set of latent continuous observations $z_{ij} \sim \mathrm{N}(\eta_{ij}, 1)$. This choice implies that the observed $y_{ij} = 1$ if $z_{ij} > 0$ and $y_{ij} = 0$ if $z_{ij} < 0$. Integrating out z_{ij} the specification of Equation (1) is obtained.

2.1 Bayesian inference via Variational Bayes

Combining the LFM with the data augmentation strategy of [1], the model can be revised in terms of a conditional conjugate exponential family. Beyond facilitating inference via Gibbs sampler, the availability of closed-form expressions for full conditional distributions also simplifies Bayesian estimation via Variation Bayes (VB) [3, 4].

Specifically, I focus on a Mean Field (MF) restriction, conducting approximate Bayesian inference via optimisation. Such a strategy aims at finding the closest distribution in KL divergence to the true posterior subject to the MF product restriction. In particular, the following restriction will be assumed for the variational family of distributions.

$$q(\beta_0, \mathbf{W}, \mathbf{z}) = q(\beta_0) q(\mathbf{W}) q(\mathbf{z}), \tag{3}$$

with $\mathbf{W} = \{w_i\}_{i=1}^{n}$, $\mathbf{z} = \{z_{ij}\}_{i=2,\dots n}^{j=1,\dots,i}$

Since the LFM falls within the class of conditionally conjugate exponential families, each full conditional distribution — available in closed form — is in the exponential family, being either multivariate Gaussian or truncated normal. This result is particularly useful since, under the MF factorisation in (3), it is possible to show [e.g. 4, Sec 2.4] that the optimal distributions $q^\star$ have closed form expressions. Indeed, the optimal distribution for each factor is in the same parametric (exponential) family of the corresponding full conditional distribution, with natural parameters replaced with variational expectations [8] which can be easily computed for both the Gaussian and the truncated Gaussian full conditionals. The optimal solution can be found iteratively, maximising each variational distribution on the basis of the current values of the remaining parameters, until convergence.

3 Results

The approach is applied to the Facebook network described in Section 1, focusing for simplicity on $H = 2$. Figure 1 reports the posterior mean of the latent coordinates w_i and the ROC curve, using the posterior predictive mean $\mathbb{E}[\pi_{ij} \mid \mathbf{Y}]$ for predicting the edge probabilities. Such a quantity can be easily obtained via Monte-Carlo integration, relying on an independent sample from the approximate posterior distribution. Results indicated a good performance for the estimated model, with an Area Under the ROC curve close to 0.84; see the right panel of Figure 1. In addition,

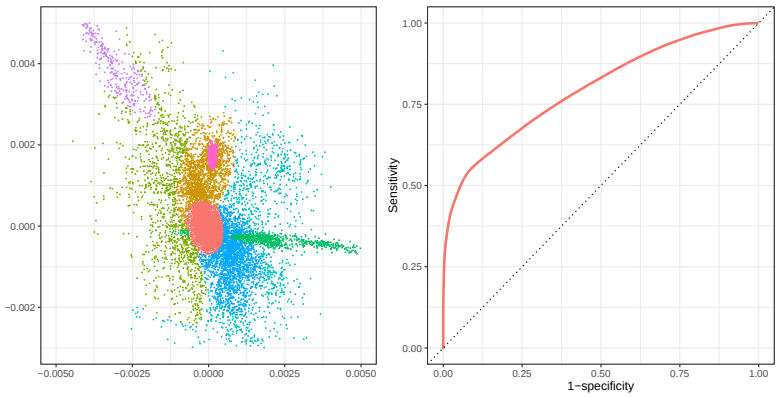


Fig. 1 Left panel: latent factors (w_{i1}, w_{i2}), $i = 1, \ldots, n$. Right panel: ROC curve.

the estimates for the latent positions (posterior means, left panel of Figure 1) indicate the presence of different communities. To improve visualisation, I performed a model-based clustering via Gaussian mixtures in the latent space. Results suggest the presence of different overlapping communities, which might correspond to Facebook pages referring to similar topics or discussing similar threads.

References

[1] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

[2] Emanuele Aliverti and Daniele Durante. Spatial modeling of brain connectivity data via latent distance models with nodes clustering. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):185–196, 2019.

[3] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[6] Peter D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664, 2008.

[7] Peter D. Hoff. Additive and multiplicative effects network models. *Statistical Sciente (to appear)*, 2019.

[8] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[9] Pall F. Jonsson, Tamara Cavanna, Daniel Zicha, and Paul A. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC bioinformatics*, 7(1):2, 2006.

[10] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[11] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding, 2019.

Bayesian nonparametric adaptive classification with robust prior information

Modello Bayesiano nonparametrico per classificazione adattiva con informazione a priori robusta

Francesco Denti, Andrea Cappozzo and Francesca Greselin

Abstract In a standard classification framework, a discriminating rule is usually built from a trustworthy set of labeled units. In this context, test observations will be automatically classified as to have arisen from one of the known groups encountered in the training set, without the possibility of detecting previously unseen classes. To overcome this limitation, an adaptive semi-parametric Bayesian classifier is introduced for modeling the test units, where robust knowledge is extracted from the training set and incorporated within the priors' model specification. A successful application of the proposed approach in a real-world problem is addressed.

Abstract Di solito, in un problema di classificazione, si costruisce una regola discriminante in base ad un insieme affidabile di unità etichettate. E' così possibile, quindi, attribuire le osservazioni di un dataset di test ad uno dei gruppi noti, presenti nel training set. Non vi è invece possibilità di rilevare classi mai viste prima. In molti contesti applicativi, tuttavia, può accadere che emergano nuove classi. Per rispondere a questa necessità, si introduce un classificatore bayesiano semi-parametrico, che estrae informazione robusta dal dataset di training, la incorpora come prior knowledge ed è in grado di includere nuove classi per modellare le unità del test set. Viene poi presentata un'applicazione dell'approccio proposto su dati reali.

Key words: Supervised classification, Unobserved classes, Bayesian adaptive learning, Bayesian mixture model, Stick-breaking prior

Francesco Denti
Department of Statistics, University of California Irvine, e-mail: fdenti@uci.edu

Andrea Cappozzo • Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: a.cappozzo@campus.unimib.it; francesca.greselin@unimib.it

1 Introduction and Motivation

The usual framework of supervised classification does not contemplate the possibility of having test units belonging to a class not previously observed in the learning phase. A classic hypothesis is that the training set contains samples for each and every group within the population of interest. Nevertheless, this strong assumption may not hold true in fields like biology, where novel species may appear and their detection is an important issue, or in social network analysis where communities continuously expand and evolve. Therefore, a classifier suitable for these situations needs to adapt to the detection of previously unobserved classes, accounting also for few extreme and outlying observations that may emerge in such evolving ecosystems. Unfortunately, standard supervised methods will predict class labels only within the set of groups previously encountered in the learning phase.
We propose a flexible procedure in a semi-parametric Bayesian framework for dealing with outliers and hidden classes that may arise in the test set. The learning process articulates in two phases. First, we infer the structure of the known components from the labeled set via standard robust procedures. Consequently, employing an Empirical Bayes rationale, the dynamic updating typical of Bayesian statistics is adopted to model the new, unlabeled dataset allowing for the detection of possibly infinite new components.
The rest of the paper is organized as follows: in Section 2 the main features of the novel model are presented. An application to the discrimination of wheat kernels varieties, under sample selection bias, is reported in Section 3. Section 4 summarizes the contributions and highlights future research directions.

2 The model

Consider a classification framework with $\{(\mathbf{x}_1,\mathbf{l}_1),\dots,(\mathbf{x}_N,\mathbf{l}_N)\}$ identifying the training set: $\mathbf{x}_n$ is a p-variate observation and $\mathbf{l}_n$ its associated group label, $\mathbf{l}_n \in \{1,\dots,G\}$ with G the number of unique observed classes. Correspondingly, let $\{\mathbf{y}_1,\dots,\mathbf{y}_M\}$ be the test set, where it is assumed, differently from the standard framework, that its associated (unknown) labels may not only belong to the set of previously observed G classes, but potentially more groups could be present within the unlabeled units. That is, there may be a number H of novel classes in the test such that the total number of groups in the population is $E = G+H$, with $H \geq 0$.
We assume that each observation in the test set is generated from a mixture of $G+1$ elements: G densities $f(\cdot|\boldsymbol{\Theta}_g)$ parametrized by $\boldsymbol{\Theta}_g$ and an extra term, called *novelty* component. In formulas:

$$\mathbf{y}_m \sim \sum_{g=1}^{G} \pi_g f(\cdot|\boldsymbol{\Theta}_g) + \pi_0 f_{nov}, \tag{1}$$

where π_g, $g = 1,\dots,G$ indicates the prior probability of observing class g (already present in the learning set), while π_0 is the probability of observing a previously

unseen class, such that $\sum_{g=0}^{G} \pi_g = 1$. Different specifications for the known components can be easily accommodated in the general formulation of (1): Gaussian distributions will be subsequently considered, in line with the application reported in Section 3. A Bayesian nonparametric approach is employed to model f_{nov}. In particular, we resort to the Dirichlet Process Mixture model [1, 4], imposing the following structure:

$$f_{nov} = \int f(\cdot|\boldsymbol{\Theta}^{nov}) G(d\boldsymbol{\Theta}^{nov}), \qquad G \sim DP(\gamma, H),$$

where $DP(\gamma, H)$ is the usual Dirichlet process with concentration parameter γ and base measure H. Note that we use the superscript *nov* to denote a parameter relative to the novelty part of the model. Adopting Sethuraman's Stick Breaking construction [7], we can express the likelihood as follows:

$$\mathscr{L}(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega}) = \prod_{m=1}^{M} \left[\sum_{g=1}^{G} \pi_g \phi\left(\mathbf{y}_m | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) + \pi_0 \sum_{h=1}^{\infty} \omega_h \phi\left(\mathbf{y}_m | \boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}\right) \right], \quad (2)$$

where $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density, parametrized by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. There are two main reasons for employing a nonparametric prior in this context. First, adopting a DP as mixing measure allows an a priori unbounded number of hidden classes and/or outlying observations. Second, it reflects our lack of knowledge about the previously unseen components. The following prior probabilities for the parameters complete the Bayesian model specification:

$$\begin{aligned} \left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) &\sim NIW\left(\hat{\boldsymbol{\mu}}_{gMCD}, \tilde{\lambda}_{tr}, \tilde{\nu}_{tr}, \hat{\boldsymbol{\Sigma}}_{gMCD}\right), \quad g = 1, \ldots, G \\ \left(\boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}\right) &\sim NIW\left(\tilde{\boldsymbol{m}}, \tilde{\lambda}, \tilde{\nu}, \tilde{\boldsymbol{S}}\right), \quad h = 1, \ldots, \infty \\ \boldsymbol{\omega} &\sim SB(\gamma) \qquad \boldsymbol{\pi} \sim Dir(\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_G). \end{aligned} \quad (3)$$

A detailed explanation of the quantities in (3) follows, where we incorporate the information contained in the training set for setting robust informative priors for the parameters of the known classes. Values $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_G$ are the hyper-parameters of a Dirichlet distribution on the known classes. The learning set can be exploited to determine reasonable values of such hyper-parameters, setting $\tilde{\alpha}_g = n_g/N$ for $g = 1, \ldots, G$ with n_g the total number of observations belonging to the g-th group in the training set. The priors for the mean vectors and the covariance matrices of both known and hidden classes are assumed to follow a conjugate Normal-inverse-Wishart distribution. Robust hyperparameters $\hat{\boldsymbol{\mu}}_{gMCD}$ and $\hat{\boldsymbol{\Sigma}}_{gMCD}$, $g = 1, \ldots, G$ are obtained via the Minimum Covariance Determinant estimator (MCD) [6] computed group-wise in the training set. Subsets of sizes $\lceil 0.95 n_g \rceil$, $g = 1, \ldots, G$, over which the determinant is minimized, are employed in the application of Section 3. In this way, outliers and label noise that may be present in the labelled units will not bias the initial beliefs for the parameters of the known groups. Lastly, with $\boldsymbol{\omega} \sim SB(\gamma)$

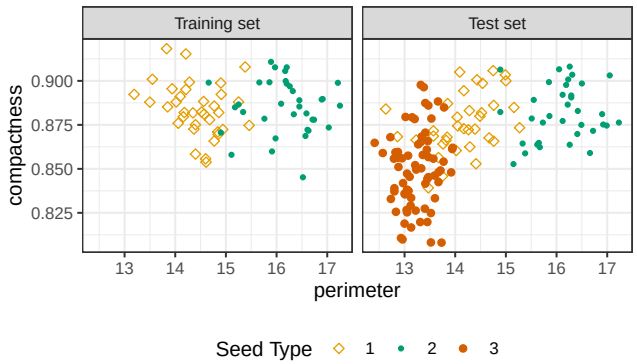


Fig. 1 Learning scenario (only `perimeter` and `compactness` variables displayed) for novelty detection of 1 unobserved wheat variety, seed dataset.

we denote the vector of Stick-Breaking weights, composed of elements defined by $w_k = u_k \prod_{l<k}(1-u_l)$, where $\forall k\ u_k \sim Beta(1,\gamma)$.
We remark that particular care is needed in choosing informative values for $\tilde{\lambda}_{tr}$ and $\tilde{\nu}_{tr}$, according to the problem at hand: inducing non-informative priors would jeopardize the robust extraction of information performed with the MCD estimator.
A blocked Gibbs sampler scheme [5] is employed for posterior computation, wherein the full conditionals for the model parameters are derived considering the following *complete likelihood*, obtained after proper reparameterization:

$$\mathscr{L}(\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{u}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\omega}) = \prod_{m=1}^{M}\Big[\pi_{\alpha_m}\mathbb{1}_{\{\alpha_m>0\cap\beta_m=0\}}\phi\left(\mathbf{y}_m|\boldsymbol{\mu}_{(\alpha_m,0)},\boldsymbol{\Sigma}_{(\alpha_m,0)}\right) + \\ + \pi_0\omega_{\beta_m}\mathbb{1}_{\{\alpha_m=0\cap\beta_m>0\}}\phi\left(\mathbf{y}_m|\boldsymbol{\mu}^{nov}_{(0,\beta_m)},\boldsymbol{\Sigma}^{nov}_{(0,\beta_m)}\right)\Big],$$

where $\alpha_m \in \{0,\dots,G\}$ and $\beta_m \in \{0,\dots,\infty\}$ are latent variables identifying the unobserved group membership for $\mathbf{y}_m$, $m = 1,\dots,M$. In detail, $\alpha_m > 0$ controls the assignation of $\mathbf{y}_m$ to one of the G training class ($\beta_m = 0$). In case $\mathbf{y}_m$ is recognized as a novel observation ($\alpha_m = 0$), it is assigned to one of the previously unseen mixture components constituting f_{nov}, according to $\beta_m > 0$.

3 Application

The methodology described in Section 2 is used to perform classification when a novelty component is present within the data units. The considered dataset contains 210 grains belonging to three different varieties of wheat. For every sample (70 units for each variety), seven geometric parameters are recorded postprocessing X-ray photograms of the kernel [3]. The obtained dataset is publicly available

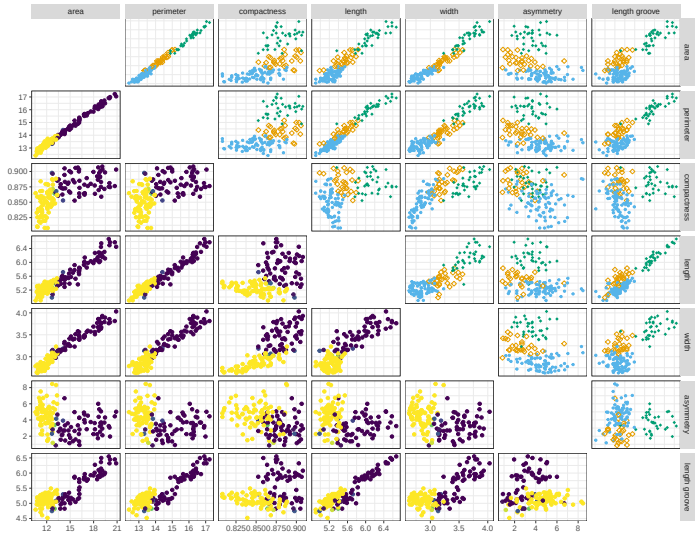


Fig. 2 Model results for experimental scenario H_3, seeds test set. Plots below the main diagonal represent the estimated posterior probability of being a novelty, according to formula (4): the brighter the color the higher the probability of belonging to f_{nov}. Plots above the main diagonal display the associated group assignments: the turquoise solid dots denote observations classified as novelties.

in the University of California, Irvine Machine Learning data repository. The study

Table 1 `Seeds dataset`. Confusion matrix between the true values (T:) and the semi-parametric Bayesian classification (C:) performed on the test set for three hyperpriors specification. The label "New" indicates units that are estimated to have arisen from the novelty component.

	$H_1: \tilde{\lambda}_{tr}=1, \tilde{\nu}_{tr}=50$			$H_2: \tilde{\lambda}_{tr}=10, \tilde{\nu}_{tr}=500$			$H_3: \tilde{\lambda}_{tr}=250, \tilde{\nu}_{tr}=1000$		
	T:1	T:2	T:3	T:1	T:2	T:3	T:1	T:2	T:3
C:1	1	35	0	27	0	3	26	0	3
C:2	34	0	70	1	35	0	0	35	0
C:New	0	0	0	7	0	67	9	0	67

involves the random selection of 70 training units from the first two cultivars, and a test set of 140 samples, including the entire set of 70 grains from the third variety: the resulting learning scenario is displayed in Figure 1. The aim of the experiment is therefore to employ the model described in Section 2 to detect the third unobserved variety, incorporating robust priors information retrieved from the training set. We employ three different hyperpriors specification ($H_1: \tilde{\lambda}_{tr}=1, \tilde{\nu}_{tr}=50$, $H_2: \tilde{\lambda}_{tr}=10, \tilde{\nu}_{tr}=500$, $H_3: \tilde{\lambda}_{tr}=250, \tilde{\nu}_{tr}=1000$) to investigate how the results change when different degrees of informativeness are adopted for the values extracted from the learning set. Model results for scenario H_3 are reported in Fig-

ure 2, where the posterior probability of being a novelty $PPN_m = \mathbb{P}[\mathbf{y}_m \sim f_{nov}|\mathbf{Y}]$, $m = 1,\dots,M$ are estimated according to the ergodic mean:

$$\hat{PPN}_m = \frac{\sum_{t=1}^{T} \mathbb{1}(\alpha_m^{(t)} = 0)}{T} \tag{4}$$

where $\alpha_m^{(t)}$ is the value assumed by the parameter α_m at the t-th iteration of the MCMC chain and T is the total number of iterations. The confusion matrices associated with the estimated group assignments for the three scenarios are reported in Table 1. When fairly and strongly informative priors for the training set information are adopted, the third group variety is effectively captured by the flexible process modeling the novel component. Notice that, whenever the novelty contains more than an extra class, its best partition can be recovered minimizing for example the Binder loss [2] or the Variation of Information [8], thus providing a way to automatically identify an infinite number of hidden classes, as well as anomalous and/or unique outlying patterns.

4 Conclusion

In the present work we have introduced an adaptive semi-parametric Bayesian classifier, capable of detecting an unbounded number of hidden classes in the test set. By means of robust procedures, prior knowledge for the known groups is reliably incorporated in the model specification. The methodology has been then effectively employed in the detection of a novel wheat variety in X-ray images of grain kernels. Future research directions will consider data-tailored extensions to the general "known classes + novelty" mixture framework introduced in this paper: a flexible specification for adaptive classification of functional data is being developed.

References

1. C. E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, nov 1974.
2. D. A. Binder. Bayesian Cluster Analysis. *Biometrika*, 65(1):31, apr 1978.
3. M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Zak. Complete gradient clustering algorithm for features analysis of X-ray images. *Advances in Intelligent and Soft Computing*, 69:15–24, 2010.
4. M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, jun 1995.
5. H. Ishwaran and L. F. James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, mar 2001.
6. P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, aug 1999.
7. J. Sethuraman. A constructive definition of Dirichlet Process prior. *Statistica Sinica*, 4(2):639–650, 1994.
8. S. Wade and Z. Ghahramani. Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

Choosing the right tool for the job: a systematic analysis of general purpose MCMC software

Scegliere lo strumento giusto: un'analisi sistematica di software MCMC di uso generale

Mario Beraha, Giulia Gualtieri, Eugenia Villa, Riccardo Vitali and Alessandra Guglielmi

Abstract We perform a throughout comparison of three of the most widely used software for Bayesian computation: JAGS, Stan and NIMBLE. We compare the performance of each software in several different settings, corresponding to four different class of models, each with three different priors and different sample sizes and dimensionality of the parameter space. By looking at several performance metrics, we are able to create a decision tree that should guide practitioners approaching Bayesian modelling into the choice of the right tool for the statistical problem under investigation.

Abstract *Confrontiamo tre dei software più utilizzati per il calcolo della distribuzione a posteriori in un modello bayesiano, e cioè JAGS, Stan e NIMBLE. In particolare confronteremo le prestazioni di ogni software in diversi esempi, corrispondenti a quattro distinte classi di modelli, ognuno con tre prior diverse e con dimensione campionarie e parametriche che variano. Dall'analisi di alcune metriche di prestazioni, saremo in grado di costruire un albero delle decisioni per guidare i ricercatori non esperti dell'argomento nella scelta dello strumento giusto per il problema statistico considerato.*

Key words: Bayesian statistics, MCMC efficiency, MCMC convergence, probabilistic programming

Mario Beraha[1, 2], Giulia Gualtieri[1], Eugenia Villa[1], Riccardo Vitali[1] and Alessandra Guglielmi[1]
[1] Department of Mathematics, Politecnico di Milano, Milano, Italy
[2] Università degli Studi di Bologna, Bologna, Italy
e-mail: {giulia.gualtieri, eugenia.villa, riccardo3.vitali}@mail.polimi.it
e-mail: {mario.beraha, alessandra.guglielmi}@polimi.it

1 Introduction

It is well known that in the Bayesian setting, the posterior distribution, i.e. the conditional law of parameters given data, is not available in closed analytic form but under simple models. From 1990's, novel techniques that go by the name of Markov Chain Monte Carlo (MCMC) sampling have been developed to simulate draws from the posterior distribution of interest. However, despite the good theoretical properties of MCMC, deriving an efficient sampler can be demanding even for successful statisticians, due both to mathematical and implementation/computational issues.

One of the main reasons behind the success, in recent years, of the Bayesian framework among practitioners, is that a number of "probabilistic" programming languages have originated. This kind of software enables the user to specify the likelihood function as well as the prior distribution of parameters using an intuitive syntax and then automatically perform MCMC sampling from the posterior. Although united by the same target, each software exploits different numerical algorithms underneath, and also employs a different syntax, hence exhibiting different behaviours.

In this work, we investigate the *performance*, under different metrics, for three types of such general purpose software, i.e. JAGS, Stan and NIMBLE. In particular, given a class of Bayesian models, such as, for example, generalized linear models or accelerated failure time models, and the size of the dataset, we aim at providing practitioners guidance in choosing the right software, that is the most efficient, for MCMC sampling.

2 Overview of the software under investigation

JAGS stands for "Just another Gibbs Sampling"[4]. It is a program for the analysis of Bayesian models using Markov Chain Monte Carlo (MCMC), like OpenBUGS, it includes a shared C library for BUGS language interpretation and MCMC computation. BUGS proceeds by creating a graph where the nodes are the parameters in the model and data, and then it chooses the "best" sampler for each node among a list of candidates. Finally, the MCMC computation is done by updating a node at a time, in a Gibbs sampler fashion. See [2] for reference to the main ideas beneath BUGS (Bayesian inference Using Gibbs Sampling) software and to the project that started it in 1989.

NIMBLE [3] extends the BUGS language, hence the same sampling strategy of JAGS is employed. However, as we shall see, the choice of the sampler for each node can be different in the two languages and thus produce different results.

Stan [1], differently from JAGS which is able to choose the best sampling method basing on the implemented model, can use only two algorithms for MCMC sam-

pling: Hamiltonian Monte Carlo (HMC) and its adaptive variant called NUTS (No-U-Turn Sampler) which is the one set by default. Being HMC based on the computation of gradients of the log joint posterior distribution with respect to parameters, it is worth remembering that Stan does not allow the direct use of discrete random parameters, while discrete response variables are supported. This can represent a limitation of Stan with respect to BUGS languages.

3 Methodology

We consider four different classes of models: linear models, generalized linear models, linear mixed effects models and accelerated failure time models. We believe that in practical applications these classes cover most of the needs of practitioners.

For each model we consider three different priors on the parameters: an hierarchical prior, an extremely non informative prior and a simpler (though less flexible) non-hierarchical prior, that is chosen as the conjugate prior when available. Hence we end up with twelve different Bayesian models, where each model is fitted using the three software we consider, and for any of them we analyze the MCMC output and the computational performance.

Here, we report detail only on the example of the homoscedastic linear model, for lack of space. The likelihood is given by

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathscr{N}(0, \sigma^2) \quad i = 1, ..., N$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the regression parameters, which may include the intercept, and $\mathbf{x}_i$ is a p–dimensional vector of covariates. We specify three priors:

Zellner's	**hierarchical**	**noninformative**
$\boldsymbol{\beta}\|\sigma^2 \sim \mathscr{N}_p(\boldsymbol{\beta}_0, \sigma^2 B_0)$	$\boldsymbol{\beta}\|\Sigma \sim \mathscr{N}_p(\boldsymbol{\beta}_0, \Sigma)$	$\beta_j \overset{\text{iid}}{\sim} \mathscr{N}(0,100)\ j = 1, \ldots, p$
$B_0 = c(X^T X)^{-1}$	$\Sigma \sim IW(S_0^{-1}, \eta_0)$	$\sigma \sim \mathscr{U}(0,100)$
$\sigma^2 \sim IG(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$	$S_0^{-1} = \sigma^2 B_0$	
$c \sim \mathscr{U}(0, 10^6)$	$\sigma^2 \sim IG(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$	

By $IG(a,b)$ we mean the inverse-gamma distribution with mean $b/(a-1)$, while $IW(S_0^{-1}, \eta_0)$ denotes the inverse-Wishart distribution, with mean $S_0^{-1}/(\eta_0 - p - 1)$ and $\mathscr{U}(a,b)$ is the uniform distribution on the real interval (a,b).

We briefly describe the second model we include here, namely the linear mixed effects model, with likelihood given by

$$Y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma}_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathscr{N}(0, \sigma^2) \quad i = 1, ..., N_j,\ j = 1, \ldots, n_{gr}$$

where n_{gr} is the number of groups, $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jq})$ is the vector of random effects parameters, and $\mathbf{z}_i$ is a q–dimensional vector of covariates. Similarly to the case of the linear regression model we consider three priors, i.e. a hierarchical prior (with exchangeable marginal priors for all the parameters in $\boldsymbol{\beta}$ and for all parameters $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{n_{gr}}$), and two priors where all the components in the vectors $\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{n_{gr}}$ are independent (in one case with noninformative marginal priors).

For each class of models we have generated multiple datasets, with different sample sizes or different sizes of the parameters (i.e. varying p in the linear model case). Data were independently generated directly from the likelihood, having fixed a "true" value of the parameters.

The metrics we consider for assessing the performance of the software are (i) the convergence to the true values, (ii) the number of iterations to reach convergence, (iii) the amount of RAM required to run the MCMC, (iv) the effective sample size, (v) the total time elapsed. We believe that these metrics provide a throughout description of the performance of each software. Of course, the convergence of the chain, measured by $\hat{R}$ is a prerequisite for each software.

We run several times the MCMC chains for each model and each software, and average the results for each metric.

For each model, changing the prior and the dimensionality of the problem, we pick a winner based on the following criteria. First of all $\hat{R}$ should be close to 1, and we decided to exclude those software that presented $\hat{R} > 1.1$. Secondly, the effective sample size should be as large as possible, so that if one software chain shows a much larger effective sample size compared to the others, it would be elected winner. In case only these two metrics were close among different software, which was the case for most of the analysed scenarios, we would go on and choose the fastest software. Finally, in case of ties of total elapsed time, we would either declare a tie or choose the software that occupies fewer RAM, but just in case that the RAM usage exceeded 2GBs.

Based on the description above, we come up with a decision tree for each model, where the nodes represent the different priors and the dimensionality of the problem, while the leaves denote the software that we recommend to choose. Figure 1 reports two of such trees, for the linear model and the linear mixed effects model.

4 Discussion and conclusion

Despite an ad-hoc discussion should be made for each model, we try to summarize our insights so far as follows. In general, Stan is more sensitive to prior selection, since the gradient computation required in the NUTS algorithm is really complex and high computationally demanding. This is why, looking at the MCMC output obtained by Stan, we have better performances with non-informative priors, since it is easy to compute the gradient of uniform densities. For the same reason, using the conjugate prior in Stan does not imply higher computational efficiency.

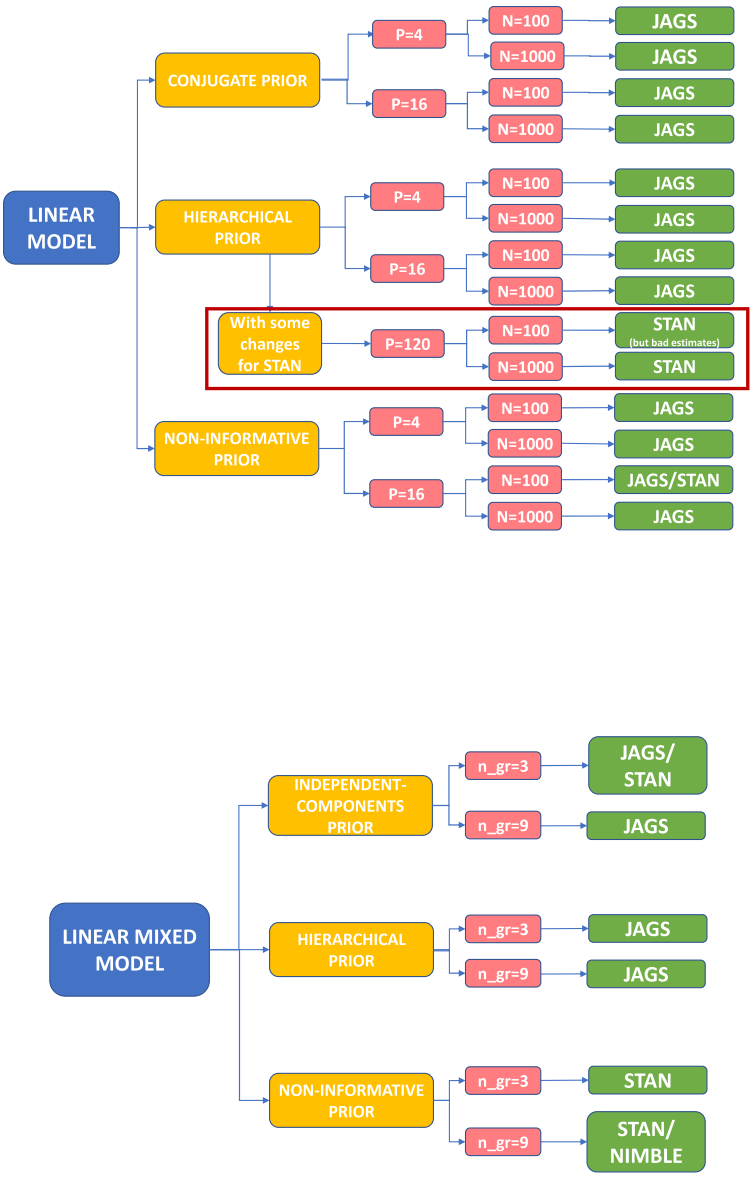


Fig. 1: The leaves denote the best choice of software for two classes of models. Top tree, linear model: p denotes the dimension of the regression parameters and N is the sample size. Bottom tree, linear mixed effects model: `n_gr` is the number of groups.

The importance of the parametric family assumed as the prior is also clear in the case of the most complex linear model (with the largest p), where, using the multivariate normal prior for $\boldsymbol{\beta}$, Stan was not able to move the chains from the initial points. In this case, the Stan manual suggests priors that are more efficient from a computational point of view, to be adopted with hyperparameters that give the prior belief available (e.g. via prior moments of the parameters). This proves that model specification is important when comparing software performance and may influence the performance itself.

It is also important to underline that, in more than one case, NIMBLE was not able to recognize conjugacy in the full conditionals, leading to an inappropriate choice of the sampling method and therefore in inefficiency when building the MCMC chains. As a general result, we have evidence that NIMBLE behaves similarly to JAGS but it is slightly less efficient in terms of effective sample size. However, when the number of data and/or parameters increases, we notice that NIMBLE becomes much faster, in terms of total elapsed time, than JAGS.

Concerning our advice in software selection, results from this procedure show to be in agreement for almost all tested models but the linear model, showing Stan as a general good choice, while JAGS should be used only for well-suited problems, namely with conjugate priors.

The most significant positive feature of Stan is its efficiency, namely the ability to create highly uncorrelated chains, thanks to the underlying algorithm. However, as a drawback, in some cases Stan requires more time to build the chains than other software, especially with specific priors. In these latter cases, such as the linear model and some examples of the linear mixed effects model, considering the ratio of effective samples size over time, one might favor JAGS over Stan, given that the total elapsed time in JAGS is much smaller.

However, when model complexity increases, under the same partial prior information (e.g. fixing the same prior moments of the parameters) if we choose the prior appropriately, Stan results again to be the most efficient software. This has clearly emerged from the analysis of the performance of the software for generalized linear models and accelerated failure time models.

References

1. B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.
2. D. Lunn, C. Jackson, N. Best, D. Spiegelhalter, and A. Thomas. *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC, 2012.
3. NIMBLE Development Team. Nimble: MCMC, particle filtering, and programmable hierarchical modeling, 2019. R package version 0.9.0.
4. M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria, 2003.

Empirical Bayes estimation for mixture models

Stima empirico-bayesiana per modelli mistura

Catia Scricciolo

Abstract Mixture models are a useful tool for analysing complex data. The Bayesian approach to inference offers a vast choice of prior laws on the mixing measure, which induce priors on the mixture density. The posterior distribution of any summary of the mixing distribution can be derived. This note reviews the current status of mixture density estimation, focussing on recent results on posterior contraction rates for the mixing distribution in the 1-Wasserstein metric. Viewing the mixing measure as a genuine prior, the single mixing parameters can be estimated exploiting the relationship, formalized by Tweedie's formula, with the marginal density.
Abstract I modelli mistura costituiscono un utile strumento per l'analisi di dati complessi. L'approccio bayesiano all'inferenza offre una vasta gamma di leggi iniziali sulla distribuzione misturante che inducono misure di probabilità a priori sulla mistura. È possibile derivare la distribuzione finale di ogni sintesi della misturante. Questa nota fornisce una rassegna sullo stato dell'arte della stima di misture di densità, focalizzando l'attenzione su recenti risultati concernenti la velocità di convergenza nella metrica di Wasserstein della legge finale della distribuzione misturante. Considerando quest'ultima come una distribuzione a priori, i singoli parametri della mistura possono essere stimati sfruttando il legame con la densità marginale stabilito dalla formula di Tweedie.

Key words: Mixture Model, Tweedie's Formula, Wasserstein Metric

1 Introduction

Mixture models

$$p_G(\cdot) = \int k(\cdot \mid \theta)\, \mathrm{d}G(\theta),$$

where k denotes a known base density and G an unknown, nonparametric mixing distribution, provide a useful tool for analysing clustering and making inference about heterogeneity. They can also be a device for density approximation or

Catia Scricciolo
Università degli Studi di Verona, Via Cantarane, 24, 37129 Verona,
e-mail: catia.scricciolo@univr.it

estimation and for performing automatic model selection. Such models are fundamental in empirical Bayes compound decision settings [11] too, see Section 3. The problem is to estimate the mixing distribution G on the base of n independent and identically distributed (iid) observations $X_1, \ldots, X_n$, or $X_{1:n}$ for short, from p_G. When θ is a location parameter, $k(\cdot \mid \theta) = k(\cdot - \theta)$ is a location model and $p_G(\cdot) = \int k(\cdot - \theta)\,\mathrm{d}G(\theta)$ a convolution, therefore we have a deconvolution problem. An alternative approach to the classical nonparametric maximum likelihood estimation of G has been recently proposed in [3], where the problem is reformulated expressing the log-derivative of G as a regression spline. Inference on the mixing distribution G has been reasonably conducted with respect to Wasserstein metrics, which have lately appeared to be natural metrics for controlling the estimation of geometric and topological features of the sampling measure and its support.

Definition 1 (Wasserstein Metrics). Given $p \geq 1$, for any two probability measures G and G' with finite pth moments, the Wasserstein distance of order p is defined as

$$\mathscr{W}_p(G, G') := \inf_{\pi} \left\{ \int \|\theta - \theta'\|^p \,\mathrm{d}\pi(\theta, \theta') \right\}^{1/p},$$

where the infimum is taken over all joint distributions with marginals G and G'.

The Wasserstein distance $\mathscr{W}_p$ is an optimal transport metric, meaning that, if $G = \sum_{j=1}^N p_j \delta_{\theta_j}$ and $G' = \sum_{j=1}^N p_j \delta_{\theta'_j}$, then, for example, $\mathscr{W}_1(G, G') = \sum_{j=1}^N p_j \|\theta_j - \theta'_j\|$ quantifies the minimum cost for moving mass from one probability measure to another one. Besides, $\mathscr{W}_p$ can be endowed with an interesting interpretation: if $\mathscr{W}_p(G_n, G) \asymp \delta_n = o(1)$ and G has a finite number of support points, then there are atoms of G_n converging to those of G at rate δ_n and atoms of G_n, which are redundant, vanishing at a faster rate than δ_n. Lower bounds on the estimation rates with respect to $\mathscr{W}_p$ for the deconvolution problem have been recently obtained in [1] for super- and ordinary smooth kernels, see Table 1.

	Gaussian kernel	Laplace kernel
Super-smooth kernel	$(\log n)^{-1/2}$	
Ordinary smooth kernel		$n^{-1/5}$

Table 1 Optimal 1-Wasserstein $\mathscr{W}_1$ estimation rates for a super-smooth kernel (Gaussian) and for an ordinary smooth kernel (Laplace).

In Bayesian inference, mixture models can be a device for density approximation or, when the kernel k is known, a tool for model-based inference. Prior laws on the mixing distribution G induce prior measures on the mixture density p_G. A common choice for the prior law on G is the Dirichlet process (DP). For DP convolution mixtures, under smoothness conditions on the sampling density p_0 (being not necessarily a mixture p_{G_0}), the posterior distribution on p_G contracts at p_0 at a minimax-optimal rate (up to a log-factor), which automatically adapts to the regularity level of p_0, see [7, 17]. For clustering, nonparametric prior laws as the Pitman-Yor process, which includes the DP as a special case, may result in inconsistent estimates of the number of clusters and mixing components, see [8, 12]. Posterior contraction rates

in Wasserstein metrics $\mathscr{W}_p$ for the latent distribution in deconvolution problems have been studied by [10, 4, 14, 6]. For DP convolution mixtures of super-smooth kernels with decreasing rate β of the Fourier transform, the speed of convergence for the posterior distribution is, up to a log-factor, equal to the optimal rate $(\log n)^{-1/\beta}$, whereas for DP convolution Laplace mixtures, the best obtained rate $n^{-1/8}$ under no assumption on the mixing distribution is sub-optimal. This rate has been recently improved to the lower bound $n^{-1/5}$ by [13] using an approximation scheme based on mixtures of Gaussian-Laplace densities, but under the assumption of Lebesgue absolute continuity of the mixing measure. For finite mixtures, two cases can be distinguished, depending on whether the number of mixture components is or not known. In the former case, using a model-based prior specification, the mixture density as well as the mixing distribution can be estimated at a nearly parametric rate $n^{-1/2}(\log n)^{\kappa}$, see [6, 15]. In the latter case, as shown in Section 2, a mixture of finite mixtures model, with a prior on the number of components, allows to retrieve the parametric rate for estimating the mixture density and the mixing distribution. By Tweedie's formula, estimation of the mixture density is useful for estimation and prediction of the single mixing parameters expressing specific features of the underlying subpopulations in compound decision problems, see Section 3.

2 Mixture Models with a Prior on the Number of Components

When the sampling density p_0 is a finite kernel mixture, with an unknown number N_0 of components, a natural way to deal with this source of uncertainty is to put a prior on it. Let the true density be

$$p_0(\cdot) \equiv p_{G_0}(\cdot) = \sum_{j=1}^{N_0} p_j^0 k(\cdot \mid \theta_j^0),$$

with mixing distribution $G_0 = \sum_{j=1}^{N_0} p_j^0 \delta_{\theta_j^0}$ for weights $\mathbf{p}_{N_0}^0 := (p_1^0, \ldots, p_{N_0}^0)$ and support points $\boldsymbol{\theta}_{N_0}^0 := (\theta_1^0, \ldots, \theta_{N_0}^0)$. Let P_0 denote the probability measure corresponding to p_0. A mixture of finite mixtures (MFM) model, see [9], with a prior on the number of components, can be thus described:

- the number of components N is a random variable (rv) with distribution ρ,
- given N, the random vectors $\mathbf{p}_N := (p_1, \ldots, p_N)$ and $\boldsymbol{\theta}_N := (\theta_1, \ldots, \theta_N)$ are independent, with $\mathbf{p}_N \sim \tilde{\pi}_N$ and $\boldsymbol{\theta}_N \sim \pi_N$,
- given N and $G = \sum_{j=1}^{N} p_j \delta_{\theta_j}$, the rv's $X_1, \ldots, X_n$ are conditionally iid according to p_G.

The overall prior is $\Pi = \sum_{N \geq 1} \rho(N)(\tilde{\pi}_N \otimes \pi_N)$. This general specification allows for other prior laws on $\boldsymbol{\theta}_N$ than the Dirichlet distribution as in [5]. Assume that

i) there exists $K > 0$ such that $\|k(\cdot \mid \theta_1) - k(\cdot \mid \theta_2)\|_1 \leq K|\theta_1 - \theta_2|$ for all θ_1, θ_2;
ii) for $\varepsilon > 0$ small enough and $c_0 > 0$, $\tilde{\pi}_N(\{\mathbf{p}_N : \sum_{j=1}^{N} |p_j - p_j^0| \leq \varepsilon\}) > \varepsilon^{c_0 N}$;
iii) the prior distribution π_N for the atoms has a continuous and positive Lebesgue density (also denoted by π_N) on an interval containing the support of G_0;
iv) the prior probability $\rho(N) > 0$ for every $N \in \mathbb{N}$.

Condition i) requires that $k(\cdot \mid \theta)$ is globally Lipschitz continuous with respect to θ. Condition ii) is satisfied for a Dirichlet prior distribution $\tilde{\pi}_N = \mathrm{Dir}(\alpha_1, \ldots, \alpha_N)$, with parameters $\alpha_1, \ldots, \alpha_N$ such that, for constants $a, A > 0$, $D \geq 1$ and $0 < \varepsilon \leq 1/(DN)$, it is $A\varepsilon^a \leq \alpha_j \leq D$, $j = 1, \ldots, N$. Condition iv) requires that every possible number of components N is assigned a priori a positive probability.

Proposition 1. *Under assumptions i)–iv) and B(1) of [6],*

- $\Pi(N = N_0 \mid X_{1:n}) \to 1$ P_0^∞*-almost surely,*
- *for* $M > 0$ *large enough,* $\Pi(G : \mathscr{W}_1(G, G_0) > M(n/\log n)^{-1/2} \mid X_{1:n}) \to 0$ *in* P_0^n*-probability.*

Proof. Apply Proposition 1 and Lemma 1 in [15] and use relationship (20) of Theorem 6.3 in [6] to conclude. □

Proposition 1 simplifies the proof of Theorem 3.1 in [5] and extends it to a larger class of prior distributions on the mixing weights.

3 Empirical Bayes Estimation of the Single Mixing Parameters

Mixture models are also fundamental in empirical Bayes compound decision problems having the following hierarchical structure, as illustrated in Figure 1:

$$\text{for each } i = 1, \ldots, n, \qquad \begin{aligned} &\theta_i \sim G \ \text{ independently,} \\ &X_i \mid \theta_i \sim k(\cdot \mid \theta_i), \end{aligned}$$

which allows for "borrowing strength" across different populations. Consider, for

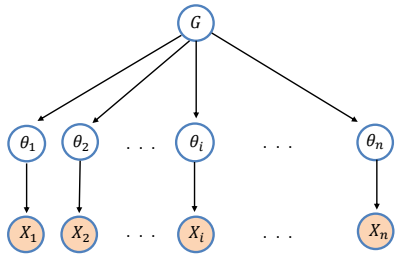


Fig. 1 Empirical Bayes compound decision problem: the parameters θ_i are sampled independently from G and the observation X_i is made within the ith group.

example, a model that allows for different disease rates in n cities, where n_i people are selected from the ith city and we observe how many people X_i have the disease being studied. We can think of the probability of disease as a random draw from some distribution G and it is of interest to estimate the parameters θ_i for each city or the overall disease rate $\int \theta \, \mathrm{d}G(\theta)$. We want to construct estimates of the θ_i's. The optimal prediction of the θ_i's, under the squared error loss, is given by the Bayes rule $\varphi_G(x) := \mathbb{E}(\theta \mid X = x)$. Clearly, $\hat{\theta}_i := \varphi_G(X_i)$ is an estimator of θ_i. For any kernel of the "Laplacian" type $k(x \mid \theta) = e^{\theta x} f(x) h(\theta)$, the following "extraordinary Bayesian estimation formula", see [2], attributed to M.C.K. Tweedie [11], holds

$$\varphi_G(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\log \frac{p_G(x)}{f(x)}\right) = \frac{f(x)}{p_G(x)} \times \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{p_G(x)}{f(x)}\right).$$

The "Laplacian" type kernel includes the exponential family $k(x \mid \theta) = e^{\theta x - \psi(\theta)} f(x)$ as a special case. In particular, for a Gaussian kernel $k(x \mid \theta) = \phi_{\theta,\sigma}(x)$, where $\phi_{\theta,\sigma}$ is the density of a normal $\mathcal{N}(\theta, \sigma^2)$, we have

$$\varphi_G(x) = x + \sigma^2 l'(x), \qquad \text{where } l'(x) := \frac{\mathrm{d}}{\mathrm{d}x} \log p_G(x) = \frac{p'_G(x)}{p_G(x)}.$$

The estimator $\hat{\theta}_i$ takes the form

$$\hat{\theta}_i = X_i + \sigma^2 l'(X_i) = \text{unbiased estimator + Bayes correction.} \tag{1}$$

As remarked by [2], the crucial advantage of Tweedie's formula is that it works directly with the marginal density $p_G(x)$. All observations $x_1, \ldots, x_n$ can then be used to obtain an estimate $\widehat{l'_n}(x)$ of $l'(x)$ yielding an empirical Bayes version of (1)

$$\hat{\theta}_i^{\mathrm{EB}} = X_i + \sigma^2 \widehat{l'_n}(X_i).$$

The "Laplacian" type kernel does not comprise the location Laplace model. Nonetheless using the representation of the Laplace density as a scale mixture of normals $e^{|x-\theta|}/2 = \int_0^{+\infty} (\phi_{\theta,\sqrt{\sigma}}(x) e^{-\sigma/2}/2)\,\mathrm{d}\sigma$ and Tweedie's formula for the Gaussian kernel, which yields $\int \theta \phi_{\theta,\sqrt{\sigma}}(x)\,\mathrm{d}G(\theta) = [x + \sigma l'(x)] p_G^{\mathrm{N}}(x)$ with $p_G^{\mathrm{N}}(x) := (G * \phi_{\theta,\sqrt{\sigma}})(x)$ and $l'(x) := (\mathrm{d}/\mathrm{d}x) \log p_G^{\mathrm{N}}(x)$, we have

$$\varphi_G(x) = \frac{1}{p_G^{\mathrm{L}}(x)} \int \frac{\theta}{2} e^{-|x-\theta|}\,\mathrm{d}G(\theta) = x + \frac{1}{p_G^{\mathrm{L}}(x)} \int_0^{+\infty} \frac{\sigma}{2} e^{-\sigma/2} \frac{\mathrm{d}}{\mathrm{d}x} p_G^{\mathrm{N}}(x)\,\mathrm{d}\sigma \tag{2}$$

and the corresponding estimator for θ_i is

$$\hat{\theta}_i = X_i + \frac{1}{p_G^{\mathrm{L}}(X_i)} \int_0^{+\infty} \frac{\sigma}{2} e^{-\sigma/2} \frac{\mathrm{d}}{\mathrm{d}x} p_G^{\mathrm{N}}(x)\Big|_{x=X_i} \mathrm{d}\sigma.$$

Since the sampling density p_G^{L} is a location mixture of Laplace densities and, for every x and σ, the density $p_G^{\mathrm{N}}(x)$ is a linear functional of the mixing distribution G, if G has Lebesgue density, then $p_G^{\mathrm{N}}(x)$ can be efficiently estimated by

$$\widehat{p_n^{\mathrm{N}}}(x) = \frac{1}{n} \sum_{i=1}^{n} \phi_{X_i,\sqrt{\sigma}}(x) \left(1 - \frac{1}{\sigma} + \frac{(x - X_i)^2}{\sigma^2}\right),$$

see [16] for the details. For any estimator $\widehat{p_n^{\mathrm{L}}}(x)$ of $p_G^{\mathrm{L}}(x)$, an empirical Bayes version of (2) is given by

$$\widehat{\varphi_G}(x) = x + \frac{1}{\widehat{p_n^{\mathrm{L}}}(x)} \int_0^{+\infty} \frac{\sigma}{2} e^{-\sigma/2} \frac{\mathrm{d}}{\mathrm{d}x} \widehat{p_n^{\mathrm{N}}}(x)\,\mathrm{d}\sigma,$$

which yields

$$\hat{\theta}_i^{\mathrm{EB}} = X_i + \frac{1}{\widehat{p_n^{\mathrm{L}}}(X_i)} \int_0^{+\infty} \frac{\sigma}{2} e^{-\sigma/2} \frac{\mathrm{d}}{\mathrm{d}x} \widehat{p_n^{\mathrm{N}}}(x) \Big|_{x=X_i} \mathrm{d}\sigma.$$

4 Concluding Remarks

We have shown that, a mixture of finite mixtures prior allows to estimate the mixing distribution of a finite kernel mixture with an unknown number of components N_0 at a nearly parametric rate, as if N_0 were known and this information could be used for the prior specification. Estimating the mixture density is also useful for estimation and prediction of the single mixing parameters or support points of the mixing distribution, which express specific features of the underlying subpopulations in compound decision problems. An empirical Bayes estimator of the single mixing parameters for Laplace convolution mixtures is proposed.

References

1. Dedecker, J., Fischer, A., Michel, B.: Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. Electron. J. Statist. **9**(1), 234–265 (2015)
2. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
3. Efron, B.: Empirical Bayes deconvolution estimates. Biometrika **103**(1), 1–20 (2016)
4. Gao, F., van der Vaart, A.: Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. Electron. J. Statist. **10**(1), 608–627 (2016)
5. Guha, A., Ho, N., Nguyen, X.: On posterior contraction of parameters and interpretability in Bayesian mixture modeling. arXiv:1901.05078 (2019)
6. Heinrich, P., Kahn, J.: Strong identifiability and optimal minimax rates for finite mixture estimation. Ann. Statist. **46**(6A), 2844–2870 (2018)
7. Kruijer, W., Rousseau, J., van der Vaart, A.: Adaptive Bayesian density estimation with location-scale mixtures. Electron. J. Statist. **4**, 1225–1257 (2010)
8. Miller, J.W., Harrison, M.T.: Inconsistency of Pitman–Yor process mixtures for the number of components. Journal of Machine Learning Research **15**, 3333–3370 (2014)
9. Miller, J.W., Harrison, M.T.: Mixture models with a prior on the number of components. Journal of the American Statistical Association **113**(521), 340–356 (2018)
10. Nguyen, X.: Convergence of latent mixing measures in finite and infinite mixture models. Ann. Statist. **41**(1), 370–400 (2013)
11. Robbins, H.: An empirical Bayes approach to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, pp. 157–163. University of California Press, Berkeley, Calif. (1956)
12. Rousseau, J., Mengersen, K.: Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(5), 689–710 (2011)
13. Rousseau, J., Scricciolo, C.: Bayesian Wasserstein deconvolution. Unpublished manuscript (2020)
14. Scricciolo, C.: Bayes and maximum likelihood for L^1-Wasserstein deconvolution of Laplace mixtures. Statistical Methods & Applications **27**(2), 333–362 (2018)
15. Scricciolo, C.: Bayesian Kantorovich deconvolution in finite mixture models. In: A. Petrucci, F. Racioppi, R. Verde (eds.) New Statistical Developments in Data Science, pp. 119–134. Springer International Publishing, Cham (2019)
16. Scricciolo, C.: Asymptotically efficient maximum likelihood estimation of linear functionals in Laplace measurement error models. Unpublished manuscript (2020)
17. Shen, W., Tokdar, S.T., Ghosal, S.: Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. Biometrika **100**(3), 623–640 (2013)

Improving ABC via Large Deviations Theory
Migliorare ABC *tramite la teoria delle Grandi Deviazioni*

Cecilia Viscardi, Michele Boreale and Fabio Corradi

Abstract *Sample degeneracy* in Approximate Bayesian Computation (ABC) is caused by the difficulty of simulating pseudo-data matching the observed data. In order to mitigate the resulting waste of computational resources and/or bias in the posterior distribution approximation, we propose to weight each parameter proposal by treating the generation of matching pseudo-data, given a "poor" parameter proposal, as a *rare event* in the sense of Sanov's Theorem. We experimentally evaluate our methodology through a proof-of-concept implementation.
Abstract *Il problema della* degenerazione del campione *in metodi ABC deriva dalla difficoltà di generare dati simili a quelli osservati. Al fine di evitare i conseguenti sforzi compuatazionali e/o distorsioni nell'approssimazione della distribuzione a posteriori, proponiamo di pesare ciascun parametro trattando la simulazione di dati uguali a quelli osservati come un evento raro nel senso del Teorema di Sanov. Si riportano i risultati di una valutazione empirica della metodologia proposta.*

Key words: ABC, Large Deviations, Sanov's Theorem, Sample Degeneracy.

1 Approximate Bayesian Computation and sample degeneracy

Let $x^n \in \mathscr{X}^n$ be a vector of observed data, which will be assumed to be drawn from a probability distribution in the family $\mathscr{F} \stackrel{\triangle}{=} \{P(\cdot|\theta) : \theta \in \Theta\}$. Suppose that our aim is to provide information about the uncertainty on θ by deriving the posterior distribution $\pi(\theta|x^n) \propto \pi(\theta)P(x^n|\theta)$ via Bayes' Theorem. When the likelihood function is analytically and numerically intractable, Approximate Bayesian Computation (ABC) allows for simulated inference by providing a conversion of samples from the prior into samples from the posterior distribution. This relies on compar-

Authors'affiliation: University of Florence - Department of Statistics and Computer Science (DiSIA) - e-mail: cecilia.viscardi@unifi.it, michele.boreale@unifi.it, fabio.corradi@unifi.it.

isons between the observed data and the pseudo-data generated from a *simulator*[1]. Algorithm 1 displays the rejection sampling scheme (R-ABC), whose origins can be traced back to [7, 4].

Algorithm 1 R-ABC

1: **for** $s = 1,...,S$ **do**
2: Draw $\theta^{(s)} \sim \Pi$
3: Generate $y \sim P(\cdot|\theta^{(s)})$
4: Accept $(\theta^{(s)}, s_y^{(s)})$ if $d(s_y^{(s)}, s_x) < \varepsilon$
5: **end for**

Algorithm 2 IS-ABC

for $s = 1,...,S$ **do**
2: Draw $\theta^{(s)} \sim q$
Generate $y \sim P(\cdot|\theta^{(s)})$
4: Set the IS weight for $\theta^{(s)}$ to

$$\omega_s = K_\varepsilon\big(d(s_y, s_x)\big) \cdot \frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}$$

end for

Samples resulting from Alg.1 are not from the exact posterior distribution since a twofold approximation is introduced by: summarizing data through a statistic $s(\cdot)$ — i.e. a function from the sample space $\mathscr{X}^n \subseteq \mathbb{R}^n$ to a lower-dimensional space $\mathscr{S}$— and assessing similarity via a distance function $d(\cdot,\cdot)$ and a tolerance threshold $\varepsilon > 0$.

Abbreviating $s(x^n)$ and $s(y^n)$ respectively as s_x and s_y, the output of the Alg.1 is a sample of pairs $(\theta^{(s)}, s_y^{(s)})$ from the following *approximated* joint posterior distribution

$$\tilde{\pi}(\theta, s_y|s_x) \propto \pi(\theta)\, P(s_y|\theta)\, \mathbb{1}\{d(s_y, s_x) \leq \varepsilon\} \tag{1}$$

where $\mathbb{1}\{d(s_y, s_x) \leq \varepsilon\}$, the indicator function assuming value 1 if $d(s_y, s_x) \leq \varepsilon$ and 0 otherwise, corresponds to the acceptance step. Marginalizing out s_y in (1), that is ignoring the simulated summary statistics, the output of the algorithm becomes a sample from the following approximated marginal posterior distribution

$$\tilde{\pi}(\theta|s_x) \propto \int_{\mathscr{S}} \pi(\theta)\, P(s_y|\theta)\, \mathbb{1}\{d(s_y, s_x) \leq \varepsilon\}\, ds_y = \pi(\theta) \cdot \Pr\big(d(s_Y, s_x) \leq \varepsilon|\theta\big). \tag{2}$$

The probability $\Pr\big(d(s_Y, s_x) \leq \varepsilon|\,\theta\big)$, where s_Y indicates $s(Y^n)$, is called the ABC *approximated likelihood.* As $\varepsilon \to 0$ the ABC likelihood converges to the true likelihood (see [3, Appendix A, p. 832]) and, whenever sufficient summary statistics for θ are chosen, $\tilde{\pi}(\cdot|s_x)$ converges to the true posterior $\pi(\cdot|x^n)$ (see [6, Ch. 1]). In practice the indicator function in (1) is often replaced by a kernel function $K_\varepsilon(\cdot)$ (e.g. triangular, Epanechnikov, Gaussian etc.) defined on the compact support $[0, \varepsilon]$ and providing a continuous decay from 1 to 0 (see e.g.[1]).

In the ABC literature a great variety of methods to sample from $\tilde{\pi}(\theta, s_y|s_x)$ have been proposed [2]. An example, is the importance sampling scheme IS-ABC reported as Alg. 2. Like the standard importance sampling, it suffers from *sample degeneracy*

[1] A simulator can be thought of as computer program taking as input a parameter value (or a vector thereof) $\theta^* \in \Theta$ and returning a sample from the distribution $P(\cdot|\theta^*)$.

[2] We refer the reader to [6, Ch 4] for an overview.

– i.e. only a small fraction of the proposed pairs has relatively high weights when the instrumental density $q(\cdot)$ is far from the target. Unlike the standard importance sampling, the IS-ABC implicitly involves a rejection when $\omega_s = 0$ — i.e. whenever a distance $d(s_y^{(s)}, s_x) > \varepsilon$ is observed [3]. Since they depend on the random variable s_Y through the distance $d(s_Y, s_x)$, when θ^* is such that $\Pr(s_Y = s_x|\theta^*)$ is close to zero, the importance weights will cause a huge number of rejections before a distance smaller than ε will be observed. This further aggravates the *sample degeneracy* issue. More sophisticated sampling schemes (e.g. MCMC-ABC, SMC-ABC, SIS-ABC, etc.) have been proposed to handle the issue of finding a good importance distribution, $q(\theta)$, but they completely ignore the effect of the kernel $K_\varepsilon(\cdot)$. In the next two sections we discuss how to define a kernel function $K_\varepsilon(\cdot)$ that improves the efficiency of ABC sampling schemes by avoiding rejections at all.

2 Large Deviations Theory in ABC

When a "poor" parameter proposal is given as an input to the generative model, simulating pseudo-data y^n such that $d(s_y, s_x) \leq \varepsilon$ can be treat as a *rare event*. This often leads to a shortage of accepted values mostly in regions of Θ with a low but positive (true) posterior density, in turn resulting in a bad approximation in the tails of the posterior distribution. A possible approach to mitigate those issues is to provide a finer estimate for the ABC likelihood allowing to avoid rejections at all. To this aim, we resort to *Large Deviations Theory* (LDT).

Let x^n be a sequence of n symbols drawn from $\mathscr{X}$ according to $P_\theta \stackrel{\triangle}{=} P(\cdot|\theta)$, say $x^n = (x_1, ..., x_n)$. The empirical distribution of x^n, written P_{x^n}, is the probability distribution on $\mathscr{X}$ defined by

$$P_{x^n}(r) \stackrel{\triangle}{=} \frac{|\{i : x_i = r\}|}{n} \quad \forall r \in \mathscr{X}. \tag{3}$$

Given a large n, observing a sequence whose empirical distribution is far from P_θ is a rare event, and its probability obeys to a fundamental result in LDT, Sanov's theorem (see [2, Th.11.4.1]).

Theorem 1 (Sanov's Theorem). *Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables on $\mathscr{X}$, with each $X_i \sim P_\theta$. Let $\Delta^{|\mathscr{X}|-1}$ be the simplex of probability distributions over $\mathscr{X}$ and let $E \subseteq \Delta^{|\mathscr{X}|-1}$. Then*

$$\Pr\left(P_{X^n} \in E \,|\, \theta\right) \leq (n+1)^{|\mathscr{X}|} 2^{-nD(P^*||P_\theta)}. \tag{4}$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence and $P^* = \operatorname{argmin}_{P \in E} D(P||P_\theta)$ is the information projection of P_θ onto E. Furthermore, if E is the closure of its interior

[3] Note that Alg. 1 is a special case of the Alg.2 where the marginal importance distribution, $q(\theta)$, is the prior distribution and the resulting importance weights are $\omega_s \in \{0, 1\}$.

$$\lim_{n\to\infty}\frac{1}{n}\log\Pr(P_{X^n}\in E\,|\,\theta)=-D(E||P_\theta)=-D(P^*||P_\theta).$$

In order to show how to make use of the Sanov's result in the approximated likelihood computation, from now on we will assume: a) $\varepsilon>0$ as a threshold; b) empirical distributions as summary statistics; c) the Kullback-Leibler divergence as distance function; d) the empirical distribution P_{x^n} to be full support. The IS-ABC generates pairs $(P^{(s)}_{y^m},\theta^{(s)})$ with $s\in\{1,...,S\}$. Each $P^{(s)}_{y^m}$ is an empirical distribution resulting from a sequence of i.i.d. random variables, $Y^m=\{Y_j\}_{j=1}^m$, distributed according to $P\big(\cdot\,|\,\theta^{(s)}\big)$. We want to stress that the length of the simulated sequence, m, need not be equal to n, the length of the observed data sequence.

Under our assumptions, each $\theta^{(s)}$ is accepted or rejected depending on the divergence $D(P_{y^m}||P_{x^n})$. Thus, we can define the following acceptance region:

Definition 1 (Acceptance region). Let $\Delta^{|\mathscr{X}|-1}$ be the simplex of probability distributions over $\mathscr{X}$ and let P_{x^n} be the empirical distribution of the observed sequence x^n. The *acceptance region* $\mathscr{B}_\varepsilon(P_{x^n})$, shortly $\mathscr{B}_\varepsilon$, is defined for any $\varepsilon\geq 0$, as

$$\mathscr{B}_\varepsilon\stackrel{\triangle}{=}\big\{P\in\Delta^{|\mathscr{X}|-1}:D(P||P_{x^n})\leq\varepsilon\big\}.$$

Sanov's result, for m large enough, allows to approximate the probability of simulating pseudo-data whose summary statistic, P_{y^m}, is in the acceptance region even when a "poor" parameter in proposed:

$$\Pr\big(P_{Y^m}\in\mathscr{B}_\varepsilon\,|\,\theta^{(s)}\big)\approx 2^{-mD(\mathscr{B}_\varepsilon||P_{\theta^{(s)}})}. \tag{5}$$

Unfortunately, the computation of the probability in (5) is still not feasible when the model $\mathscr{F}=\{P_\theta:\theta\in\Theta\}$ is unknown, as we do not know how to compute $D(\mathscr{B}_\varepsilon||P_{\theta^{(s)}})$. However, one we can prove that

$$\lim_{m\to\infty}D(\mathscr{B}_\varepsilon||P_{Y^m})=D(\mathscr{B}_\varepsilon||P_\theta)\quad a.s. \tag{6}$$

According to (5) and (6) we propose the following kernel function:

$$K_{\varepsilon,m}\big(P_{y^m}\big)=\begin{cases}1 & \text{if } D(P_{y^m}||P_{x^n})\leq\varepsilon\\ 2^{-mD(\mathscr{B}_\varepsilon||P_{y^m})} & \text{if } D(P_{y^m}||P_{x^n})>\varepsilon\end{cases} \tag{7}$$

By resorting to this kernel the joint and the marginal posterior distributions are characterized by the following equations:

$$\tilde{\pi}(\theta,P_{y^m}|P_{x^n})\propto\pi(\theta)\,K_{\varepsilon,m}\big(P_{y^m}\big)\;P_\theta(P_{y^m}) \tag{8}$$

$$\tilde{\pi}(\theta|P_{x^n})\propto\pi(\theta)\sum_{P_{y^m}\in\mathscr{P}^m}K_{\varepsilon,m}\big(P_{y^m}\big)\;P_\theta(P_{y^m}) \tag{9}$$

where $\mathscr{P}^m\subset\Delta^{|\mathscr{X}|-1}$ denotes the set of the empirical distributions with denominator m. Accordingly, the ABC likelihood is defined as follows

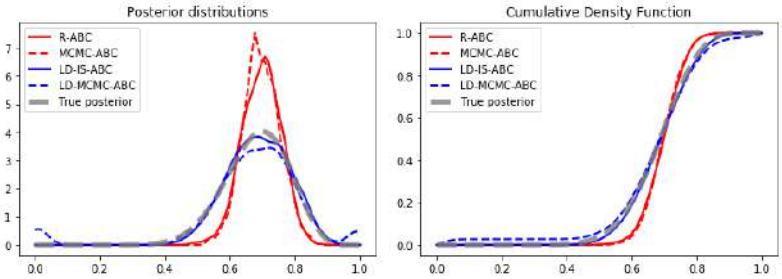


Fig. 1 Parameter posterior distributions (LHS) and posterior cumulative density functions (RHS) derived through IS and MCMC schemes with the uniform kernel and with the proposed kernel.

$$\tilde{\mathscr{L}}_{\varepsilon,m}(\theta;P_{x^n}) \stackrel{\triangle}{=} \sum_{P_{y^m}\in\mathscr{P}^m} K_{\varepsilon,m}\left(P_{y^m}\right) P_\theta\left(P_{y^m}\right). \tag{10}$$

Note that, now Alg.2 gives a positive weight to each $\theta^{(s)}$. More precisely, the weight equals 0 only when $D(\mathscr{B}_\varepsilon||P_{y^m}) = \infty$. In the next section we empirically demonstrate the improvements achieved by resorting to the proposed kernel.

3 A toy example

Let x^{20} be a sample from i.i.d. Bernoulli random variables with parameter θ. Suppose that x^{20} has empirical distribution $P_{x^n} = [0.3, 0.7]$. Assuming an uniform prior distribution, the posterior distribution, $\pi(\theta|x^{20})$, is a Beta distribution with parameters $\alpha = 15$ and $\beta = 7$.

We ran $S = 10000$ iterations of IS-ABC both with the uniform kernel and the proposed kernel. Note that in the first case the algorithm corresponds to a R-ABC. We also implemented the MCMC-ABC sampling scheme (see [6, Ch. 4]). For the sake of simplicity we adopt the abbreviation LD, standing for Large Deviations, to indicate that the employed kernel function is (7).

Fig.1 shows the posterior distributions and cumulative density functions (CDF) approximated by each algorithm. As it is apparent, the LD algorithms (blue lines) approximate better the true posterior (dashed grey line). Looking at the CDF's, we can see that using the uniform kernel (red lines) results in a worse approximation in the tails.

We evaluate the posterior mean point estimates and the posterior density estimates through the Squared Error and the Integrated Squared Error respectively. We also consider the Effective Sample Size as a measure of the degree of sample degeneracy. From Tab.1, we can see that, despite the quality of the point estimations is almost the same, the proposed kernel function leads to clear improvements in terms of density estimations and *ESS*, both for the IS-ABC and for the MCMC-ABC.

Table 1 Squared Errors, Integrated Squared Errors and Effective Sample Sizes with $\varepsilon = 0.01$, $m = 100$.

Algorithm	*SE*	*ISE*	*ESS*
R-ABC	0.0002	0.6428	1279
LD-IS-ABC	0.0003	0.0096	3060
MCMC-ABC	0.0002	0.8597	655
LD-MCMC-ABC	0.0002	0.041	1929

4 Conclusions

We have put forward an approach to address sample degeneracy in ABC. Our proposal consists in the definition of a convenient kernel function which, via Large Deviations Theory, takes into account the probability of rare events. Being defined on a non-compact support, the proposed kernel allows to avoid rejections, thus mitigating the effects of the sample degeneracy. We have also evaluated our methodology on a simple example, showing that it provides a better approximation of the posterior density and increases the Effective Sample Size.

References

1. Beaumont, M. A., Zhang, W., Balding, D. J.: Approximate Bayesian computation in population genetics. Genetics,**96**(4), 2025–2035 (2002).
2. Cover, T.M and Thomas, J. A. Element of information theory, John Wiley & Sons (2006).
3. Prangle, D., Everitt, R. G., Kypraios, T.: A rare event approach to high-dimensional approximate Bayesian computation. Statistics and Computing, **28**(4), 819–834 (2018).
4. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., Feldman, M. W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular biology and evolution, **16**(12), 1791–1798 (1999).
5. Rubin, D.B. : Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics, 1151–1172 (1984).
6. Sisson, S. A., Fan, Y., & Beaumont, M. Handbook of approximate Bayesian computation. Chapman and Hall& CRC (2018).
7. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. : Inferring coalescence times from DNA sequence data. Genetics, **145**(2), 505–518 (1997).

Learning Bayesian Networks for Nonparanormal Data

Apprendimento di reti bayesiane per dati non parametrici

Abstract In the literature, structural learning procedures for selecting the directed acyclic graph of a Bayesian network are increasingly explored and specified according to the analyzed data typology. With respect to data drawn from a Gaussian Copula model, the Rank PC algorithm, based on Spearman rank correlation, has been introduced. Moreover, we recently proposed a modified version of the well known Grow-Shrink algorithm, the Copula Grow-Shrink one, based on the Spearman rank correlation and the Copula assumption. Here, we show a simulation study to verify the robustness of our Copula Grow-Shrink algorithm and we discuss the performance results in comparison with the baseline and the Rank PC algorithm.

Abstract *In letteratura, le procedure di apprendimento strutturale per la stima di un grafico diretto aciclico di una rete bayesiana sono sempre più esplorate e dettagliate in base alla tipologia di dati analizzati. Per quanto riguarda i dati derivanti da un modello di copula gaussiana l'algoritmo Rank PC, basato sulla correlazione di Spearman, è stato proposto in letteratura. Inoltre, abbiamo recentemente proposto una versione modificata del noto algoritmo Grow-Shrink, l'algoritmo Copula Grow-Shrink, basato sulla correlazione tra ranghi di Spearman e sull'assunzione di Copula. Qui, mostriamo uno studio di simulazione per verificare la solidità del nostro algoritmo Copula Grow-Shrink e discutiamo i risultati delle prestazioni rispetto agli algoritmi Grow-Shrink e Rank PC.*

Key words: joint normal copula, Copula Grow-Shrink algorithm, simulation study, diagnostic measures

1 Introduction

Statistical multivariate data modeling is increasingly carryed out through Bayesian networks, (BN, [2]) that depict the multivariate probability distribution of a set of

variables by a graphical representation of independencies encoded in a directed acyclic graph (DAG). A DAG is a finite set of nodes, standing for random variables, and directed edges, arranged never producing cycles, that point out direct relevance of one variable to another. In a DAG, a parent node has an outcoming arrow pointing to another node namely child; every node is associated with a conditional distribution given its parents and the joint distribution can be factorized according to the DAG.

In this context, a common issue concerns the DAG structural elicitation. When the dependencies are unknown or partially known, DAG structure has to be estimated directly from data. Most often, researchers wish to maximize the learning power respecting the typology of managed data. For nonparanormal data some structural learning algorithms have been discussed in the literature; we recently proposed the Copula Grow-Shrink [1], a modified version of the Grow-Shrink algorithm, based on the recovery of the Markov blanket of the nodes and on the Spearman correlation. The paper, aiming at evaluating the robustness of the our proposal, is organized as follows: nonparanormal graphical models are briefly recalled in Section 2; the Grow-Shrink and the Copula Grow-Shrink algorithms are discussed in Section 3; the simulation study and preliminary results are addressed in Section 4.

2 Nonparanormal Graphical models and their estimations

Nonparanormal data modeling by graphical models has been studied in the literature. Generally speaking, a nonparanormal graphical model is a semiparametric extentions of a Gaussian graphical model useful when the analysed continuous variables follow a Gaussian graphical model only if trasformed by unknown smooth monotone functions preserving the dependencies structure of the underlying multivariate normal distribution. According to [7]:

> Definition 1. Let $f = (f_v)_{v \in V}$ a collection of strictly increasing functions $f_v : R \to R$ and $\Sigma \in R^{V \times V}$ be a positive definite correlation matrix. The nonparanormal distribution NPN(f, Σ) is the distribution of the random vector $(f_v(Z_v))_{v \in V}$ for $(Z_v)_{v \in V} \sim N(0, \Sigma)$.

> Definition 2. The nonparanormal graphical model NPN(G) associated with a DAG G is the set of all distributions NPN(f, Σ) that are Markov with respect to G.

The function f_v realizes a deterministic transformation on Z_v preserving the same dependence structure of the underlying latent multivariate normal distribution also in the nonparanormal model.
If $X \sim NPN(f, \Sigma)$ and $Z \sim N(0, \Sigma)$, for any triple of pairwise disjoint set $A, B, S \subset V$, then $X_A \perp\!\!\!\perp X_B | X_S \Leftrightarrow Z_A \perp\!\!\!\perp Z_B | Z_S$.
For two nodes (u, v) and a separating set S we have $X_u \perp\!\!\!\perp X_v | X_S \Leftrightarrow \rho_{uv|S} = 0$.

A trigonometric transformation on Spearman rank correlation (r) produces latent Normal correlation coefficients accurate estimators. Reference [5] show that if (X,Y) are bivariate normal with $Corr(X,Y) = \rho$, it yelds:

$$P(|2sin(\frac{\pi}{6}\widehat{r}) - \rho|) > \varepsilon) \leq 2exp(-\frac{2}{9\pi^2}n\varepsilon^2) \tag{1}$$

Since $\widehat{r}$ depends on the observations *via* their ranks that are preserved under strictly increasing functions, (1) still holds for nonparanormal graphical models with Pearson correlation $\rho = \Sigma_{xy}$ in the underlying latent bivariate normal distribution. On the basis of the previous result ρ is estimated as:

$$\hat{\rho} = 2\sin(\frac{\pi}{6} \cdot \hat{r}) \tag{2}$$

The same transformation still holds for the partial correlation coefficients.

3 Bayesian Netoworks Structural Learning

Bayesian netoworks structural learning methods are mainly *scoring and searching* techniques or *constraint-based* algorithms; they estimate and depict the unknown independencies relations among variables by a DAG. The most spread algorithm is the PC algorithm [9] that proceeds along three steps: (i) the skeleton identification by testing marginal and conditional independencies by Pearson correlation test for Gaussian data, (ii) the v-structures identification standing for conditional dependence between two nodes given a third and (iii) the orientation of the remaining links without producing additional v-structures and/or directed cycles. If variables are not Gaussian, a PC algorithm rank version named Rank PC (RPC) algorithm is available [7]. RPC algorithm tests conditional independence between two variables given a separating set by computing the rank-based partial correlation estimates (Eq. 2). The RPC algorithm consistency is proved by [7], under some non-strict assumptions. It is shown that RPC works at the same strength of PC algorithm for normal data but considerably better for non-normal data under the *strong* assumption of joint distribution following a normal copula model. The RPC algorithm could be implemented using the `pcalg R` package [4].
A competitive algorithm to these common choices is the Grow-Shrink algorithm (GS, [6]) based on the intuitive concept of the Markov blanket (MB) of a variable, *i.e.* the set of all parents, children and parents of children of the variable of interest, say X. Moreover, the MB(X) d-separates variable X from any other variable outside its Markow blanket. The GS algorithm focuses on the recovery of the MB(X) based on pairwise independence tests by two phases: the growing phase, where, from MB(X) empty set denoted by S, the procedure adds variables to S as long as they are associated with X given the current contents of S; the shrinking phase identifies and removes variables not really belonging to MB(X) eventually added to S. Our Copula GS algorithm (CGS) [1] has the same logical structure as GS but the

marginal and partial correlations coefficients used in the statistical test for independence are computed trough (2). The GS is implemented in `bnlearn R` package [8] so that our proposal is developed in `R` as well.

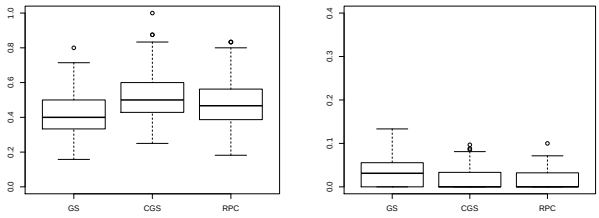
4 Main Results and Conclusions

With the aim to explore the robustenss, a simulation study has been carried out. According to the procedure in [3] and to the simulation plan in [7], we simulated 200 random DAGs with sparsity parameter $s = 0.3$ and we sampled from a Gaussian Copula distribution faithful to them. Considered sample sizes are $n = 50$ and $n = 1000$. On every training set, we performed the structural learning GS, CGS and RPC algorithms with a significance level of 0.05. Algorithm performances have been compared in terms of sensitivity, specificity and precision.
In details, the diagnostic measure *true positive rate* (TPR) is the proportion of edges correctly estimated on the true edges; the closer the value is to 1, the better is the sensitivity. The *false positive rate* (FPR) is the proportion of edges incorrectly found over the number of true gaps; the closer the value is to 0, the better is the specificity. The *true discovery rate* (TDR) is the proportion of edges correctly found on the total number of estimated edges; the closer the value is to 1, the better is the precision. The performance measure distributions from simulations are displayed in the following boxplot (see Figures 1 e 2).

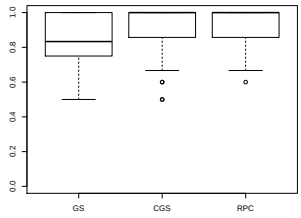
For small sample size $n = 50$ (see Figure 1) the sensitivity of CGS algorithm outperforms the GS and the RPC ones denoting a better capacity to catch the real structure. The specificity of CGS is still better with respect to GS and only slightly more variable in comparison to RPC. Also in terms of TDR, the CGS outperforms the GS and works the same as the RPC. For large sample size $n = 1000$ (see Figure 2) the CGS outperforms the GS and works slightly better than the RPC in terms of specificity and precision but gains a stronger sensitivity.
According to these simulation results the algorithm we propose represents a better choice to estimate a DAG in case of nonparanormal data. Since $1 - TPR$ is equal to the False Negative Rate (FNR), it means that the CGS algorithm prevents from bias in the model due to the absence of a "true" arc. We argue that, as the RPC one, also the CGS algorithm reduces the risk of an overparametrization of model since the FPR is smaller than that of GS for both sample sizes.

References

1. author: (year)
2. Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer, New York (1999)
3. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the pc-algorithm. J. Mach. Learn. Res. **8**, 613–636. (2007)

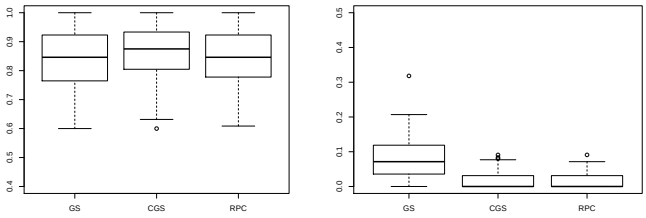


(a) TPR (b) FPR

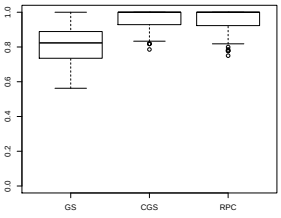


(c) TDR

Fig. 1 Boxplot of diagnostic measures for n=50



(a) TPR (b) FPR



(c) TDR

Fig. 2 Boxplot of diagnostic measures for n=1000

4. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. J. Stat. Softw. **47**(11), 1–26 (2012). URL http://www.jstatsoft.org/v47/i11/.
5. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: High-dimensional semiparametric gaussian copula graphical models. The Annals of Statistics **40**(4), 2293–2326 (2012). DOI 10.1214/12-AOS1037. URL http://dx.doi.org/10.1214/12-AOS1037
6. Margaritis, D.: Learning bayesian network model structure from data. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA (2003). Technical Report CMU-CS-03-153
7. Naftali, H., Drton, M.: Pc algorithm for nonparanormal graphical models. J. Mach. Learn. Res. **14**, 3365–3383. (2013)
8. Scutari, M.: Learning bayesian networks with the bnlearn r package. J. Stat. Softw. **35**(3), 1–22 (2010). URL http://www.jstatsoft.org/v35/i03/
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT press, Cambridge, Massachusetts (2000)

Measuring well-being combining different data sources: a Bayesian networks approach

Misurare il benessere combinando diverse fonti di dati: un approccio basato sulle reti bayesiane

F. Cugnata, S. Salini, E. Siletti

Abstract In this paper we focus on a Bayesian networks approach to combine traditional survey and social networks data to evaluate subjective well-being. Bayesian networks permit to use data with not the same geographical levels, provincial and regional, and with different time frequencies, quarterly and annual. Moreover, we remark that in this proposal we combine both categorical and continuous data. The application, referred to Italy from 2012 to 2017, has been performed using ISTAT's survey data, some covariates, from official statistics or related to the considered time period, and a composite index of well-being obtained by Twitter data.

Abstract *In questo lavoro ci focalizziamo su un approccio basato sulle reti bayesiane che, combinando dati provenienti da sondaggi tradizionali e da social networks, permette di valutare il benessere soggettivo. Le reti bayesiane consentono di utilizzare dati con livelli geografici non uguali, provinciali e regionali, e con frequenze temporali diverse, trimestrali e annuali. Inoltre, sottolineiamo che in questa proposta combiniamo dati sia categorici che continui. L'applicazione, riferita all'Italia dal 2012 al 2017, è stata eseguita utilizzando i dati provenienti da sondaggi ISTAT, alcune covariate, provenienti da statistiche ufficiali o relative al periodo considerato, e un indice composito di benessere ottenuto dai dati Twitter.*

Key words: Bayesian networks, big data, well-being, life satisfaction, sentiment analysis

---

Federica Cugnata
University Centre of Statistics for Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, e-mail: cugnata.federcia@hsr.it

Silvia Salini
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, e-mail: silvia.salini@unimi.it,

Elena Siletti
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, e-mail: elena.siletti@unimi.it

1 Introduction and background

Following the Stiglitz Commission's suggestions, which in summary advised to build a complementary system focused on social well-being, suitable for measuring sustainability, and which takes into account also subjective assessment [17], new measures have been proposed. In these new indices, the subjective dimensions are traditional investigated by ad hoc surveys, nevertheless these data are not free of issues, we briefly remember those related to the survey structure or plan, which despite all the efforts made [11], still show some methodological drawbacks [4, 5], and those according their poor geographical disaggregation or low time frequency.

Moreover, an objective evaluation of the indices currently proposed cannot fail to note a limited and undersized presence of the subjective and perceived dimension. Trying to fill this information gap, since 2012, with the aim of finding complementary info, a new subjective and perceived Italian well-being index deepening social networks data has been proposed (Subjective Well-being Index - SWBI, [8]). This composite index, providing a measure across the same domains considered by the New Economic Foundation for its Happy Planet Index [14], is the result of a human supervised sentiment analysis (Integrated Sentiment Analysis - iSA, [2]) on Twitter data. Recently, [9] suggest to use SWBI to provide a measurement of subjective well-being available at Italian sub-national levels and at different moments of time.

However, despite social networks data has been defined the largest available focus group in the world [12, 7], they permit to cover several topics, are continuously updated, and are free or cheap, it is undeniable that also these data are not free of disadvantages. Even if the social media users are always increasing [1], not all are users; hence, one of the main issue concerns sampling bias. To not renounce such an intense source of information, scholars are still working to the solution for these issues. Especially, [10] suggested a new Italian well-being measure mashing-up official statistics with Twitter data using a weighting procedure combined with a Small Area Estimation (SAE) model to consider precisely sampling bias.

Following in this practice, with this contribute we combine, in a Bayesian networks (BNs) approach, traditional survey and social networks data to evaluate subjective well-being. Adopting this approach we can use both categorical and continuous data with different geographical area levels and time frequencies. Section 2 reports a brief presentation of BN, while in section 3 there is a first example referring to Italian data from 2012 to 2017.

2 Method

BN are both mathematically rigorous and intuitively understandable data analytic tools. They implement a graphical model structure, known as a directed acyclic

[1] "Digital in 2019", `http://wearesocial.com`: from Jan 2018 to Jan 2019 the world growth is +9.1% for the Internet accesses and +9% for the social media active accounts.

graph (DAG), that is popular in statistics, machine learning and artificial intelligence. They enable an effective representation and computation of a joint probability distribution (JPD) over a set of random variables.

The DAG's structure is defined by a set of nodes, representing random variables and plotted by labeled circles, and a set of arcs, representing direct dependencies among the variables and plotted by arrows. Thus, an arrow from X_i to X_j indicates that a value taken by variable X_j depends on the value taken by variable X_i. Node X_i is then referred to as a "parent" of X_j and, similarly, X_j is referred to as the "child" of X_i. An extension of these genealogical terms is often used to define the sets of "descendants", i.e., the set of nodes from which the node can be reached on a direct path. The DAG guarantees that there is no node that can be its own ancestor (parent) or its own descendent. Such a condition is of vital importance to the factorization of the joint probability of a collection of nodes. Although the arrows represent direct causal connection between the variables, under causal Markov condition, the reasoning process can operate on a BN by propagating information in any direction. A BN reflects a simple conditional independence statement, namely that each variable, given the state of its parents, is independent of its non-descendants in the graph. This property is used to reduce, sometimes significantly, the number of parameters that are required to characterize the JPD. This reduction provides an efficient way to compute the posterior probabilities given the evidence in the data [13, 15]. In addition to the DAG structure, which is often considered to be the qualitative part of the model, one estimates the quantitative parameters applying the Markov property, where the conditional probability distribution at each node depends only on its parents.

More formally, BNs are defined by a network structure, a DAG $G = (\mathbf{V};A)$, in which each node $v_i \in \mathbf{V}$ corresponds to a random variable X_i; a global probability distribution, $\mathbf{X}$, which can be factorised into smaller local probability distributions according to the arcs $a_{ij} \in A$ in the graph. The main role of the network structure is to express the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorisation of the global distribution:

$$\prod_{i=1}^{p} P(X_i|\Pi_{X_i}) \qquad \text{where } \Pi_{X_i} = \text{parents of } X_i.$$

The choice of the probability distribution $P(X)$ should such that the BN can be learned efficiently from data. This distribution is flexible, so the assumptions should not be too strict, and it is easy to query to perform inference [16]. The three most common choices in the literature, are:

- Discrete BNs (DBNs): X and $X_i|\Pi_{X_i}$ are multinomial;
- Gaussian BNs (GBNs): X is multivariate normal and $X_i|\Pi_{X_i}$ are normal;
- Conditional Linear Gaussian BNs (CLGBNs): X is a mixture of multivariate normal and $X_i|\Pi_{X_i}$ are either multinomial, normal or mixtures of normal.

It has been proved that exact inference is possible in these three cases, hence their popularity. In this work we consider CLGBNs: some variables are categorical

and some other are numerical. BNs have already been used to the analysis of multidimensional well-being [1], taking into account the correlation among dimensions. The same authors applied also multivariate statistical techniques to ISTAT's BES index [3]. Moreover, [18] employed BNs in the context of subjective well being, focusing on the ability to predict it by material living conditions and deprivation, using the European Quality of Life Study data (2011) in four Central European countries. This proposal combines, in a BNs approach, well-being data from social networks [8] and surveys.

3 Application: Data and Preliminary Results

As survey data we consider some variables from the "Aspect of daily life" report. This ISTAT sample survey collects fundamental details on Italian individual and household daily life, focusing on several thematic areas on different social aspects useful to study well-being, this data are adopted also for the subjective dimension in the ISTAT's well-being index BES (Benessere Equo e Sostenibile). This is an annual survey that consider people aged 14 and over, and data, with a regional aggregation, are available free of charge (http://dati.istat.it/).

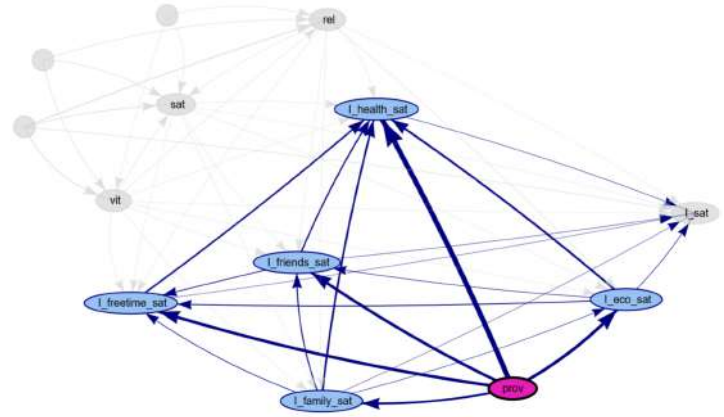
The considered variable are: (`I_sat`) the average rating of *satisfaction with life as a whole* (with a 1 to 10 scale), the percentage of people very or fairly satisfied with the *economic situation* (`I_eco_sat`), with *health* (`I_health_sat`), with *family relationships* (`I_family_sat`), with their *friendships* (`I_friends_sat`), and with their *free time* (`I_freetime_sat`).

Moreover we consider, as covariate, other social and demographic data and some variables related to the considered time period:

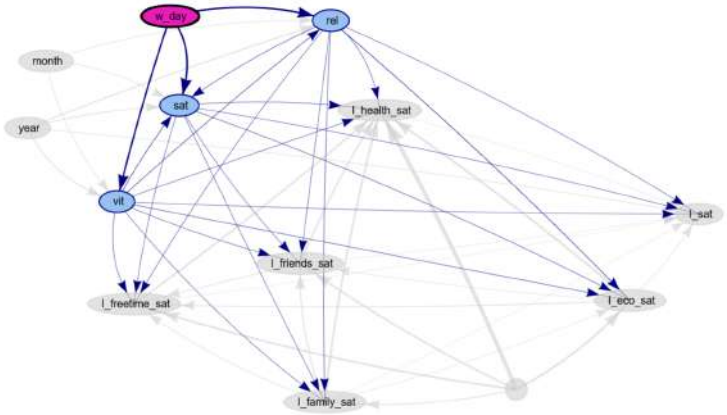
- the quarterly regional *unemployment rate* (`t_unemployment`), *number of inactive people* (`inactives`), *number of employed people* (`employed`), and *labor force* (`laborforce`);
- the quarterly provincial *broadband coverage* proportion (`broadband`);
- the annual regional *percentage of trusting people* (`trust`);
- the *week day* (`w_day`), *month* (`month`), and the *year* (`year`).

Focusing on SWBI, it is defined by eight domains, all measured in a scale $0-100$, according to the *personal well-being*, the *social well-being*, and the *well-being at work*, with daily frequency for all the Italian provinces, we consider:

- *Life satisfaction* (`sat`): having positive assessment of the overall life satisfaction;
- *Vitality* (`vit`): having energy, feeling well-rested and healthy, and being active;
- *Resilience and self-esteem* (`res`): a measure of individual psychological resources, optimism, and the ability to deal with stress;
- *Trust and belonging* (`tru`): trusting other people, feeling treated fairly and respectfully while experiencing sentiments of belonging;
- *Relationships* (`rel`): the degree and quality of interactions in close relationships with family, friends and others who provide support;



(a)



(b)

Fig. 1 Bayesian network obtained with the Hill-Climbing algorithm with BIC score functions. Line widths are proportional to the strength of each arc.

- *Quality of job* (`wor`): feeling satisfied with job, with work-life balance, and evaluating the emotional experiences of work, and work conditions.

The data are integrated at daily level using for annual and quarterly data repeated values.

A preliminary analysis was performed taking into account six survey variables (`I_sat`, `I_eco_sat`, `I_health_sat`, `I_family_sat`, `I_friends_sat`, and `I_freetime_sat`), three SWBI variables (`sat`, `vit` and `rel`), and four covariates. The analysis was performed using the R statistical language, using the `bnlearn` [16] and the `BNViewer` library [6]. In this plot it is possible to highlight

some nodes (in purple), their children (blue), the other descendants (gray); nodes not descendent are transparent (empty gray). Fig.1 shows the BN obtained with the Hill-Climbing algorithm with the BIC score functions. Line widths are proportional to the strength of each arc. Fig.1(a) highlights the descendent nodes of the day of week. As expected, the day of week has a direct impact on the indexes obtained by Twitter data, `sat`, `vit` and `rel`. These measures have, in turn, an effect on the ISTAT's satisfaction indices. Fig.1(b) highlights the descendent nodes of the province. This variable has a direct impact on the ISTAT survey data.

In the next future, we will focus on developing the analysis considering also the other variables. In addition, further research will be devoted to investigate how the BNs approach might be helpful for the assessment of hypothetical scenarios.

References

1. Ceriani, L., Gigliarano, C.: Multidimensional well-being: A Bayesian Networks approach. ECINE Society for the Study of Economic Inequality. **WP: 399**, (2016)
2. Ceron, A., Curini, L., Iacus, S.M.: iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. Inform.Sciences. **367-368**, 105–124 (2016)
3. Chelli, F.M., Ciommi, M., Emili, A., Gigliarano, C., Taralli, S.: Assessing the Equitable and Sustainable Well-Being of the Italian Provinces. Int.J.Uncertain Fuzz. **Sup.1**, 39–62 (2016)
4. Deaton, A.: The Financial Crisis and the Well-Being of America. University of Chicago Pres. Chicago, IL (2012)
5. Feddersen, J., Metcalfe, R., Wooden, M.: Subjective wellbeing: why weather matters. J.R.Stat.Soc.A Stat. **179(1)**, 203–228 (2016)
6. Fernandes, R.: `bnviewer`: Interactive Visualization of Bayesian Networks. R package version 0.1.4. `https://CRAN.R-project.org/package=bnviewer`
7. Hofacker, C.F., Malthouse, E.C., Sultan, F.: Big Data and consumer behavior: imminent opportunities. Ital.J.Consum.Mark. **33(2)**, 89–97 (2019)
8. Authors: Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. ArXiv **1512.01569** (2015)
9. Authors: Social Networks Data and Subjective Well-Being. An Innovative Measurement for Italian Provinces. Ital.J.Reg.Stu. **18 S/2019**, 667–678 (2019)
10. Authors: Controlling for Selection Bias in Social Media Indicators through Official Statistics: A Proposal. J.Off.Stat. (2020 Forthcoming)
11. Kahneman, D., Krueger, A.B.: Developments in the measurement of subjective well-being. J.Econ.Perspect. **20(1)**, 3–24 (2006)
12. Kwong, B.M., McPherson, S.M., Shibata, J.F.A., Zee, O.T.: Facebook: Data mining the world's largest focus group. Graziadia Bus.Rev. **15**, (2012)
13. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. J.R.Stat.Soc.B **50(2)**, 157–224 (1988)
14. NEF: The Happy Planet Index: 2012 Report. A global index of sustainable well-being. New Economics Foundation. (2012)
15. Pearl, J.: Causality: Models, Reasoning, and Inference, Cambridge University Press, Cambridge (2000)
16. Scutari, M.: Learning Bayesian networks with the`bnlearn` R package. J.Stat.Softw. **35(3)** (2010)
17. Stiglitz, J., Sen, A., Fitoussi, J.P.: Report by the Commission on the Measurement of Economic Performance and Social Progress. INSEE (2009)
18. Svorc, J., Vomlel, J.: Employing Bayesian Networks for Subjective Well-being Prediction, Proceedings of the 11^{th} WUPES'18. 189–204 (2018)

Penalising the complexity of extensions of the Gaussian distribution

Penalizzazione della complessità relativa alle estensioni della distribuzione normale

Diego Battagliese and Brunero Liseo

Abstract The Gaussian distribution has ever been the most popular and usable device in the field of statistics. Even in the context of penalised complexity (PC) priors, the normal density has a particular meaning, especially because we can consider it as a base model which could be extended both in terms of tail thickness and skewness. We derive the numerical PC prior for the shape parameter of the skew-normal density and the analytical PC prior for the degrees of freedom of the t-distribution. We also perform an approximation of the Kullback-Leibler divergence (KLD) in the the skew-normal model.

Abstract *La distribuzione normale ha sempre ricoperto un ruolo fondamentale in statistica. Anche nel caso delle PC prior essa riveste un ruolo importante, giacché può essere estesa sia per via di una componente di curtosi sia per una di asimmetria. Qui deriviamo la PC prior numerica per il parametro di forma di una normale asimmetrica e l'espressione analitica della PC prior per i gradi di libertà di una t di Student. Inoltre, proponiamo un'approssimazione della KLD quando l'estensione è in termini di asimmetria.*

Key words: PC priors, Skew-normal distribution, Student t-distribution, Kullback-Leibler divergence.

Diego Battagliese
Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161, Rome, e-mail: diego.battagliese@uniroma1.it

Brunero Liseo
Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161, Rome, e-mail: brunero.liseo@uniroma1.it

1 Introduction

In many practical statistical works, datasets reveal departures from symmetry, hence something more flexible than the normal model is needed. The skew-normal distribution [1] extends the normal one by introducing in the cumulative distribution function a perturbation parameter that accounts for skewness. The probability density function of a scalar skew-normal random variable X is of the form

$$f(x;\mu,\sigma,\lambda) = \frac{2}{\sigma}\,\phi\Big(\frac{x-\mu}{\sigma}\Big)\,\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big), \quad x\in\mathbb{R}, \quad \lambda\in(-\infty,+\infty), \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard Gaussian pdf and CDF respectively . Also the t-distribution extends the Gaussian in terms of robustness. Penalised complexity priors have been proposed in [4] and are based on the KLD between the simpler and the complex models. For a review of the principles behind the construction of a Penalised Complexity prior, see [4].

2 PC prior for the shape parameter in the skew-normal model

We can look at the skew-normal model as a flexible version of the normal distribution, where the latter represents the base model. In fact, for a particular value of λ, i.e. $\lambda = 0$, the density in (1) boils down to the normal density as $\Phi(0) = 1/2$. An important feature of the PC prior for λ is the invariance with respect to the location and scale parameters.

Proposition 1 (Invariance wrt location-scale) *Let $X_1 \sim SN(\mu,\sigma^2,\lambda)$ and $Y_1 \sim N(\mu,\sigma^2)$ be the skew-normal and normal densities respectively, with the same location and scale parameters. Furthermore, let $X_2 \sim SN(0,1,\delta)$ and $Y_2 \sim N(0,1)$ be the standard versions of the above densities. The Kullback-Leibler divergence between X_1 and Y_1 does not differ from the one between X_2 and Y_2.*

$$\int_{\mathscr{X}} \frac{2}{\sigma}\frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big\}\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big)\log\Big\{2\Phi\Big(\lambda\frac{x-\mu}{\sigma}\Big)\Big\}dx, \tag{2}$$

can be written as

$$\int_{\mathscr{T}} 2\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}t^2\Big)\Phi\Big(\lambda t\Big)\log\Big\{2\Phi\Big(\lambda t\Big)\Big\}dt, \tag{3}$$

where $t = \frac{x-\mu}{\sigma}$ and $dt = \frac{dx}{\sigma}$.

In other words, the resulting PC prior for λ does not depend on μ and σ. Suppose $X \sim SN(0,1,\lambda)$, the distance in terms of δ is

$$d(\delta) = \sqrt{2\mathrm{KLD}(\delta)} = \sqrt{2\int_{\mathscr{X}} 2\;\phi(x)\;\Phi\{\lambda(\delta)x\}\log[2\;\Phi\{\lambda(\delta)x\}]\;dx}, \tag{4}$$

where $\delta = \delta(\lambda) = \frac{\lambda}{\sqrt{1+\lambda^2}}$, $\delta \in (-1,1)$. The distance function in (4) is symmetric around 0, as well as the KLD. The minimum is at 0, where $d(0) = 0$, while the maximum is attained at the boundary values. The distance is exponentially distributed. We must be careful in making the change of variable to get the prior for δ, because we have to handle each monotone curve separately. In particular, the function $d(\delta)$ is monotone on $(-1,0)$ and on $(0,1)$. Then, the pdf for δ is

$$\pi(\delta) = \begin{cases} \sum_{i=1}^{2} \pi\{d_i(\delta)\} \left| \frac{\partial d_i(\delta)}{\partial \delta} \right| & \text{if } d(\delta) \in \Theta \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\Theta = (0,\infty)$ and we make use of Leibniz's Rule to numerically compute the derivative of the distance. The PC prior for δ is

$$\pi^{PC}(\delta|\theta) = \frac{\theta}{2} e^{-\theta\sqrt{2\mathrm{KLD}(\delta)}} \frac{|\mathrm{KLD}'(\delta)|}{\sqrt{2\mathrm{KLD}(\delta)}}, \tag{6}$$

where θ regulates the shrinkage of the prior mass towards the base model. The higher θ the more the shrinkage.

2.1 Approximation of the KLD

The prior above has not a closed form. To this aim we perform an approximation of the KLD based on the moments of the skew-normal distribution. The approximation works pretty good, but not so good on the tails, especially when the parameter θ is small as the probability mass spreads out at the boundaries. Here, we approximate the logarithm of the normal CDF by means of a quintic polynomial regression. The amazing fact is that the intercept α gets closer and closer to $-\log 2$ as we increase the degree of the polynomial regression, and this is crucial to have the $\mathrm{KLD}(\lambda = 0) = 0$. It is not convenient to consider more moments as the quintic approximation seems to work very well. Given $Y \sim SN(\lambda)$, the KLD can be written as

$$\mathbb{E}_Y[\log\{2\Phi(\lambda Y)\}] = \log 2 + \mathbb{E}_Y(\alpha + \beta\lambda Y + \xi\lambda^2 Y^2 + \gamma\lambda^3 Y^3 + \varepsilon\lambda^4 Y^4 + \eta\lambda^5 Y^5), \tag{7}$$

where α, β, ξ, γ, ε and η are the coefficients of the polynomial regression. So, the KLD can be approximated by the first five moments of the skew-normal distribution

$$\log 2 + \alpha + \beta\lambda\sqrt{\frac{2}{\pi}}\delta + \xi\lambda^2 + \gamma\lambda^3\sqrt{\frac{2}{\pi}}(3\delta - \delta^3) + 3\varepsilon\lambda^4 + \eta\lambda^5\sqrt{\frac{2}{\pi}}(15\delta - 10\delta^3 + 3\delta^5). \quad (8)$$

In this way, we would be able to derive an analytical PC prior for λ or δ.

2.2 Bayesian inference for the skew-normal model

We check out the frequentist properties of our PC prior and we compare it to the Jeffreys' prior in [3], in order to see if there could be a certain value of the parameter θ that can be interpreted as objective. We perform a simulation study for different values of the shape parameter, for various sample sizes and for several values of the shrinkage parameter, θ. For any combination we calculate the MSE of the posterior median, the coverage probabilities and the Bayes factor. The posterior median is a reasonable choice, especially for samples where the MLE is infinite, because this entails the non finiteness of the posterior mean, see [2]. The simulation study confirms that large values of θ are quite useless, in the sense that they produce more biased estimates, especially in samples where the true $\lambda \neq 0$. A large value of θ works well only when the true $\lambda = 0$. Anyhow, the gap with respect to a small value of θ vanishes for large sample sizes. In the current work we are interested to find a particular θ that can be interpreted as objective. The simulation study shows that for θ approximately equal to 0.5, the PC prior approaches the estimates produced by the Jeffreys' prior. So, if we had to choose a noninformative value for θ we would say approximately 0.5.

2.3 Bayesian hypothesis testing

We use our PC prior for a Bayesian hypotheses test and we compare it to the Jeffreys' prior in [3], namely a $t(\lambda|\mu = 0, \sigma = \pi/2, \nu = 1/2)$. The proposition stated in Sect. 2 is very important as it allows us to write the Bayes factor in a simplified manner, i.e. without considering the joint prior distribution over the location and scale parameters. The Bayes factor for testing

$$H_0 : \lambda = 0 \qquad \text{vs} \qquad H_1 : \lambda \neq 0$$

can be written as

$$\mathrm{BF}_{01}(x) = \frac{\prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)|_{\lambda=0}}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} 2\phi(x_i)\Phi(\lambda x_i)\pi^{PC}(\lambda|\theta)d\lambda}. \quad (9)$$

We use a uniform importance distribution. Indeed by using a standard Monte Carlo we would draw directly from the PC prior for λ and consequently we could obtain samples that produce negligible values of the likelihood function, for instance if the parameter θ is small and the asymmetry is close to 0. However, choices of small or large θ are suboptimal, in terms of convergence of the Bayes factor towards the true model. If we draw from a PC prior with a parameter θ too small, it is more likely to get extreme values of λ. Then, the marginal likelihood will be close to 0, as long as λ and x_i will have opposite signs. Suppose to draw values of λ from a PC prior with $\theta \to 0$, then for a generic x_i

$$\text{if} \begin{cases} \lambda \to \infty \begin{cases} x_i & \text{is positive} \implies \text{BF}_{01} \approx 0.5 \\ x_i & \text{is negative} \implies \text{BF}_{01} \approx \infty \end{cases} \\ \lambda \to -\infty \begin{cases} x_i & \text{is positive} \implies \text{BF}_{01} \approx \infty \\ x_i & \text{is negative} \implies \text{BF}_{01} \approx 0.5 \end{cases} \end{cases}.$$

On the other hand, the Bayes factor gives no evidence for the true model when the PC prior has a large θ. It doesn't matter what the true model is, and for $\theta \to \infty$ it will be exactly equal to 1. For $\theta \to \infty$ the PC prior becomes a Dirac centered at 0. Then

$$\text{BF}_{01}(x) = \frac{f(x|\lambda)|_{\lambda=0}}{\int_{-\infty}^{\infty} f(x|\lambda)\mathbb{I}_{\{\lambda=0\}} d\lambda} = 1, \tag{10}$$

where $\mathbb{I}_{\{\lambda=0\}}$ denotes the Dirac distribution and $f(x|\lambda)$ is the likelihood function. Simulations seem to favor a $\theta = 2$. The comparison with the Jeffreys' prior encourages the use of our prior.

3 PC prior for the degrees of freedom of the *t*-distribution

For the Gaussian base model, the Kullback-Leibler divergence can be resorted in terms of entropy and second moment of the more complex model. We use the following result

Theorem 1 (Alternative KLD for the Gaussian base model) *Suppose to have a standard normal variate whose density function is f, and a random variable, Y, with a more flexible distribution, g. Then, the KLD between any model that is built up by adding a component to the standard normal base model and the standard normal distribution itself can be expressed as*

$$\text{KLD}(g\|f) = -H(Y) + \frac{1}{2}\left\{\mathbb{E}(Y^2) + \log(2\pi)\right\}, \tag{11}$$

where $H(\cdot)$ stands for the entropy.

Proof.

$$\begin{aligned}\mathrm{KLD}(g\|f) &= \int g \log\left(\frac{g}{f}\right) dy \\ &= \int g \log g \, dy - \int g \log\left\{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)\right\} dy \\ &= -H(Y) - \left\{-\frac{1}{2}\int y^2 g \, dy + \log\left(\frac{1}{\sqrt{2\pi}}\right)\right\} \\ &= -H(Y) + \frac{1}{2}\mathbb{E}(Y^2) + \log(\sqrt{2\pi}).\end{aligned}$$

□

We exploit the theorem above to derive the PC prior for the degrees of freedom, ν, of a t-distribution. The base model for the t-distribution is the Gaussian, which occurs when $\nu = \infty$. In [4] there is an approximation of the KLD, whilst Theorem 1 allows us to derive an analytical expression for the KLD and consequently for the PC prior. In addition, once again the prior is invariant with respect to the location-scale structure. Therefore

$$\mathrm{KLD}(\nu) = -\frac{\nu+1}{2}\left\{\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right\} - \log\left\{\sqrt{\nu}\, B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right\} + \frac{1}{2}\frac{\nu}{\nu-2} - \log\left(\frac{1}{\sqrt{2\pi}}\right), \quad (12)$$

where Ψ is the digamma function and B is the beta function.
The resulting prior is defined only for $\nu > 2$ since the second moment of the Student t-distribution exists only for more than two degrees of freedom. Then

$$\pi(\nu) = \theta e^{-\theta\sqrt{\mathrm{A}(\nu)}} \frac{\left|\frac{1}{4}\left\{-\frac{2}{\nu} - \frac{4}{(\nu-2)^2} + (\nu+1)\Psi^{(1)}\left(\frac{\nu}{2}\right) - (\nu+1)\Psi^{(1)}\left(\frac{\nu+1}{2}\right)\right\}\right|}{\sqrt{\mathrm{A}(\nu)}}, \quad (13)$$

where $A(\nu) = 2\mathrm{KLD}(\nu)$ and $\Psi^{(1)}$ is the trigamma function.

References

1. Azzalini, A.: A class of distributions which includes the normal ones. Scand. J. Statist. **12**, 171–178 (1985)
2. Liseo, B.: La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. Statistica **50**, 59–70 (1990)
3. Liseo, B., Loperfido, N.: A note on reference priors for the scalar skew-normal distribution. J. Stat. Plan. Infer. **136**, 373–389 (2006)
4. Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H.: Penalising model component complexity: A principled, practical approach to constructing priors (with Discussion). Stat. Sci. **32**, 1–28 (2017)

Predictive discrepancy of credible intervals for the parameter of the Rayleigh distribution

Una misura predittiva di discrepanza tra intervalli di credibilità per il parametro della distribuzione di Rayleigh

Fulvio De Santis and Stefania Gubbiotti

Abstract The two most commonly used methods for Bayesian set estimation of an unknown one-dimensional parameter are equal-tails and highest posterior density intervals. The resulting estimates may be numerically different for specific observed samples but they tend to become closer and closer as the sample size increases. In this article we consider a pre-posterior measure of the progressive overlap between these two types of intervals and relationships with the skewness of the posterior distribution. We illustrate the implementation of the method for the Rayleigh model that is often used in the context of reliability and survival analysis.

Abstract *In ambito bayesiano la stima intervallare di un parametro scalare viene comunemente effettuata mediante intervalli di credibilità "equal-tails" e "highest posterior density". Le stime ottenute per particolari campioni osservati possono essere numericamente differenti ma tendono a coincidere all'aumentare della numerosità campionaria. In questo lavoro viene considerata una misura predittiva della discrepanza tra questi due tipi di intervalli, in relazione all'asimmetria della distribuzione a posteriori. La metodologia proposta viene illustrata per il parametro di un modello di tipo Rayleigh che viene solitamente applicato nell'analisi della affidabilità e della sopravvivenza.*

Key words: Bayesian inference, pre-posterior analysis, sample size, skewness.

Fulvio De Santis
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro n. 5, 00185 Roma, e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro n. 5, 00185 Roma, e-mail: stefania.gubbiotti@uniroma1.it

1 Introduction

Equal-tails (ET) and highest posterior density (HPD) intervals are the two most widely used methods for Bayesian set estimation of a scalar parameter. Let us assume that the posterior distribution of the parameter is a density function with a unique mode internal to the parameter space. In many common models ET and HPD intervals do not coincide and show complementary properties and drawbacks. ET intervals are (a) easy to obtain, (b) invariant under (monotonic) parameter transformation; (c) not of minimal length among credible intervals of given credibility level. HPDs are, typically, (a) not easy to compute, (b) not invariant under parameter transformation but (c) of minimal length among intervals of given credibility. Despite the differences, numerical values of the bounds of ETs and HPDs tend to be extremely close for sufficiently large sample sizes, as a consequence of the progressive normalization of the posterior density. Dealing with credible intervals the following two questions may typically arise:

- how do we measure the discrepancy between ET and HPD intervals?
- how do we determine a sample size sufficiently large to guarantee that ET and HPD intervals are close enough?

In this article we propose a method for determining the minimal sample size such that we have good chances of obtaining an ET interval fairly close to the HPD set. We consider a measure of discrepancy between ET and HPD intervals based on probability tails of HPDs originally proposed by [2] and the corresponding preposterior criterion for the selection of the sample size. For illustration we consider the Rayleigh model, often used in reliability and survival analysis (see [1]).

2 Methodology

Assume that $\boldsymbol{X_n} = (X_1, X_2, \ldots, X_n)$ is a random sample from $f_n(\cdot|\theta)$, where $\theta \in \Theta$ is an unknown scalar parameter and $\pi(\theta)$ its prior density. Given an observed sample $\boldsymbol{x_n} = (x_1, x_2, \ldots, x_n)$, let $\pi(\theta|\boldsymbol{x_n})$ be the posterior density of θ with a unique mode $\tilde{\theta}(\boldsymbol{x_n})$, not on the boundary of Θ. Let $C^E_{1-\gamma}(\boldsymbol{x_n}) = [\ell^E_n(\boldsymbol{x_n}), u^E_n(\boldsymbol{x_n})]$ and $C^H_{1-\gamma}(\boldsymbol{x_n}) = [\ell^H_n(\boldsymbol{x_n}), u^H_n(\boldsymbol{x_n})]$ denote the ET and HPD $(1-\gamma)$-credible intervals, respectively. More specifically, $C^E_{1-\gamma}(\boldsymbol{x_n})$ is defined by setting $\ell^E_n(\boldsymbol{x_n}) = q_{\frac{\gamma}{2}}(\boldsymbol{x_n})$ and $u^E_n(\boldsymbol{x_n}) = q_{1-\frac{\gamma}{2}}(\boldsymbol{x_n})$, where $q_\varepsilon(\boldsymbol{x_n})$ is the ε-quantile of $\pi(\theta|\boldsymbol{x_n})$; the HPD bounds are found as roots of the inequality $\pi(\theta|\boldsymbol{x_n}) \geq k_\gamma$, where k_γ is the constant such that $\mathbb{P}[\Theta \in C^H_{1-\gamma}(\boldsymbol{x_n})|\boldsymbol{x_n}] = 1-\gamma$. If $\pi(\theta|\boldsymbol{x_n})$ is symmetric with respect to its mode, $\tilde{\theta}(\boldsymbol{x_n})$, $C^E_{1-\gamma}$ and $C^H_{1-\gamma}$ coincide. If $\pi(\theta|\boldsymbol{x_n})$ is skewed, $C^E_{1-\gamma} \neq C^H_{1-\gamma}$. The degree of posterior skewness depends on the basic inputs of Bayesian analysis: the likelihood (i.e. model and observed data) and the prior distribution. However, under standard regularity conditions, as n increases, both likelihood and posterior tend to normalize

(Bayesian Central Limit Theorem) and any measure of discrepancy between $C^E_{1-\gamma}$ and $C^H_{1-\gamma}$ becomes more and more negligible.

A way to quantify the discrepancy between $C^E_{1-\gamma}$ and $C^H_{1-\gamma}$ proposed in [2] is based on the differences between their respective tail-probabilities:

$$|F_n(\ell^H_n|\boldsymbol{x_n}) - \gamma/2| + |1 - F_n(u^H_n|\boldsymbol{x_n}) - \gamma/2| = |2F_n(\ell^H_n|\boldsymbol{x_n}) - \gamma|, \tag{1}$$

where $F_n(\cdot|\boldsymbol{x_n})$ is the posterior cumulative distribution function (cdf) of Θ and where the equality follows recalling that $F_n(u^H_n|\boldsymbol{x_n}) - F_n(\ell^H_n|\boldsymbol{x_n}) = 1-\gamma$. The quantity in (1) takes values in $[0,\gamma)$. A relative measure of discrepancy is then given by

$$T_n = \frac{|2F_n(\ell^H_n|\boldsymbol{x_n}) - \gamma|}{\gamma}. \tag{2}$$

The values of T_n depend on the overall skewness of the posterior density: the smaller the skewness of $\pi(\theta|\boldsymbol{x_n})$, the smaller the values of T_n.

Before observing the data, the values of T_n can be summarized by its expected value, computed with respect to the sampling distribution of the data $f_n(\cdot|\theta)$. More specifically, let $e^T_n = \mathrm{e}_d[T_n(\boldsymbol{X_n})]$ be the expected value of $T_n(\boldsymbol{X_n})$ computed with respect to $f_n(\cdot|\theta_d)$, where θ_d is a hypothetically true value of θ. We assume that, as n increases, the random sequence $(T_n(\boldsymbol{X_n}),\ n \in \mathbb{N})$ converges in probability to zero – with respect to $f_n(\cdot|\theta_d)$ – and that the numerical sequence $(e^T_n, n \in \mathbb{N})$ converges to zero. If we want to obtain a small discrepancy between HPD and ET intervals, we select the minimum n such that e^T_n is sufficiently small, i.e.

$$n^\star = \min\left\{n \in \mathbb{N} :\ e^T_n < \lambda\right\}, \tag{3}$$

where $\lambda \in (0,1)$ is a chosen threshold.

3 Example: Rayleigh model

In order to illustrate the ideas sketched above we consider the Rayleigh model. Let $X_i|\theta \sim Rayleigh(\theta)$. Given an observed sample $\boldsymbol{x_n} = (x_1,...,x_n)$, the likelihood function is $L(\theta;\boldsymbol{x_n}) = \theta^{-2n}\left(\prod_{i=1}^n x_i\right)\exp\left\{-\frac{s^2}{2\theta^2}\right\}, x_i \geq 0, \theta > 0,$ where $s^2 = \sum_{i=1}^n x_i^2$. Following [1] let us consider a conjugate square root inverse gamma prior for θ, i.e. $\theta \sim SqInvGa(\alpha/2,\beta)$. This yields *a posteriori* $\theta|\boldsymbol{x_n} \sim SqInvGa(\bar{\alpha}/2,\bar{\beta})$, where $\bar{\alpha} = \alpha + s^2$ and $\bar{\beta} = \beta + n$.

In this model HPD intervals do not have a closed-form expression. In the following examples we find the HPD credible bounds numerically and then we compute T_n and e^T_n.

Figure 1 shows the behaviour of e^T_n as a function of n for several values of the prior parameter α, which is the prior parameter that affects the skewness of $\pi(\theta|\boldsymbol{x_n})$. For comparison we also plot the expected value of the posterior skewness (denoted

by e_n^A). Table 1 reports the optimal sample sizes based on criterion (3) for different choices of γ and θ_d.
Here are the main comments.

- Overall, the values of e_n^T decrease as n increases. As a consequence a smaller value of the threshold λ in (3) yields larger values of the optimal sample size.
- The values of e_n^T are affected by the different choices of α at least for small values of the sample size n, i.e. when the prior is not overwhelmed by the data; whereas e_n^T results totally insensitive with respect to β.
- The choice of θ_d does not seem to affect the value of the optimal sample sizes.
- As the credibility level $1-\gamma$ decreases, the values of e_n^T are uniformly smaller for each given sample size n. In fact, when the value of $1-\gamma$ is smaller, the credible interval is defined by a narrower neighborhood of the posterior mode in which the posterior density is substantially symmetric with respect to the mode.

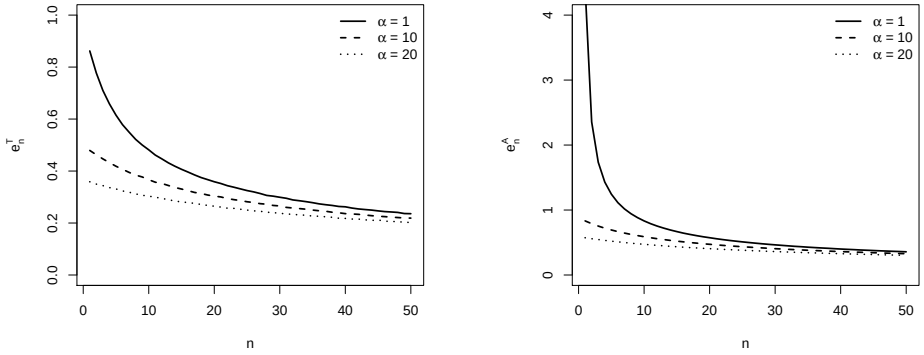


Fig. 1 Values of e_n^T and of e_n^A as functions of n, for different values of the prior parameter α.

$\lambda = 0.3$	θ_d		
γ	0.5	1	10
0.1	33	33	33
0.2	26	26	26
0.3	22	22	22

$\lambda = 0.2$	θ_d		
γ	0.5	1	10
0.1	113	112	116
0.2	64	64	64
0.3	48	47	48

Table 1 Optimal sample sizes for several choices of γ, θ_d and λ, given $\alpha = 3$, $\beta = 4$.

4 Concluding remarks

In this article we consider a measure of discrepancy between HPD and ET intervals based on their tail probabilities. We summarize this quantity, $T_n(\boldsymbol{X_n})$, by its predictive expected value, e_n^T. This quantity overlooks variability of $T_n(\boldsymbol{X_n})$. An alternative summary of the distribution of $T_n(\boldsymbol{X_n})$ is the predictive probability that $T_n(\boldsymbol{X_n})$ is less than λ, which takes into account not only its predictive expected value but also its variability.

As a second extension, we could consider alternative measures of discrepancy, such as the absolute differences between lower and upper limits of HPD and ET intervals. However, an advantage of our proposal is that it provides a relative measure of discrepancy whereas it is not in general straightforward to determine the range of the measure based on the difference between HPD/ET bounds.

In our contribution we limited the analysis to standard conjugate proper priors, which makes derivation of HPD and ET sets straightforward. This assumption can be easily dropped. A first alternative is to consider standard non-informative priors, as long as they lead to proper posteriors. For instance, it can be checked (see [4]) that the Jeffreys' prior is proportional to $1/\theta$ and yields a posterior still within the squared root inverse gamma model. The analysis of the progressive overlap between HPD and ET sets is typically interesting when non-informative priors are used, since these priors tend to induce higher posterior skewness compared to the proper priors in the conjugate family. This is what happens, for instance, in the Poisson-Gamma model (see [2]).

Of course one can easily consider non-conjugate priors. The additional difficulty in this case would be the necessity of implementation of numerical techniques in order to obtain HPD and ET intervals.

Finally, one further topic to explore more deeply is the relationship between the discrepancy T_n and indexes of posterior asymmetry of the posterior distribution as it is done, for instance, in the Poisson-Gamma model by [2].

References

1. Dey, S., Dey, T.: Bayesian estimation and prediction intervals for a Rayleigh distribution under a conjugate prior. Journal of Statistical Computation and Simulation, **82**(11): 1651-1660 (2012).
2. De Santis, F., Gubbiotti, S.: A note on the progressive overlap of two alternative Bayesian intervals. Communications in Statistics – Theory and Methods. *In press* (2019).
3. Ferentinos, K. K., and Karakostas K. X.: More on shortest and equal tails confidence intervals. Communications in Statistics – Theory and Methods, **35**(5): 821–9 (2006).
4. Ahmed, A., Ahmad, S.P. and Reshi, J.A.: Bayesian analysis of Rayleigh distribution. International Journal of Scientific and Research Publications, **3**(10): 1–9 (2013).

Small-area statistical estimation of claim risk

Stima per piccole aree del rischio di sinistri: un approccio statistico

Francesca Fortunato, Fedele Greco and Pierpaolo Cristaudo

Abstract When dealing with *spatial* data, it is crucial to account for the spatial dependency amongst the areas in order to provide reliable estimates of the relative risks. In this work, we borrow techniques from epidemiology and disease mapping to assess the spatial pattern of motor claims and to identify high-risk areas. In particular, we perform a spatial Bayesian modelling approach, combined with the Integrated Nested Laplace Approximation for inference. A LASSO regression is primarily used for the choice of the geographic covariates to include in the model. The dataset at hand consists of 63132 geo-referenced motor claims occurred in the Municipality of Rome throughout 2018 and a set of spatially referenced variables provided by the ISTAT.

Abstract *Quando ci si occupa di dati geo-referenziati, è essenziale tener conto della dipendenza spaziale tra le aree così da fornire stime attendibili del rischio relativo. In questo lavoro, si prende spunto dai modelli epidemiologici per valutare il pattern spaziale dei sinistri auto e identificare le zone a più alto rischio. Nello specifico, si utilizza un approccio di tipo Bayesiano, combinato con l'approssimazione di Laplace per l'inferenza sui parametri. Come step iniziale, si implementa una regressione LASSO per selezionare le variabili da includere nel modello. Il dataset utilizzato contiene informazioni su 63132 sinistri auto accaduti nel Comune di Roma durante il 2018 e su alcune variabili spaziali fornite da ISTAT.*

Key words: Bayesian hierarchical regression models, integrated nested Laplace approximation, spatial regression, motor insurance, ...

Francesca Fortunato
Department of Statistical Sciences and CRIF S.p.A., Bologna, Italy, e-mail: francesca.fortunato3@unibo.it

Fedele Greco
Department of Statistical Sciences, Via delle Belle Arti 41, Bologna, Italy, e-mail: fedele.greco@unibo.it

Pierpaolo Cristaudo
CRIF S.p.A., Via M. Fantin, 1-3, Bologna, Italy, e-mail: p.cristaudo@crif.com

1 Introduction

For a long time, insurance companies have computed the premium for a car insurance just based on some *apriori* factors such as the age and sex of the policyholder, the characteristics of the car he uses and the geographical zone where he lives or gravitates. Recently, companies have realized that also *micro-geography*, i.e. geographical information on a small and detailed scale, can be advantageously taken into account to price contracts at best. Models for premium rating on a small-area level (see [5] and [6]) are borrowed from spatial epidemiology, where the aim is to study the distribution of a disease given the number of occurrences at each location. The basic philosophy of this approach relies upon the belief that events happening in areas that are geographically close will likely display some spatial dependence after accounting for marginal/confounding effects. In practical terms, such dependence structure is enabled by allowing the risk of each location to depend on that of the neighbouring areas. In so doing, the underlying spatial pattern is expected to be correctly recovered and used to provide unbiased estimates of the risk rates.

In the statistical and actuarial literature, some recent works have focused on assessing the relative insurance risk while using the information on the *residence address* of the policyholders (see, for example, [7] and [11]). In this work, we propose to change the point of view and to involve in the analysis the *exact location at which each claim occurs*, rather than the policyholder's residence address. The model is specified in a Bayesian framework by simply extending the idea of hierarchy so as to account for similarities based on the neighbourhood or the distance. The computational burden is made feasible with the INLA approach, i.e. the use of Integrated Nested Laplace Approximation for inference.

We analyse a dataset consisting of 63132 geo-referenced motor claims occurred in the Municipality of Rome throughout 2018 and a set of spatially referenced variables provided by the ISTAT. The aim of the study is to provide insight knowledge about the risk associated to each small-area, while outlining a unified picture of the claim risk based on some relevant demographic and socio-economic variables.

2 Claim risk mapping: a hierarchical Bayesian approach

The most simple method to deal with discrete counts (e.g. number of diseases or claims that occur at a specific location) consists in assuming that these data follow a Poisson distribution. Specifically, given y_i the number of observed cases and θ_i the claim rate corresponding to area i ($i = 1, \ldots n$), then

$$y_i|\theta_i \sim \text{Poisson}(E_i\theta_i), \tag{1}$$

where E_i represents the 'expected' cases in the i-th area according to the reference population. This procedure is known as indirect standardization. The Standardized Incidence Ratio (SIR) is given by $SIR_i = y_i/E_i$, and is the maximum likelihood

estimator (MLE) of the area-specific relative risk θ_i. Although simple, this approach does not consider the spatial dependence among adjacent areas and, thus, it could provide biased estimates of the risk rates.

In order to cope with this issue, methods that *smooth* the SIRs across 'neighboring' areas have been introduced in this context. These methods include two random-effects: a spatial component that accounts for the nearby regions information and an unstructured component that favours shrinkage towards the global mean. If we assume that the Poisson model described in (1) still holds, then a log-linear relationship in the form of a *latent Gaussian*[1] *model* is specified on the linear predictor

$$\log(\theta_i) = \eta_i = \beta_0 + \mathbf{x}_i^T \beta + u_i + v_i, \tag{2}$$

where β_0 is the intercept, quantifying average risk rate in the entire study region, $\mathbf{x}_i$ and β are the area-specific covariates and the regression parameters, respectively, $v_i \sim N(0, \frac{1}{\tau_v})$ represents the spatial *unstructured* area-specific random effect and u_i is the spatial *structured* area-specific random effect.

Several structures can be used to model the spatial correlation $\mathbf{u} = (u_1, \ldots, u_n)$; one of the most popular approaches is the conditional autoregressive (CAR) model introduced by Besag in [1]. Following this approach, the conditional distribution for u_i is

$$P(u_i|\mathbf{u}_{-i}) = P(u_i|u_{j,j\in\mathcal{N}_i}) = N\left(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} u_j, \frac{1}{|\mathcal{N}_i|\tau_u}\right), \tag{3}$$

where $\mathcal{N}_i$ denotes the set of neighbours to area i (i.e. the set of regions that share a common border with i), τ_u is a precision parameter (i.e. $\tau_u = 1/\sigma_u^2$, where σ_u^2 is the variance of the structured random effect) and the conditional variance is inversely proportional to the number of neighbours. A common choice for the priors of $\tau = (\tau_u, \tau_v)$ are the independent gamma distributions (see [4]), so as to enable the data to speak for themselves.

Due to the non-positive definiteness of the covariance matrix, there is no proper joint distribution for $\mathbf{u}$, as it is invariant to the addition of any constant. This issue can be fixed by imposing a sum-to-zero constraint such that $\sum_{i=1}^n u_i = 0$. The CAR specification, along with the exchangeable random effect described in (2), originate the so-called Besag-York-Mollié (BYM) model presented in [2]. The condition described in (3) ensures that the model described in (2) represents a Gaussian Markov Random Field (GMRF) [9], that is a GMRF characterized by a sparse precision matrix $\mathbf{Q}$.

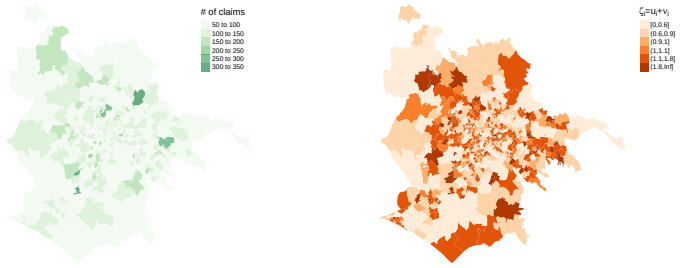
The most challenging aspect of Bayesian inference for spatial models rests in its high computational cost, given the added complexity due to the spatial structure. A number of strategies have been proposed in the literature (see [8] for a detailed review), but many of them (e.g. Monte Carlo Markov Chain, MCMC, algorithms and their variations) suffer from slow convergence and poor mixing for large and complex models. In order to solve these issues, Rue *et al.* [10] introduced an algorithm

[1] The Gaussian part refers to the fact that Gaussian priors are assigned to the vector of latent (nonobservable) components.

for full Bayesian inference of GMRF, based on INLA. Differently from MCMC methods that sample from the posterior distribution of parameters, INLA combines modern numerical techniques for sparse matrices with the Laplace method so as to provide accurate results in a shorter computing time. A very fast implementation of the INLA approach is available in the `R-INLA` package [3] that copes with a large variety of models.

3 Claim data

The dataset at hand consists of $N = 63132$ geo-referenced motor claims requiring the intervention of the traffic police and occurred in the Municipality of Rome throughout 2018[2]. In order to obtain insights about the association of claim frequency and some risk factors, $p = 92$ covariates are added to the dataset. These demographic and socio-economic variables are supplied by the Italian National Statistical Agency (ISTAT) and are available for each census area (i.e. the smallest territorial entity characterized by homogeneous both environmental and socio-economic characteristics). Since we believe that the census division provided by the ISTAT may be too crisp for our purposes (e.g. the variability present in some micro-samples might be not enough to identify meaningful patterns), we decide to consider (small) aggregations of these districts as the unit levels for our spatial regression. In the end, $n = 713$ sections of different shape are involved in the study. Covariates and claim counts are coherently added up, as depicted in Figure 1a, where the number of claims observed in each area is displayed.



(a) Map of the observed number of claims occurred in the Municipality of Rome (2018).

(b) Map of the posterior mean for the area-specific relative risks $\zeta_i = u_i + v_i$.

In order to explore the impact of the areal characteristics on the claim frequency, count data are initially modeled as pure *Generalized Linear Models* (GLMs). Namely, any spatial and unstructured effect is excluded from the model specification with the aim to identify a suitable subset of covariates that provides good

[2] These data are open and available at https://dati.comune.roma.it/catalog/dataset/d655

predictive performances. An effective algorithm for this purpose is the LASSO regression [12], implemented through the `glmnet` R-package with $\alpha = 1$. To avoid the randomness involved in training, a cross-validation procedure is run 100 times. Then, the cross-validated minimum error rates and the corresponding optimal λs (i.e. the regularization parameter) are averaged and the best λ is identified. The LASSO results in the choice of 28 different covariates; this number is further reduced by removing those variables showing a mean absolute correlation larger than 0.75. The spatial structured random-effect is finally included in the linear predictor (see Equation 2) and the model is iteratively fitted until all the β-coefficient show a significant effect. In the end, only just 12 geographic variables are involved in the analysis.

If exponentiated, the fixed effects can be directly interpreted as relative risks, i.e. as an increase or decrease in the probability of a claim according to a unit variation in the exposure factor, *ceteris paribus*. The risk rates provided by the BYM and GLM models are illustrated in Figure 2 and they seem to be in line with our expectations. Nevertheless, these estimates can be further improved if a more suitable strategy for the computation of the "expected" cases is employed. Namely, the use of information on the road or traffic density rather than a procedure that calculates the number of expected claims proportionally to the areal extension (i.e. our approach) would probably contribute to provide more accurate estimates. In this sense, the strong decrease in claim risk (about 60% for any unitary increase) associated to the "% of rural area within 500m" should be carefully interpreted as we improperly *expect* a "non-zero" number of claims in areas with few or no roads.

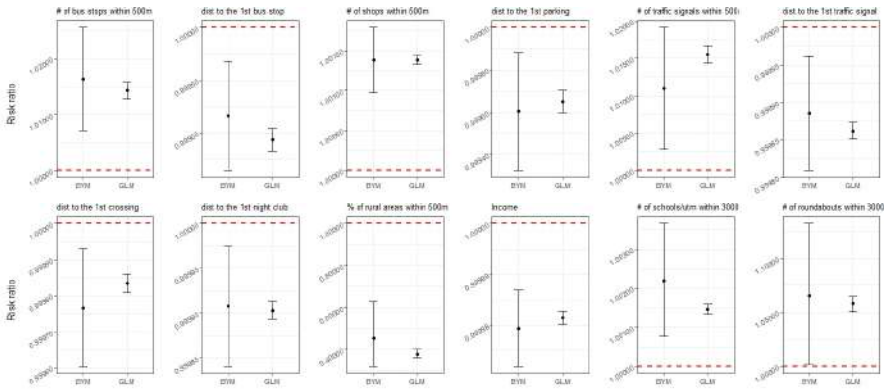


Fig. 2: Posterior mean and 95% credibility interval for the fixed effects on the original scale. The horizontal dashed line highlights a risk rate equal to 1, which corresponds to the absence of effect.

The area-specific relative risks are shown in the map of Figure 1b. They can be interpreted as the residual relative risks for each area (compared to the whole Municipality of Rome) after the fixed effects (i.e. the 12 covariates) are considered.

The analysis of the Deviance Information Criteria (DIC, i.e. the sum of the posterior mean of the deviance and the effective number of parameters) reported in Table 1 and associated to different model specifications suggests that the BYM model with covariates described in (2) is the most suitable for the data at hand[3]. This result implies that it is important to account for both the spatial dependence among the areas and some geographical variables to explain the claim risk in the Municipality of Rome.

Model	Linear predictor	DIC
IID	$\eta_i = \beta_0 + \mathbf{x}_i^T \beta + v_i$	5857
BYM (without covariates)	$\eta_i = \beta_0 + u_i + v_i$	5865
BYM (with covariates)	$\eta_i = \beta_0 + \mathbf{x}_i^T \beta + u_i + v_i$	**5854**

Table 1: DIC for different model specifications. The best model is in bold.

Aknowledgements. This paper is based upon work supported by CRIF group s.p.a and Emilia-Romagna Region.

References

1. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological) **36**(2), 192–225 (1974)
2. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Annals of the institute of statistical mathematics **43**(1), 1–20 (1991)
3. Bivand, R.S., Gómez-Rubio, V., Rue, H.: Spatial data analysis with R-INLA with some extensions. Journal of Statistical Software **63**(20), 1–31 (2015). URL http://www.jstatsoft.org/v63/i20/
4. Blangiardo, M., Cameletti, M.: Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons (2015)
5. Boskov, M., Verrall, R.J.: Premium rating by geographic area using spatial models. ASTIN Bulletin **24**(1), 131–143 (1994). DOI 10.2143/AST.24.1.2005085
6. Brouhns, N., Denuit, M., Masuy, B.: Ratemaking by geographical area: A case study using the boskov and verrall model. Publications of the Institut de statistique, Louvain-la-Neuve pp. 1–26 (2002)
7. Haringa, M.: Small-area statistical analyses of claim frequency in motor insurance. Master's thesis, Amsterdam School of Economics (2016)
8. Heaton, M.J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., Lindgren, F., et al.: Methods for analyzing large spatial data: A review and comparison. arXiv preprint arXiv:1710.05013 **22** (2017)
9. Rue, H., Held, L.: Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC (2005)
10. Rue, H., Martino, S., Chopin, N.: Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) **71**(2), 319–392 (2009)
11. Tufvesson, O.: Spatial statistical modeling of insurance risk. Master's thesis, Lund University (2017)
12. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) **67**(2), 301–320 (2005)

[3] Note that the smaller the DIC, the better is the trade-off between model fit and complexity.

Subject-specific Bayesian Hierarchical model for compositional data analysis

Modello subject-specific gerarchico bayesiano per l'analisi di dati di composizione

Matteo Pedone and Francesco C. Stingo

Abstract We consider the problem of modeling association between continuous covariates and count tables. We propose a hierarchical model which takes into account the complex structure of interactions, imposes sparsity for covariate selection and allows for heterogeneity in the covariate effects through a subject specific regression and still it is computationally tractable. We illustrate the proposed approach through a simple example.

Abstract *Sviluppiamo un metodo generale per l'analisi delle associazioni tra covariate continue e conteggi. Proponiamo un modello gerarchico che tenga conto della complessa struttura delle interazioni, imponga sparsità per la selezione di variabili, che consideri l'eterogeneità nell'effetto delle covariate attraverso una regressione* subject-specific *e che sia computazionalmente efficiente. Infine esponiamo le performance del metodo proposto con un semplice esempio.*

Key words: Bayesian Variable Selection, Dirichlet Multinomial, Microbiome data, Subject-Specific Regression

1 Introduction

Recent studies suggest that the knowledge of microbiome composition and its function has a huge potential as diagnostic tool (Zhu et al., 2010). Motivated by the

Matteo Pedone
Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni,
viale Morgagni, 59 - 50134 Firenze,
e-mail: matteo.pedone@unifi.it

Francesco C. Stingo
Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni,
viale Morgagni, 59 - 50134 Firenze,
e-mail: francescoclaudio.stingo@unifi.it

need for adequate methods for the analysis of such peculiar data, that feature zero inflation and overdispersion, we developed a hierarchical method able to tackle these issues, yet preserving flexibility. Within the Dirichelet-multinomial regression framework, we propose a Bayesian hierarchical model that accounts for the complex structure of the association between continuous covariates and count tables. Moreover, as relationships between the predictor variables can be assumed to be non-zero, first order interactions are considered, leading to a high-dimensional scenario. In this framework effects can be hard to estimate, and this issue can be addressed via variable selection or regularization.

The regression formulation adopted conveniently allows us to construct regression that are subject-specific, which is often more appropriate than population-level models (Ni et al., 2019). To this end, the regression coefficients are allowed to change with the covariates; the model effectively allows the effects of the covariates on the counts to be heterogeneous even when the sample size is small. Inference is conducted through a Markov Chain Monte Carlo algorithm.

The remainder of the manuscript is organized as follows. Section 2 describes the proposed method. Section 3 reports a simulation study and we conclude with a brief discussion of the results and the forthcoming development in Section 4.

2 Model

We let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$, for $i = 1, \ldots, n$, represent a $J-$dimensional response vector of counts, where n is the number of available observations, J the number of categories and Y_{ij} denotes the count of category j in sample i. Let $\mathbf{X}_i = (X_{i1} \ldots, X_{iP})$, $p = 1, \ldots, P$, be the vector of independent covariates, where P is the number of continuous covariates.

For sample i we model the counts with a Multinomial distribution: $\mathbf{y}_i \mid \boldsymbol{\phi}_i \sim \mathit{Multinomial}(y_i^+, \boldsymbol{\phi}_i)$ with $y_i^+ = \sum_j y_{ij}$ being the sum of all counts in the $i-$th vector. The $\boldsymbol{\phi}_i$ parameter vector is defined on the $J-$dimensional simplex. Imposing a Dirichlet prior on $\boldsymbol{\phi}$, that is $\boldsymbol{\phi} \sim \mathit{Dirichlet}(\boldsymbol{\gamma})$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$, $\gamma_j > 0 \; \forall \; j$, and being the Dirichlet distribution a conjugate distribution to the multinomial distribution, $\boldsymbol{\phi}$ can be integrated out and we obtain a compound distribution, the Dirichlet-Multinomial (DM) distribution. The probability mass of a $J-$category count vector $\mathbf{y}$ over $y_i^+ = \sum_j y_{ij}$ trials under Dirichlet-multinomial with parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$, $\gamma_j > 0$, is:

$$f(\mathbf{y} \mid \boldsymbol{\gamma}) = \frac{\Gamma(\sum_j \gamma_j)}{\Gamma(\mathbf{y}^+ + \sum_j \gamma_j)} \times \prod_{j=1}^{J} \frac{\Gamma(\mathbf{y}^+ + \gamma_j)}{\Gamma(\gamma_j)}. \tag{1}$$

Also known as the Dirichlet Compound Multinomial (DCM) or the Pólya distribution, it is a multivariate extension of the beta-binomial distribution. DM is a flexible model for count data, hence may be used when the Multinomial or Poisson model are not appropriate, due to observed overdispersion in the data.

Covariates are included through a log-linear regression model (Wadsworth et al., 2017), imposing dependence of the DM parameters on external covariates: $\zeta_j = \log(\gamma_j)$. Subject-specific regression is adopted, because is more appropriate than single (population-level) regression in some contexts (Ni et al., 2019).

$$\zeta_j = \mu_j \mathbb{1}_n + \sum_{p=1}^{P} \beta_{pj}(\mathbf{X}) \odot \mathbf{X}_p,\ \beta_{pj}(\mathbf{X}) = \theta_{pj}\mathbb{1}_n + h\left(\sum_{k>p}^{P} b_{pkj}\mathbf{X}_k, t\right) \tag{2}$$

where $\odot$ denotes the Hadamard product and $h(\cdot,\cdot)$ is a thresholding function such that $h(x,t) = xI_{[|x|>t]}$, with $I_{[\cdot]}$ denoting the characteristic function. Note that, being the coefficients $\{\beta_{pj}\}$ function of the covariates, we let effects vary similarly for similar covariate values, obtaining similar coefficients for close realizations of covariates.

The model depicted in Equations (2), even for modest dimension data leads rapidly to a high-dimensional framework. The population-level structure needs $J + (((P-1)\times P)/2)\times J$ parameter: for example, considering $J = 10$ and $P = 10$, we need to estimate 460 population-level parameters. Especially in the case of small sample size, inducing sparsity is vital for the correct estimation of the model. This is accomplished through *spike-and-slab* mixture prior for the main effects $\{\theta_{pj}\}$ (George and McCulloch, 1997) and *horseshoe prior* (Carvalho et al., 2010) for the interaction terms $\{b_{pkj}\}$. Since horseshoe prior encourages shrinkage of regression coefficients close to the null value, it is coupled with a soft-thresholding mechanism that allows to set exactly to zero effects shrunk toward small and negligible values.

Spike-and-slab mixture prior is a common approach for the selection of significant associations between the covariates and the response (George and McCulloch, 1997). It assumes a discrete mixing distribution with an atom at zero and one at a non-zero value. Spike-and-slab priors are placed on the main effects $\{\theta_{pj}\}$:

$$\theta_{pj} \sim \xi_{pj} Normal(0, \tau_j^2) + (1-\xi_{pj})\delta_0(\theta_{pj}), \tag{3}$$

where $\delta_0(\cdot)$ is a Dirac function and $\xi_{pj} \sim \text{Bin}(1, \omega_{pj})$. Moreover, we assume a beta hyper-prior for $\omega_{pj} \sim \text{Beta}(a,b)$, that automatically adjusts for multiplicity (Scott and Berger, 2010). This hierarchy is equivalent to placing a beta mixed binomial distribution on ξ_{pj}, which is a valuable property, as the prior expected mean of the beta distribution reflects the proportion of associations to be selected *a priori* (Wadsworth et al., 2017).

The horseshoe prior (Carvalho et al., 2010), as opposed to spike-and-slab, assumes an absolutely continuous mixing distribution and belongs to the class of global-local scale mixtures of normals. It is adopted for the coefficients that model the interactions among covariates, namely b_{pkj}:

$$b_{pkj} \sim Normal(0, \lambda_{pkj}^2\tau_{kj}^2),\ \lambda_{pkj} \sim C^+(0,1),\ \tau_{kj} \sim C^+(0,1), \tag{4}$$

where $C^+(\cdot)$ is a truncated Cauchy distribution and the matrix $\mathbf{b}_j$ is strictly triangular to ensure identifiability of the parameters. Parameters λ_{pkj}'s are called *local* shrink-

age parameters, while τ_{kj}'s are the *global* ones. They create an absolutely continuous mixture, though the discrete mixture approach is the underlying methodological foundation. The relationship between the two parameters of the scale mixture is such that the global parameter takes action on the noise, while the local one (which is an outlier to the global one) discloses the signals. This class of priors encourages shrinkage of regression coefficients, but it induces sparsity in a weaker sense. All coefficients will be nonzero, nonetheless only the significant associations will have large values -due to the heavy tails of the prior- compared to negligible values of the non significant ones, severely shrunk because the prior has an infitely tall spike in a neighborhood of zero. An advantage of horseshoe priors, with respect to spike-and-slab priors, is their computational efficiency.

To obtain truly sparse solutions, the horseshoe prior has been combined with the soft thresholding mechanism proposed and discussed in Ni et al. (2019), where the threshold is interpreted as minimum effect size. The threshold is not fixed *a priori*, and a prior distribution is imposed, in particular we assume $t \sim U(a,b)$. This modeling strategy implies that both the magnitude and the variability of the regression coefficients are taken into account in the selection mechanism.

2.1 Algorithm

We present the posterior inference procedure for the purpose of estimating the proposed model, via MCMC algorithm approach. The MCMC sampler goes as follows:

1) **Update** μ_j: by Metropolis-Hastings step with a symmetric random walk proposal;
2) **Update** θ, ξ: these parameters are jointly updated by Gibbs scan and a Metropolis acceptance step, with an adaptive proposal, as detailed in Wadsworth et al. (2017).
3) **Update b**: these parameters are updated with a Metropolis step, using an adaptive proposal;
4) **Update** λ, τ: these parameters are updated with Slice Sampler, following the procedure detailed in the Supplemental Material of Polson et al. (2014).;
6) **Update** t: by Metropolis-Hastings step with a symmetric random walk proposal.

3 Simulation Study

We perform a simulation study to evaluate the variable selection performance of our method. We simulate $n = 100$ samples, considering $P = 5$ covariates and $J = 5$ categories, being $P^\star$ the relevant covariates with each being associated with $J^\star$ relevant categories. The design matrix is obtained using a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, $\Sigma_{i,j} = \rho^{|i-j|}$. For each covariate, the association coefficient θ_{pj} for the significant categories are sampled from the interval $[1.5, 3.5]$ with alternate sign.

Considering the exponential growth model defined in Chen and Li (2013) we obtained the proportion of the $j-$th category in the $i-$th sample. The overdispersion is regulated by the parameter $\psi \in [0,1]$, when $\psi \to 0$ the simulated values approximate Multinomial distribution, while the larger ψ, the more overdispersed are the values. To evaluate the performance of the method on different ground of sparsity, we generated the data with different number of non-zero principal effect, hence $p^\star = P^\star \times J^\star / P \times J$.

Table 1 Simulated data: performance assessment for 12 scenarios, characherized by different levels of sparsity in the main effects, different values of the dispersion parameter ψ and the assumption on the interaction generative mechanism.

	Strong Heredity			Weak Heredity		
	$p^\star = 0.08$	$p^\star = 0.12$	$p^\star = 0.16$	$p^\star = 0.08$	$p^\star = 0.12$	$p^\star = 0.16$
$\psi = 0.01$						
MCC	1.0000	1.0000	1.000	0.7158	1.0000	0.7103
FPR	0.0000	0.0000	0.0082	0.0000	0.0000	0.0000
FNR	0.0000	0.0000	0.0000	0.0376	0.0000	0.0525
AUC	1.0000	1.0000	1.000	0.9812	1.0000	0.9738
$\psi = 0.1$						
MCC	1.0000	0.6853	0.7021	1.0000	1.000	0.6548
FPR	0.0000	0.0000	0.0755	0.0000	0.000	0.3333
FNR	0.0000	0.3333	0.0000	0.0000	0.000	0.1551
AUC	1.0000	0.9167	0.9623	1.0000	1.000	0.7577

Values are rounded averages over ten replicates. Results for Matthews' Correlation Coefficient, False Positive Rate and False Negative Rate are based on the median probability model.

Prior independence between the regression coefficients could often be a misleading assumption, so first order interactions are considered. Given the main effects, there are several assumptions that could be made on the interactions generative mechanism. We will consider two forms of heredity principle to this end: *strong heredity*, states that an interaction can only be included if both main effects are included, while *weak heredity* states that an interaction can be included if at least one main effect is included (Griffin et al., 2017). The values for these parameters are obtained sampling uniformly on the set $\{0, 1.0, 3.0\}$. Moreover, a priori, the sign of the interaction is not related to the signs of the main effects. We choose to take into account the kind of heredity of the interaction, but note that in principle any criterion could be used. In particular:

$$\text{sgn}(b_{pqj}) := \begin{cases} \text{sgn}(\theta_{pj} \cdot \theta_{qj}) \text{ if } \theta_{pj}, \theta_{qj} \neq 0; \\ \text{sgn}(u), u \sim U(-1,1) \text{ otherwise.} \end{cases}$$

In order to assess the performance of the proposed method and explore its capability to recover true associations and interaction, a simulation study is conducted. The selection is evaluated on the ground of the Matthews' Correlation Coefficient, False Positive Rate, False Negative Rate and the Area Under the Curve. The selection is compared under strong heredity e weak heredity assumption for a different level of sparsity in the main effects. The method is tested for low and moderate level of overdispersion in the data. The results are reported in Table (1).

4 Discussion

We presented a simple strategy for modeling associations between continuous covariates and counts, that can take into account *subject-specific* interactions. Our goal is to develop an approach for the selection of relevant associations between microbiome taxa and covariates. Several extensions are currently under investigation: for example, non-linear effects could be included to let covariates effects vary *smoothly* with the covariates.

References

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

J. Chen and H. Li. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics*, 7(1), 2013.

E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

J. Griffin, P. Brown, et al. Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12(1):135–159, 2017.

Y. Ni, F. C. Stingo, and V. Baladandayuthapani. Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197, 2019.

N. G. Polson, J. G. Scott, and J. Windle. The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):713–733, 2014.

J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010.

W. D. Wadsworth, R. Argiento, M. Guindani, J. Galloway-Pena, S. A. Shelburne, and M. Vannucci. An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics*, 18(1):94, 2017.

B. Zhu, X. Wang, and L. Li. Human gut microbiome: the second genome of human body. *Protein & cell*, 1(8):718–725, 2010.

Wasserstein consensus for Bayesian sample size determination

Consenso di Wasserstein per la determinazione Bayesiana della dimensione campionaria

Michele Cianfriglia, Tullia Padellini and Pierpaolo Brutti

Abstract The sample size determination problem deals with the selection of the optimal number of subjects to be enrolled in a study in order to achieve a pre–specified inferential goal. While this problem can of course be approached from a frequentist viewpoint, often the Bayesian paradigm is preferred as it allows to blend and balance the strength of the observed empirical evidence with the available prior knowledge. In this work, we focus on the case of a "community of priors" representing, for example, different expert opinions. Within this setup, we are interested in selecting the smallest sample size that guarantees "agreement" between these, possibly conflicting, opinions, having formalized the loose idea of "agreement" in terms of the Wasserstein distance between posteriors stemming from different priors.

Abstract *Il problema della determinazione della dimensione campionaria concerne la selezione del numero ottimo di soggetti da considerare in uno studio al fine di raggiungere un prestabilito livello di accuratezza inferenziale. In quest'ambito, l'approccio Bayesiano è in genere preferito perché permette di bilanciare i dati osservati con le informazioni a priori disponibili. Scopo di questo lavoro è partire da una"famiglia di a priori" che rappresentano, ad esempio, le opinioni di un pool di esperti, ed arrivare ad individuare il numero minimo di osservazioni necessario a garantire un accordo tra le diverse posizioni considerate. Il concetto di accordo è formalizzato attraverso la distanza di Wasserstein tra le distribuzioni a posteriori ottenute, ad esempio, a partire dalle diverse opinioni a priori degli esperti coinvolti.*

Key words: Bayesian consensus, sample size determination, Wasserstein distance, clinical trials.

Michele Cianfriglia · Pierpaolo Brutti
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: {michele.cianfriglia}, {pierpaolo.brutti}@uniroma1.it

Tullia Padellini
Department of Epidemiology and Biostatistics, Imperial College London
e-mail: t.padellini@imperial.ac.uk

1 Introduction

In the design of a statistical experiment a fundamental issue is the determination of the optimal sample size (SSD). Focusing, for example, on clinical trials, SSD is particularly critical as there is an immediate interpretation of the optimal sample size in terms of the minimum number of subjects to be enrolled in order to achieve a pre–specified inferential target. Historically, while both frequentist and Bayesian solutions have been provided for this problem, the Bayesian paradigm is often preferred as it allows to combine observed empirical evidence with all the prior knowledge already available, including historical data. Among others, examples of this approach can be found in [3, 7, 8, 10] and [4].

In their general formulation, common Bayesian criteria used to determine the optimal sample size need to control the inferential performance of specific functionals of interest of the posterior distribution, e.g. posterior mean or credible interval. For this reason the optimal sample size is usually defined as the minimum sample size for which the aforementioned functionals achieve some inferential goal on average or with large enough (predictive) probability.

Our goal in this work is to address the SSD problem in case where we are dealing with more than one prior distribution for the parameter of interest, a "community of priors". Indeed this is a not an uncommon setting as it can be used, for example, to model multiple scenarios we want to account for, or even different expert opinions. A possible approach to this problem detailed in [2], consists in aggregating different beliefs through the use of finite mixture models. Here however we adopt a different definition of consensus, which is not based on aggregation but rather on *distance minimization*. To this end, our work heavily relies on the *Wasserstein metric*, a well known distance between probability distributions that nowadays is gaining momentum in the statistics and machine learning communities.

Hence, once the loose idea of "agreement" has been formalized in terms of the Wasserstein distance between posteriors stemming from different prior distributions, we propose a Bayesian SSD procedure for the selection of the smallest sample size that guarantees "agreement" between possibly conflicting prior opinions. We detail our proposal in the case of the ubiquitous conjugate *Normal model* for which the Wasserstein distance admits an exact expression neatly depending on the posterior means and standard deviations of the distributions involved.

This paper is organized as follows. In the next section we introduce the new definition of consensus. In Section 3 two main Bayesian sample size criteria, that is the Predictive Expectation Criterion and the Predictive Probability Criterion, are redefined in terms of the Wasserstein distance, and an application to real data is reported. Finally Section 4 shows some motivating results based on the data analyzed.

2 A new perspective of consensus: Wasserstein consensus

In designing a new clinical trial it is quite common to take into account the results provided by previous, related, medical studies or to consult several experts. Since the information provided by these variety of sources can happen to be quite conflicting, once the uncertainty surrounding them is captured in terms of a suitable probability distribution, we are left with the statistical problem of working with a "community of priors" elicited over the same parameter of interest.

In the literature, conflicting prior opinions are typically handled by *robust methods* – such as finite mixture or ε–contaminated priors – as, for example, in [1, 2, 4]. There are two main issues with this approach: (i) we lose in "resolution" since we start the Bayesian machinery from a single, encompassing, prior built to cover the range, from optimistic to pessimistic, spanned by the pool of elicited distributions; and consequently (ii) we also lose the opportunity to address the SSD problem, or any other design problem for what matters, with the explicit goal of achieving any sort of "agreement" between conclusions stemming from different scenarios.

Broadly speaking, there are of course many ways to define "agreement". Within the Bayesian framework we are working with, this loose idea naturally translates into guaranteeing *posterior* "agreement" between possibly conflicting prior opinions. To date, the only work which addresses the issue of determining the optimal sample size in order to guarantee posterior consensus is [6].

In this work, we formalize Bayesian consensus via the Wasserstein distance between posterior distributions stemming from different priors. To this end, taking inspiration from the main Bayesian sample size criteria, we want to find the minimum number of subjects which ensures that a suitable posterior summary of the Wasserstein distance is lower than a pre–specified threshold.
The use of the Wasserstein distance has several advantages, for instance:

- we easily define "agreement" between experts: two experts "agree" if their inferential conclusions, namely their posterior distributions, are "close" enough;
- it "metricizes" convergence in distribution: if two distinct distributions are close with respect to the Wasserstein distance, then they are probabilistically similar;
- it tells us why distributions differ: the Wasserstein distance is a transportation distance and comes with a transportation plan that detail how to move the mass of a distribution to morph it into another.

3 Our proposal: Wasserstein based criteria

Before describing our proposal, let us recall the basic ideas and definitions behind the two main Bayesian sample size criteria. Generally speaking, these criteria are defined in terms of suitable summaries of the posterior distribution, that is

$$\rho_\pi(\theta|\mathbf{x}_n) = \int g(\theta)\pi(\theta|\mathbf{x}_n)\,\mathrm{d}\theta,$$

where different choices of $g(\theta)$ lead to different summaries of $\pi(\theta|\mathbf{x}_n)$. Denoting by $m(\cdot)$ the prior predictive distribution, we are now in position to define

- *Predictive Expectation Criterion (PEC)*: let e_n be the expected value of $\rho_\pi(\theta|\mathbf{x}_n)$ with respect to $m(\cdot)$. Given a suitable threshold η_e the optimal sample size is

$$n_e^\star = \min\{n \in \mathbb{N} : e_n > \eta_e\};$$

- *Predictive Probability Criterion (PPC)*: let p_n be the probability with respect to $m(\cdot)$ that $\rho_\pi(\theta|\mathbf{x}_n)$ is bigger than a constant γ. Given a suitable threshold $\eta_p \in (0,1)$ the optimal sample size is

$$n_p^\star = \min\{n \in \mathbb{N} : p_n > \eta_p\}.$$

A few comments: (i) these criteria are called *predictive* because they use the marginal predictive distribution $m(\cdot)$; (ii) in defining the Bayesian criteria we usually consider two distinct prior distributions, specifically π_A (*analysis prior*) – which is employed in determining the posterior distribution – and π_D (*design prior*) – which induces the marginal distribution of the data.

We can now define the Wasserstein based criteria. In the following, d_W denotes the 2–Wasserstein distance, based on the L_2 ground metric, between two posterior distributions derived by two different analysis priors.

- *Predictive Expectation Criterion (PEC)*: let e_n^W be the expected value of d_W with respect to $m(\cdot)$. Given a suitable threshold η_e the optimal sample size is

$$n_{e,W}^\star = \min\{n \in \mathbb{N} : e_n^W < \eta_e\}; \tag{1}$$

- *Predictive Probability Criterion (PPC)*: let p_n^W be the probability with respect to $m(\cdot)$ that d_W is bigger than a constant γ. Given a suitable threshold $\eta_p \in (0,1)$ the optimal sample size is

$$n_{p,W}^\star = \min\{n \in \mathbb{N} : p_n^W < \eta_p\}. \tag{2}$$

Please notice that the optimal sample sizes based on these criteria have a *particular* interpretation because we are searching for the smallest sample size that guarantees "agreement" between possibly conflicting opinions, therefore it is natural to reverse the inequality of the Bayesian criteria as done in (1) and (2). In addition, under a Normal conjugate model, the functions e_n^W and p_n^W admit a closed form expression making the computation of the Wasserstein based criteria extremely efficient. This is quite crucial since the computation of the Wasserstein distance becomes quite expensive and time consuming when multidimensional posterior distributions are involved.

4 A real data example

We conclude by showing how the Wasserstein based–criteria perform on a real data application. More in details, we revisit an example reported in [9], where results of a meta–analysis are reinterpreted from a Bayesian perspective. The setup is the following: a series of small randomized trials was conducted to assess a proactive effect of intravenous magnesium sulphate after acute myocardial infarction. Even if many studies were already conducted, further investigation was suggested and the massive ISIS-4 trial started. Unfortunately, this study did not provide evidence of any benefit contradicting previous conclusions (see [5] and Figure 1 for details).

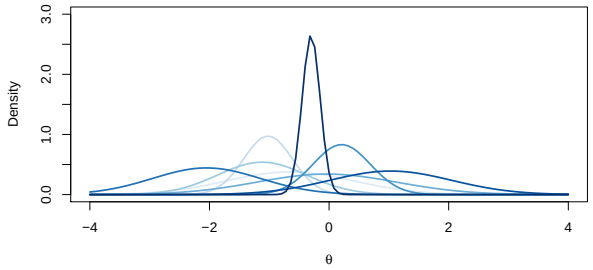


Fig. 1 Comparison between Normally distributed priors encoding evidence from previous studies. The parameter of interest θ is the log odds ratio of intravenous magnesium with respect to placebo. As suggested in [9], the variance of the analysis prior is σ^2/n_0, where $\sigma^2 = 4$ is fixed and n_0 represents the *effective number of events* in the study considered.

Table 1 shows the optimal sample size $n^{\star}_{e,W}$ obtained using our version of the PEC with $\eta_e = 0.05$. The marginal predictive distribution is based on a Normally distributed design prior with mean 0.058 and variance $\sigma^2 = 4/n_D$. Intuitively, different choices of the design prior sample size n_D reflect different strength of the skepticism toward the magnesium treatment: as n_D increases it becomes easier and easier to bring consensus between different parties since we have more and more information about the phenomenon. From a mathematical point of view, this aspect can be understood looking at the behavior of e^W_n as function of n_D.

n_D	$n^{\star}_{e,W}$	$n^{\star}_{e,M}$
4319	361	498
432	371	509
43	468	190

Table 1 Comparison between optimal sample sizes. The first column contains the prior sample size of the design prior, whereas the other two columns contain, respectively, optimal sample sizes associated with the Wasserstein and mixture versions of the PEC.

Finally, in order to highlight the need for "consensus–assessment", we compare our optimal sample sizes with those obtained by [2], where a finite mixture version of the standard PEC is used. Note that the two methods are *not* directly comparable as their inferential goals are different.

Nevertheless we stress the fact that consensus does not come automatically and "for free": the more the parties disagree, the higher is the number of subjects to be enrolled in the study in order to achieve inferential agreement.

References

1. Brutti, P., De Santis, F., Gubbiotti, S.: Robust bayesian sample size determination in clinical trials. Statistics in Medicine **27**(13), 2290–2306 (2008)
2. Brutti, P., De Santis, F., Gubbiotti, S.: Mixtures of prior distributions for predictive bayesian sample size calculations in clinical trials. Statistics in medicine **28**(17), 2185–2201 (2009)
3. Brutti, P., De Santis, F., Gubbiotti, S.: Bayesian–frequentist sample size determination: a game of two priors. Metron **72**(2), 133–151 (2014)
4. De Santis, F.: Using historical data for bayesian sample size determination. Journal of the Royal Statistical Society: Series A (Statistics in Society) **170**(1), 95–113 (2007)
5. Higgins, J.P., Spiegelhalter, D.J.: Being sceptical about meta-analyses: a bayesian perspective on magnesium trials in myocardial infarction. International journal of epidemiology **31**(1), 96–104 (2002)
6. Joseph, L., Bélisle, P.: Bayesian consensus-based sample size criteria for binomial proportions. Statistics in medicine (2019)
7. Joseph, L., Bélisle, P., Du Berger, R.: Bayesian and mixed bayesian/likelihood criteria for sample size determination. Statistics in medicine **16**(7), 769–781 (1997)
8. Joseph, L., Du Berger, R., Wolfson, D.B.: Some comments on bayesian sample size determination. Journal of the Royal Statistical Society: Series D (The Statistician) **44**(2), 167–171 (1995)
9. Spiegelhalter, D.J., Abrams, K.R., Myles, J.P.: Bayesian approaches to clinical trials and health-care evaluation, vol. 13. John Wiley & Sons (2004)
10. Wang, F., Gelfand, A.E., et al.: A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. Statistical Science **17**(2), 193–208 (2002)

Biostatistics

A comparison of the CAR and DAGAR spatial random effects models with an application to diabetics rate estimation in Belgium

Confronto tra i modelli CAR e DAGAR per effetti spaziali latenti con un'applicazione alla stima del tasso di diabetici in Belgio

Vittoria La Serra, Christel Faes, Niel Hens and Pierpaolo Brutti

Abstract When hierarchically modelling an epidemiological phenomenon on a finite collection of sites in space, one must always take a latent spatial effect into account in order to capture the correlation structure that links the phenomenon to the territory. In this work, we compare two autoregressive spatial models that can be used for this purpose: the classical CAR model and the more recent DAGAR model. Differently from the former, the latter has a desirable property: its ρ parameter can be naturally interpreted as the average neighbor pair correlation and, in addition, this parameter can be directly estimated when the effect is modelled using a DAGAR rather than a CAR structure. As an application, we model the diabetics rate in Belgium in 2014 and show the adequacy of these models in predicting the response variable when no covariates are available.

Abstract *Quando si vuole modellare un fenomeno epidemiologico su un insieme finito di siti, bisogna sempre tenere conto della presenza di un effetto spaziale latente, necessario per spiegare la struttura di correlazione che lega il fenomeno al territorio. In questo lavoro confrontiamo due modelli autoregressivi spaziali atti allo scopo: il CAR e il DAGAR. A differenza del primo, il secondo risulta naturalmente parametrizzabile in termini di correlazione media tra siti vicini, caratteristica questa correttamente stimabile allorché l'effetto venga modellato mediante un DAGAR anziché un CAR. Oltre che attraverso uno studio di simulazione, il comportamento dei due modelli viene testato su un'applicazione a dati reali relativi al tasso di diabetici in Belgio nel 2014.*

Key words: Spatial Correlation, Autoregressive models, CAR, DAGAR, Epidemiology, Diabetes.

Vittoria La Serra, Pierpaolo Brutti
Sapienza Università di Roma, e-mail: vittoria.laserra@uniroma1.it, pierpaolo.brutti@uniroma1.it

Christel Faes, Niel Hens
Hasselt University, e-mail: christel.faes@uhasselt.be, niel.hens@uhasselt.be

1 Introduction

Disease mapping is an area of spatial epidemiology whose purpose is to estimate the spatial pattern of disease rates over some geographical region of interest, often taken as a finite collection of sites, like cities or counties. Spatial units are usually organized in graph representations where each unit is a node and two nodes are connected by an edge if they can be considered "neighbors". In spatial epidemiology, for example, it is common to assume that two sites are neighbors if they share a boundary.

When observing an epidemiological phenomenon at a specific site, we can imagine it as being influenced by its behaviour in neighboring sites; this effect is usually modelled by adopting a hierarchical approach involving a suitable spatial autoregressive model for the so-called *latent spatial effect*, that is, a random process whose covariance structure captures information on the local spatial dependency that links the phenomenon to the territory.

2 CAR and DAGAR models

Among the available autoregressive models for a latent spatial effect – typically a zero-mean, structured–covariance multivariate process – one of the most used is the Conditional Autoregressive (CAR) model. This model adopts a *symmetrical* neighborhood structure, meaning that for a couple of nodes $\{i, j\}$ with neighborhood sets $\{N(i), N(j)\}$ we have:

$$j \in N(i) \Leftrightarrow i \in N(j). \tag{1}$$

The implied model on the latent spatial effect is then specified as

$$w_i | w_{-i} \sim \mathrm{N}\left(\sum_{j \in N(i)} \frac{\rho'}{n_i} \cdot w_j, \tau_w \cdot n_i \right), \forall i, \tag{2}$$

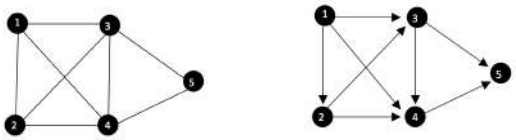
where n_i denotes the cardinality of $N(i)$, τ_w is a precision parameter and ρ' is the spatial correlation taken to be in $(0, 1)$ for easy of comparison.

A more recent addition to the disease mapping toolkit is the Directed Acyclic Graph Autoregressive (DAGAR) model introduced in [1]. Differently from the CAR, this model is based on an *asymmetric* neighborhood structure. More specifically, in order to get some desirable properties, to build a DAGAR neighborhood $N(i)^\star$, we do start from a symmetrical one $N(i)$, but we then impose a *fictional*[1] ordering on the nodes, which is used to "prune down" the original neighborhood $N(i)$ to get the new, simplified one, as follows:

[1] In practice, the sites of a specific territory rarely come with a natural order attached. Hence the ordering we need to introduce is typically fictional and subjective (for instance, based on longitude or latitude). In [1] the Authors proved the robustness of the results to changes in the ordering.

$$N(i)^\star = \{j \in N(i) \text{ such that } j < i\}, \forall i. \tag{3}$$

Using this constraint, we can turn an undirected graph into a directed one, as exemplified in Figure (1), together with the neighborhood reduction.



(a) An undirected graph.

(b) An associated directed graph.

Fig. 1: Focusing on node 3: (a) $N(3) = \{1,2,4,5\}$. (b) By Eq.(3), $N(3)^\star = \{1,2\}$.

For each node i, we want to specify a model for $w_i|w_{-i}$ and we do it by assuming an AR(1) model with parameter $\rho \in (0,1)$ on the local spanning tree made by the nodes $\{i, N(i)^\star\}$ and the edges connecting them. In this case, ρ^d is clearly the correlation between any two variables whose indices are d edges apart.[2] So we write:

$$\left(w_i, w_{N(i)^\star}\right) \sim \text{MVN}\left(0, \begin{bmatrix} 1 & v_i^\top \\ v_i & \Sigma_i \end{bmatrix}\right), \tag{4}$$

where v_i is a vector with all entries equal to ρ, whereas Σ_i is a square matrix with all ones on the diagonal and all ρ^2 outside. Assuming Eq.(4) and asking for $\mathbb{E}\left(w_i|w_{N(i)^*}\right)$ to be a linear combination of $\{w_j, j \in N(i)^\star\}$, we obtain the following conditional model:

$$w_i|w_{-i} \sim \text{N}\left(\sum_{j \in N(i)^\star} \frac{\rho}{1+(n_i^\star-1)\rho^2} \cdot w_j, \tau_w \cdot \frac{1+(n_i^\star-1)\rho^2}{1-\rho^2}\right), \forall i, \tag{5}$$

where $n_i^\star$ is the cardinality of $N(i)^\star$.

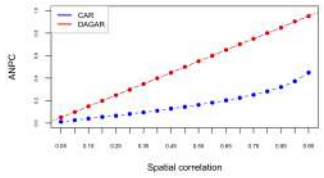
With this model specification, we see that the ρ parameter can be naturally interpreted as the correlation between two neighboring[3] nodes; the same cannot be said for ρ' in the CAR model. To check it empirically, we consider two types of territory characterized by a regular and an irregular grid of counties; on each of the two, we build a symmetrical neighborhood structure and we simulate 1000 times from a CAR model; we then impose a longitude-based ordering on the nodes, we reduce the neighborhood structure using Eq.(3) and we simulate 1000 times from a DAGAR model. Based on these simulations, we can approximate the Average Neighbor Pair Correlation (ANPC) under the two models as:

[2] We want $\rho \in (0,1)$ for the matrix ρ^D to be positive definite, if D is the distance matrix.

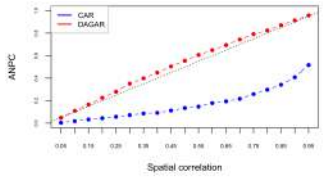
[3] It makes sense for such a correlation to be positive since we expect a "concordant" behaviour of the phenomenon in neighboring sites.

$$\text{ANPC} = \frac{\sum_{i\sim j} \mathbb{C}\text{or}\,(w_i, w_j)}{\sum_{i\sim j} n_i}, \text{ where } i \sim j \text{ means they are neighboring sites.} \quad (6)$$

This is repeated for different values of the spatial correlation parameter in $(0,1)$, respectively ρ' and ρ, holding τ_w fixed and equal under both models. The results are collected in Figure (2). As we can see, in both the regular and the irregular case, the ρ parameter in the DAGAR model almost equals the ANPC, whereas the CAR model is associated to an ANPC that systematically underestimates ρ'.



(a) 20×20 1^{st}-order regular grid. In the symmetrical structure, two nodes are neighbors if they are connected by an edge of the grid.



(b) Irregular grid made of the 44 municipalities of Limburg (BE). The symmetrical structure is built as an extension of a $k = 4$ NN one.

Fig. 2: Average correlation plots for neighbor pairs.

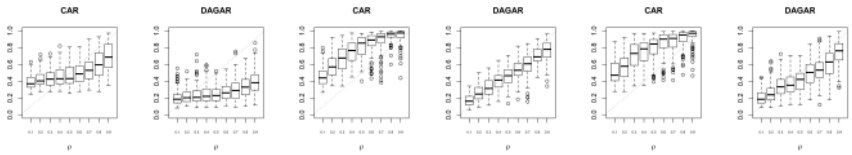
3 Simulation analysis

In this section we detail the results of the following simulation study: taking ρ as the spatial correlation parameter in each model, we fix it as a value in the set $\{0.1, \ldots, 0.9\}$, then we simulate the spatial effect w from one of these models: the CAR, the DAGAR and a Gaussian Process that ensures the true ρ to be equal to the ANPC.[4] We then simulate the response variable as $Y = X\beta + w + \varepsilon$ after having generated the covariates $X \sim \text{MVN}(0, \mathbb{I})$ and the error $\varepsilon \sim \text{MVN}(0, \tau_e \cdot \mathbb{I})$ for a fixed set of precision parameters $\{\tau_w, \tau_e\}$ and regression coefficients β. We fit a hierarchical model on the data: we give β an MVN prior, ρ a $\text{Unif}(0,1)$ prior, whereas the precision parameters τ_w and τ_e are given Gamma priors. We write the model in Nimble [2]. Posterior estimates of the parameters in the model are obtained as MCMC posterior means; the used algorithm is *Metropolis*.

Our results show that if the true ρ equals the ANPC, as in Figures 3(b) and 3(c), by fitting a DAGAR model on the latent spatial effect we are able to correctly recover both ρ and the correlation from posterior simulations. The same considerations can-

[4] Given ρ, simulate $w \sim \text{MVN}(\text{Mean} = 0, \text{Sigma} = \text{e}^{D \cdot \log(\rho)})$ where D is a scaled distance matrix; this process' average neighbor pair correlation is equal to ρ.

not be made for the CAR model. In fact, in the same settings, the CAR posterior estimates of ρ seem to systematically overestimate the true parameter value. Even when the effect is actually simulated from a CAR model, from Figure 3(a) we see that the posterior estimates for ρ do not seem to capture the true values of the parameter.



(a) Simulations from CAR (b) Simulations from DAGAR (c) Simulations from GP

Fig. 3: Posterior estimates for ρ. Each boxplot is made out of 100 posterior means.

On the other hand, other results (not shown here) suggest that both models can accurately estimate the regression parameters showing only small over/underestimation problems when dealing with the precision parameters.

4 Real data analysis

In this final section we consider an application to real data concerning the rates[5] of diabetics observed in 2014 in each of the 589 Belgian municipalities.[6]

The log-rates are shown in Figure 4(a). Here we can clearly see strong dissimilarities between two areas of Belgium that are notoriously very different: Flanders (North) and Wallonia (South). Qualitatively speaking, it seems that spatial correlation among neighbors would be very strong in both areas taken separately, but not over Belgium as a whole. Knowing the meaning of ρ in a DAGAR model (spatial correlation between two neighbors), fitting a single model on the whole area would lead to misleading estimates for this parameter. Therefore, in the following, we will work on the two areas separately.

In each area, we assume a fairly simple (no covariates) model for the log-rate $Y = \mu + w + \varepsilon$ and we fit two models, one with $w \sim$ CAR and one with $w \sim$ DAGAR, both having a Normal-distributed μ and a Unif$(0,1)$-distributed spatial correlation parameter, ρ' and ρ respectively; the symmetrical neighborhood structure is built from a sharing–a–boundary type of graph and the fictional ordering of the municipalities is based on their longitude.

[5] The rate in each municipality is computed as the ratio between the observed and the expected number of cases, marginalized over age and gender.

[6] Count data about the people who received diabetics-related medications in 2014 have been collected by the InterMutualische Agentschap (www.ima-aim.be).

In Table (1), we see some results of the model fitting procedure; estimates are obtained as MCMC posterior means.

Table 1

Area	Model	Spatial correlation	Error $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$
North	CAR	0.666	0
	DAGAR	0.989	0.00119
South	CAR	0.751	0
	DAGAR	0.987	0.00084

As it is apparent, both models produce posterior estimates for the spatial correlation that are large in both areas, confirming our original guess – keep in mind that the estimates from the DAGAR can actually be interpreted as ANPC. In reconstructing the diabetics log-rate, we see from both Table (1) and Figure (4) that the two models worked very well, the CAR actually overfitting the data.

Overall, the two models appear to be very competitive, each with its pros and cons. Both worked reasonably well on the considered dataset, producing expected and hoped–for results even without additional information provided by covariates.

(a) Observed log-rates (b) CAR-estimated (c) DAGAR-estimated

Fig. 4: Diabetics log-rates in Belgium

References

1. Datta, A., Banerjee, S., Hodges, J.S., Gao L.: Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. Bayesian Analysis. **14**(4), 1221–1244 (2019)
2. de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., Bodik, R.: Programming with models: writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics. **26**, 403–413 (2017)

A functional approach to study the relationship between dynamic covariates and survival outcomes: an application to a randomized clinical trial on osteosarcoma

Un approccio funzionale per studiare la relazione tra covariate dinamiche e sopravvivenza: un'applicazione ad uno studio clinico randomizzato sull'osteosarcoma

Marta Spreafico, Francesca Ieva and Marta Fiocco

Abstract In clinical research, associating dynamic time-varying covariates (e.g. biomarkers or drug assumption) with an event-time outcome represents a challenging task that could be tackled exploiting Functional Data Analysis (FDA). In particular, FDA techniques can be used to represent dynamic time-varying covariates in terms of functions, which can be plugged into a Cox-type regression model to investigate the effect on survival outcomes. Data from MRC BO06/EORTC 80931 randomised controlled trial for treatment of osteosarcoma were analysed. Time-varying covariates related to alkaline phosphatase levels and chemotherapy dose during treatment were considered.

Abstract *Nella ricerca clinica, associare covariate tempo-dipendenti dinamiche (biomarcatori o assunzione di farmaci) ad un evento di interesse rappresenta un problema stimolante che può essere affrontato tramite l'analisi dei dati funzionali (FDA). In particolare, tecniche di FDA possono essere utilizzate per rappresentare le covariate tempo-varianti come funzioni del tempo. Tali funzioni possono poi essere inserite in un modello di regressione di Cox per studiare l'effetto che esse hanno sulla sopravvivenza. Abbiamo analizzato i dati relativi allo studio clinico randomizzato MRC BO06/EORTC 80931 per il trattamento dell'osteosarcoma, considerando come variabili tempo-dipendenti il livello di fosfatasi alcalina nel tempo e la dose di chemioterapia assunta durante il trattamento.*

Key words: Functional Data Analysis, Survival Models, Osteosarcoma

Marta Spreafico[1,2,3] Francesca Ieva[1,3,4] Marta Fiocco[2,5,6]
[1]MOX – Department of Mathematics, Politecnico di Milano, Milan 20133, Italy
[2] Mathematical Institute, Leiden University, Leiden, The Netherlands
[3]CHRP, National Center for Healthcare Research and Pharmacoepidemiology, Milan 20126, Italy
[4]CADS, Center for Analysis Decisions and Society, Human Technopole, Milan 20157, Italy
[5] Dept. of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[6] Trial and Data Center, Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
e-mail: marta.spreafico@polimi.it francesca.ieva@polimi.it m.fiocco@math.leidenuniv.nl

1 Introduction

Osteosarcoma is a malignant bone tumour mainly affecting children and young adults. Multidisciplinary management including neoadjuvant and adjuvant chemotherapy with aggressive surgical resection [1] or intensified chemotherapy [2] has improved clinical outcomes but the overall 5-year survival remained unchanged in the last 40 years at 60–70%.

Depending on patient's health status or development of toxicity, biomarker values evolve and chemotherapy treatment is modified by delaying a course or reducing the dose intensity. Alkaline phosphatase (ALP) has been identified and reported as prognostic marker [3]: high ALP level is associated with poor overall or event-free survival and presence of metastasis [4, 5]. ALP levels have always been included in survival models as baseline time-fixed covariate. Similarly, chemotherapy is usually modelled by different allocated regimens (Intention To Treat - IIT analysis), without considering changes in drug assumptions over time. It has been shown that there is mismatch between target and achieved dose of chemotherapy [6] whose impact on patient's survival is still unclear. Therefore, the most appropriate and realistic way to look at both drug consumption and biomarkers is to model them as time-varying variables. In this way it is possible to investigate the dynamic effect of chemotherapy over time and to carry out information that may be related to the survival.

In clinical research, associating dynamic time-varying covariates with an event-time outcome related to a specific event of interest (e.g. death) represents a challenging task that could be tackled by exploiting Functional Data Analysis (FDA)[7]. In recent years FDA has been increasingly used in healthcare research, since it provides a novel modelling and prediction approach. There is a great potential for many applications in public health and biomedical fields [8]. In this work we focus on time-varying covariates related to ALP levels and chemotherapy dose during treatment in order to study how they influence patients' long-term survival.

2 Dataset

Data from MRC BO06/EORTC 80931 Randomised Controlled Trial (RCT) for treatment of osteosarcoma were analysed. Patients were randomized at baseline between Conventional (*Reg-C*) or Dose-Intense (*Reg-DI*) regimens, with identical anticipated cumulative dose but different duration. Chemotherapy was administered for six cycles C_j with $j \in \{1,...,6\}$ (a cycle is a period of either 2 or 3 weeks depending on the allocated regimen), before and after surgical removal of the primary osteosarcoma. In both the *Reg-C* and *Reg-DI* arms, doxorubicin (DOX: 75 mg/m^2) plus cisplatin (CDDP: 100 mg/m^2) were given over six cycles. Surgery to remove the primary tumour was scheduled at week 6 after starting treatment in both arms. Figure 1 shows the trial design. Laboratory tests, such as ALP test, were performed before each cycle to monitor patient's health status. Delays or chemotherapy dose

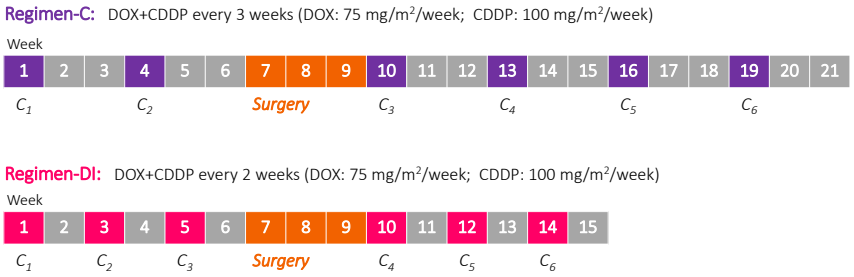


Fig. 1 Patients are randomized at baseline to one of the two regimens, with identical anticipated cumulative dose but different duration.

reductions during treatment were applied in the presence of toxicity. Details can be found in the primary analysis of the trial [2].

3 Methods

In this study, each patient is followed from date of randomization. As shown in Figure 2, the study period is divided in two intervals: a *pre-defined period*, used for the reconstruction of the time-varying covariates, and a *follow-up period*, used for the survival analysis. The *pre-defined period* starts from the date of randomization (T_R in Figure 2) and it is used to computed both longitudinal and functional representations of ALP biomarker and chemotherapy dose. The *follow-up period* is then used for the time-to-event analysis in which the event of interest is patient's death for any cause. Overall Survival (OS) is measured from the end of the *pre-defined period*, i.e. $T_0 = T_R + 180$ days, to the date of death or last visit.

The methodology applied for the analysis can be divided into the following steps:

1. select a proper cohort of patients, i.e. patients who completed the chemotherapy protocol within 180 days after randomization;
2. reconstruct the longitudinal trajectories of ALP biomarker and dose of chemotherapy;
3. use FDA techniques to reconstruct the functional data related to dynamic ALP biomarker and dose of chemotherapy in the interval $[T_R; T_0]$ starting from the longitudinal representations;
4. use Functional Principal Component Analysis (FPCA)[7] to summarise each function to a finite set of covariates, i.e. the FPC scores;
5. fit a Cox's regression model with FPC scores and baseline characteristics in order to quantify the association between time-varying processes and long-term survival [9].

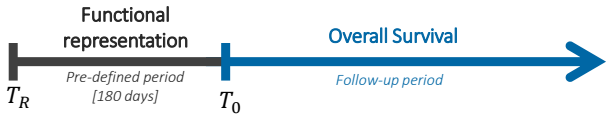


Fig. 2 Study design for a patient in the study cohort. The *pre-study period* is used to compute the functional representations of time-varying covariates. T_R is the randomization time and $T_0 = T_R + 180$ [days] is the time from which Overall Survival (OS) is computed. The *follow-up period* is used for survival analysis.

4 Application and Results

In the study cohort 376 patients that completed the chemotherapy within 180 days from randomization were included. Median age at randomization was 15 years (51.1% patients with age ≥ 15) with a percentage of male patients equal to 60.6% (228 patients). The median value of baseline White Blood Count (WBC) was $7.65 \times 10^9/L$ ($IQR = [6.30; 9.13]$) and regimen Dose-Intense was allocate in 52.3% of the patients (197). Median follow-up time, computed using the reverse Kaplan-Meier method [10], was 46.09 months ($IQR = [22.77; 76.91]$) and 296 patients (78.7%) were alive at the last follow-up visit.

The values of ALP biomarkers $y_i(t)$ were measured through laboratory tests, usually performed before each cycle of chemotherapy. For each patient i, a vector $\boldsymbol{y}_i = \{y_i(t_{il}), l = 1, ..., n_i\}$ of longitudinal values of ALP measurements was reconstructed, where t_{il} is the time of the l-th ALP test and n_i is the number of different laboratory tests for patient i. The time-varying standardized cumulative dose of drugs $\delta_i(t)$ was defined as the cumulative assumed dose of $DOX + CDDP$ up to time t divided by the total target dose over the six cycles. For each patient i and cycle j, a vector $\boldsymbol{\delta}_i = \{\delta_i(t_{ij}), j = 1, ..., 6\}$ of longitudinal values of standardized cumulative dose of chemotherapy over time was defined.

FDA techniques [7] were used to convert the longitudinal values into functional representations. In both cases B-spline basis functions were considered, and the functional data objects were expressed using a general functional form and a monotone functional form, for ALP biomarker and chemotherapy dose respectively. Data were then smoothed by (penalized) regression analysis, constraining functions to be positive and upper bounded for clinical indication. A graphical representation of functional ALP biomarker curves $\tilde{x}_i^{(ALP)}(t)$ (left panel) and of functional standardized cumulative dose curves $\tilde{x}_i^{(\delta)}(t)$ (right panel) are shown in Figure 3, in which each curve represents the functional predictor of a patient. Exploiting FDA to obtain a functional representation of the time-varying processes of interest is an effective exploratory technique to highlight trends and variations in the shape of processes over time.

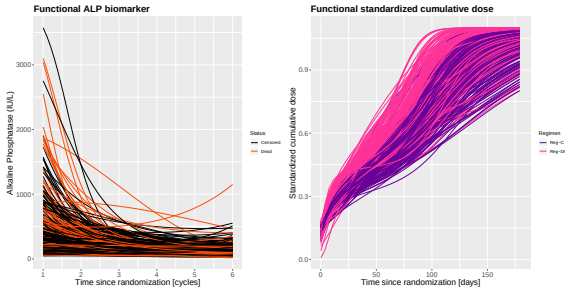


Fig. 3 Left panel: functional representations of ALP biomarker over cycles coloured by event status (black: *Censored*, red: *Dead*). Right panel: functional representations of standardized cumulative dose of chemotherapy over time coloured by allocated regimen (pink: *Reg-DI*, purple: *Reg-C*).

Once computed the functional predictors $\tilde{x}_i^{(ALP)}(t)$ and $\tilde{x}_i^{(\delta)}(t)$, to summarize each function into a finite set of covariates two Functional Principal Component Analyses (FPCAs) were performed. In both cases, we found that two principal components were enough to consider at least 95% of the explained variance, hence the finite set of covariates related to FPC scores was $\left\{f_{i1}^{(ALP)}, f_{i2}^{(ALP)}, f_{i1}^{(\delta)}, f_{i2}^{(\delta)}\right\}$.

To assess the role of available functional covariates with respect to the overall survival time of a patient, a multivariate functional Cox regression model adjusting for baseline characteristics $\boldsymbol{\omega}_i = (WBC_i, gender_i, age_i)$ was estimated:

$$h_i\left(t|\boldsymbol{\omega}_i, \tilde{x}_i^{(\delta)}(t), \tilde{x}_i^{(ALP)}(t)\right) = h_0(t)\exp\left\{\gamma_1 WBC_i + \gamma_2 gender_i + \gamma_3 age_i + \alpha_1^{(\delta)} f_{i1}^{(\delta)} + \alpha_2^{(\delta)} f_{i2}^{(\delta)} + \alpha_1^{(ALP)} f_{i1}^{(ALP)} + \alpha_2^{(ALP)} f_{i2}^{(ALP)}\right\}. \quad (1)$$

Gender, level of WBC at randomization and the score related to the first PC of alkaline phosphatase $f_{i1}^{(ALP)}$ were statistically significant at confidence level $\alpha = 5\%$. In particular, males experienced the event of interest 1.7-times faster than females. The higher the value of WBC at baseline, the lower the survival. The higher the value of the score related to the first PC of ALP biomarker $f_{i1}^{(ALP)}$, the lower the survival. This suggested that patients with higher trajectories ALP levels with respect to the mean curve had poor survival probability. FPC scores related to functional chemotherapy dose were not statistically significant. This suggest that even more informative representation of dose-intense profiles, do not show a beneficial effect on survival.

5 Conclusion

The novelty of this approach can take into account important information about the dynamic behaviours of the generating processes that underpin the data, represent-

ing interpretative and forecasting tools in osteosarcoma research. High ALP levels during time reflected poor overall survival and dose-intense profiles were not associated with good survival. However, functional representations were able to capture the individual realisations of the intended treatment, suggesting that other peculiar aspects of chemotherapy, such as latent accumulation of toxicity [11], should be taken into account.

The complexity of the phenomenon asks for the developments of new methodologies. This study showed that working in this direction is a difficult but profitable approach, which could lead to new improvements for subject-specific predictions and personalised treatment.

Acknowledgements The authors thank Medical Research Council for sharing the dataset used in this work and Prof.dr. Hans Gelderblom (Department of Medical Oncology, Leiden University Medical Center, Leiden, The Netherlands) for the clinical suggestions.

References

1. Ritter, J., Bielack, S.S.: Osteosarcoma. In: Ann. Oncol. **21**(suppl 7), vii320– vii325 (2010).
2. Lewis, I.J. et al.: Improvement in Histologic Response But Not Survival in Osteosarcoma Patients Treated With Intensified Chemotherapy: A Randomized Phase III Trial of the European Osteosarcoma Intergroup. In: JNCI **99**(2), 112–128 (2007).
3. Kim, S.H. et al.: Reassessment of alkaline phosphatase as serum tumor marker with high specificity in osteosarcoma. In: Cancer Med. **6**(6), 1311–1322 (2017).
4. Hao, H. et al.: Meta-analysis of alkaline phosphatase and prognosis for osteosarcoma. In: Eur. J. Cancer Care **26**(5), e12536 (2017).
5. Ren, H.Y. et al.: Prognostic Significance of Serum Alkaline Phosphatase Level in Osteosarcoma: A Meta- Analysis of Published Data. In: BioMed Res. Int., Article ID 160835 (2015).
6. Lancia, C. et al.: Method to measure the mismatch between target and achieved received dose intensity of chemotherapy in cancer trials: a retrospective analysis of the MRC BO06 trial in osteosarcoma. In: BMJ open **9**(5) (2019).
7. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics. Springer New York, 2005.
8. Ullah, S., Finch, C.F.: Applications of functional data analysis: A systematic review. In: BMC Med. Res.Methodol. **13**(43) (2013).
9. Kong, D. et al.: FLCRM: Functional linear cox regression model. In: Biometrics **74**(1), 109–117 (2018).
10. Schemper, M. and Smith, T.L.: A note on quantifying follow-up in studies of failure time. In: Control. Clin. Trials **17**(4), 343–346 (1996).
11. Lancia, C. et al.: Marginal structural models with dose-delay joint-exposure for assessing variations to chemotherapy intensity. In: Stat. Methods Med. Res. **28**(9), 2787-–2801 (2019).

A Statistical Approach to the Alignment of fMRI Data

Un approccio statistico al problema dell'allineamento di dati fMRI

Angela Andreella, Ma Feilong, Yaroslav Halchenko, James Haxby, and Livio Finos

Abstract Multi-subject functional Magnetic Resonance Image studies are critical. The anatomical and functional structure varies across subjects, so the image alignment is necessary. We define a probabilistic model to describe functional alignment. Imposing a prior distribution, as the matrix Fisher Von Mises distribution, of the orthogonal transformation parameter, the anatomical information is embedded in the estimation of the parameters, i.e., penalizing the combination of spatially distant voxels. Real applications show an improvement in the classification and interpretability of the results comparing to various functional alignment methods.

Abstract *Gli studi di risonanza magnetica funzionale multi-soggetto sono critici. La struttura anatomica e funzionale del cervello varia tra i soggetti, l'allineamento delle immagini è indispensabile. In questo lavoro, si definisce un modello probabilistico per descrivere l'allineamento funzionale. Definendo una distribuzione a priori, come la Fisher Von Mises, per le matrici ortogonali, le informazioni anatomiche sono incorporate nel processo di stima, penalizzando la combinazione*

Angela Andreella
Department of Statistical Sciences, University of Padua, Italy,
e-mail: angela.andreella@phd.unipd.it

Feilong Ma
Center for Cognitive Neuroscience, Dartmouth College, NH, United States
e-mail: Feilong.Ma@dartmouth.edu

Yaroslav Halchenko
Center for Cognitive Neuroscience, Dartmouth College, NH, United States
e-mail: Yaroslav.O.Halchenko@Dartmouth.edu

James Haxby
Center for Cognitive Neuroscience, Dartmouth College, NH, United States
e-mail: James.V.Haxby@dartmouth.edu

Livio Finos
Department of Developmental Psychology and Socialization, University of Padua, Italy
e-mail: livio.finos@unipd.it

di voxel spazialmente distanti. Applicazioni reali mostrano un miglioramento nella classificazione e nell'interpretazione dei risultati rispetto ad altri metodi.

Key words: fMRI data; Hyperalignment; Fisher Von Mises; Generalized Procrustes Analysis

1 Introduction

Multi-subjects functional Magnetic Resonance Images (fMRI) studies permit to compare analysis across subjects. However, the anatomical and functional structure of brains varies across subjects, even in response to identical sensory inputs. The images must be aligned to a common spatial space. The alignment process could use anatomical or functional features of the brains or both. [10] proposed a spatial anatomical normalization, but it doesn't account for idiosyncratic functional topography. For that, [8] suggested a functional alignment, called hyperalignment, using sequential Procrustes transformations. The activations of different subjects are mapped into a common high dimensional model, representing a linear combination of voxel activations.

After [8], various modifications of hyperalignment appeared; however, these methods analyze only the huge computational effort without emphasizing the final interpretation of the transformed high dimensional space.

The method proposed shall implement an approach that permits to restrict the range of possible transformations used to map the neural response into a common high dimensional space. The restriction is based on the idea that the anatomical information could give a direction of the most plausible transformation, that match as best as possible the subject's data. We implement an alignment technique that exploits the anatomical features in addition to the correlation among voxels' time series. The objective of the Procrustes problem is to find an orthogonal transformation that minimizes the pairwise differences between two or more matrices, i.e., images. This minimization can be considered as a least-squares problem, allowing a definition of a statistical model describing the data generating process. Therefore, the set of transformation solutions can be shrinkage using a prior distribution for the orthogonal matrix transformation. The use of priors based on neurological information permits to improve the interpretability of the transformation matrix founded, rather than the hyperalignment method that doesn't impose any constraint to the orthogonal transformation solution. The prior distribution proposed is the Matrix Fisher Von Mises distribution. The specification of the location matrix parameter permits to penalizes spatially distant voxels in the Procrustes transformation. This penalization is expressed in terms of the idea that the rotation loadings of contiguous voxels are similar, and the rotation loadings of distant voxels are probably less similar. So, the location matrix parameter is defined as a similarity matrix derived from the Euclidean distance of the three-dimensional brain voxels coordinates. The

proposed method considers both anatomical and functional data features, in contrast to the hyperalignment that is anatomy free.

In the following, in Section 2 is introduced the Procrustes problem and the main technical contribution. Section 3 outlines the performance evaluation of the method proposed in the multi-subject classification framework.

2 Methodology

In this Section, we briefly revisit some results of [9] and [5] to set the stage and introduce notation.

2.1 Generalized Procrustes Analysis

The data are represented in a matrix $X_i \in \mathbb{R}^{t \times v}$, $i = 1, \ldots, m$, one for each subject. The t rows of X_i represent the response activation of v voxels during a stimulus, and the v columns represent the time series of activation for each v voxel. The matrices are ordered consistently across all subjects considering the rows because the stimuli are time-synchronized; this does not apply to the columns. For that, the alignment step is essential to perform inter-subject analysis.

Procrustes methods are useful for assessing the distance between matrices. When X_1 is transformed into the space of X_2 by an orthogonal transformation R, the Procrustes problem is called *Orthogonal Procrustes Problem* (OPP), i.e.:

$$\min_R ||X_1 - X_2 R||_F^2 \quad \text{subject to} \quad R^T R = R R^T = I_v \,. \tag{1}$$

A first solution of Eq. 1 was proposed by [9], the minimum is given by $\hat{R} = UV^T$, where U and V are derived from the singular value decomposition of the matrix $T = X_2^T X_1 = U \Sigma V$.

Analysing $m \geq 2$ subjects, the *Generalized Procrustes Analysis* (GPA), defined by [4], is considered, i.e.:

$$m \min_{R_i} \sum_{i=1}^{m} ||X_i R_i - M||_F^2 \quad \text{subject to} \quad R_i^T R_i = R_i R_i^T = I_v \qquad \forall i = 1, \ldots, m \,. \tag{2}$$

where $M = m^{-1} \sum_{i=1}^{m} X_i R_i$, called configuration average matrix. Eq. 2 hasn't a closed solution, [5] proposed an iterative procedure to estimate R_i .

The hyperalignment method proposed by [8] is a sequential approach of OPP (Eq. 1); therefore, the results depend on the order of the subjects. Also, It doesn't reach the global minimum imposed by the GPA (Eq. 2).

The next Section outlines the approach proposed, which is an extension of GPA (2), our estimation of R_i is guided by prior anatomical information.

2.2 Procrustes problem using prior information

The OPP is based on the least-squares criterion to find the optimal transformation. However, the solution founded by the GPA is not unique. Every random rotation of the founded solution is still a valid solution, i.e. the set of solutions equals $\left\{\hat{R}S : \forall S \in \mathcal{O}(m)\right\}$.

For that, we use prior information about the structure of R_i that leads the optimization step. The Procrustes problem is then analyzed from a likelihood perspective as proposed by [3], defining a statistical model called the perturbation model.

The minimization problem defined in Eq. 2 can be formulated as

$$X_i = MR_i^\top + E_i \qquad \text{subject to} \quad R_i^\top R_i = R_i R_i^\top = I_v \qquad \forall i = 1, \ldots, m \tag{3}$$

where E_i is the matrix with dimension $t \times v$ of the residual terms, and M is the configuration average matrix. E_i follows the multivariate normal matrix distribution [7], where each rows of $E_i \sim N(0, \sigma^2 I_v)$. The $M \in \mathbb{R}^{\mathrm{t}\times\mathrm{v}}$ matrix is the configuration average matrix. The X_i is then described as a randomly Gaussian perturbation of M. The R_i permits to represent each X_i in the shared space defined by M .

Therefore, we define a prior distribution for R_i, that takes values in the Stiefel Manifold $V_v(R^v)$. [1] proposed the matrix Fisher Von Mises distribution, lately used by [6]. It is described as:

$$f(R_i) \sim C(F,k)\exp(\mathrm{tr}(kF^T R_i)) \tag{4}$$

where $C(F,k)$ is a normalizing constant, $F \in \mathbb{R}^{\mathrm{v}\times\mathrm{v}}$ is the location matrix parameter and $k \in \mathbb{R}_{>0}$ the concentration parameter. The specification of the matrix F is the focal point. We think that closer voxels have similar rotation loadings while voxels very far each other should have less probably similar loadings. For that, the location matrix parameter F is constructed as a Euclidean similarity matrix using the three-dimensional coordinates of the voxels.

Without a prior assumption, [2] found that the least-squares solution to Eq. 3 is the Procrustes solution founded by [9]. Having the Fisher Von Mises prior distribution, we found that the solution proposed is a slight modification of the solution founded by [9] in the case of one subject, and a slight modification of the solution given by [4] in the case of multiple matrices. The modification is focused on considering the singular values decomposition of $X_i^\top M + kF$ instead of $X_i^\top M$.

Being a minor modification of the Generalized Procrustes [4], the method proposed can find the global minimum, unlike the hyperalignment method. Also, the hyperalignment and GPA methods are completely anatomy free. In contrast, the method proposed preserves the anatomical meaning of the space behaviour of the voxel in the brain or the region of interest. Also, if the hyperalignment method converges to the GPA, it is immediate to prove that the solution to the minimization is not unique. If not, the results still depend on the order of the subjects, leading to a replicability problem. So concluding, in the method proposed, the transformations are unique and also follow an anatomical structure.

3 Results

The protocol to evaluate the performance of the method follows the one used in [8]. Because the results of hyperalignment depend on the order of subjects, we permuted it 100 times and used the distribution of classification accuracy. Figure 1 shows that the accuracy computed by the method proposed is higher than the median result given by the hyperalignment in both analyses. Also, the new method generates unique transformations, simplifying the interpretation of the results.

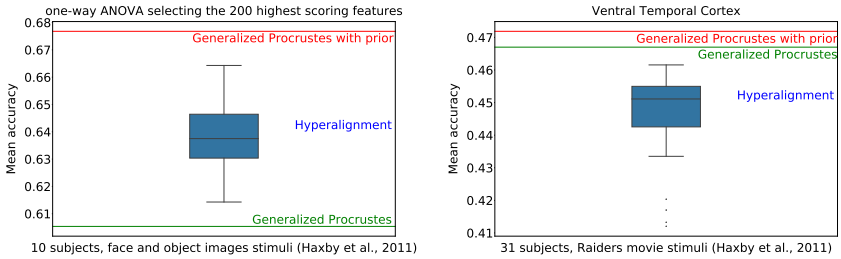


Fig. 1 Box-plots showing the classification mean accuracy permuting 100 times the order of the subjects, analyzing the Faces and Objects dataset by the Multi-class Linear Support Vector classifier (left) and the Raiders dataset by the one nearest neighbor classifier (right)

Figure 2 upper panel represents one of the classifier's coefficients used in the object recognition analysis with data aligned by the method proposed, while the lower panel by hyperalignment. The spatial regularization, imposed on the estimation of the orthogonal transformation, leads to a smoother map of classifier coefficients. It can capture the spatial information of the voxels and choose between the infinite possible solutions, the orthogonal transformation more reasonable. Also, the pairwise distance (measured as the norm of difference matrices) between subjects is lower, considering data aligned by the technique proposed (36.28) than data aligned by the hyperalignment method (444479.13). Thus, the brain images after hyperalignment result to be more different between each other.

4 Conclusions

Rephrasing the Procrustes problem as a statistical model, we can shrink the set of transformation solutions using a prior distribution, i.e., the Matrix Fisher Von Mises distribution, for the orthogonal matrix parameter. This assumption allows embedding anatomical information in the estimation of parameters, by penalizing the similarity of spatially distant voxels when we estimate the common reference space.

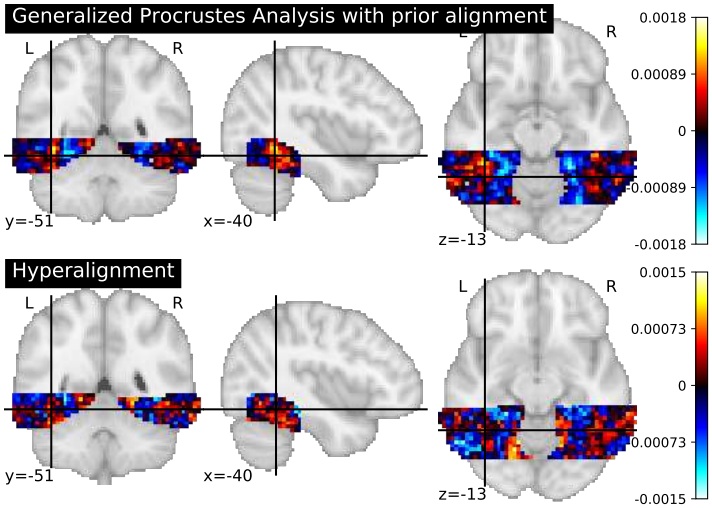


Fig. 2 Coefficients of the Multi-class Support Vector Machine considering the monkey face versus male face classifier analyzing the Faces and Objects data aligned by the Generalized Procrustes Analysis with prior information and by the hyperaligment method.

The method returns a unique solution of the rotation matrices, having anatomical meaning, so more interpretable. Using real data, we found that the alignment proposed improves the classification analysis between-subjects in comparison to hyperalignment. It also permits reaching the global minimum imposed by the GPA, and it doesn't depend on the order of the subjects, giving smaller between-subject distance.

References

1. Downs, D. T.: Orientation Statistics. Biometrika. **59**, (3), 665–676 (1972)
2. Dryden, I.L., Mardia, K.V.: Statistical shape analysis. Wiley, Chichester (1998)
3. Goodall, C.: Procrustes Methods in the Statistical Analysis of Shape. J. R. Statist. Soc. B. **53**, (2), 285–339 (1991)
4. Gower, J.C.: Generalized procrustes analysis. Psychometrika. **40**, (1), 33–51 (1975)
5. Gower, J.C., Dijksterhuis, B.G.: Procrustes Problems. Oxford University Press, (2004)
6. Green, P.J., Mardia, K.V.: Bayesian alignment using Hierarchical Models with Applications in Protein Bioinformatics. Biometrika. **93**, (2), 235–254 (2006)
7. Gupta, A.K., Nagar, D.K.: Matrix Variate Distributions. CRC Press (1999)
8. Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., Ramadge, P.: A common high-dimensional model of the representational space in human ventral temporal cortex. Neuron. **72**, (1), 404–416 (2011)
9. Schonemann, P.H.: A generalized solution of the orthogonal procrustes problem. Psychometrika. **31**, (1), 1–10 (1966)
10. Talairach, J., Tournoux, P.: Co-Planar Stereotaxic Atlas of the Human Brain.3-Dimensional Proportional System: An Approach to Cerebral Imaging. G. Thieme, 1988

Adaptive clinical trials: Bayesian decision-theoretic and frequentist approaches for cost-effectiveness analysis

Prove cliniche adattive: una comparazione tra approcci Bayesiani e frequentisti per l'analisi costi-efficacia

Martin Forster and Marco Novelli

Abstract We consider some recently published Bayesian and frequentist models of sequential experimentation and their use for conducting a cost-effectiveness analysis of health technologies alongside a clinical trial. Simulations compare a range of operating characteristics and are used to propose an agenda for future research.
Abstract In questo lavoro vengono comparati alcuni modelli bayesiani e frequentisti per la sperimentazione clinica sequenziale finalizzata all'analisi costi-efficacia delle tecnologie sanitarie. Attraverso simulazioni vengono approfondite le caratteristiche operative dei due approcci al fine di evidenziarne pregi e difetti e delineare alcune possibili sviluppi per la ricerca futura.

Key words: Clinical trials, Health economics, Bayesian inference, Response-Adaptive Randomization, Sequential sampling

1 Introduction

There is growing interest in the use of adaptive clinical trials to assess the effectiveness of health technologies [3]. For example, in the frequentist literature, [14] combine sequential monitoring with response-adaptive randomisation in a design which offers potential for increased power and earlier stopping, given a predefined type I error rate. [12] argue that the Bayesian approach provides a formal way to monitor accumulating data and decide whether to stop a trial. [10] and [5] present Bayesian decision-theoretic approaches to stopping a two-armed sequential trial.

Martin Forster
Department of Economics and Related Studies, University of York, Heslington, York YO10 5DD,
e-mail: mf8@york.ac.uk
Marco Novelli
Department of Statistical Science, University of Bologna, Bologna,
e-mail: marco.novelli4@unibo.it

We consider some recent frequentist and Bayesian approaches using simulations based on the case study presented by [10], who considered a two-armed sequential clinical trial evaluating the cost-effectiveness of a new health technology relative to a known standard when there exists zero, or negligible, delay in observing the patient-level outcomes and treatment costs. Reflecting growing interest in the literature [6], our approach focuses on the cost-effectiveness of the technologies under consideration, rather than their effectiveness alone. The operating characteristics which we assess explicitly account for benefits which accrue to the population to benefit from the adoption decision, the proportion of patients allocated to each treatment during the trial, and the research costs of the trial itself.

2 Bayesian and frequentist designs

The designs address a common problem similar to those of [10] and [5]. A cost-effectiveness analysis of two health technologies, N ('new') and S ('standard', the incumbent technology), is carried out alongside a clinical trial in which patients are allocated sequentially and outcomes on effectiveness and treatment cost are observed without delay. The outcome of interest is the difference between the net monetary benefit (NMB) of technology N and technology S. NMB is defined as $Y_i = \lambda E_i - C_i$, $i \in \{N,S\}$, where E is a random variable denoting effectiveness and C denotes the known treatment cost. λ is the monetary valuation of one unit of effectiveness, a parameter required to convert health outcome data (such as a measure of quality of life) into monetary units.[1] Incremental net monetary benefit (INMB) is the difference between the NMB of N and S, that is, $X = Y_N - Y_S$. It is assumed that X has a normal distribution in the population, with unknown expected value μ_X, and variance σ_X^2, assumed known. The trial may recruit a maximum of M patients. Sequential monitoring of the estimate of INMB takes place. Interest lies in deciding when to stop the trial to choose the better of the two technologies on cost-effectiveness grounds. We consider two approaches.

1. Bayesian decision-theoretic designs of [10] (and of [5], when the delay in observing outcomes is set equal to zero): These methods build on the work of [4] and previous, related, works. Define W as a random variable representing beliefs about the unknown μ_X. We assume that it has a normal prior distribution, with expected value μ_0 and variance σ_0^2. Following, e.g., [4], we assume that pairs of patients are randomised sequentially to treatments N and S and that the resulting data are generated by a normal likelihood function. Each pairwise allocation costs $c > 0$ in research monies and the technology adoption decision benefits $P > 0$ patients. We consider the case of no discounting of benefits and costs. Given that the prior distribution is conjugate to the normal likelihood, Bayes' rule may be used to update the prior distribution sequentially and the posterior mean is used to estimate the unknown μ_X. The problem is to choose the sample size at which to stop recruitment to

[1] For example, in the United Kingdom's National Health Service, λ may be chosen to be equal to £20,000 for one Quality Adjusted Life Year [8].

the trial, $0 < T \leq M/2$, and pick the superior technology for treating the P patients. If the posterior mean is greater than zero, it is optimal to pick N, for incremental net benefit equal to $P\mu_X$; if it is less than or equal to zero, it is optimal to pick S, for incremental reward equal to 0. T maximises:

$$V(\mu_0, t_0) = \mathbb{E}\left[\sum_{i=1}^{T} -c + \max_{d_T \in \{S,N\}} \{0, PW\} \middle| \mu_0, t_0\right], \tag{1}$$

where d_T is the technology adoption decision at stopping and $t_0 = \sigma_X^2/\sigma_0^2$. The choice of a conjugate normal prior permits standard expressions to be obtained for the posterior mean and variance. The method of [10] may be applied to obtain an optimal policy for whether to continue the trial or to stop as a function of the posterior mean and the number of pairwise allocations. An approximation to the policy is obtained by recasting the model in continuous time and solving what is termed a 'free boundary problem', using the methods of [4].

2. Frequentist designs: In contrast to the Bayesian approach, which assumes pairwise randomisation to each arm of the trial, the frequentist models monitor the proportions randomised to N and S as the trial progresses, as well as the accumulating evidence on outcomes and costs. Adaptive randomization seeks to achieve a target allocation to N and S which is either a function of the unknown model parameters or the treatment costs. Choice of which technology to select is based on a test of the null hypothesis of equality of treatment cost-effectiveness, combined with the [11] and [9] boundaries to control the type-I error probability. We consider both response-adaptive (RA) and allocation-adaptive (AA) randomisation rules. As a RA design, we adopt the Doubly Adaptive Biased Coin Design (DBCD) [14] with randomization parameter 2, combined with the Normal target suggested in [1]. Here, the target is updated as the evidence from the trial accumulates, by combining the estimate of incremental net monetary benefit with a tuning parameter which manages the degree of imbalance between the two arms. For the AA rule, we take into account the completely randomized design (CRD) equipped with either a balanced allocation, or an optimal target allocation aimed to minimize the total study cost, subject to a minimal required level of statistical power [13].

3 Methods

We use as an application the case study presented by [10, section 4], who considered the use of azithromycin extended release (N) versus amoxicillin and clavulanate potassium (S) for treatment of acute bacterial sinusitis. This is suited to our models because follow-up of the cost-effectiveness data was almost immediate (average time to symptom resolution was about one week) and the treatment cost data were known. Parameter values are based on those reported in [10, Table 1], adapted slightly to illustrate ideas (see Table 1). The main change is to increase the research cost, c, ten-fold, building on the results from [7].

Parameter	Definition	Value
Common to both designs		
C_S	Cost of one course of treatment with AC	\$31.99
C_N	Cost of one course of treatment with AZ-ER	\$55.68
$\bar{E_S}$	Mean days to symptom resolution for AC	8.12
$\bar{E_N}$	Mean days to symptom resolution for AZ-ER	7.55
λ	Estimate of willingness to pay for a symptom free day	\$73.2
σ_X†	Standard deviation of individual NIMB	\$300.00
N	Maximum sample size (number of patients)	4000
Bayesian decision-theoretic design only		
c†	Marginal cost of sampling	\$1000
σ_0†	Standard deviation of prior distribution concerning μ_X	\$300.00
P	Number of patients to be treated	1,000,000

Table 1 Parameter values and sources used for the simulations (based on [10]; †adjusted).

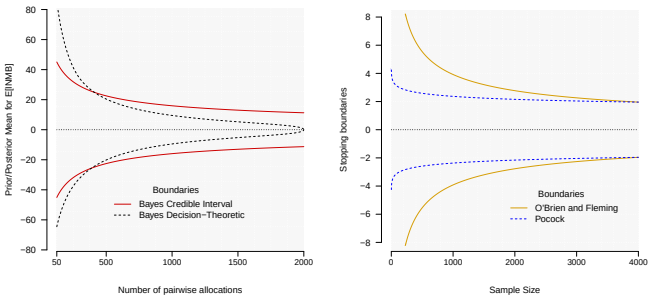


Fig. 1 Bayesian and frequentist stopping boundaries

To compare with existing Bayesian sequential models, we add a Bayesian design based on a 'credible interval' approach. This stops the trial when the probability that one of the two technologies is superior reaches 5% (see e.g. [12]). Matlab code [10] and R code is used. Monte Carlo simulation is used to obtain 10,000 paths for the posterior mean (Bayesian models) and the Wald test statistic, with $\alpha = 5\%$, (frequentist models), assuming that $\mu_X = \$18.2$ (obtainable from Table 1 data as follows: $\mu_X = -\$73.2 \times (7.55 - 8.12) - (55.68 - 31.99) \approx \$18.2$). For each design, the appropriate stopping boundary is used to stop each path and the new technology adopted, for reward equal to $P \times \$18.2$, or not adopted, for reward equal to zero.

4 Results and discussion

Results are dependent on the parameter values selected and should not be considered to hold in general. In particular, there is likely to be uncertainty around the willingness to pay parameter, λ, as well as σ_X^2. In addition, note that we focus on

	Frequentist		Bayesian	
	Pocock	O'Brien and Fleming	Credible interval	Decision-theoretic
Average sample size to first crossing (total number of patients (SD))				
DBCD	2052 (1612)	3314 (832)	-	-
CRD	1804 (1439)	2939 (888)	-	-
CRD 1:1	1510 (1362)	2736 (892)	-	-
1:1	-	-	1120 (1114)	1090 (648)
Proportion of correct decisions				
DBCD	0.69	0.54	-	-
CRD	0.82	0.75	-	-
CRD 1:1	0.88	0.83	-	-
1:1	-	-	0.91	0.98
Average incremental net benefit to patients				
DBCD	1.25×10^7	9.88×10^6	-	-
CRD	1.50×10^7	1.37×10^7	-	-
CRD 1:1	1.61×10^7	1.51×10^7	-	-
1:1	-	-	1.66×10^7	1.78×10^7
Average cost of trial				
DBCD	1.03×10^6	1.66×10^6	-	-
CRD	9.02×10^5	1.47×10^6	-	-
CRD 1:1	7.55×10^5	1.37×10^6	-	-
1:1	-	-	5.60×10^5	5.45×10^5
Average net benefit of trial				
DBCD	1.15×10^7	8.22×10^6	-	-
CRD	1.41×10^7	1.22×10^7	-	-
CRD 1:1	1.53×10^7	1.37×10^7	-	-
1:1	-	-	1.60×10^7	1.72×10^7

Table 2 Operating characteristics of the designs. *Average sample size to first crossing*: sample size at which boundary is first crossed; *Proportion of correct decisions*: proportion of times the new technology is selected; *Average incremental net benefit to patients*: $\delta_{\mu>0} \times P \times \mu_X + (1 - \delta_{\mu_X>0}) \times P \times 0$, where $\mu_X = 18.2$ δ_F is an indicator function equal to one if F is true; *Average cost of trial*: sample size at first crossing / 2 multiplied by $c = \$1000$; *Average net incremental benefit of trial* = Average incremental net benefit to patients minus Average cost of trial.

drawing from a population distribution with a fixed mean. An extension to a fully Bayesian analysis is left for future research.

The stopping boundaries are shown in Figure 1 and the operating characteristics in Table 2. A direct comparison of the Bayesian and frequentist stopping boundaries in Figure 1 is not possible because the scales are not common. Table 2 shows that the frequentist designs report higher expected sample sizes to stopping and lower proportions of correct decisions than do both Bayesian designs. This implies that Bayesian designs exhibit both lower trial costs and higher average IMB to patients, which in turn leads to a higher overall expected net benefit of the trial. Within the frequentist designs, the CRD outperform the DBCD with the CRD 1:1 being the best performing and the CRD close behind. This is mostly due to the fact that the balanced allocation maximises the power of the Wald test. As expected, the OBF rules are more conservative than the Pocock. Comparison of the standard deviations for sample sizes relative to their respective averages suggests that the Bayesian de-

signs are stopping some paths very early, a phenomenon that has also been observed in [7], who raised the concern that low sample sizes may hinder adoption of the preferred technology. What has not been accounted for in these rules, however, are the benefits accruing to trial participants, something that the DBCD and CRD designs seek to address through their choice of target allocation, but which the CRD 1:1 and both Bayesian designs do not.

Future research should explore more the commonalities and differences between the approaches, noting that, differently from the frequentist designs, the Bayesian ones currently do not implement allocation-adaptive and response-adaptive randomisation, although they do account for the size of the post-trial population to benefit. Further, although the frequentist designs' hypothesis tests include treatment costs, research costs are not currently considered. Research should also investigate further the variability of stopping times, a fully Bayesian approach to the simulation and, finally, credibility for deployment of the methods in health care systems.

References

1. Atkinson, A. C., Biswas A.: Bayesian adaptive biasedcoin designs for clinical trials with normal responses. *Biometrics* **61**, 118-125 (2005)
2. Baldi Antognini, A., Giovagnoli, A.: Adaptive designs for sequential treatment allocation. Chapman and Hall/CRC (2015)
3. Berry, D. A.: Interim analyses in clinical trials: classical vs. Bayesian approaches. *Statistics in medicine* **4**, 521-526(1985)
4. Chernoff, H., Petkau J.: Sequential medical trials involving paired data. *Biometrika 68*, 119-132 (1981)
5. Chick, S., Forster M., Pertile P.: A Bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society: Series B*, **79**, 1439-1462 (2017)
6. Flight, L., Arshad, F., Barnsley, R., Patel, K., Julious, S., Brennan, A., Todd, S. A review of clinical trials with an adaptive design and health economic analysis. *Value in Health*, **22**, 391-398 (2019)
7. Forster, M., Brealey, S., Chick, S., Keding, A., Corbacho, B., Alban, A., Rangan, A. Cost-effective clinical trial design: application of a Bayesian sequential stopping rule to the ProFHER Pragmatic Trial. 1/19, Department of Economics, University of York (2019)
8. National Institute for Health and Care Excellence: Guide to the Methods of Technology Appraisal. National Institute for Health and Care Excellence, London, 2013
9. O'Brien P. C., Fleming T. R.: A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556 (1979)
10. Pertile P., Forster M., La Torre D.: Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies, *Journal of the Royal Statistical Society, Series A*, **177**, 419–438 (2014)
11. Pocock S. J., Hughes M. D.: Practical problems in interim analyses, with particular regard to estimation, *Controlled Clinical Trials*, **10**, 209S–221S (1989)
12. Spiegelhalter D. J., Freedman L.S., Parmer M.K.B.: Bayesian approaches to randomised trials. *Journal of the Royal Statistical Society, Series A*, **157**, 357–416 (1994)
13. Sverdlov O., Ryeznik Y.: Implementing unequal randomization in clinical trials with heterogeneous treatment costs, *Statistics in Medicine*, **38**, 2905–2927 (2019)
14. Zhu H., Hu F.: Sequential monitoring of response-adaptive randomized clinical trials. *The Annals of Statistics*, **38**, 2218–2241 (2010)

Bootstrap corrected Propensity Score: Application for Anticoagulant Therapy in Haemodialysis Patients

Propensity Score aggiustato via Bootstrap: un'applicazione a dati di terapia anticoagulante su pazienti in emodialisi

Maeregu W. Arisido, Fulvia Mecatti and Paola Rebora

Abstract The inverse propensity score weighting (IPSW) has been often used to estimate causal effects of treatments for observational data. However, IPSW requires strong assumptions, in which their misspecifications may severely bias estimated treatment effect. We present a bootstrap based bias-correction to adjust the propensity score weights in case of misspecifications of one of the main assumption. We showed, using simulation, the approach performs well in correcting biases due to model misspecifications in various contexts.The method was also illustrated using a real data based on end-stage renal disease.

Abstract *Il metodo del Propensoty Score è ben noto e largamente impegato per la stima di effetti causali su dati osservazionali. Tuttavia, l'efficacia del metodo è condizionata da assunzioni stringenti che, qualora non verificate in pratica, possono condurre a severe distorsioni nella stima dell'effetto del trattamento. In questo lavoro presentiamo una soluzione bootstrap per "aggiustare" il metodo del propensity score nel caso di mancata specificazione di una delle principali assunzioni. L'efficacia di tale "aggiustamento" bootstrap è illustrata mediante un esercizio di simulazione che utilizza dati reali di pazienti all'ultimo stadio della malattia renale.*

Key words: bias-correction, bootstrap, Propensity score, simulation.

Maeregu W. Arisido
University of Milan-Bicocca, Department of Sociology and Social research, Via Bicocca degli Arcimboldi, 8 - 20126 Milan, e-mail: maeregu.arisido@unimib.it

Fulvia Mecatti
University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 Milan, e-mail: fulvia.mecatti@unimib.it

Paola Rebora
University of Milan-Bicocca, Via Cadore, 48 20900 Monza, e-mail: paola.rebora@unimib.it

1 Introduction

When the goal is to establish the effectivness of a treatment, randomized controlled trials (RCTs) are the gold standard [2]. Due to various reasons, observational studies are used. The present work is inspired by an observational study where the aim was to explore the use of oral anticoagulant treatment (OAT) in reducing mortality due to atrial fibrillation in patients with end-stage renal disease (ESRD) [4]. A defect of such studies is that treatment allocation is influenced by baseline characteristics, which results in a bias affecting the estimated treatment effect. IPSW has been increasingly used to deal with such bias [8]. IPSW uses weights based on the propensity score to balance baseline covariates between the treated and control groups [6]. A key issue of IPSW is that strong assumptions are required to successfully balance the covariates and avoid the bias from the estimated treatment effect. The assumption of ignorability, which implies that there is no unmeasured covariates that affect both the outcome and treatment simultaneously, is a fundamental aspect of IPSW [8, 6]. Violation of this assumption deters the performance of IPSW.

We present a bootstrap-corrected IPSW (BC-IPSW) to address bias of estimated treatment effect when the propensity score weights are used for the estimation of treatment effect. The bootstrap has been commonly used for the estimation of standard errors of estimated treatment effect [e.g., 7]. However, the use of bootstrap together with the propensity score to avoid or minimize bias of the average treatment effect has been particularily rare. Recently, a simulation study by [5] showed that the bootstrap can be beneficial to reduce bias from the maximum likelihood estimator of the average treatment effect on a Gaussian outcome. The BC-IPSW firstly estimates a treatment effect using IPSW, based on multivariable logistic regression model. We then apply a bootstrap bias correction on the estimated treatment effect. The goal in this paper is twofold: (a) to evaluate whether OAT could lead to a reduction of mortality for the ESRD data, (b) to conduct simulations to examine the impact of IPSW misspecifications and assess to what extent the BC-IPSW reduces bias of the misspecification in various contexts. Section 2 below discusses the ESRD application data, section 3 presents the methodologies and simulation protocols. We present the results in section 5 and section 6 provides concluding discussion.

2 Motivating Real Data Application

The motivating application is based on determining the effectivness of Oral anticoagulant treatment (OAT) in reducing mortality due to atrial fibrillation in patients with end-stage renal disease (ESRD). OAT is the main treatment in ESRD, although presently evidences emerge that OAT leads to the risk of bleeding [4]. The data is coming from a prospective cohort study of 290 patients with AF and received OAT therapy in ten Italian haemodialysis centers. The patients were followed-up for 4 years from 31 October 2010 to 31 October 2014. At recruitment, 134 patients (46.2%) were taking OAT, in which $72(53\%)$ died. Among the 156 non treated patients, $98(63\%)$ died. There were 9 covariates measured at the begining of the four-

year study: age (≤ 65), gender, type of atrial fibrillation (AF), bleeding, antiplatelet therapy, hypertension, diabetes mellitus, ischaemic stroke, heart failure.

3 Methods

In this section, we outline the methods that permits to address the two aims introduced in the introduction. Let T_i^* be the failure time of subject i $(i = 1,\ldots,n)$ in a cohort of size $n = 290$. We are interested to estimate the true effect of OAT on mortality, denoted by β. Because of censoring, the observed survival time is defined as $T_i = \min(T_i^*, C_i)$, where C_i is the right censoring time and $\delta_i = I(T_i^* \leq C_i)$ is the event indicator, indicating whether the survival or the censoring time is observed.

3.1 The Naive Average Treatment Effect Estimate

To investigate how the problem of treatment bias impacts the estimate of the average OAT effect on mortality for the ESRD data, we first perform a *naive* analysis which ignores the risk of bias due to confounding between the treatment status and covariates using the Cox proportional hazards model [3] in the form

$$h_i(t) = h_0(t)\exp(\beta_o z_i + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_9 x_9) \tag{1}$$

where $h_0(t)$ denotes the unspecified baseline hazard, β_o is the log hazard measuring the association between the treatment z_i and the hazard $h_i(t)$, $(x_1, x_2 \ldots, x_9)$ denotes the vector of 9 baseline covariates and $(\alpha_1, \alpha_2, \ldots, \alpha_9)$ denotes their corresponding regression coefficients. The hazard ratio HR $= \exp(\beta_o)$ is interpreted as the relative change in the hazard at any time t for a subject who received OAT relative to a subject in the control group in the presence of confounding covariates.

3.2 Inverse Propensity Score Weighted Treatment Effect

IPSW allows to estimate weights based on propensity score to match the probability distributions of baseline covariates between the treated and control patients. Formally, let $p(x|z = 1)$ be the distribution of a covariate x for the treated patients and $p(x|z = 0)$ be the distribution of x for the control patients. Under RCT, we expect these two distributions be similar. When there is a lack of randomization, IPSW is used to estimate a weight $\omega(x)$ so that $p(x|z = 1) = \omega(x)p(x|z = 0)$. The weight $\omega(x)$ is unknown, and can be estimated by the logistic regression by postulating z_i as the probability of treatment assignment conditional on baseline covariates $x_1 - x_9$. We then obtain the maximum likelihood estimates of $\left(\alpha_1, \ldots, \alpha_9\right)$ and the weight ω is the predicted probability given as

$$\hat{\omega} = \frac{\exp(\hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \cdots + \hat{\alpha}_9 x_9)}{1 + \exp(\hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \cdots + \hat{\alpha}_9 x_9)} \tag{2}$$

Then, IPSW adjusted average treatment effect is estimated by $h(t) = h_0(t)\exp(\beta_p z_i)$, where the estimate of β_p represente IPSW adjusted average treatment effect on

mortality and obtained by maximizing the weighted partial likelihood. The estimate of β_p may still be biased due to the misspecification of the ignorability assumption.

3.3 Bias-Correction of the Inverse Propensity Score Weights

To reduce the bias of $\hat{\beta}_p$ due to the misspecification of the IPSW, we now focus on bootstrap bias correction of the IPSW (BC-IPSW). A classic nonparametric iid bootstrap is applied to the original ESRD data. The total number of bootstrap samples was set $B = 2000$ such that the size of each bootstrap sample is equal to that of the ESRD data ($n = 290$). In each bootstrap sample, the treatment effect $\hat{\beta}_p^*$ adjusted by IPSW is replicated.The bias of $\hat{\beta}_p$ as an estimator of the true β is $\text{bias}(\hat{\beta}_p) = E(\hat{\beta}_p) - \beta$. Then a bootstrap estimate for such bias is

$$\widehat{\text{bias}}_B(\hat{\beta}_p) = E(\hat{\beta}_p^*) - \hat{\beta}_p, \quad \text{where} \quad E(\hat{\beta}_p^*) = \frac{1}{B}\sum_{b=1}^{B} \beta_p^*(b) \tag{3}$$

Then, the bias-corrected estimate of β is given by

$$\bar{\beta}_c = \hat{\beta}_p - \widehat{\text{bias}}_B(\hat{\beta}_p) = 2\hat{\beta}_p - \hat{\beta}_p^* \tag{4}$$

4 Simulation Study

We conducted a simulation study to broadely compare the performances of each method. $M = 1000$, each with a sample size of $n = 290$ as in ESRD data, with follow-up time for 4 years ($t = 0, \ldots, 4$). A treatment status was generated from the Bernoulli distribution $z_i \sim \text{Bernoulli}(\hat{p}_i)$, where the probability $\hat{p}_i$ is the propensity score as estimated in (2). The survival time T_i^* was generated by evaluating the proportional hazard model of (1), where the Weibull hazard $h_0(t) = \lambda\rho t^{\rho-1}$ with the rate and shape parameters were $\lambda = 0.1$ and $\rho = 1.4$, respectively. The censoring time C_i was generated according to a uniform distribution in $(0,4)$. For each M simulatation, a BC-IPSW procedure based on $B = 2000$ bootstrap runs has been performed. To evaluate the impact of violating the ignorability assumption (i.e. unaccounted covariate) from the treatment response model, we consider two scenarios as in [1]: a *mild* misspecification, where covariate with weak impact assumed missing, and a *gross* misspecification, where a covariate with strong impact was missing. We summarize the simulation results using quantities such as bias, absolute percentage bias, mean-squared error and the 95% coverage probabilities.

5 Results

We first discuss the estimate of OAT effect on mortality for the ESRD dataset. The estimated hazard ratio (HR) from each method indicated that OAT treatmen has the benefit of reducing mortality as the HRs$<$1, although the impact is different for each method. The HR of 0.90(95% CI:0.60-1.35) from the naive analysis is less than the HR of 0.79(95% CI:0.46-1.16) obtained from IPSW model. This shows that the difference between the hazard of the treated and the control groups is quite small

for the naive case. Further, these two method estimated a statistically non-significant OAT effect, but IPSW produced a narrow 95% confidence interval. The BC-IPSW estimated a HR=0.65(95% CI: 0.42-0.99) which implies a statistically significant effect of OAT on mortality. This HR indicates that the rate of mortality for the OAT treated group decreases by 35% compared with the rate in the control group. Table 1 shows the results of the simulation study with the true $\beta \in (-0.3, 0.4)$, corresponding to the scenarios of a negative and a positive association between treatment and mortality. In both scenarios, the naive method estimated higher bias compared with the biases in the other two approaches. For $\beta = -0.3$, the naive method estimated a percentage bias of 5% and this bias was lowerd to 3.33% when IPSW was used. The BC-IPSW resulted in a further lowered percentage bias of 0.67%. In terms of accuracy, the three approaches were comparable as the MSEs are relatively similar, indicating that the treatment selection bias affects the estimated treatment effect with the impact less pronounced in precision and coverage properties.

Table 1 Results of the simulation study based on naive, IPSW and BC-IPSW, with the true $\beta \in (-0.3, 0.4)$. The shown are estimate, bias, absolute percentage bias (%Bias), Emperical Monte Carlo standard error (ESE), mean squared error (MSE) and 95% coverage probabilities (CP)

True β	Method	Estimate	Bias	%Bias	MSE	ESE	CP
	Naive	-0.315	-0.015	5.000	0.041	0.202	95
-0.3	IPSW	-0.290	0.010	3.333	0.039	0.196	95
	BC-IPSW	-0.298	0.002	0.667	0.041	0.203	95
	Naive	0.417	0.017	4.250	0.035	0.185	96
0.4	IPSW	0.385	-0.015	3.750	0.035	0.185	95
	BC-IPSW	0.403	0.003	0.750	0.033	0.181	96

Table 2 reports misspecification of the ignorability assumption. The impact of the misspecification is clearly observed in the estimates of the IPSW and BC-IPSW. For $\beta = -0.3$, we observe a percentage bias of 11.33% for the IPSW under slight misspecification. This corresponds to an extra 8% bias as compared with the same scenario without misspecification (see, Table 1). The extra bias could be as large as 13% under gross misspecification. When the BC-IPSW is applied, the percentage bias was reduced to 5% and 7.33% for the mild and gross misspecifications, respectively. A similar bias reducing performance was observed for $\beta = 0.4$.

6 Discussion

We presented a bootstrap based bias correction mechanism when treatment effect is estimated via inverse propensity score weights (BC-IPSW) for observational studies. Application of the method on a real ESRD data indicated an improved estimate of the OAT effect, in which the benefit of OAT implied by a statistically significant reduction of mortality for patients receiving the treatment. We showed, by the simulation study, the BC-IPSW performs well in correcting biases due to model misspecifications. When the misspecification is due to failure of the ignorability

Table 2 Simulation results under the gnorability assumption misspecification. The shown are bias, absolute percentage bias (%Bias), mean squared error (MSE) and 95% coverage probabilities (CP).

		slight misspecification				gross misspecification			
β	Method	Bias	%Bias	MSE	CP	Bias	%Bias	MSE	CP
	Naive	-0.015	5.000	0.041	95	-0.015	5.000	0.041	95
-0.3	IPSW	0.034	11.333	0.038	95	0.049	16.333	0.040	93
	BC-IPSW	0.015	5.000	0.040	95	0.022	7.333	0.040	94
	Naive	0.012	3.000	0.038	93	0.012	3.000	0.038	93
0.4	IPSW	0.035	8.750	0.036	94	0.059	14.750	0.036	93
	BC-IPSW	0.019	4.750	0.036	96	0.055	13.750	0.036	94

assumption, the BC-IPSW method substantially reduces the bias. This is apparent when the misspecification is due to missing a weak covariate from the treatment response model. In the event of a gross misspecification, the BC-IPSW never provided estimates that are less biased than the estimate obtained by the naive analysis.

References

[1] Arisido, M.W., Antolini, L., Bernasconi,D.P., Valsecchi, M.G. and Rebora, P.: (2019). Joint model robustness compared with the time-varying covariate Cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint. BMC Medical Research Methodology, **19**, doi:10.1186/s12874-019-0873-y

[2] Austin, P. C.: (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, **46**, 399–424

[3] Cox D.R.: (1972). Regression models and life tables. Journal of the Royal Statistical Society, **34**, 187–220

[4] Genovesi, S., Rossi, E., Gallieni, M., Stella, A and other.: (2014). Warfarin use, mortality, bleeding and stroke in haemodialysis patients with atrial fibrillation. Nephrology Dialysis Transplantation, **30**, 491–498

[5] Gubhinder K, P. R., and Marcel, V.: (2018). Bootstrap Bias Correction For Average Treatment Effects With Inverse Propensity Weights. Journal of Statistical Research, **52**, 187–200

[6] Joffe, M. M., Ten Have, T. R., Feldman, H. I. and Kimmel, S. E.: (2004). Model selection, confounder control, and marginal structural models: review and new applications. The American Statistician, **58**, 272-279

[7] Peng, X. and Jing, P.: (2011). Bootstrap confidence intervals for the estimation of average treatment effect on propensity score. Journal of Mathematics Research, **3**, 52–58

[8] Rosenbaum, P. R., and Rubin, D. B.: (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, **70**, 41–55

Combining multiple sources to overcome misclassification bias in epidemiological database studies

Integrazione di fonti diverse per correggere la distorsione da errata classificazione negli studi epidemiologici su basi di dati

Francesca Beraldi, Rosa Gini, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini

Abstract In epidemiological database studies, the true value of the outcome of interest is unknown, and it is measured with error through indicators implemented by appropriate algorithms. Differential misclassification of the study outcome over the exposure groups may severely bias the results of database studies. In this work, we address misclassification bias by proposing an estimator of the risk ratio that exploits two indicators from different sources: a main indicator with high Positive Predictive Value (PPV) and an auxiliary indicator with high sensitivity. We also develop a bootstrap test for differential sensitivity. We conduct a simulation study to explore the properties of the proposed methods under multiple scenarios.

Abstract *Negli studi epidemiologici su basi di dati il vero valore della variabile di risultato non è noto, ma è misurato con errore attraverso indicatori implementati da opportuni algoritmi. In questo contesto, se l'errore di misura è differenziale per strati di esposizione si possono produrre distorsioni rilevanti. Proponiamo di risolvere il problema dell'errore di classificazione differenziale tramite uno stimatore del rapporto tra rischi basato su due indicatori derivati da fonti diverse: un indicatore principale con un alto valore predittivo positivo (PPV) e un indicatore ausiliario con alta sensibilità. Sviluppiamo, inoltre, un test bootstrap per la sensibilità differenziale. Le proprietà dei metodi proposti sono valutate in vari scenari tramite uno studio di simulazione.*

Key words: bootstrap test, risk ratio, measurement error, validity indices

Francesca Beraldi, Emanuela Dreassi, Leonardo Grilli, Carla Rampichini
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy e-mail: francesca.beraldi@stud.unifi.it, emanuela.dreassi@unifi.it, leonardo.grilli@unifi.it, carla.rampichini@unifi.it

Rosa Gini
Agenzia regionale di sanità della Toscana, Florence, Italy e-mail: rosa.gini@ars.toscana.it

1 Background

One of the major aims of epidemiology is to assess the relationship between an exposure E and an outcome Y. Here we assume that both the exposure and the outcome are binary and we focus on estimating the *risk ratio* $RR_Y = \pi^E/\pi^{\overline{E}}$, where $\pi^E = Pr(Y = 1|E = 1)$ and $\pi^{\overline{E}} = Pr(Y = 1|E = 0)$. In database studies, E and Y are measured with error by means of indicators derived through specific algorithms from archives that are typically designed for other purposes. Here we assume that E is known without error, whereas Y is measured with error by an indicator M_Y. Therefore, the observed risk ratio based on M_Y, denoted with RR_{M_Y}, is different from the true risk ratio RR_Y. The bias of the observed risk ratio depends on the validity indices: sensitivity $SE = Pr(M_Y = 1|Y = 1)$; specificity $SP = Pr(M_Y = 0|Y = 0)$; positive predictive value $PPV = Pr(Y = 1|M_Y = 1)$; negative predictive value $NPV = Pr(Y = 0|M_Y = 0)$. Each validity index can be defined with respect to an exposure group: for example, $SE^{\overline{E}} = Pr(M_Y = 1|Y = 1, E = 0)$ is the sensitivity for non-exposed individuals. The misclassification error is said to be *non-differential* if it does not depend on the exposure group, namely if $SE^E = SE^{\overline{E}}$ and $SP^E = SP^{\overline{E}}$.

The relationship between the true risk ratio and the observed risk ratio [1] is

$$RR_Y = RR_{M_Y} \frac{PPV^E}{PPV^{\overline{E}}} \frac{SE^{\overline{E}}}{SE^E} \tag{1}$$

The common approach to overcome the measurement error problem is to assume that the sensitivity is non-differential across exposure groups ($SE^E = SE^{\overline{E}}$), so that the last factor in equation (1) disappears. This assumption is motivated by the large sample sizes usually needed to estimate the sensitivities. Then the naif approach is to assume that the PPV is the same in the two exposure groups, so that $RR_Y = RR_{M_Y}$. Typically, to justify this assumption, the current practice is to search for an indicator with PPV=1, which is hard to achieve; moreover, an indicator with high PPV is more likely to have differential sensitivity, thus increasing the bias of the risk ratio. Therefore, a better approach is to devise a validation study to estimate the PPV in the two exposure groups in order to adjust the observed risk ratio [2]. We refer to this procedure as the *single-indicator strategy*. Note that this strategy rests on the assumption of non-differential sensitivity, which is not tested in current epidemiological practice. Unfortunately, differential sensitivity is plausible in many settings. For example, in safety studies persons exposed to a drug with a suspect adverse outcome may be more easily diagnosed and recorded, and thus correctly classified by a specific indicator.

2 Proposed strategy

In database studies, there usually are several possible indicators. Let us consider a situation where the set of available indicators includes a *main indicator A* with high PPV and an auxiliary indicator B with high sensitivity. The role of the auxiliary indicator is to define a *composite indicator* $A \cup B$ to be exploited for estimating the risk ratio. To this end, the auxiliary indicator B should have high sensitivity so that the composite indicator $A \cup B$ has sensitivity equal to 1. In fact, the prevalence in a given exposure group $g \in \{E, \overline{E}\}$ can be written as

$$\pi^g = \frac{P_A^g PPV_A^g + P_B^g PPV_B^g - P_{A\cap B}^g PPV_{A\cap B}^g}{SE_{A\cup B}^g} \tag{2}$$

where $P_{M_Y}^g$ is the proportion of cases with $M_Y = 1$ in group g (observed prevalence), which is always available. We assume that the composite indicator has unit sensitivity ($SE_{A\cup B} = 1$), which can be attained by appropriately choosing the auxiliary indicator B. Under this assumption, the only unknown quantities in equation (2) are the PPVs, which can be estimated from validation studies on A and B. Then we make the simplifying assumption that A and B have no intersection ($A \cap B = \emptyset$), so equation (2) reduces to $\pi^g = P_A^g PPV_A^g + P_B^g PPV_B^g$. The ratio of the estimates of π^E and $\pi^{\overline{E}}$ is the proposed estimate of the risk ratio. We refer to this procedure as the *multiple-indicator strategy* or *component strategy*. This strategy directly estimates the risk ratio instead of trying to remove the bias using equation (1). The advantage is to avoid the assumption of non-differential sensitivity; the drawback is the need to estimate the PPV of the auxiliary indicator B in addition to the PPV of the main indicator A, implying a loss of accuracy in the estimation of the risk ratio.

The choice between the single-indicator strategy and the multiple-indicator strategy should be based on whether the sensitivities of A are equal across the exposure groups, namely whether the hypothesis $H_0 : SE_A^E = SE_A^{\overline{E}}$ is true. Therefore, we build a test for this hypothesis, which we show to be equivalent to

$$H_0 : \frac{P_B^E PPV_B^E}{P_A^E PPV_A^E} - \frac{P_B^{\overline{E}} PPV_B^{\overline{E}}}{P_A^{\overline{E}} PPV_A^{\overline{E}}} = 0 \tag{3}$$

In this setting, the multiple-indicator strategy with indicators A and B is not used for direct estimation of the risk ratio, but it is exploited to test the assumption of non-differential sensitivity, which is crucial for an effective adjustment of the bias in the single-indicator approach. The test statistic is obtained by replacing the values of PPV in equation (3) with the estimates from validation studies. The sampling distribution is intractable, thus we build a bootstrap test. Specifically, we take 1000 bootstrap samples from the validation sample of A and similarly for B, then we compute the 95% percentile confidence interval for the test statistic and reject the null hypothesis when the interval does not contain zero.

3 Simulation study

We devise a simulation study to evaluate the performance of two competing strategies: use the main indicator A to compute the observed risk ratio, then adjust it with the PPV of A in the exposure groups, under the assumption of non-differential sensitivity (RR1, *single-indicator strategy*); exploit the auxiliary indicator B to construct the composite indicator $A \cup B$, which allows to estimate the risk ratio without assuming non-differential sensitivity (RR2, *multiple-indicator strategy*).

In the simulation study we generate a population of $N = 10000$ individuals, with a proportion of exposed equal to 0.2, and prevalences set to $\pi^E = 0.10$ and $\pi^{\overline{E}} = 0.05$. We consider two indicators such that $SE_{A\cup B} = 1$ and $A \cap B = \emptyset$, with the following specificities: $SP_A^E = 0.95$, $SP_A^{\overline{E}} = 0.90$, $SP_B^E = 0.20$, $SP_B^{\overline{E}} = 0.15$. For the main indicator A we devise two scenarios: *a*) non-differential sensitivity with $SE_A^E = SE_A^{\overline{E}} = 0.40$; *b*) differential sensitivity with $SE_A^E = 0.80$ and $SE_A^{\overline{E}} = 0.45$. The sensitivities of B derive from the previous values. Finally, the PPVs are estimated with validation samples of size n_A and n_B, which are initially set to 100 and then increased.

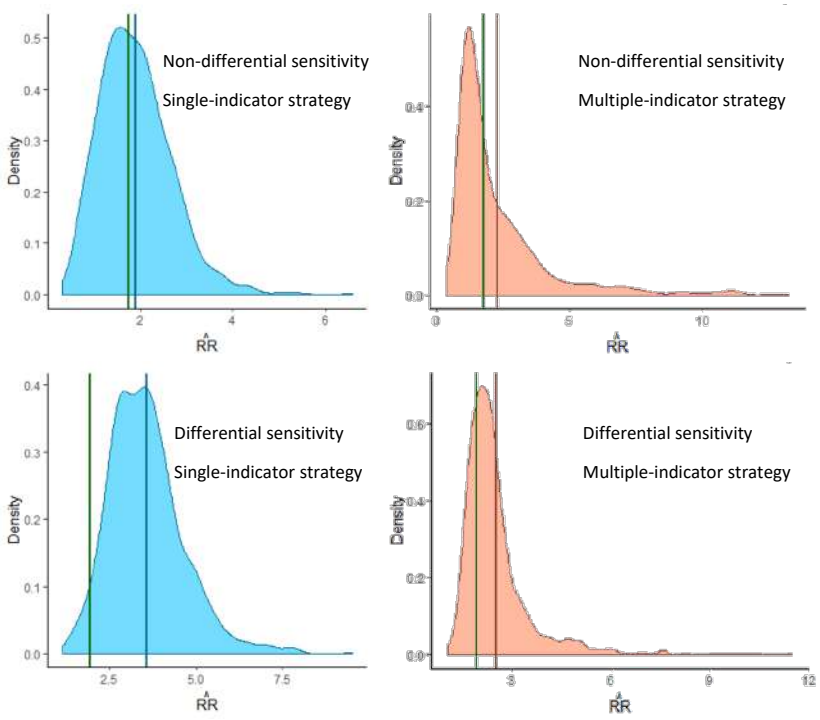


Fig. 1 Kernel density plots of the Monte Carlo distributions of the competing estimators of the RR under the two scenarios. Green vertical line: true value; other vertical line: Monte Carlo mean. PPVs estimated with validation samples of size $n_A = n_B = 100$.

We evaluate the performance of the competing approaches RR1 and RR2 through the Monte Carlo sampling distributions obtained by repeating 1000 times the extraction of the validation samples. Figure 1 reports kernel density plots of the Monte Carlo distributions of the competing estimators of the RR under the two scenarios.

The simulation results are in line with our expectation: in scenarios where the assumption of non-differential sensitivity is violated, the multiple-indicator strategy works well and it is clearly superior to the single-indicator strategy. On the other hand, under non-differential sensitivity the single-indicator strategy is preferable due to a lower variability, even if this is relevant only for small validation studies: in fact, the variances of the two estimators RR1 and RR2 become similar with larger validation studies (e.g. $n_A = 200$ and $n_B = 500$). In the simulation study we also evaluate the proposed bootstrap test for non-differential sensitivity ($H_0 : SE_A^E = SE_A^{\overline{E}}$). It turns out that the test works properly, but the power is low (about 0.5) with validation studies of size $n_A = 100$ and $n_B = 100$. In order to achieve a power of 0.8 the required sizes are $n_A = 200$ and $n_B = 500$, which are higher than the sizes usually adopted in validation studies.

Acknowledgements This work has received support from the EU/EFPIA Innovative Medicines Initiative [2] Joint Undertaking ConcePTION grant n. 821520.

References

1. Brenner, H., Gefeller, O.: Use of positive predictive value to correct for disease misclassification in epidemiologic studies. American journal of epidemiology. **138**, 1007–1015 (1994)
2. Gini, R., Dodd, C., Bollaerts, K., Bartolini, C., Roberto, G., Huerta, C., Martín-Merino, E., Duarte-Salles, T., Picelli, G., Tramontan, L., Danieli, G., Correa, A., McGee, C., Becker, B., Switzer, C., Gandhi-Banga, S., Bauwens, J., Maas, N., Spiteri, G., Sturkenboom, M.: Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project. Vaccine. (2019) doi 10.1016/j.vaccine.2019.07.045

Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics

Selezione Variabili tramite Deep Sparse Autoencoders per validazione di SNPs in Radiogenomica del Cancro alla Prostata

Michela Carlotta Massi[1,2], Francesca Ieva[1,2,3], Anna Maria Paganoni[1,2,3], Andrea Manzoni[1], Paolo Zunino[1], Nicola Rares Franco[1], Tiziana Rancati[4], and Catharine West[5,6]

Abstract Prostate cancer is the most diffused cancer affecting the male population. As therapies improve their effectiveness, surviving patients might be affected by complications induced by radiotherapy in the long run. To predict the onset of such rare late toxicities, because of the failure of phenotypic characteristics, the attention is shifting towards identifying specific genetic locations (Single Nucleotide Polimorphisms, or SNPs) associated with them. Because of the complexity of the problem, SNPs identified in a study are rarely validated on a different cohort of patients. In this case study we apply a novel approach for feature selection (namely a Deep Sparse Autoencoder-based Feature Selection method), to validate SNPs associated with radiotherapy-induced late toxicity causing urinary frequency variation (UFV).

Abstract *Il cancro alla prostata è il più diffuso tra la popolazione maschile. Nonostante il miglioramento nei trattamenti, i pazienti comunque essere affetti da complicazioni indotte dalla radioterapia nel lungo periodo. Per predire l'emergere di queste rare tossicità tardive, visto il fallimento nell'utilizzare caratteristiche fenotipiche dei pazienti, l'attenzione si sta spostando sull'identificare loci genetici (SNPs) a loro associate. Per la complessità del problema, le SNP individuate in uno studio sono raramente validate su una coorte differente di pazienti. In questo caso studio applichiamo un nuovo metodo di selezione delle variabili (un metodo basato su Deep Sparse Autoencoders), per validare le SNPs associate con la variazione tardiva della frequenza urinaria.*

Key words: Radiogenomics, Feature Selection, Autoencoders, Deep Learning, SNPs

---

[1]MOX Laboratory, Math Department, Politecnico di Milano, Milan, Italy
[2]CADS-Center for Analysis, Decisions and Society, Human Technopole, Milan, Italy
[3]CHRP-National Center for Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy
[4]Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[5]Translational Radiobiology Group, Division of Cancer Sciences, University of Manchester
[6]Manchester Academic Health Science Centre, Christie Hospital, UK

1 Introduction

Prostate cancer is the most diffused cancer affecting the male population in Europe. According to the American Cancer Society, about 1 american man in 9 will be diagnosed with prostate cancer during his lifetime. Because of the recent advancements in treatments, survival rates are high, but patients might suffer from debilitating complications resulting from therapies in the long run (radio-therapy induced late toxicity) [1, 2].

Traditional methods based on patients' phenotypic characteristics and treatment details fail in stratifying the treated population and in predicting the onset of such negative, but still very rare, side-effects. For this reason, the attention is shifting towards investigating possible relations between the genotype and the adverse outcomes in the so called 'precision medicine' approach, driving the need for novel statistical methods to address this question.

This case study was conducted with the support of Fondazione IRCSS Istituto Nazionale dei Tumori, the Italian National Cancer Research Institute, that provided us with data regarding the REQUITE [3] cohort of prostate cancer patients, with the aim of validating some specific genetic markers (in the form of Single Nucleotide Polymorphisms, SNPs) that could be predictive for late toxicity. The identification and validation of predictive biomarkers is an objective of paramount importance in a setting such as Genome-Wide Association Studies (GWAS), as the complexity of the problem, the rarity of the *traits* (or negative outcome) and the numerosity of the genetic traits to evaluate makes it extremely complex and rare for different studies to recognize similar patterns in data.

In this short paper we present a novel approach to SNPs validation, exploiting a Deep Sparse Autoencoder-based (DSAE) feature selection method to identify relevant SNPs associated with radiotherapy-induced late Urinary Frequency Variation (UFV). The task at hand requires us to identify predictive features for an extremely small minority class in a setting characterized by complex non-linear interactions among genetic loci, small sample size, several confounding factors, noisy data and the need for results interpretability to drive real clinical research.

For this reason, the work presented in this study exploits a feature selection method tailored to identify relevant features to discriminate the minority class from the majority class, and improve minority class classification accuracy. The adopted methodology for this case study was developed in a previous work in [4], where a detailed description of the algorithm can be found. For this reason, in Section 2 we will provide only a brief description of the main concepts, while the rest of the paper will be devoted to the case study.

The benefits of this Deep Learning (DL) model for our objective are several: it is a non-linear and stratified model, allowing to learn complex and hyerarchical relationships in data; additionally, the model deals well with large numbers of features, and has the capability of autonomously ignore noise.

2 Methods

Deep Sparse AutoEncoders (DSAE). An AutoEncoder (AE) is a neural network whose output provides a reconstruction of the input (Hinton and Salakhutdinov, 2006). The network can be seen as constituted by two parts: an encoder and a decoder.

The encoder function $\mathbf{h}_i = f(\mathbf{W}\mathbf{x}_i + b)$, encodes each input vector $\mathbf{x}_i$ into an encoded version of itself of size H. Here $f : \mathrm{R}^{\mathrm{J}} \rightarrow \mathbb{R}^{\mathrm{H}}$ is usually non-linear, $\mathbf{W}_{H\times J}$ is called *weight matrix* and $\mathbf{b}$ is an H-dimensional *bias* vector.

The decoder maps back the encoded vector to the J-dimensional space in most cases using a squashing non-linear function $\hat{\mathbf{x}}_i = g(\mathbf{W}'\mathbf{h}_\mathbf{i} + \mathbf{b}')$, $g : \mathbb{R}^{\mathrm{H}} \rightarrow \mathbb{R}^{\mathrm{J}}$ with parameters $\mathbf{W}'$ and $\mathbf{b}'$. The model is trained through gradient descent of the loss function $L(\mathbf{x}, \hat{\mathbf{x}})$; where L is typically the Mean Squared Reconstruction Error (RE), i.e. the mean squared Euclidean distance between the input values and the reconstructed values for each observation.

To force the model to learn more useful representaitons of the input data, one approach is to force sparsity in the central hidden layer. A sparse representation can be obtained adding a penalty term that penalizes the L_1 norm of the vector $\mathbf{h}_i^{(l)}$ of activation of the hidden nodes (where (l) indicates the layer the hidden nodes belong to, and it should be considered the most internal layer in case of a Deep AE), for each observation i, controlled by the parameter λ, i.e.:

$$L_i = L(\mathbf{x}_i, \hat{\mathbf{x}}_i) + \lambda |\mathbf{h}_i^{(l)}|. \tag{1}$$

The parameter λ can be optimized through grid search or can be arbitrarily chosen in the design phase of the model.

DSAE for Minority Class Feature Selection. The main idea behind the choice of a DSAE as a mean to perform feature selection is that the model trained to reconstruct *normal* observations only (majority class, or *healthy* patients) would make higher errors by trying to reconstruct anomalous patterns in *outliers* (minority class, or *unhealthy* patients showing late toxicity). Indeed, it is on the analysis of the average RE performed by the model on each feature for each class that we identify those that could discriminate better between the two classes. A detailed description of the proposed methodology can be found in our previous work [4]. In Figure 1 we propose a schema of the algorithm after the trained DSAE is tested on both healthy and unhealthy patients. Note that as a result the algorithm provides a subset of features which dimension depends on a parameter ($\delta \in [0,1]$) set by the user: the closer the δ value is to 1, the smaller the subset.

3 Urinary Frequency: Case Study Setting

From the original dataset, we selected a cohort of 1,296 patient, among which 55 (4.2%) belonged to the class of *cases* (y=1, i.e. the patients reported radiotherapy-induced late UVF), while 1241 (95.8%) belonged to the *controls*'s class. Each pa-

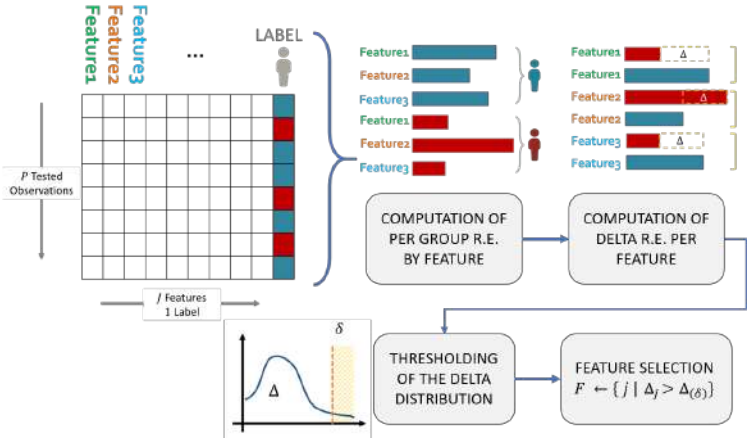


Fig. 1 Schema of the proposed algorithm: after training the DSAE on healthy patients only, the model is supplied with a test set composed of healthy and unhealthy patients. The passages in this schema depict all the steps from the collection of the RE to the feature selection based on δ. More details can be found in [4].

tient was characterized by 43 genetic traits (SNPs), among which 9 were identified in previous studies as predictive biomarkers for this endpoint. In Table 1 we list the biomarkers to validate. The unbalancing of the classes and the complexity of the problem (there potentially exist complex non-linear relations among biomarkers affecting the outcome [6]) makes this field of application an interesting fit for the peculiarities and potentials of our proposed model [4]. We performed the training and testing of the DSAE 50 times, extracting from the 1,296 the training set (1,186 observations, i.e. all *controls* except the 55 included in the test set) and test set (110 observations, half *cases* and half *controls*). The algorithm was implemented in Python. The DSAE had one input layer with 43 nodes, and three encoding hidden layers (with 40, 30 and 20 nodes respectively), followed by three decoding layers (30, 40, 43 nodes respectively). The training of each DSAE was performed for 400 hundred epochs, with a batch size of 10 observations, and the whole procedure of sampling, training and testing took on average (over the 50 repetitions) 3.22 minutes to complete. Note that the training time of the whole algorithm highly depends on the number of repetitions of sampling and training, and could be highly reduced in case less repetitions are needed to capture the most relevant variations between the two classes, or the number of minority class observations is sufficiently large to require a smaller number of sampling procedures on the healthy

SNP	Reference
rs17599026	Kerns et al. (2016) [5]
rs342442	Kerns et al. (2016) [5]
rs8098701	Kerns et al. (2016) [5]
rs7366282	Kerns et al. (2016) [5]
rs10209697	Kerns et al. (2016) [5]
rs4997823	Kerns et al. (2016) [5]
rs7356945	Kerns et al. (2016) [5]
rs6003982	Kerns et al. (2016) [5]
rs10101158	Kerns et al. (2016) [5]

Table 1 SNPs previously identified in literature as associated with late UVF

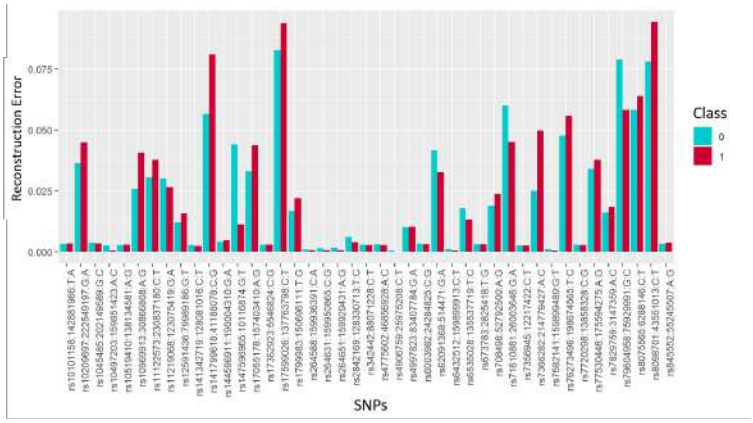


Fig. 2 Reconstruction Error by SNP and by Group (cases in red and controls in blue)

population to guarantee a robust comparison.
Once the AE was trained to reconstruct the training set of the healthy population, the test set (composed of healthy and unhealthy observations) was supplied to the model, collecting the Reconstruction Error (RE).
As in the process depicted in Figure 1, we obtained a matrix where each patient (row) that belonged to the test set at least once was associated with an outcome and a set of REs for each feature (SNP). This allowed us to group patients w.r.t. the enpoint, and estimate the average RE for each SNP for *cases* and *controls*.

4 Results

In Figure 2 we report the results of the procedure just described. The bars in the barplot represent the reconstruction error for each group (cases in red, controls in blue). On the x-axis one can read all the 43 SNPS, each one with two associated bars. To validate the SNPs in a robust way, we selected different values for the δ threshold (δ equal to 0.75, 0.8, 0.85 and 0.9). In Table 2 are reported the 9 SNPs previously identified in literature (already mentioned in Table 1) as predictive for the onset of late UFV after radiotherapy. As shown in Table 2, for δ=0.75 the model identifies as relevant (thus validating) 4 out of 9 SNPs coming from literature. Interestingly, the four identified SNPs present the highest odds ratio w.r.t. the outcome, according to the study that first mentioned them. Note that the study in [5] was performed on a different cohort of patients. The fact that the identified SNPs are those most evidently related to the outcome on different data is both a proof of our methodology to identify the most discriminative features, and of the generalizability of its results. Unfortunately, we do not have access to the data from the mentioned study to cross-validate our method on that cohort. Another notable aspect of our results, is that the four identified SNPs remain relevant almost for all δ values, except for one that is excluded after the last threshold (0.9).

ODDS RATIO	**THRESHOLD**			
	0.75	0.8	0.85	0.9
3,2	**rs7366282**	**rs7366282**	**rs7366282**	**rs7366282**
3,12	**rs17599026**	**rs17599026**	**rs17599026**	**rs17599026**
2,66	**rs10209697**	**rs10209697**	**rs10209697**	rs10209697
2,41	**rs8098701**	**rs8098701**	**rs8098701**	**rs8098701**
1,8	rs10101158	rs10101158	rs10101158	rs10101158
1,74	rs7356945	rs7356945	rs7356945	rs7356945
0,51	rs342442	rs342442	rs342442	rs342442
0,51	rs6003982	rs6003982	rs6003982	rs6003982
0,49	rs4997823	rs4997823	rs4997823	rs4997823
TOTAL SNPS	43	43	43	43
TOTAL SELECTED	11	9	7	5
TOTAL IDENTIFIED	4	4	4	3
PERCENTAGE IDENT/SEL	36.36%	44.44%	57.14%	60.00%
PERCENTAGE SEL/TOT	25.58%	20.93%	16.28%	11.63%

Table 2 Results of the SNPs validation for UFV. SNPs validated by our methodology are in bold, for different threshold values. The first column reports the ORs associated with these SNPs in the study in [5].

5 Conclusion

In this paper we presented a novel approach to SNPs validation through the use of a DSAE-based feature selection method to select relevant minority class features. We applied the methodology to a case study that required us to validate SNPs previously identified in literature as predictive for the onset of radiotherapy-induced late UFV. Despite the complex unsupervised setting does not allow us to compare our results with a *ground truth*, the robustness of the identified SNPs and the height of the Odds Ratio associated to them on another cohort of patients support our results.
Using a DL approach in a GWAS seems therefore to be a viable strategy to tackle the peculiar complexities of this setting, and opens the venue for relevant future research.

References

1. M. J. Zelefsky, A. Pinitpatcharalert, *et al.*, "Early tolerance and tumor control outcomes with high-dose ultrahypofractionated radiation therapy for prostate cancer," *European urology oncology*, 2019.
2. T. Rancati and C. Fiorino, "Predicting toxicity in external radiotherapy: A critical summary," in *Modelling Radiotherapy Side Effects*, pp. 337–363, CRC Press, 2019.
3. P. Seibold, A. Webb, *et al.*, "Requite: A prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer," *Radiotherapy and Oncology*, vol. 138, pp. 59–67, 2019.
4. M. C. Massi, F. Ieva, F. Gasperoni, and A. M. Paganoni, "Minority class feature selection through semi-supervised deep sparse autoencoders," *MOX Report 38/2019*, 2019.
5. S. L. Kerns, L. Dorling, L. Fachal, *et al.*, "Meta-analysis of genome wide association studies identifies genetic markers of late toxicity following radiotherapy for prostate cancer," *EBioMedicine*, vol. 10, pp. 150–163, 2016.
6. B. Liu, Y. Wei, Y. Zhang, and Q. Yang, "Deep neural networks for high dimension, low sample size data.," in *IJCAI*, pp. 2287–2293, 2017.

Graphical models for count data: an application to single-cell RNA sequencing

Modelli grafici per dati di conteggio: un'applicazione a dati di RNA a singola cellula

Nguyen Thi Kim Hue, Monica Chiogna, and Davide Risso

Abstract Biological processes involve complex interactions between genes. Such interactions may be represented employing a conditional independence graph. Here, we propose the use of a PC-stable algorithm to estimate the neighbourhood of each node in the graph, thus learning its structure. We apply our algorithm to two single-cell RNA sequencing datasets, demonstrating that we can identify important genes as hub nodes in the networks.

Abstract *Molti processi biologici sono caratterizzati da complesse interazioni tra geni. Queste interazioni possono essere rappresentate da un grafo di indipendenza condizionale. In questo lavoro proponiamo un algoritmo PC-stable per la stima della struttura del grafo. Applichiamo l'algoritmo proposto a due dataset di sequenziamento di RNA a singola cellula, dimostrando la sua abilità nel trovare geni importanti come nodi hub della rete.*

Key words: Graphical models, undirected graphs, structure learning, count data, gene expression, single cell, RNA sequencing

1 Introduction

Biological processes underlying the basic functions of a cell involve complex, and most of the times unknown, interactions between genes.

From a technical point of view, these interactions can be represented through a graph where genes and their connections are, respectively, nodes and edges. Es-

Nguyen Thi Kim Hue
Dept. of Statistical Sciences, University of Padova, e-mail: nguyen@stat.unipd.it

Monica Chiogna
Dept. of Statistical Sciences, University of Bologna, e-mail: monica.chiogna2@unibo.it

Davide Risso
Dept. of Statistical Sciences, University of Padova, e-mail: davide.risso@unipd.it

timation of the structure of such graphs from observed data is, generally speaking, a challenging task. State-of-the-art inference procedures assume that the data arise from a multivariate Gaussian distribution. However, high-throughput omics data, such as that from next generation sequencing, often violate this assumption. Such violations are particularly evident in single-cell RNA sequencing (RNA-seq), a technology that has enabled researchers to sequence the RNA of individual cells. In contrast to established sequencing techniques, which require the aggregation of thousands of cells, single-cell RNA-seq permits the study of gene expression at single cell resolution, allowing researchers to answer previously inaccessible questions (e.g., how stem cells develop into mature cells [1]). The resulting data are discrete, usually highly skewed, possibly zero inflated. In these contexts, the assumption of normality is far from being met.

In order to accommodate the distributional features observed in single cell gene expression, McDavid et al (2019) [2] proposed a Hurdle model, equivalent to a finite mixture of singular Gaussian distributions. The interesting aspect of the Authors' proposal is the possibility of accounting for zero-inflation. Nevertheless, a continuity of the data is still implicitly assumed.

In this work, we propose an algorithm for learning the structure of graphs for count data. The algorithm exploits an iterative testing procedure, which can easily account for sparsity of the data. A biological validation of the algorithm, performed by applying it to two case studies, shows its abilities to retrieve biologically relevant information.

2 Our proposal

A probabilistic graphical model requires the definition of a pair, $(G, \mathscr{F})$ say. Here, $G = (V, E)$ represents an undirected graph, where V is the set of nodes, and $E = \{(v,t) : v \neq t\}$ represents the set of undirected edges. Moreover, $\mathscr{F}$ represents a family of probability measures for the random vector X_V, indexed by the set V, on the sample space $\mathscr{X}_V$. Each node in the graph corresponds to a random variable $X_v, v \in V$; the existence of an edge (v,t), $v,t \in V$, indicates the dependency of the random variables X_v and X_t.

In conditional independence graphs, a connection is established between the graph G and a set of independence assumptions for the variables in X_V. The local Markov property assures that a variable X_v depends on the other variables only through its neighbors. This connection opens to the possibility of factorizing members of $\mathscr{F}$ over the graph G.

Here, we assume that the distribution of each variable X_v, conditional to all possible subsets of variables in $V \setminus v$, is a Poisson distribution:

$$X_v | x_{K\setminus\{v\}} \sim P\big(\exp\big\{\sum_{t\in K\setminus\{v\}} \theta_{vt|K} x_t\big\}\big), \quad \forall v \in K \subset \{1,\ldots,p\}, \tag{1}$$

A missing edge between node v and node t corresponds to the condition $\theta_{vt} = \theta_{tv} = 0$. On the other hand, one edge between node v and node t implies $\theta_{vt} \neq 0$ and $\theta_{tv} \neq 0$.

A conditional independence graph $G = (V, E)$ on X_V can be estimated by estimating, for each node $v \in V$, its neighborhood. Hence, one can proceed by estimating the conditional distribution of $X_v | X_{V \setminus v}$ and fixing the neighbourhood of v to be the index set of variables X_w on which the conditional distribution depends.

To estimate the neighbourhood of each node, we employ the PC-stable algorithm, a variant of the PC algorithm first proposed by [3]. The PC algorithm starts with a complete graph on V. Marginal independences for all pairs of nodes are tested, and edges removed when marginal independences are found. Then, for every pair of linked nodes, independence is tested conditional to all subsets of cardinality one of the adjacency sets of the two nodes. This testing procedure is iterated, increasing in turn the size of the conditioning sets, until this reaches its maximum limit, or a limit imposed by the user. Reasons for choosing the PC algorithm are many, spanning from its consistency (assuming no latent confounders) under i.i.d. sampling [3], to its ability to deal with a large number of variables and only moderately large sample sizes. The variant that we employ, PC-stable [4], allows to control instabilities due to the order in which the conditional independence tests are performed. To perform the tests, Wald-type test statistics are employed, for which a normal asymptotic distribution can be obtained by standard asymptotic theory (for details, we refer readers to [5]).

The chosen learning strategy has some advantages over alternative approaches based on sparse regressions (see [5]). Sparsity can be easily implemented by a control on the conditional set size, instead of a control on parameter magnitudes, which can lead to over-shrinkage. Moreover, it offers computational advantages, especially when sparse networks are the target of inference. Finally, hypothesis testing is scale-invariant, i.e., is not affected by scale transformations of regressors. For a discussion of convergence of the algorithm in the Poisson setting, see [5].

3 Single-cell RNA sequencing: two case studies

We consider two sets of single cell RNA-seq data, from the same biological system, but with two different technologies, with a particular focus on the transcription factor (TF) genes of the neuronal and sustentacular cell lineages. The first dataset, called `data1`, obtained using the Fluidigm C1 system, consists of 634 cells and 850 TF genes. The second, called `data2`, obtained using the 10X Genomics platform, consists of 634 cells and 877 TF genes. As measurements of both data were zero-inflated and highly skewed, standard preprocessing was applied to the data (see [5, 6]). In particular, we selected the top 20% variables with the highest mean; normalized the data by 75% quantile matching; selected the top 50% most variable genes across the data; and used a power transform X^{α} (with $\alpha = 0.219$ for the first data;

and $\alpha = 0.5$ for the second data). The distributions of the normalized data in the two datasets for four representative genes are shown in Figure 1.

4 Results

The biological system under study is the mouse olfactory epithelium (OE), which is often employed to study tissue regeneration by adult stem cells. Following an injury, OE stem cells generate, through a series of intermediate states, olfactory neurons and sustentacular cells, following two distinct developmental lineages, called the neuronal lineage and the sustentacular lineage, respectively [1]. We applied our algorithm to both groups individually (i.e., considering the neuronal lineage and sustantacular cell lineage separately), and to all data at once.

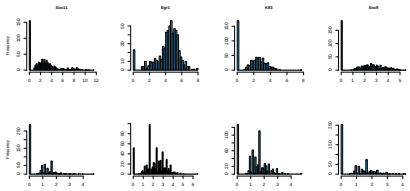
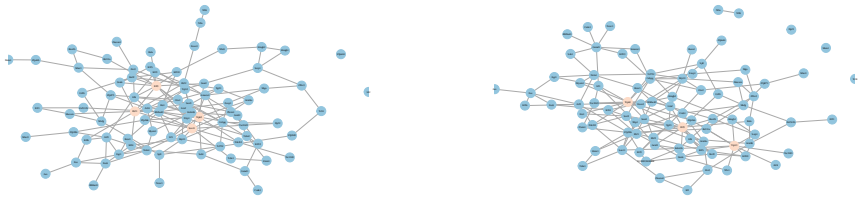


Fig. 1: Distribution of RNA-Seq read counts for four genes: `data1` (top), `data2` (bottom).

Figure 2 shows the networks estimated using our algorithm at a significance level of 5% on `data1` for each lineage. In the neuronal lineage, the TF genes Sox11, Ebf1, Elf3, and Trp63 are hub nodes (i.e., nodes with at least 9 edges). All these genes are known to be important in neurogenesis or stem cell differentiation. For instance, Tpr63 is a gene essential to maintain the quiescent state of stem cells. In fact, by knocking out this gene stem cells will differentiate into mature cell types [1].



(a) Neuronal lineage

(b) Sustentacular cell lineage

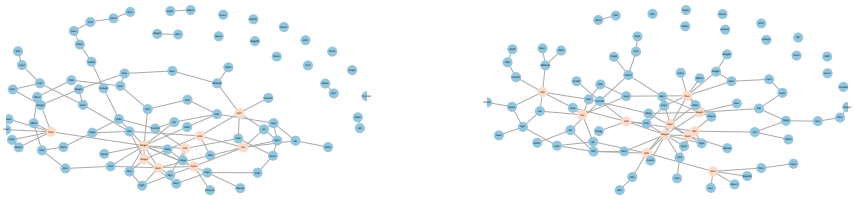
Fig. 2: Olfactory gene networks estimated by our algorithm on `data1` (hub nodes coloured orange).

For the sustantacular cell lineage, three hub nodes are identified: Trp53, Elf3, Trp63 (see Figure 2b). In addition to Trp63 (discussed above), the identification of Elf3 is of note: while there is no specific indication in the literature that it plays

an important role in the olfactory epithelium, this gene is known to be important in other epithelial systems [7] and is a good candidate for follow up studies.

The networks estimated from `data2` are showed in Figure 3. Here, we identified nine hub nodes (i.e., nodes with at least 5 edges) in the neuronal lineage: Ybx1, Hmga1, Hmga2, Fos11, Atf5, Klf4, Trp53, Fos, Atf3 (see Figure 3a); and eleven hub nodes in the sustantacular cell lineage: Ybx1, Hmga1, Hmga2, Fos, Atf3, Xbp1, Myc, Elf5, Sox2 (see Figure 3b).

Again, many of these genes are known in the field to be important (some of which in common with `data1`). For instance, Klf4 is expressed in neural stem cells and controls axonal regeneration [8]. Moreover, Sox2 is essential for maintaining the self-renewal property of stem cells [9]. It is interesting to note that, for each dataset, gene networks obtained from different lineage trajectories are consistent, i.e., many common interactions between two TF genes are found by applying the algorithm to each group individually.



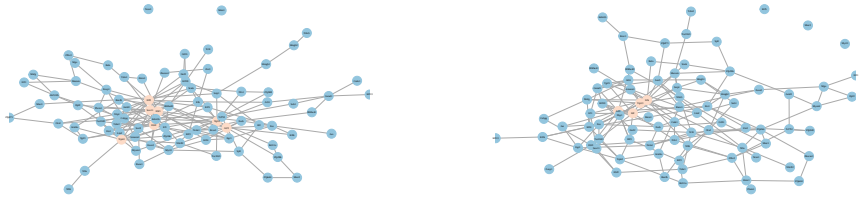
(a) Neuronal lineage (b) Sustentacular cell lineage

Fig. 3: Olfactory gene networks estimated by our algorithm on `data2` (hub nodes coloured orange).

We then applied our approach to each full dataset (the two lineages together). Figure 4a shows that, for `data1`, the network obtained in this way is consistent with the networks obtained from each lineage separately. However, this is not the case for `data2`. An explanation for this could be the `data2` contains more zeros, due to the nature of the platform (see Figure 1). Hence, our algorithm based on a Poisson model is less accurate in this case.

5 Discussion

In our work, we have inferred gene networks from two OE datasets using a PC-like algorithm. However, structure learning of graphical models for single cell gene expression is difficult due to the lack of detectable expression of many genes, giving rise to zero-inflated expression patterns. In the joint analysis of `data2`, our algorithm was unable to reconstruct the full network possibly because of zero inflation, which the Poisson is not able to model. The Poisson assumption limits the poten-



(a) `data1` (b) `data2`

Fig. 4: Olfactory gene networks estimated by our algorithm on `data1` and `data2` (hub nodes coloured orange).

tial of principled graphical modelling in various applications. Future work can thus focus on developing new algorithms using different distributions that are more sensitive to zero-inflation and overdispersion, such as the zero-inflated Poisson or the (zero-inflated) negative binomial.

References

1. Gadye, L. *et al.* Injury activates transient olfactory stem cell states with diverse lineage capacities. *Cell Stem Cell* **21,** 775–790 (2017).
2. McDavid, A., Gottardo, R., Simon, N. & Drton, M. Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics* **13,** 848 (2019).
3. Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, prediction, and search* (MIT press, 2000).
4. Colombo, D. & Maathuis, M. H. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* **15,** 3741–3782 (2014).
5. Hue Nguyen, T. K. & Chiogna, M. Structure learning of undirected graphical models for count data. *ArXiv e-prints.* arXiv: `1810.10854 [stat.ME]` (Oct. 2018).
6. Allen, G. & Liu, Z. A local Poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on Nanobioscience* **12,** 189–198 (2013).
7. Tymms, M. J. *et al.* A novel epithelial-expressed ETS gene, ELF3: human and murine cDNA sequences, murine genomic organization, human mapping to 1q32. 2 and expression in tissues and cancer. *Oncogene* **15,** 2449 (1997).
8. Qin, S. & Zhang, C.-L. Role of Krüppel-like factor 4 in neurogenesis and radial neuronal migration in the developing cerebral cortex. *Molecular and Cellular Biology* **32,** 4297–4305 (2012).
9. Mercurio, S., Serra, L. & Nicolis, S. K. More than just Stem Cells: Functional Roles of the Transcription Factor Sox2 in Differentiated Glia and Neurons. *International Journal of Molecular Sciences* **20,** 4540 (2019).

Interregional mobility, socio-economic inequality and mortality among cancer patients

Mobilità inter-regionale, disuguaglianze socio-economiche e mortalità in pazienti oncologici

Claudio Rubino, Mauro Ferrante, Antonino Abbruzzo, Giovanna Fantaci, Salvatore Scondotto

Abstract This paper investigates 3-years mortality after discharge in patients residing in Sicily (Italy) diagnosed with cancer among: colon, stomach, liver, and lungs, between 1/1/2010 - 31/12/2011. The effect of mobility and socio-economic status on mortality is evaluated through survival analysis approach. Results shows that out-of-region hospitalization is associated with higher survival time; no association of mortality with socio-economic status appears. The extent of patients' mobility, and its relation with mortality raises regional policy considerations.

Abstract *Il presente articolo analizza la mortalità a tre anni dalla dimissione in pazienti residenti in Sicilia ricoverati tra il* 1/1/2010 - 31/12/2011 *con diagnosi di tumore, quali: colon, stomaco, fegato e polmoni. Gli effetti della mobilità interregionale e dello status socio-economico sulla mortalità vengono studiati attraverso modelli di sopravvivenza. I risultati mostrano un'associazione tra mobilità e tempi di sopravvivenza, mentre non appaiono effetti significativi dello status socioeconomico. La rilevanza del fenomeno della mobilità e la sua relazione con la mortalità, sollevano riflessioni utili per la programmazione sanitaria regionale.*

Key words: patients' mobility, health outcome, survival analysis, socio-economic inequalities

---

Claudio Rubino, Mauro Ferrante (*corresponding author*)
Department of Culture and Society, University of Palermo e-mail: claudio.rubino@unipa.it; mauro.ferrante@unipa.it

Antonino Abbruzzo
Department of Economics, Business and Statistics, University of Palermo e-mail: antonino.abbruzzo@unipa.it

Giovanna Fantaci
ASP Trapani & Dipartimento Attività Sanitarie e Osservatorio Epidemiologico Regione Sicilianae-mail: giovanna.fantaci@asptrapani.it

Salvatore Scondotto
Dipartimento Attività Sanitarie e Osservatorio Epidemiologico Regione Siciliana e-mail: salvatore.scondotto@regione.sicilia.it

1 Introduction

Interregional mobility is a relevant phenomenon, due to its implications under the financial and the equity perspective. In publicly funded healthcare systems covering the whole population, patients are able to choose any hospital that best meets their needs. In these systems, since tariffs are generally fixed at the national level, patient mobility can act as a stimulus for hospital competition [1]. However, it can be also an issue for the patient that has to look out of her region to receive better quality of care [8]. Especially in case of severe health needs, patients' mobility might indicates a threat to equity in the healthcare system at the country level [2].

In Italy the phenomenon is remarkable: out-region hospitalizations are almost 1 over 10, involving about 730 thousands patients in 2018 [9]. The greatest pattern of patient mobility involves southern regions, thereby generating additional amounts of financial flows in favor of central–northern regions [3]. Previous studies [7] have shown an impact of socioeconomic inequality on interregional mobility, with a a highest propensity for being hospitalized outside the region among people residing in less deprived areas compared to those living in areas with highest deprivation. In the current analysis, patient-level data regarding in- and out-region hospitalizations in cancer patients residing in Sicily (Italy) are used to evaluate 3-year mortality. The analysis is performed by considering four different cancer sites, namely: stomach, colon, liver, and lungs.

2 Data and Methods

In this study, patient-level data on hospital discharge records (SDO) of patients residing in Sicily were made available from the Epidemiology Department of Sicily Region. The analysis included all people residing in Sicily who had been diagnosed with one of four selected cancer site, among: stomach (ICD-9-CM: 151), colon (ICD-9-CM: 153, 154, or 1590), liver and intrahepatic bile ducts (ICD-9-CM: 155, 156), trachea bronchus and lung (ICD-9-CM: 162), hospitalized in and beyond Sicily between 1 January 2010 - 31 December 2012. Further selection criteria were that patients had not been hospitalized for cancer in the previous seven years, and that each patient had only one type of cancer. SDO dataset was used also to determine whether the hospitalization occurred in- or out-region, as well as other patient-level characteristics, including gender, age, and Charlson comorbidity index [5]. Information on mortality was derived from the Regional register of causes of death (ReNCaM). Finally, the availability of information on the census tract of residence for those patients residing in municipalities with more than 10,000 inhabitants, allowed for the inclusion of the deprivation index [4] of the census tract of residence. The cohort under analysis comprises 8937 incident episodes of hospitalization of patients with a diagnosis of colon ($n = 3912$), liver and intrahepatic bile ducts ($n = 1483$), stomach ($n = 787$), and trachea, bronchus and lung ($n = 2755$). The phenomenon of interregional mobility concerned approximately 5.5% of the

patients under investigation, and it ranges from a minimum of 5.1% for colon, to a maximum of 6% for trachea, bronchus and lung. 3-years mortality is approximately 55.7%, and it ranges from 34.4% for colon to 76.3% for trachea, bronchus and lung.

Four separate Cox models [6] (one for each cancer site) have been estimated to evaluate the effect of several risk factors (reported in Table 1) on the survival time. It is worth noting that, given the multilevel structure of the socioeconomic status, the inclusion of random effects were initially considered. However, the census tract of residence did not proved to be a relevant source of heterogeneity with respect to survival time; therefore, the simpler non RE model was finally considered. The Cox model allows us to examine how specified factors influence the rate of a particular event happening (e.g., survival time for patients with stomach cancer) at a particular point in time. This rate is commonly referred as the hazard rate. The Cox model is expressed by the hazard function denoted by $h(t)$:

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \tag{1}$$

where t represents the survival time, $h(t)$ is the hazard function determined by a set of p covariates $(x_1, x_2, \ldots, x_p)$, the coefficients $(\beta_1, \beta_2, \ldots, \beta_p)$ measure the impact (i.e., the effect size) of covariates. The hazard function can be interpreted as the risk of dying at time t. The term h_0 is called the baseline hazard. It corresponds to the value of the hazard if all the x_i are equal to zero (the quantity $\exp(0)$ equals 1). The t in $h(t)$ reminds us that the hazard may vary over time. It is assumed that the linear predictor does not depend on time, hence the hazard ratio (HR) for any pair of individuals with different values of the covariates vector (e.g. $\mathbf{x}_1$ and $\mathbf{x}_2$) is time constant (proportional hazard assumption) and it is given by $\exp((\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta})$. Proportional hazard assumption was checked by looking at the correlation between survival time and the scaled Schoenfeld residuals. When proportional hazard assumption was not met, the relevant categories were included in the model as time-dependent coefficients [10]. The estimated time-varying coefficients are decreasing (logarithmic) functions of time.

3 Results

The distribution of factors potentially associated with patient mortality, along with corresponding mortality rate for the cancer sites considered are reported in Table 1. For all the cancer sites considered, three-years mortality increases with age and Charlson Index category. It is lower for patients who have been hospitalized outregion compared to those who do not. Mortality doesn't seem to be influenced by gender, with the exception of trachea, bronchus and lung (0.70 for females vs 0.78 for males). In general, an increase in the level of deprivation appears to be slightly associated with mortality rates. Fig. 1 shows Kaplan-Meier estimates of the survival functions both for patients who experienced mobility and for patients who do not, for each cancer site; dashed lines indicate the median survival times. P-values refer

Table 1 Distribution of patient-level characteristics and 3-years mortality according to cancer site.

	Colon		**Liver**		**Stomach**		**Trachea**	
	n	Mortality	*n*	Mortality	*n*	Mortality	*n*	Mortality
Sex								
Female	1917	0.34	661	0.68	332	0.67	631	0.70
Male	1995	0.34	822	0.67	455	0.67	2124	0.78
Age								
< 60	794	0.23	238	0.53	169	0.58	569	0.67
$(60, 80]$	2366	0.31	964	0.67	450	0.68	1791	0.77
> 80	752	0.56	281	0.81	168	0.76	395	0.86
Deprivation								
Low	2128	0.33	797	0.67	392	0.65	1421	0.75
Mid	1467	0.36	559	0.68	320	0.69	1055	0.78
High	317	0.38	127	0.72	75	0.72	279	0.78
Charlson								
$[0, 2]$	2081	0.21	427	0.63	335	0.58	794	0.65
3	594	0.30	385	0.61	127	0.59	468	0.65
$[4, 6]$	347	0.43	328	0.65	88	0.74	355	0.80
$[7, 10]$	864	0.65	321	0.83	228	0.82	1066	0.87
> 11	26	0.81	22	0.95	9	1.00	72	0.89
Mobility								
No	3713	0.35	1402	0.69	743	0.69	2589	0.77
Yes	199	0.21	81	0.49	44	0.43	166	0.61

to the Log-Rank tests of difference between the two groups, which resulted significant for all the sites considered. Table 2 shows the results of the Cox models for all the four cancer sites considered. The results of the models show that the relative risk of 3-years mortality for all the cancer sites considered is significantly lower for patients who experienced interregional mobility compared with to those who do not. The effect of gender was significant only for trachea, bronchus and lung site, with males being the highest risk group. For all the sites considered, the risk of death significantly increases with age and according to the Charlson index category, whereas the effect of deprivation on survival time turned out to be not statistically significant.

4 Conclusion

Our analysis shows that interregional mobility is associated with a lower risk of mortality for all the cancer sites considered. These results raise concerns on the equity in the provisioning of healthcare services at the national level. Despite its potential action as stimulus for competition, patient mobility is generally seen as a phenomenon that should be reduced [9]. In line with the National Outcomes Programme (PNE), the provision of indicator of treatment quality represents a fundamental tool to guide patient choice, and for the monitoring of quality of care.

Table 2 Model results for colon, liver and intrahepatic bile ducts, stomach, and trachea, bronchus and lung.

	Estimate	exp (est)	SE (est)	P-value
Colon				
Mobility Yes (Ref: No)	-0.516	0.597 (0.437, 0.816)	0.159	0.001
Age (60,80] (Ref: <= 60)	0.376	1.456 (1.236, 1.714)	0.083	0.000
Age > 80	1.795	6.020 (3.933, 9.213)	0.217	0.000
Charlson 3 (Ref: [0,2])	0.321	1.379 (1.157, 1.643)	0.089	0.000
Charlson [4,6]	0.610	1.840 (1.521, 2.225)	0.097	0.000
Charlson [7,10]	1.533	4.630 (4.081, 5.253)	0.064	0.000
Charlson >11	1.968	7.154 (4.608, 11.107)	0.224	0.000
Deprivation Mid (Ref: Low)	0.106	1.112 (0.992, 1.246)	0.058	0.067
Deprivation High	0.139	1.149 (0.947, 1.394)	0.099	0.160
Age > 80 : log(Time)	-0.111	0.895 (0.829, 0.965)	0.039	0.004
Liver and intrahepatic bile ducts				
Mobility Yes (Ref: No)	-0.516	0.597 (0.434, 0.821)	0.162	0.001
Age (60,80] (Ref: <= 60)	0.410	1.507 (1.244, 1.827)	0.098	0.000
Age > 80	1.644	5.176 (3.301, 8.115)	0.229	0.000
Charlson 3 (Ref: [0,2])	-0.104	0.902 (0.755, 1.076)	0.090	0.251
Charlson [4,6]	0.011	1.011 (0.844, 1.210)	0.092	0.905
Charlson [7,10]	1.404	4.071 (2.684, 6.174)	0.213	0.000
Charlson >11	1.290	3.633 (2.323, 5.683)	0.228	0.000
Deprivation Mid (Ref: Low)	0.013	1.013 (0.888, 1.156)	0.067	0.848
Deprivation High	0.129	1.138 (0.911, 1.422)	0.114	0.255
Age > 80 : log(Time)	-0.162	0.850 (0.777, 0.930)	0.046	0.000
Charlson [7,10] : log(Time)	-0.160	0.853 (0.781, 0.930)	0.044	0.000
Stomach				
Mobility Yes (Ref: No)	-0.760	0.468 (0.295, 0.741)	0.235	0.001
Age (60,80] (Ref: <= 60)	1.742	5.707 (2.021, 16.113)	0.530	0.001
Age > 80	2.373	10.733 (3.522, 32.705)	0.568	0.000
Charlson 3 (Ref: [0,2])	-0.069	0.933 (0.713, 1.222)	0.138	0.617
Charlson [4,6]	0.338	1.402 (1.055, 1.864)	0.145	0.020
Charlson [7,10]	0.774	2.168 (1.771, 2.655)	0.103	0.000
Charlson >11	1.673	5.327 (2.698, 10.516)	0.347	0.000
Deprivation Mid (Ref: Low)	0.094	1.099 (0.917, 1.316)	0.092	0.307
Deprivation High	0.120	1.127 (0.839, 1.514)	0.151	0.426
Age (60,80] : log(Time)	-0.259	0.772 (0.639, 0.932)	0.096	0.007
Age > 80 : log(Time)	-0.327	0.721 (0.586, 0.886)	0.106	0.002
Trachea, bronchus and lung				
Mobility Yes (Ref: No)	-0.422	0.656 (0.537, 0.801)	0.102	0.000
Sex Female (Ref: Male)	-0.236	0.790 (0.710, 0.879)	0.054	0.000
Age (60,80] (Ref: <= 60)	0.277	1.320 (1.177, 1.480)	0.059	0.000
Age > 80	1.381	3.979 (2.706, 5.851)	0.197	0.000
Charlson 3 (Ref: [0,2])	-0.064	0.938 (0.814, 1.082)	0.073	0.380
Charlson [4,6]	0.264	1.302 (1.123, 1.510)	0.076	0.000
Charlson [7,10]	0.689	1.992 (1.787, 2.220)	0.055	0.000
Charlson >11	1.496	4.462 (2.152, 9.254)	0.372	0.000
Deprivation Mid (Ref: Low)	0.081	1.084 (0.990, 1.188)	0.047	0.082
Deprivation High	0.043	1.044 (0.902, 1.209)	0.075	0.562
Age > 80 : log(Time)	-0.137	0.872 (0.807, 0.943)	0.040	0.001
Charlson >11 : log(Time)	-0.154	0.857 (0.731, 1.005)	0.081	0.057

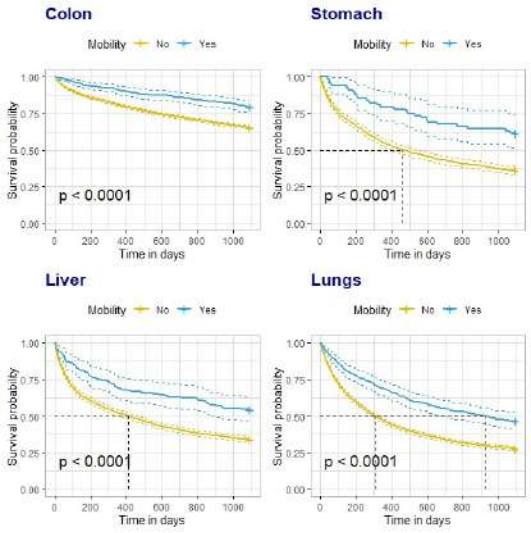


Fig. 1 Kaplan-Meier estimates of the survival functions for patients who experienced mobility and for those who do not, for each cancer site; P-values refer to the Log-Rank tests; dashed lines indicate median survival times.

References

1. Aggarwal, A.K., Sujenthiran, A., Lewis, D., Walker, K., Cathcart, P., *et al.*: Impact of patient choice and hospital competition on patient outcomes after prostate cancer surgery: A national population-based study. Cancer **125**(11), 1898–1907 (2019).
2. Balia, S., Brau, R., Moro, D.: Choice of hospital and long-distances: Evidence from Italy. Reg. Sci. Urban Econ., **81**, 103502 (2020).
3. Berta, P., Guerriero, C., Levaggi, R.: Hospitals' Strategic Behaviours and Patient Mobility: Evidence from Italy. No. 555. Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy (2020).
4. Caranci, N., Biggeri, A., Grisotto, L., Pacelli, B., Spadea, T., Costa, G.: The Italian deprivation index at census block level: definition, description and association with general mortality. Int. J. Health Policy Manag., **4**(6), 363–372 (2015).
5. Charlson, M., Szatrowski, T. P., Peterson, J., Gold, J.: Validation of a combined comorbidity index. J. Clin. Epidemiol., **47**(11), 1245–1251 (1994).
6. Cox, D.R.: Regression models and life-tables. J R Stat Soc Ser B Methodol., **34**(2),187–202 (1972).
7. Ferrante, M., Sciuto, V., Pollina-Addario, S., Parroco, A. M.: Individual and contextual determinants of inter-regional mobility in cancer patients. Electron. J. Appl. Stat. Anal., **12**(1), 14–25 (2019).
8. Levaggi, R., Zanola. R.: Patients' migration across regions: the case of Italy. Appl. Econ., **36**(16), 1751–1757 (2004).
9. Ministero della Salute: Rapporto annuale sull'attività di ricovero ospedaliero. Dati SDO 2018. Direzione generale della programmazione sanitaria. Roma (2020).
10. Therneau, T.M., Grambsch, P.M.: Modeling Survival Data: Extending the Cox Model. New York, NY, Springer-Verlag (2000).

PET radiomics-based lesions representation in Hodgkin lymphoma patients

Rappresentazione delle lesioni di pazienti affetti da linfoma di Hodgkin basata su radiomica PET

Lara Cavinato, Martina Sollini, Margarita Kirienko, Matteo Biroli, Francesca Ricci, Letizia Calderoni, Elena Tabacchi, Cristina Nanni, Pier Luigi Zinzani, Stefano Fanti, Anna Guidetti, Alessandra Alessi, Paolo Corradini, Ettore Seregni, Carmelo Carlo-Stella, Arturo Chiti, Francesca Ieva

Abstract As medical image analysis has been proven to entail tumor-specific information, the so-called radiomics paradigm holds the promise to characterize the disease and infer long term outcomes of chemotherapy. In this work, we propose an insightful framework for disease characterization in Hodgkin lymphoma which could inform future research. Particularly, an intra-patient similarity index (ISI) was built to represent the homogeneity of the patients' disease, while a radiomics-based fingerprint was create for local lesion description. Through descriptive statistics and classification algorithms, ISI-weighted fingerprint has been showed to be discriminatory between responders and relapsing patients.

Abstract *Recentemente, l'informazione derivante da immagini mediche è stata introdotta in maniera massiva nell'analisi quantitativa delle lesioni, volta alla creazione di modelli diagnostici e prognostici. Questo paradigma, in particolare ma non limitatamente al contesto oncologico, prende il nome di radiomica. Nel presente lavoro, proponiamo un metodo di profilazione dei pazienti affetti da linfoma di Hodgkin in termini di omogeneità paziente-specifica della malattia (ISI) e descrizione locale dellle lesioni, basata su caratteristiche radiomiche. Attraverso statistiche descrittive e algoritmi di classificazione, tale metodo si è rivelato essere discriminante dei pazienti che hanno risposto e quelli refrattari alla prima linea di chemioterapia.*

Lara Cavinato
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy
e-mail: lara.cavinato@polimi.it

Francesca Ieva
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy
CADS – Center for Analysis, Decision and Society, Human Technopole, Milan, Italy
e-mail: francesca.ieva@polimi.it

Key words: Hodgkin Lymphoma, PET/CT, Radiomics, Similarity index, Precision medicine, Metabolic tumor volume, Silhouette, Unbalanced classification

1 Introduction

Hodgkin's lymphoma (HL) is one out of a group of blood cancers that develop from lymphocytes. Although most of the cases end up with long term outcome [3], up to 30% of patients with early or advanced stage HL can become refractory or relapsing [1], bringing the first-line therapy into failure. For this reason, detecting cases at high risk of event recurrence at baseline would inform and significantly impact on HL patients therapy trial. However, at this stage, according to the gold standard of the International Prognostic Score, stratification and therapeutic strategies are based on clinical risk factors [1]. The staging system for patients with HL relies on the number of involved lesion sites, the severeness of lesions (i.e. bulky), the nodal or extra nodal nature of the disease, the presence of typical systemic symptoms (B symptoms) and lymph nodes stage (on one or both sides of the diaphragm). Early prognostic factors are often investigate the presence of a large mediastinal mass, an elevated sedimentation rate, the involvement of multiple nodal sites, including extra-nodal, age $\geq$ 50 years, or massive splenic disease [6].

Over the last years, research has moved forward to a more quantitative approach. Indeed, the value of image analysis applied to Positron Emission Tomography (PET)/Computer Tomography (CT) for response evaluation and treatment monitoring has been suggested and showed as a promising strategy [2]. Specifically, standardized texture feature extraction in PET/CT images, i.e. radiomics, quantifies the heterogeneity of tracer uptake within a metabolically active region of interest, i.e. tumor lesions. In combination with patients' information, such data are fed into statistical models developed for both research and clinical purposes, such as diagnosis or prognosis. Accordingly, radiomics-derived lesion description of refractory/relapsing HL has been supposed to differ from the one of long term responders [2]. Although straightforward, this workflow might suffer from several limitation and critical issues, such as the enormous amount of parameters that may be computed within image regions, the high co-linearity between features and the lack of biological interpretation which can be inferred from the analyses. The present study aimed at developing a methodological framework in HL for radiomics feature reduction and selection in order to locally describe lesions, different in size and nature, while combining their characterization at patient's level though an intra-patient similarity index. The ultimate perspective of such patient representation lies on further modeling responding/refractory phenotypes in first-line chemotherapy response assessment.

2 Materials and Methods

In accordance with the Declaration of Helsinki, this observational retrospective study was approved by the local ethics committee of all the three centers involved and the signature of a specific informed consent was waived.

Data have been collected from 85 patients with pathological diagnosis of HL undergoing a pre-treatment PET/CT scan. In particular, two categories of patients have been analyzed: non-relapsing/refractory (non-R/R), i.e. long term responders, and relapsing/refractory (R/R) subjected to more than one chemotherapy line and candidate to immunotherapy. Patients with extravasation at injection site, no clinical data availability or having only one tumor lesion have been excluded.

In order to assess imaging data, pre-treatment (baseline) [18F]FDG-PET/CT and the study before immunotherapy have been analyzed for non-R/R HL and R/R HL patients, respectively. PET/CT images have been acquired according to standard institution-specific procedure protocols and images have been retrieved and qualitatively evaluated with LIFEx package (LIFEx website) [4]. Specifically, HL [18F]FDG-avid lesions have been identified, semi-automatically segmented by clinical experts and labeled as lymph nodal or extra-nodal. Fifty-two radiomic features have been computed with respect to every region of interest from histograms of grey levels, geometric factors, co-occurrence and higher order zone-length and run-length matrices. After z-score normalization, a feature reduction has been performed as described by the framework below where the pairwise distances were calculated according to the Euclidean distance definition. An intra-patient similarity index (ISI) defined by the silhouette has been built and treated as a proxy of lesions' homogeneity within patients. Specifically, high values of silhouette implies high similarity within patient's lesions and, viceversa, low values were interpreted as high heterogeneity among patient's tumor sites.

Feature reduction and similarity computation framework:

- Step 1 Selection of all the lesions meeting the inclusion criteria of the current analysis, grouped by patients;
- Step 2 Volume-based grouping of radiomics variables: whether covariates show significant correlation (p-value of the chi-squared test > 0.8) or uncorrelation (p-value of the chi-squared test < 0.0001) with respect to lesion volume, they are grouped into two set of features, representing volume-related (*VOL_set*) and heterogeneity-related (*NOVOL_set*) information respectively;
- Step 3 Application of principal component analysis on both *VOL_set* and *NOVOL_set* of covariates, resulting in two new set of features describing 95% of the total variability;
- Step 4 Juxtaposition of *VOL_set* and *NOVOL_set* of covariates to form a representative fingerprint of each lesion in the cohort;

Step 5 Computation of the silhouette (equation 1) within each patient as the comparison between cohesion (equation 2) and separation (equation 3) of his/her lesions:

$$s(P_i) = \frac{b(P_i) - a(P_i)}{maxa(P_i), b(P_i)} \tag{1}$$

where

$$a(P_i) = \frac{1}{|P_i| - 1} \sum_{i,j \in P_i, i \neq j} d(i,j) \tag{2}$$

$$b(P_i) = \min_{k \neq i} \frac{1}{|P_k|} \sum_{i \in P_i, k \in P_k} d(i,k) \tag{3}$$

Step 6 Evaluation of similarity indexes in different groups.

3 Results

First, ISI values have been computed and analyzed within each group, non-R/R and R/R, independently. Specifically, homogeneity across subjects has been assessed including only nodal and both nodal and extra nodal lesions. While accounting for nodal lesions of the two groups, probability density functions were compared; on contrary, when extra nodal lesions were added upon the nodal ones, overall-ISI results have been evaluated in terms of variation with respect to nodal-ISI: increments of values suggest an increasing in the homogeneity of lesions coming from different sites, while decrements are a proxy of heterogeneity in lesions of diverse nature. The analysis shows the different profiles of distributions of non-R/R (0.11 ± 0.42 - 18/26 positive) and R/R (0.24 ± 0.45 - 38/49 positive). Overall, the comparison between distributions suggest a higher intra-patients lesions similarity in the R/R dataset than in the non-R/R one. In the non-R/R dataset, distribution of ISI built upon nodal lesions is centered in -0.01 ± 0.46, increasing up to 0.12 ± 0.61 after adding extra-nodal lesions: positive ISI increases from 4/8 to 6/8. Similarly, the distribution of ISI in R/R dataset evolves from from 0.13 ± 0.46 to 0.42 ± 0.43 as extra-nodal lesions are included beside the nodal ones: positive ISI rise from 12/19 to 17/19. Overall, these results support a higher intra-patients lesions similarity in the R/R dataset than in the non-R/R.

In order to test ISI discrimination power between the two groups, a Logistic Regression (LR) model has been built upon the only ISI value per patient, resulting in a significant odds ratio of 1.85. Basing on this evidence, a classification model has been performed on the ISI-weighted fingerprint covariates: indeed, each lesion's vector of covariates enclosed both radiomic fingerprint - as described above - and its patient's ISI value. ISI thus represents a grouping/weighting factor, suitable for

mantaining the hierarchical inherent nature of data. Specifically, from the original dataset comprising 85 patients and 543 observations, 115 non-R/R and 255 R/R lesions have been randomly sampled to form the training set, while the test set included 57 non-R/R and 116 R/R lesions. Since class imbalance represents a diriment issue to be addressed in medical applications, a Random Undersamplig Boosting of Tree Ensemble (RUBTE) was used for classification purposes [5]. The RUBTE performance has been evaluated in terms of accuracy (82%), sensitivity (70%) and specificity (88%). Furthermore, results at lesion level were aggregated at patient's level through majority voting. Accordingly, true positive rate rose to 89%, although accuracy slightly fell to 73% and recall dropped to 38%.

4 Discussion and conclusion

The present work underlines the different radiomics information entailed on non-R/R and R/R lesions. R/R patients showed higher intra-patients lesion similarity with respect to non-R/R ones, behavior further confirmed while introducing extra-nodal lesions to nodal ones. Indeed, in the non-R/R group, the addition of extra-nodal lesions ones had a minor effect on similarity. Physicians speculated that non-R/R group, naïve from any treatment, including either long term responders, and primary refractory patients is keen on being the most heterogeneous one. Conversely, R/R patients would be biologically more homogeneous, since repeated treatments might result in resistant clones selection.

Despite the current approach suffer from several limitations due to the retrospective design of the study and the reduced sample size, the goal of defining a methodological feature reduction framework to demonstrate the potential predictive value of radiomics in HL has been achieved. Further efforts will be focused on testing our ISI-weighted fingerprint as representative for any kind of lesion at baseline.

Acknowledgements We thank Martina Sollini, Margarita Kirienko, Matteo Biroli and Francesca Ricci for collecting imaging data and performing the radiomc feature extraction at Humanitas Research Hospital in Rozzano; professor Arturo Chiti and professor Carmelo Carlo-Stella for supervising clinical evaluation; Letizia Calderoni, Elena Tabacchi, Cristina Nanni, Pier Luigi Zinzani and Stefano Fanti for providing external validation data fromn Policlinico S. Orsola in Bologna; Anna Guidetti, Alessandra Alessi, Paolo Corradini and Ettore Seregni for gathering and sharing data from Fondazione IRCCS Istituto Nazionale dei Tumori in Milan.

References

1. Ansell, S. M.: Hodgkin lymphoma: 2018 update on diagnosis, risk-stratification, and management. American journal of hematology. **93.5**, 704–715 (2018).
2. Kirienko, M., Sollini, M., Chiti, A.: Hodgkin lymphoma and imaging in the era of anti-PD-1/PD-L1 therapy. Clinical and Translational Imaging. **6.6**, 417–427 (2018).

3. Mottok, A., Steidl, C.: Biology of classical Hodgkin lymphoma: implications for prognosis and novel therapies. Blood. **131.15**, 1654–1665 (2018).
4. Nioche, C., et al.: LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer research **78.16**, 4786–4789 (2018).
5. Seiffert, C., et al.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions On Systems. Man, And Cybernetics-Part A: Systems And Humans. **40.1** (2010).
6. Tubiana, M., et al. Toward comprehensive management tailored to prognostic factors of patients with clinical stages I and II in Hodgkin's disease. The EORTC Lymphoma Group controlled clinical trials. 47–56 (1989).

Prediction of late radiotherapy toxicity in prostate cancer patients via joint analysis of SNPs sequences

Predizione della radiotossicità tardiva per pazienti con tumore alla prostata basata sull'analisi congiunta di sequenze di SNPs

Nicola Rares Franco[1], Michela Carlotta Massi[1,2], Francesca Ieva[1,2,3], Anna Maria Paganoni[1,2,3], Andrea Manzoni[1], Paolo Zunino[1], Tiziana Rancati[4], and Catharine West[5,6]

Abstract Over the past few decades, oncological research and radiotherapy techniques have experienced massive improvements. However, many treatments that are currently used suffer from mild to severe side-effects. Focusing on the case of radiotherapy for prostate cancer patients, we assess the already suggested possibility of predicting long-term radiotoxicity through genomic factors. We construct and validate a radiosensitivity indicator that is based on the analysis of SNPs (Single Nucleotide Polimorphism) sequences. Differently from previous works, we directly account for interactions among SNPs and provide a final score that can be easily incorporated into larger models that allow for additional predictors.

Abstract *Negli ultimi decenni, la ricerca oncologica e le tecniche di radioterapia hanno visto miglioramenti sostanziali. Tuttavia, molti dei trattamenti utilizzati soffrono al momento di effetti collaterali blandi e severi. Concentrandoci sul caso della radioterapia per pazienti con cancro alla prostata, indaghiamo la possibilità, già suggerita da ricerche precedenti, di predirre la radiotossicità tardiva attraverso fattori genetici. Quindi, costruiamo e validiamo un indicatore di radiosensibilità basato sull'analisi di sequenze di SNPs (Single Nucleotide Polimorphism). Diversamente da altri lavori in letteratura, teniamo esplicitamente conto di possibili interazioni tra SNPs e forniamo uno score finale facilmente incorporabile in modelli con predittori aggiuntivi.*

Key words: SNPs, late toxicity, radiotherapy, prostate cancer, score, patterns

[1]MOX Laboratory, Math Department, Politecnico di Milano, Milan, Italy
[2]CADS-Center for Analysis, Decisions and Society, Human Technopole, Milan, Italy
[3]CHRP-National Center for Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy
[4]Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[5]Translational Radiobiology Group, Division of Cancer Sciences, University of Manchester
[6]Manchester Academic Health Science Centre, Christie Hospital, UK

1 Introduction

Radiotherapy (RT) is a widely used treatment for prostate cancer [6] which has the benefit of performing well in terms of patient's survival [9]. On the other hand, it is acknowledged that a significant percentage of men who undergo RT report late toxicity side-effects [1].

Currently the risk of severe toxicity is kept under control as modern treatments have become more and more accurate [4], but mild to moderate effects are still very common (up to 50% of the treated patients) [6]. These issues can significantly impact patients' lives and play an important role when men consider treatment options [2], which is why new suitable prevention tools are needed.

During the last decade, many authors have investigated the factors that determine radiosensitivity (such as age, abdominal surgeries etc. [8]), as well as the possibility of making accurate predictions before the treatment. An emerging approach is that of including genetic risk factors, which is mainly motivated by the evidence that radiosensitivity is a heritable human trait [12]. Several works have been conducted in this direction (e.g. [5, 3, 7, 9]), highlighting deep connections between the RT-caused late toxicity and certain genetic mutations. In particular, each mutation was associated with a different SNP, which means that it involved variations of a single nucleotide at a specific position in the genome.

Our work, which is conducted in collaboration with Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, Italy), falls in this branch of research and can be seen as a natural continuation of [9]. There, by combining machine learning and statistical techniques, the authors were able to identify the SNPs associated with long-term toxicity through the use of a deep sparse autoencoder [10]. Nevertheless, the deep learning approach adopted in [9] had the benefit of accounting for possible interactions among SNPs.

Starting from here, we move towards the problem of making reliable predictions. At this purpose, we propose a radiosensitivity indicator that is based on genomic information encoded in sequences (or patterns) rather than standalone SNPs. We validate our model on real data and discuss possible variants as well as future applications.

2 Models and methods

We assess the problem of predicting the expected presence (or absence) of a certain long-term RT side-effect O, in terms of a -fixed- group of p SNPs $\left\{S^1,..,S^p\right\}$. In full generality, this corresponds to studying a model of the type

$$\mathbf{E}\left[O|S^1,..,S^p\right] = f\left(S^1,..,S^p\right)$$

and the goal is to properly estimate f. In practice though, determining f through real data is an ill-posed problem up until one reduces the search to a suitable family

of functions [11]. For instance, if one requires f to be of the form $f(x_1,..,x_p) = \sigma(\beta_0 + \sum \beta_i x_i)$, where $\sigma(x) := \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{x}{2}\right)$ is the sigmoid while the β's are the parameters to be estimated, then everything reduces to the well known case of classical logistic regression; however, since we want to account for interactions among SNPs, we avoid this approach.

Our idea is to consider models of the form:

$$f\left(S^1,..,S^p\right) = \sigma\left(\mu_0 + \alpha|\mathscr{R}|^{-1}\sum_{R\in\mathscr{R}}\mathbf{1}_R\left(S^1,..,S^p\right) + \beta|\mathscr{P}|^{-1}\sum_{P\in\mathscr{P}}\mathbf{1}_P\left(S^1,..,S^p\right)\right) \tag{1}$$

where $\sigma : \mathbf{R} \to [0,1]$ is a fixed link function (e.g. the sigmoid), while $\mathscr{R}$ and $\mathscr{P}$ are suitable collections of SNPs sequences, whereas μ_0, α and β are scalar coefficients to be estimated during the training phase.

2.1 SNPs patterns

Following the classical SNPs encoding, each SNP S^j is a tricotomic variable taking values in $\{0,1,2\}$. Consequently, we may define the collection of all possible SNPs patterns (or sequences) to be

$$\mathscr{S} = \left\{A_1 \times ... \times A_p \quad | \quad A_i = \{0\},\ \{1\},\ \{2\},\ \{0,1,2\}\right\}$$

For a pattern $\mathbf{S} = A_1 \times ... \times A_p \in \mathscr{S}$, we define its length as

$$L(\mathbf{S}) := |\{A_i \neq \{0,1,2\}\}|$$

For instance, if $p = 3$ then $\mathbf{S} := \{0\} \times \{0,1,2\} \times \{2\}$ is a pattern of length 2, where the first SNP equals 0 while the third one equals 2 (in fact, by definition one has $\{(S^1,\, S^2,\, S^3) \in \mathbf{S}\} \iff \{S^1 = 0,\, S^3 = 2\}$). As the cardinality of $\mathscr{S}$ grows exponentially with p, it is clear that a model based on such sequences will -somewhere- need a feature selection procedure.

2.2 Radiosensitivity score

We are given a dataset $\left\{(s_i^1,..,s_i^p,o_i)\right\}_{i=1}^N$ consisting of N i.i.d. observations, each reporting the values of the p SNPs and the presence/absence (respectively 1 or 0) of the late RT side-effect in a different individual.

The first step is to extract the inferable SNPs patterns. This can be done by defining a subset $\mathscr{S}_0 \subseteq \mathscr{S}$ consisting of patterns that show up with a sufficient frequency in both the group of healthy $\{o_i = 0\}$ and diseased $\{o_i = 1\}$ patients. If p is large,

the search of patterns can become computationally heavy: at this purpose it can be convenient to consider only patterns having length less or equal than a fixed $L_0 < p$.

Next we study the effect of each single pattern $\mathbf{S} \in \mathscr{S}_0$ over the outcome O. One way of doing this is determining the odds-ratio $OR(\mathbf{S})$, which in practice corresponds to studying the joint behavior of $\mathbf{1}_{\mathbf{S}}\left(S^1,...,S^p\right)$ and O over the sample. This allows us to split the patterns into two groups

$$\mathscr{R}_0 = \{\mathbf{S} \mid OR(\mathbf{S}) > 1\}, \qquad \mathscr{P}_0 = \{\mathbf{S} \mid OR(\mathbf{S}) < 1\}$$

respectively the risk and the protective patterns. In general, due their combinatorial nature, the cardinality of both $\mathscr{R}_0$ and $\mathscr{P}_0$ will be huge. As a remedy, we propose a feature selection algorithm based on an elbow analysis and a suitable metric over the space of patterns: at the end of this procedure, we obtain the two subsets $\mathscr{R} \subset \mathscr{R}_0$ and $\mathscr{P} \subset \mathscr{P}_0$.

Finally, according to equation (1), we estimate the coefficients μ_0, α and β through logistic regression. We define the radiosensitivity score of the ith individual as

$$RS_i = \alpha|\mathscr{R}|^{-1} \sum_{R\in\mathscr{R}} \mathbf{1}_R\left(s_i^1,..,s_i^p\right) + \beta|\mathscr{P}|^{-1} \sum_{P\in\mathscr{P}} \mathbf{1}_P\left(s_i^1,..,s_i^p\right)$$

Experiments over real data show that the *RS* score can perform very well in terms of predicting long-term radiotoxicity (for some side effects we obtain classifiers having AUCs greater than 0.75). Furthermore, the coefficients α and β seemingly always happen to be statistically significant and have the expected signs, i.e. $\alpha > 0$ and $\beta < 0$: this is of particular interest as it allows us to interpret the *RS* score as a weighted sum of a risk and a protection score.

3 Future steps

We propose a novel approach for predicting radiosensitivity that, unlike other works in the literature, accounts for joint effects of multiple SNPs. We believe that our promising results can have a significant impact in our understanding of how genetic factors are encoded through SNPs and how they affect late RT toxicity.

As future steps we wish to investigate possible variants of our algorithm, as well as the possibility of incorporating the *RS* score into larger models where the late toxicity is infered through a mix of genetic and classical factors.

References

1. Bentzen, S.M., Constine, L.S., Deasy, J.O., et al (2010). Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *Int. J. Radiat. Oncol. Biol. Phys.* 76, S3–S9.

2. Davison, B.J., Gleave, M.E., Goldenberg, S.L., et al (2002). Assessing information and decision preferences of men with prostate cancer and their partners. *Cancer Nurs.* 25, 42–49.
3. Fachal, L., Gómez-Caamaño, A., Barnett, G.C., Peleteiro, P., Carballo, A.M., Calvo-Crespo, P., et al (2014). A three-stage genome-wide association study idenfities a susceptibility locus for late radiotherapy toxicity at 2q24. 1. *Nature genetics* 46, 891.
4. Garibaldi, C., Jereczek-Fossa, B.A., Marvaso, G., et al (2017). Recent advances in radiation oncology. *Ecancermedicalscience*. 2017;11:785.
5. Kerns, S.L., Stock, R.G., Stone, N.N., et al (2013). Genome-wide association study identifies a region on chromosome 11q14. 3 associated with late rectal bleeding following radiation therapy for prostate cancer. *Radiotherapy and Oncology* 107, 372-376.
6. Kerns, S.L., Dorling, L., Fachal, L., et al (2016). Meta-analysis of Genome Wide Association Studies Identifies Genetic Markers of Late Toxicity Following Radiotherapy for Prostate Cancer. *EBioMedicine* 10, 150–163.
7. Kerns, S.L., Fachal, L., Dorling, L., et al (2019). Radiogenomics consortium genome-wide association study meta-analysis of late toxicity after prostate cancer radiotherapy. *JNCI: Journal of the National Cancer Institute*, Volume 112, Issue 2, February 2020, Pages 179–190
8. Landoni, V., Fiorino, C., Cozzarini, C., et al (2016). Predicting toxicity in radiotherapy for prostate cancer. *Physica Medica* 32, 521-532.
9. Massi, M.C., Gasperoni, F., Ieva, F., et al (2020). A Deep Learning approach for validating the effect of SNPs on late RT toxicity for prostate cancer patients. *Frontiers in Oncology*, submitted.
10. Massi, M.C., Ieva, F., Gasperoni, F. and Paganoni, A.M. (2019). Minority Class Feature Selection through Semi-Supervised Deep Sparse Autoencoders. *MOX-Report* no. 38/2019, Dipartimento di Matematica, Politecnico di Milano, 2019. [Online] https://www.mate.polimi.it/biblioteca/add/qmox/38-2019.pdf
11. Poggio, T. and Smale, S. (2005). The Mathematics of Learning: Dealing with Data. *Notices Am Math Soc.* 50. PL-5. 10.1109/ICNNB.2005.1614546.
12. West, C.M. and Barnett, G.C. (2011). Genetics and genomics of radiotherapy toxicity: towards prediction. *Genome Med.* 3, 52.

Predictive versus posterior probabilities for phase II trial monitoring

Confronto tra l'uso di probabilità predittive e a posteriori per il monitoraggio di prove cliniche di fase II

Valeria Sambucini

Abstract Bayesian monitoring of clinical trials is typically based on posterior or predictive probabilities. In the first case, the decision rules are based on the posterior probability that the experimental treatment shows the required performance, given the interim data. In the second case, the idea is to evaluate the predictive probability of observing a positive result if the trial were to continue to its pre-specified maximum sample size. In this paper, we compare the two strategies when applied to a single-arm phase II trial based on binary efficacy and toxicity endpoints.

Abstract *Le procedure Bayesiane per monitorare gli studi clinici si basano tipicamente sul computo di probabilità a posteriori o predittive. Nel primo caso, le regole decisionali si basano sulla probabilità a posteriori che il trattamento sperimentale mostri le prestazioni richieste, date le informazioni ad interim. Nel secondo caso, si valuta la probabilità predittiva di osservare un risultato positivo al termine della prova clinica, quando la dimensione campionaria massima prestabilita è stata raggiunta. Oggetto di querto lavoro è il confronto tra le due strategie quando applicate a uno studio di fase II a braccio singolo basato su due endpoint binari di efficacia e tossicità.*

Key words: Bayesian monitoring, binary bivariate endpoint, phase II clinical trials, posterior probabilities, predictive probabilities.

1 Introduction

The Bayesian approach is particularly suitable to interim monitoring of clinical trials, since it inherently allows to incorporate the evidence from current data in

Valeria Sambucini
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: valeria.sambucini@uniroma1.it

the interim rules. The methodologies typically exploit either posterior probabilities or predictive probabilities.

In general terms, let us assume that a phase II single-arm trial with a maximum of N patients is conducted to test the null hypothesis of no treatment effect $H_0 : \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_1$. Here, θ denotes the parameter of interest that measures the "goodness" of the experimental treatment. When the interim data d_n have been observed among the first n patients, the futility stopping rule based on posterior probabilities consists in terminating the trial, declaring the new treatment not enough promising, if the posterior probability assigned to the alternative hypothesis is not sufficiently high, that is if

$$Pr(\theta \in \Theta_1 | d_n) < \lambda,$$

where λ is a pre-specified threshold.

As an alternative to this procedure, there are decision rules which exploit accumulated data to evaluate the posterior predictive probability of claiming that the experimental treatment is promising at the scheduled end of the trial. More specifically, let us denote by D_{N-n} the random future data, whose posterior predictive distribution conditional on the current information is $m(\cdot|d_n)$. We also assume that at the end of the study the treatment will be declared promising if the posterior quantity $Pr(\theta \in \Theta_1 | d_n, D_{N-n})$ will exceed a desired threshold λ. Then, the futility decision rule at the interim look is to stop the trial early if we obtain low values for the predictive probability

$$\mathbb{P}_{m(\cdot|d_n)}\Big(Pr(\theta \in \Theta_1 | d_n, D_{N-n}) > \lambda\Big),$$

where $\mathbb{P}_{m(\cdot|d_n)}$ denotes the probability measure associated with $m(\cdot|d_n)$.

Monitoring procedures based on both the approaches for phase II trials that consider a single binary response variable have been widely discussed in the literature. In this context, Yin [5] (section 5.6.5) provides a comparison between using posterior and predictive probabilities for trial monitoring: the main difference that comes to light is that the predictive rules, differently from the posterior ones, are able to compromise the current information and the amount of future data.

In this paper, we focus on two binary endpoints of interest. As a typical application, we can consider a single-arm phase II trial based on efficacy and toxicity binary outcomes. Several reasons why phase II trials should formally incorporate also toxicity outcomes are provided in [1]. Different decision rules to monitor both the endpoints by exploiting posterior probabilities have been presented since the 90s (see [4] and [6], among others), whereas the use of the predictive approach has been proposed more recently (see [2] and [3]). In line with Yin [5], we aim at comparing the two approaches by highlighting the main differences of the decision rules based on the same scenarios assumed to be observed at the interim look.

The outline of the article is as follows. In Section 2 we formalize the problem and describe the Bayesian monitoring procedures based on posterior and predictive

probabilities. In Section 3 a numerical comparison is provided. Finally, Section 4 contains some concluding remarks.

2 Bayesian monitoring rules based on two binary endpoints

When efficacy and toxicity binary outcomes are of interest, there are 4 exclusive events that each patient may experience: $1 = (E,T)$, $2 = (E,T^C)$, $3 = (E^C,T)$ and $4 = (E^C T^C)$, where E and T denote *efficacy* and *toxicity*, whereas E^C and T^C indicate *no efficacy* and *no toxicity*, respectively.

Let N be the maximum achievable sample size. For $k = 1,2,3,4$ and given n current patients, let us denote by x_k the number of subjects who experienced event k and by θ_k the probability that event k occurs, with $\sum_{i=1}^{4} x_i = n$ and $\sum_{i=1}^{4} \theta_i = 1$. The vector of interim data, $\mathbf{x} = (x_1,x_2,x_3,x_4)$, comes from a multinomial distribution with parameter $\theta = (\theta_1,\theta_2,\theta_3,\theta_4)$. From standard conjugate analysis, if we introduce a Dirichlet prior distribution for θ, $\theta \sim \text{Dir}(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$, the corresponding posterior distribution is

$$\theta|\mathbf{x} \sim \text{Dir}(s_1,s_2,s_3,s_4),$$

where $s_k = \alpha_k + x_k$ for $k = 1,2,3,4$.

The most used approach to monitor both the bivariate endpoints consists in exploiting marginal models. In practice, the stopping rules are based on two separated conditions that involve $\theta_E = \theta_1 + \theta_2$ and $\theta_T = \theta_1 + \theta_3$, i.e. the marginal probabilities of efficacy and toxicity, whose posterior densities are

$$\theta_E|\mathbf{x} \sim \text{Beta}(s_1 + s_2,\ s_3 + s_4)$$
$$\theta_T|\mathbf{x} \sim \text{Beta}(s_1 + s_3,\ s_2 + s_4).$$

One possible way of proceeding is to specify two target values of interest, θ_E^* and θ_T^*, and focus on the posterior probabilities that $\theta_E > \theta_E^*$ and $\theta_T < \theta_T^*$, in order to define the treatment success both at the interim look or at the end of the study.

2.1 Predictive probabilities vs posterior probabilities

At the interim stage where n patients have been observed, the rules based on posterior probabilities establish that the trial continues and additional patients need to be enrolled if

$$Pr(\theta_E > \theta_E^*|\mathbf{x}) > \gamma_E \qquad \text{and} \qquad Pr(\theta_T < \theta_T^*|\mathbf{x}) > \gamma_T, \tag{1}$$

where γ_E and γ_T denote two pre-specified thresholds. In practice, it is required to obtain from the current data sufficiently high posterior probabilities that the new treatment is enough effective and that it is not highly toxic.

To implement the procedures based on predictive probabilities, we need to consider the future efficacy and toxicity data that will be obtained among the remaining $N-n$ patients, $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$. We have that the posterior predictive distribution of $\mathbf{Y}$, conditional on $\mathbf{x}$, is the Dirichlet-Multinomial distribution,

$$m_{N-n}(\mathbf{y}|\mathbf{x}) = \text{DirMultinom}(N-n, \mathbf{s}), \tag{2}$$

where $\mathbf{s} = (s_1, s_2, s_3, s_4)$. If we observe the future data $\mathbf{Y} = \mathbf{y}$ at the end of the trial, the posterior distributions of θ_E and θ_T will be

$$\begin{aligned} \theta_E|\mathbf{x}, \mathbf{y} &\sim \text{Beta}(h_1 + h_2,\ h_3 + h_4) \\ \theta_T|\mathbf{x}, \mathbf{y} &\sim \text{Beta}(h_1 + h_3,\ h_2 + h_4), \end{aligned}$$

where $h_k = \alpha_k + x_k + y_k$ for $k = 1, 2, 3, 4$. Moreover, the treatment will be declared promising if

$$Pr(\theta_E > \theta_E^*|\mathbf{x}, \mathbf{y}) > \lambda_E \qquad \text{and} \qquad Pr(\theta_T < \theta_T^*|\mathbf{x}, \mathbf{y}) > \lambda_T, \tag{3}$$

where λ_E and λ_T denote two probability cutoffs, reasonably chosen as high values. Therefore, we can compute the predictive probability of a successful conclusion should the trial be conducted to the maximum planned sample size by summing the predictive probabilities of all the possible future outcomes that simoultaneuously satisfy the conditions in (3). Thus, the predictive probability of interest is

$$PP = \sum_{\mathbf{y}} m_{N-n}(\mathbf{y}|\mathbf{x}) I\Big\{ Pr(\theta_E > \theta_E^*|\mathbf{x}, \mathbf{y}) > \lambda_E \Big\} I\Big\{ Pr(\theta_T < \theta_T^*|\mathbf{x}, \mathbf{y}) > \lambda_T \Big\}, \tag{4}$$

where $I\{\cdot\}$ is the indicator function. Then, the futility rule establishes that the trial continues if PP exceeds a desired threshold θ_L, generally chosen as a value less than 0.2. Otherwise the trial terminates. In other words, additional patients are enrolled if it is not unlikely that the experimental treatment will show the desired performance in terms of efficacy and safety at the conclusion of the study.

3 Numerical comparison

To make a comparison between using posterior and predictive probabilities for trial monitoring in the presence of two binary efficacy and toxicity outcomes, we consider a hypothetical scenario observed *ad interim*. More specifically, we assume that 9 responses $(x_1 + x_2)$ and 5 toxicities $(x_1 + x_3)$ have been observed among $n = 20$ current patients. By setting $\theta_E^* = 0.45$, $\theta_T^* = 0.25$ and $\alpha = (0.25, 0.25, 0.25, 0.25)$, we have that the posterior probabilities of interest are $Pr(\theta_E > \theta_E^*|\mathbf{x}) = 0.503$

and $Pr(\theta_T < \theta_T^*|\mathbf{x}) = 0.482$, whatever are the maximum planned sample size and the joint counts of the current data. On the contrary, the predictive probability in (4) strongly depends on N and on the joint structure of the interim data. This is because it involves the posterior predictive distribution in (2), that provides different probability values for different combinations of the element of $\mathbf{x} = (x_1, x_2, x_3, x_4)$ compatible with the same marginal number of total responses and total toxicities.

In the specific case we are considering, we can express all the possible current data consistent with 9 responses and 5 toxicities in terms of the number of observed responses without toxicities (i.e. x_2). In Figure 1 we show the behaviour of PP as a function of x_2 for different values of N, when $\theta_E^* = 0.45$, $\theta_T^* = 0.25$, $\lambda_E = \lambda_T = 0.9$ and $\alpha = (0.25, 0.25, 0.25, 0.25)$. As x_2 increases, the current data provide more support to the experimental treatment and, as a consequence, the predictive probability of declaring the treatment success at the planned end of the study increases. Moreover, the higher N, the larger PP. This is because, given the current information, for low values of the number of remaining patients it is unlikely that the posterior probabilities of a high efficacy and a low toxicity will reach the thresholds 0.9 at the end of the study. On the contrary, if $N - n$ is large, the strength of the interim data is less relevant and the predictive probability of interest grows. In Table 1 the variations of PP as a function of x_2 and N are provided for different couples of values of θ_E^* and θ_T^*. The computations of the posterior probabilities of interest are also shown at the bottom of the table. As expected, both the predictive and the posterior probabilities decreases as the target values become harder to reach (i.e. θ_E^* increases and/or θ_T^* decreases).

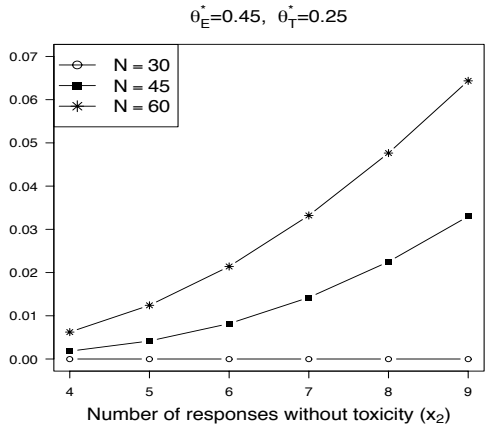


Fig. 1 Behaviour of PP as a function of all possible values of x_2 consistent with the fixed number of total responses ($x_1 + x_2 = 9$) and total toxicities ($x_1 + x_3 = 5$) for different values of N, when $n = 20$, $\lambda_E = \lambda_T = 0.9$ and $\alpha = (0.25, 0.25, 0.25, 0.25)$.

Table 1 Variations of PP for different values of x_2, θ_E^*, θ_T^* and N, when $\lambda_E = \lambda_T = 0.9$ and $\alpha = (0.25, 0.25, 0.25, 0.25)$. The values of the posterior probabilities in (1) are also provided.

	$\theta_E^* = 0.5,\ \theta_T^* = 0.2$			$\theta_E^* = 0.45,\ \theta_T^* = 0.25$			$\theta_E^* = 0.4,\ \theta_T^* = 0.3$		
x_2	$N = 30$	$N = 45$	$N = 60$	$N = 30$	$N = 45$	$N = 60$	$N = 30$	$N = 45$	$N = 60$
4	0.0000	0.0000	0.0005	0.0000	0.0018	0.0062	0.0022	0.0249	0.0622
5	0.0000	0.0001	0.0012	0.0000	0.0042	0.0124	0.0048	0.0415	0.0904
6	0.0000	0.0002	0.0027	0.0000	0.0082	0.0214	0.0092	0.0616	0.1203
7	0.0000	0.0006	0.0053	0.0000	0.0142	0.0332	0.0156	0.0846	0.1513
8	0.0000	0.0012	0.0093	0.0000	0.0225	0.0476	0.0242	0.1104	0.1835
9	0.0000	0.0021	0.0148	0.0000	0.0330	0.0644	0.0347	0.1387	0.2170
	$Pr(\theta_E > 0.5\|\mathbf{x}) = 0.328$			$Pr(\theta_E > 0.45\|\mathbf{x}) = 0.503$			$Pr(\theta_E > 0.4\|\mathbf{x}) = 0.680$		
	$Pr(\theta_T < 0.2\|\mathbf{x}) = 0.275$			$Pr(\theta_T < 0.25\|\mathbf{x}) = 0.482$			$Pr(\theta_T < 0.3\|\mathbf{x}) = 0.677$		

4 Conclusion

Similarly to the case of a single binary endpoint, also when dealing with two binary outcomes, monitoring rules based on predictive probabilities represent a compromise between interim data and future sample size. By contrast, the rules based on posterior probabilities are not affected by the remaining number of subjects in the study.

More interestingly, if the definition of a successful treatment is based on conditions imposed on the marginal probabilities of efficacy and toxicity, monitoring strategies based on predictive probabilities enables to account for the joint structure of the interim data. Thus, the current information is fully exploited, taking in account also the association between the two endpoints.

References

1. Petroni, G.R.,Conaway, M.R.: Designs Based on Toxicity and Response. In: Handbook of Statistics in Clinical Oncology. Third edition. Crowley J, Hoering A, editor. Chapman and Hall/CRC. (2012)
2. Sambucini,V.: Bayesian predictive monitoring with bivariate binary outcomes in phase II clinical trials. Comput. Stat. Data Anal., **132**, 18–30 (2019)
3. Sambucini,V.: Efficacy and toxicity monitoring via Bayesian predictive probabilities in phase II clinical trials. Submitted, (2020)
4. Thall, P.F., Simon, R., Estey, E.H.: Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat. in Med., **14**, 357–379 (1995).
5. Yin, G.: Clinical Trial Design: Bayesian and Frequentist Adaptive Methods. Wiley, Hoboken. (2012)
6. Zhou, H., Lee, J.J.,Yuan, Y.: BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. Stat. in Med., **36**, 3302–3314 (2017)

Profile networks for precision medicine

Reti profilo per la medicina di precisione

Andrea Lazzerini · Monia Lupparelli · Francesco C. Stingo

Abstract We propose a class of profile graphical models to model the effect of an external factor on the dependence structure of a multivariate set of variables. The main aim is to provide a joint representation based on a single graph of the probability distribution of a multivariate random vector given different levels of an external factor. In particular, we explore the marginal dependence structure by using the subclass of bi-directed profile graphical models and we show that the selected graphical model is compatible with a two block regression graph. An application is discussed based on protein networks in various subtypes of acute myeloid leukemia.
Abstract *Proponiamo una classe di modelli grafici profilo per studiare l'effetto di un fattore esterno sulla struttura di dipendenza di un insieme di variabili. Lo scopo è rappresentare congiuntamente in un unico grafo la distribuzione di probabilità congiunta di un vettore casuale di variabili, dati diversi livelli di un fattore esterno. In particolare, si esplora la struttura di dipendenza marginale utilizzando la sotto classe di modelli grafici profilo bi-direzionali e si mostra la compatibilità del modello grafico selezionato con un grafo di regressione a due blocchi. Si discute un'applicazione su reti di proteine per diversi sottotipi di leucemia mieloide acuta.*

Key words: multiple graphs, regression graphs.

Andrea Lazzerini
University of Bologna, Via Zamboni, 33 - 40126 Bologna, e-mail: andrea.lazzerini2@unibo.it

Monia Lupparelli
University of Florence, Piazza di San Marco, 4 - 50121 Firenze, e-mail: monia.lupparelli@unifi.it

Francesco C. Stingo
University of Florence, Piazza di San Marco, 4 - 50121 Firenze, e-mail: francescoclaudio.stingo@unifi.it

1 Introduction

We propose a novel class of graphical models to study the effect of an external factor on the marginal dependence structure of a multivariate random vector of outcomes. Exploring marginal independencies in a multivariate setting may provide relevant insights when the association structure is induced by ignoring the effect of latent/unobservable variables.

Let $Y_V = (Y_i)_{i \in V}$ be a random vector of outcomes indexed by the finite set V with $p = |V|$. Also, let us consider the categorical variable X representing an external factor for the random vector Y_V. The variable X takes level $x \in \mathscr{X}$ with $q = |\mathscr{X}|$. For any $x \in \mathscr{X}$, let $Y_V(x)$ be a *profile outcome vector*, that is the random vector $Y_V|\{X = x\}$ conditioned on a specific profile x of the factor X, and let $P[Y_V(x)]$ be the corresponding *profile outcome distribution*, that is the conditional probability distribution $P(Y_V|\{X = x\})$. Then, for a given multivariate random vector Y_V and an external factor X, let $Y_{V|\mathscr{X}} = [Y_V(x)]_{x \in \mathscr{X}}$ be the finite set of all profile outcome vectors and let $P(Y_{V|\mathscr{X}}) = [P(Y_V(x))]_{x \in \mathscr{X}}$ be the corresponding set of all profile outcome distributions. Given a partition $A, B, C \subseteq V$, the *profile marginal independence* $Y_A(x) \perp\!\!\!\perp Y_B(x)$ corresponds to the factorization

$$P[Y_A(x), Y_B(x)] = P[Y_A(x)] \times P[Y_B(x)], \quad x \in \mathscr{X}. \tag{1}$$

of the joint profile distribution $Y_V(x)$.

Let $B = (V, E_B)$ be a *bi-directed graph* defined by a set of vertices $i \in V$ and a set of edges $(i, j) \in E_B$, drawn as bi-directed edges, joining pairs of vertices $i, j \in V$. Vertices $i \in V$ are associated to variables Y_i of a random vector Y_V and, under suitable Markov properties, missing edges in the graph corresponds to marginal independencies for the joint probability distribution $P(Y_V)$. For the pairwise Markov property, a missing edge $(i, j) \notin E_B$ corresponds to the marginal independence $Y_i \perp\!\!\!\perp Y_j$, for any $i, j \in V$. For technical aspects see [4]. A *regression graph* $R = [\{R_T\}_{T \in \mathscr{T}}, E_R]$ is defined by a set of vertices partitioned in components R_T, for any T in a finite set $\mathscr{T}$ and a set of edges E_R. Vertices i, j within any component R_T can be joined by bi-directed edges and vertices between components, $i \in R_T$ and $j \in R_{T'}$ are joined by directed edges preserving the same direction such that cycles are not allowed, for any $T, T' \in \mathscr{T}$, with $T \neq T'$. Regression graph models allow to explore the dependence structure of Y_V given X and the effect of X on each $Y_i \in Y_V$. The corresponding Markov properties and more details are given in [1]. An example of bi-directed and regression graph can be found in Fig. 1 and Fig. 2 respectevely (ignore for the moment the first graph on the left inside of both figures). Consider the bi-directed graph $B(0) = (V, E_{B_0})$ where $V = \{a, b, c, d\}$ and $E_{B_0} = \{(a, b), (b, c), (b, d)\}$. For the pairwise Markov property it holds for instance that $Y_a \perp\!\!\!\perp Y_d$. Consider now the regression graph $R = [\{R_T\}_{T \in \mathscr{T}}, E_R]$ where $\mathscr{T} = \{\{X\}, \{a, b, c, d\}\}$ and $E_R = \{(X, a), (X, b), (a, b), (a, c), (b, c), (b, d)\}$. For the pairwise Markov property it hold for example that $X \perp\!\!\!\perp Y_d$ and $Y_c \perp\!\!\!\perp Y_d | X$.

The drawback of regression graph models is that they do not provide information on how the joint dependence structure of Y_V may considerably vary for any

level $x \in \mathscr{X}$. From this perspective, useful insights are given by the use of multiple graphical models. A collection of *multiple bi-directed graphs* $B_{V|\mathscr{X}} = \{B(x) = (V, E_B(x))\}_{x \in \mathscr{X}}$ associated to the profile outcome distributions $P(Y_{V|\mathscr{X}})$ represents a class of independence models for any profile outcome vector $Y_V(x)$, with $x \in \mathscr{X}$. Basically, any graph $B(x) \in B_{V|\mathscr{X}}$ represents the marginal independence structure of the outcomes Y_V given a profile $x \in \mathscr{X}$ of the factor X. These graphs may reasonably have different skeletons for any $x \in \mathscr{X}$, so that the outcome independence structure is not invariant with respect to different factor levels. On the other side, the drawback of multiple graphs is that they do no provide information about the effect of the external factor X on single outcomes $Y_i \in Y_V$; for further details see [2, 6].

In essence, our idea is to provide a single graph able to embed, at the same time, information about the effect of an external factor on the values of a set of outcomes, as a regression graph does, and on the dependence structure of each profile outcome, as a collection of multiple graphs does. Regression graphs and multiple graphs are special cases of the proposed profile graphical models.

2 Profile bi-directed graphical models

A *profile bi-directed graph* $\mathscr{G}_B = (V, \mathscr{E}_B)$ is defined by the full set V of vertices and a $(p \times p)$ matrix $\mathscr{E}_B$ of $\mathscr{Z}$-labelled edges, drawn as bi-directed line, which are labelled according to any subset $\mathscr{Z} \subseteq \mathscr{X}$. Let $(i, j)^{\mathscr{Z}}$ be the generic element of $\mathscr{E}_B$ associated to any couple $i, j \in V$, where the presence or non-presence of the edge between i and j is univocally determined by the subset $\mathscr{Z}$ of the state space $\mathscr{X}$. Given $(i, j)^{\mathscr{Z}} \in \mathscr{E}_B$ for any pair $i, j \in V$, if $\mathscr{Z} = \mathscr{X}$, vertices i and j are not coupled by any edge, if $\mathscr{Z} \subset \mathscr{X}$, we have that vertices i and j are coupled by a $\mathscr{Z}$-labelled edge. In particular, if $\mathscr{Z}$ is a nonempty proper subset of $\mathscr{X}$, $\mathscr{Z} \subset \mathscr{X}$ and $\mathscr{Z} \neq \emptyset$, vertices i and j are joined by a dotted $\mathscr{Z}$-labelled edge, if $\mathscr{Z} = \emptyset$, vertices are joined by a full edge and, for sake of simplicity, the $\emptyset$-label is removed by the edge.

Under suitable Markov properties the profile graph $\mathscr{G}_B$ provides an independence model for the joint distributions of a random vector $Y_{V|\mathscr{X}}$ of profile outcomes. In particular, a missing edge in $\mathscr{G}_B$ corresponds to a profile marginal independence for every level $x \in \mathscr{X}$, and, therefore, to a marginal independence for the entire vector $Y_{V|\mathscr{X}}$. A $\mathscr{Z}$-labelled dotted edge in $\mathscr{G}_B$ corresponds to profile marginal independencies holding only for the levels $x \in \mathscr{Z}$, with $\mathscr{Z} \subset \mathscr{X}$. For instance, under the *profile bi-directed pairwise Markov property* we have that (i) a missing edge for a pair $i, j \in V$ of vertex implies $Y_i(x) \perp\!\!\!\perp Y_j(x)$, for any $x \in \mathscr{X}$; (ii) a $\mathscr{Z}$-labelled dotted edge for a pair $i, j \in V$ of vertex implies $Y_i(x) \perp\!\!\!\perp Y_j(x)$, for any $x \in \mathscr{Z}$.

Finally, we assume a partition of the vertex set $V = V_\square \cup V_\bigcirc$ in order to specify the set of vertices which are independent or dependent, respectively, of the external factor X. Then, any vertex $i \in V_\square$ is drawn as a square vertex, while any $i \in V_\bigcirc$ is drawn as a circled vertex. In the class of profile graphs we consider, any pair $i, j \in V_\square$ are not allowed to be joined by a $\mathscr{Z}$-labelled dotted edge, for any nonempty $\mathscr{Z} \subset \mathscr{X}$. Also, given the partition $V = V_\square \cup V_\bigcirc$, for any $i \in V_\square$, we have that $Y_i \perp\!\!\!\perp X$.

We give an illustrative example. Figure 1 shows a profile bi-directed graph associated with the profile outcome vector $Y_{V|\mathscr{X}}$, with $V = \{a,b,c,d\}$, and the corresponding induced class of multiple bi-directed graphs $B(x)$, $\forall x \in \mathscr{X} = \{0,1,2\}$. Figure 2 shows the same profile graph of 1 with square vertices according to the compatible regression graph R. Under the pairwise Markov property, for example, we have that $Y_a(x) \perp\!\!\!\perp Y_c(x)$ for $x = 0$ because the two nodes are joined by a dotted $\{0\}$-labelled edge, while $Y_c(x) \perp\!\!\!\perp Y_d(x)$ for any $x \in \mathscr{X}$ because there is no edge between the two nodes. There is no independence between b and d since they are joined by a full edge. Furthermore, the partition $V = V_\square \cup V_\bigcirc$ of the vertex set provides information about the dependence of each vertex $Y_i \in Y_V$ with respect to X. For instance, we have that $Y_c \perp\!\!\!\perp X$. Notice that multiple graphs imply all the profile marginal independencies, however they do not provide information about the dependence of any $Y_i \in Y_V$ with respect to X while regression graphs imply the latter but not the profile marginal independencies when $\mathscr{Z} \subset \mathscr{X}$, $\mathscr{X} \neq \emptyset$. Conversely, profile graphs involve them all.

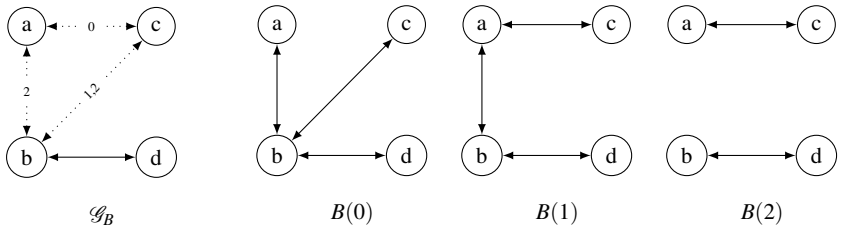


Fig. 1 A profile bi-directed graph with the induced class of multiple bi-directed graphs.

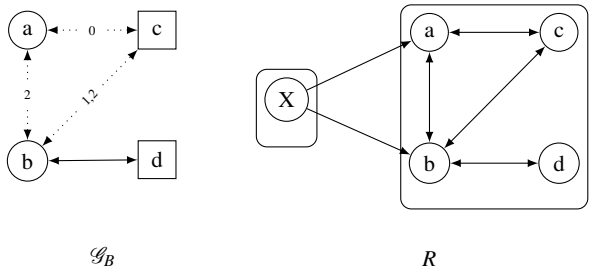


Fig. 2 A profile bi-directed graph with a compatible regression graph.

3 Protein networks

The proposed method is used to analyze protein expression data from patients affected by acute myeloid leukemia (AML). The goal of this analysis is to reconstruct a protein network for each subtype of the disease; comparing the networks for these groups provides insight into the differences in protein signaling that may affect whether treatments for one subtype will be effective in another.

A set of protein levels is observed in a sample of 213 newly diagnosed AML patients. Data are available upon request at the MD Anderson Department of Bioinformatics and Computational Biology. Protein expression were collected using the reverse phase protein array (RPPA) technology. For previous work on inference of protein networks from RPPA data see [7] . Patients are classified by subtype according to the French-American-British (FAB) classification system. The subtypes, based on criteria including cytogenetics and cellular morphology, show varying prognosis. Then, it is reasonable to expect a different protein interactions in the subtypes. A set of 18 proteins is considered and they are known to be involved in apoptosis and cell cycle regulation according to the KEGG database [3]. We consider 4 AML subtypes, for which a reasonable sample size is available: M0 (17 subjects), M1 (34 subjects), M2 (68 subjects), and M4 (59 subjects).

Our interest is modelling the effect of the AML subtype on the joint dependence structure of the protein levels and, in particular the interest is exploring via a graphical modelling approach how this structure may change under different AML subtype. We are also interested in changes of expression across subtypes. Profile graphical models are an encompassing tool that coherently and jointly performs all inferential tasks of interest. Therefore, considering the $p = 18$ protein levels following a multivariate Gaussian distribution and $q = 4$ different subtypes of AML, where the levels $x \in \mathscr{X} = \{0,1,2,3\}$ denote the subtypes M0, M1, M2, M4 respectively, we estimate and select the profile bi-directed graphical model represented in Figure 3, with a suitable profile Lasso approach in a neighborhood selection scheme [5].

Let us give a short comment on the results obtained. The selected graph shows, for instance, that $Y_{\text{BAD.p155}}(x) \perp\!\!\!\perp Y_{\text{GSK3}}(x)$, for any $x \in \mathscr{X}$. Also, it shows, for instance, that $Y_{\text{AKTp.308}}(x) \perp\!\!\!\perp Y_{\text{AKTp.473}}(x)$, for any $x \in \{0,3\}$. The single levels of the proteins AKT1, BAD.p112, BAD.p112, BAD.p155, BAX, GSK3, GSK3.p and XIAP are independent of the subtypes of AML. Pairwise associations between the proteins AKT1, BAD.p112, BAX, GSK3.p, PTEN and XIAP are independent of the AML subtypes. The remaining proteins have at least one pairwise association that varies with the AML subtype.

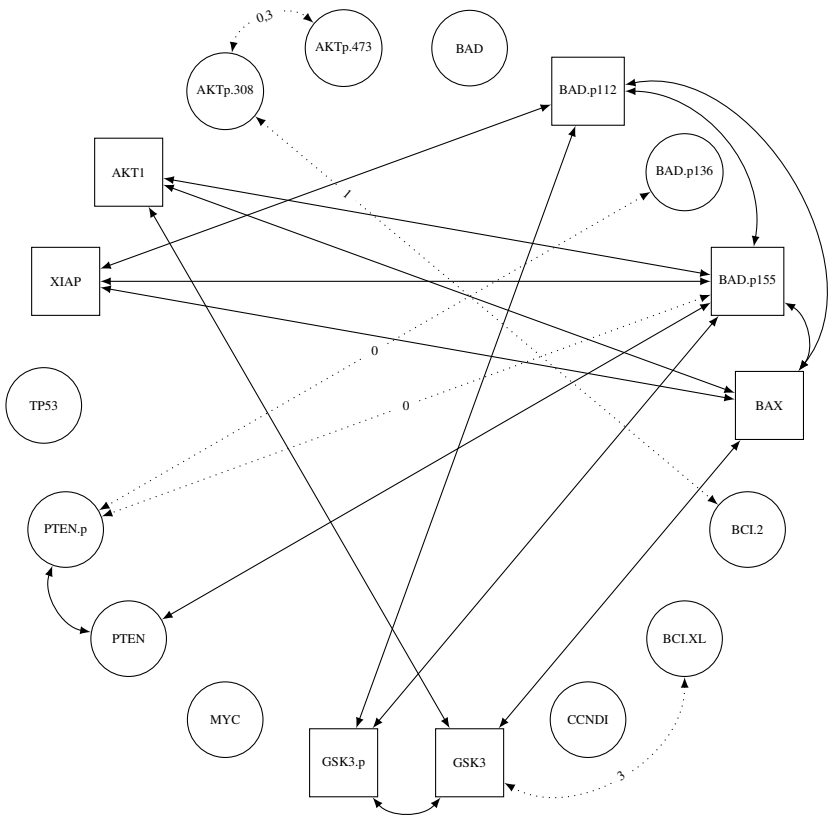


Fig. 3 The selected profile bi-directed graph model for protein data

References

1. Cox, D. R., Wermuth, N.: Linear dependencies represented by chain graphs. Stat. Sci. **8**, 204–218 (1993)
2. Guo, J., Levina, E., Michailidis, G., Zhu, J.: Joint estimation of multiple graphical models. Biometrika **98**, 1–15 (2011)
3. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: Kegg for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. **40**, 109–114 (2012)
4. Kauermann, G.: On a dualization of graphical Gaussian models. Scand. J. Stat. **23**, 105–116 (1996)
5. Meinshausen, N., Buhlmann P.: High-dimensional graphs and variable selection with the lasso. Ann. Stat. **34**, 1436–1462 (2006)
6. Peterson, C. B., Stingo F. C., Vannucci, M.: Bayesian inference of multiple Gaussian graphical models. J. Am. Stat. Assoc. **110**, 159–174 (2015)
7. Telesca, D., Müller, P., Kornblau, S., Suchard, M., Ji, Y.: Modeling protein expression and protein signaling pathways. J. Am. Stat. Assoc. **107**, 1372–1384 (2012)

Proton-Pump Inhibitor Provider Profiling via Funnel Plots and Poisson Regression

Profilazione dei Distributori di Inibitori della Pompa Protonica con Funnel Plot e Regressione di Poisson

Dario Delle Vedove, Francesca Ieva, and Anna Maria Paganoni

Abstract Proton-pump inhibitors (PPI) are one of the most prescribed medication worldwide. Even if they are largely recognized as a safe drug, long-term use can lead to an increased risk of collateral disease. Funnel plots have been proposed as a useful tool for assessing and visualizing surveillance data about PPI prescriptions. We studied a way to operate funnel plot analyses on both raw PPI patient rates and adjusted rates considering the effects of risk factors. Taking as a case study the German region of Saarland during 2010, 2013 and 2016, we showed that funnel plots of raw rates and of risk adjusted rates lead to partially different conclusions that can be proficiently compared and discussed.

Abstract *Gli inibitori della pompa protonica (PPI) sono una delle categorie di farmaci più prescritte al mondo. Nonostante siano largamente riconosciuti come farmaci sicuri, un uso prolungato potrebbe condurre ad un aumentato rischio di effetti collaterali. Abbiamo proposto i funnel plot come utile strumento di valutazione e visualizzazione per le prescrizioni di PPI. Abbiamo progettato un modo per applicare i funnel plot sia ai tassi di pazienti a cui sono stati prescritti PPI, sia a tali tassi corretti tenendo in considerazione fattori di rischio. Prendendo come caso studio la regione tedesca del Saarland durante gli anni 2010, 2013 e 2016, abbiamo mostrato che i funnel plot sui tassi semplici e sui tassi corretti conducono a*

Dario Delle Vedove
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan,
CADS – Center for Analysis, Decision and Society, Human Technopole, Milan, Italy
e-mail: dario.dellevedove@polimi.it

Francesca Ieva
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, via Bonardi 9, Milan, Italy e-mail: anna.paganoni@polimi.it

conclusioni parzialmente diverse, che possono però essere confrontate e discusse produttivamente.

Key words: Funnel Plot, Poisson Regression, Proton-Pump Inhibitor, Public Health, Provider Profiling, Over-Dispersion, Outliers

1 Introduction

Proton-Pump Inhibitors (PPI) are among the most prescribed class of medicines in wide areas of the world [4], due to their effectiveness for long-lasting reduction of stomach acid production [12]. However, in the last few years, some issues related to the PPI use for chronic or inappropriate conditions raised: even if it is a class of drugs which is considered to be well tolerated, increasing concerns are emerging about their abuse. Some studies recognize a significant association between PPI long-term use and an increased risk of gastric cancer and harmful interaction with other drugs [1, 10]. For these reasons, it is worth monitoring the prescribing phenomenon to avoid not only the spread of diseases related to the overprescription of PPIs, but also to save large amount of money for one of the most consumed class of medications that may be reallocated somewhere else.

In order to understand the goodness of resource distribution, valuable insights can be gathered by comparing performances of medical care providers, or, in other words, "profiling" them. It is a worthwhile process for targeting improvement strategies [11]. A challenge of the quality of healthcare research is to account varying population and hospital characteristics while developing statistical procedures to distinguish under and over performing medical providers. In particular, we aim to detect practices with a proportion of patients who received PPI prescriptions significantly larger than the average and detect them. In other words, we want to understand if there are prescribers providing PPI with laxed standards which are possibly misusing healthcare resources.

2 Materials and Methods

2.1 Data

Data were provided by Central Research Institute of Ambulatory Health Care, called Zi, from Germany. It is a non-profit foundation under private law, financed by the 17 Associations of Statutory Health Insurance Physicians (called KVs).

Data are aggregated to practice-level and concern number of patients, costs, daily defined doses, age, gender, and an analogous of all these information about PPI prescription and patients. For each observation, we have also if the practice has one

or more specialization (among general medicine, cardiology and neurology or psychiatry), the KV of the practice, the year (2010, 2013, 2016) and the quarter. Comprehensively, the dataset counts 702,353 rows, and between 50 and 65 thousands observations each quarter.

2.2 Statistical methods

Funnel plot is a valuable statistical method introduced in meta-analyses studies that allows to visually detect outliers [3]. The funnel plots use normal Gaussian control limits approximating binomial distribution, due to the fact that our objective variable is the rate given by the proportion between patients who received PPI prescriptions and the total amount of patients who received a prescription. These limits are given by

$$\hat{\theta}_0 \pm \Phi(1-\frac{\alpha}{2})\sqrt{\frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\rho}} \tag{1}$$

where $\hat{\theta}_0$ is fixed at the overall provincial rate as estimated from the data, $\Phi()$ is the cumulative inverse normal distribution evaluated for $1-\alpha\%$ control limits (we set $\alpha = 0.05$) and ρ is the number of observations to compute a rate. It is a precision parameter, determining the accuracy with which the indicator is being measured [13].

Funnel plots are based on the assumption that the null distribution fully expresses the variability of the in-control units. If we are not in this case, observations in funnel plot may result 'over-dispersed' around the target. In general, this issue raises when unmeasured covariates that are not taken into account in any risk-adjustment method. To deal with this behaviour, we can inflate a factor $\sqrt{\hat{\phi}}$, where $\hat{\phi} = \frac{1}{I}\sum_i z_i^2$, I is the institution number and z_i is the standardized Pearson residual of the i-th practice, so the control limits become $\hat{\theta}_0 \pm \Phi(1-\frac{\alpha}{2})\sqrt{\frac{\hat{\phi}\hat{\theta}_0(1-\hat{\theta}_0)}{\rho}}$. In order to minimize the influence of outlying case, it may be useful to compute the inflation factor with Winsorised z-score [9].

Risk adjustment was operated choosing a priori covariates that may explain practice rate variability. We fitted a Poisson regression against PPI patient number with a $\log(total\ patients)$ offset. Demographic factors (age, gender) and economic factors (cost) were successively included in the model.

The adjusted rate is computed as the product of the average PPI patients rate and the ratio of observed to expected values from the Poisson regression model. These models are compared with Pearson goodness-of-fit statistics [2].

3 Results

In order not to introduce effect modifications due to geographical location and to the nature of prescribers, we decided to limit the analyses to practices located in Saarland and specialized only in general medicine (more than 80% of the practices in the KV). Saarland was chosen as the smallest German territorial KV for practical reasons, but also because of its representativeness in terms of rural and urban areas for the whole country [8]. On the other hand, depending on the quarter, 82-89% of patients receiving a PPI prescription in Saarland is treated by a general medicine practice.

It is interesting to notice how the proportion of PPI prescriptions had a constant growth in this subset of practices: from 4.9% of the total amount of prescriptions at the beginning of 2010 to 6.1% in the end of 2016, reflecting a similar national trend, but sharper.

We are interested in explaining how the amount of patients who received at least a PPI prescription varies. We considered the average age of patients, since the proportion of the population prescribed PPI therapy may increase with age [6]; the gender proportion in each practice, since gender differences may exist in response to PPI [7, 14]; average cost of PPI definded daily doses; and the total amount of patients receiving a prescription.

Modeling the probability of a patient who received a PPI prescription as binomial, we produced a funnel plot of PPI patient rate vs patient number, but significant overdispersion was observed. Inflating into the variance a multiplicative factor computed with Winsorised z-scores (in this case, we used 5% most extreme top and bottom z-scores), it was possible to solve this issue. We found 4 to 8 % of outliers per quarter, but only 2-3% of the practices have a suspiciously high number of PPI patients. Figure 1(a) shows a funnel plot referring to the 4-th quarter of 2016. It is interesting to notice that 7 out of 803 of the practices are always high outliers over quarters, when we have information about them.

On the other hand, we considered the PPI patient count coming from a Poisson distribution. In searching for a model with a better fit to the data, we adjusted for age mean, sex percentage and average cost of a PPI daily defined dose. With the Pearson chi-square goodness of fit it was shown that risk adjusted Poisson regression improves the model fit significantly with respect to the unadjusted one, i.e., computed with the only total patient count as covariate. Also in this case we checked that there is overdispersion, so it is worth to introduce a multiplicative factor into the control limits, which was computed again with 10% Winsorized z-scores. Again 4-7% of practices outside the 95% limits are detected, 2-3% of which are high outliers. It can be observed in figure 1(b). In this setting, 6 out of 803 of the practices are always high outliers over quarters.

Comparing the two methodologies, we noticed that that there are at least two providers that result to be high outliers in both the settings in any quarter. The funnel plot for raw rate is more conservative with high outliers, but less with low outliers compared with the funnel plot for adjusted rates. Moreover, comparing out-of-bound practices found with the different funnel plots allows us to understand if high rates

can be justified with explanatory variables or not. For example, in the 4-th quarter of 2016, a practice with 336 has a PPI patient rate of 0.39, resulting to be a high outlier. Correcting for risks, it comes out that the adjusted rate is not so surprising after having accounted for our covariates.

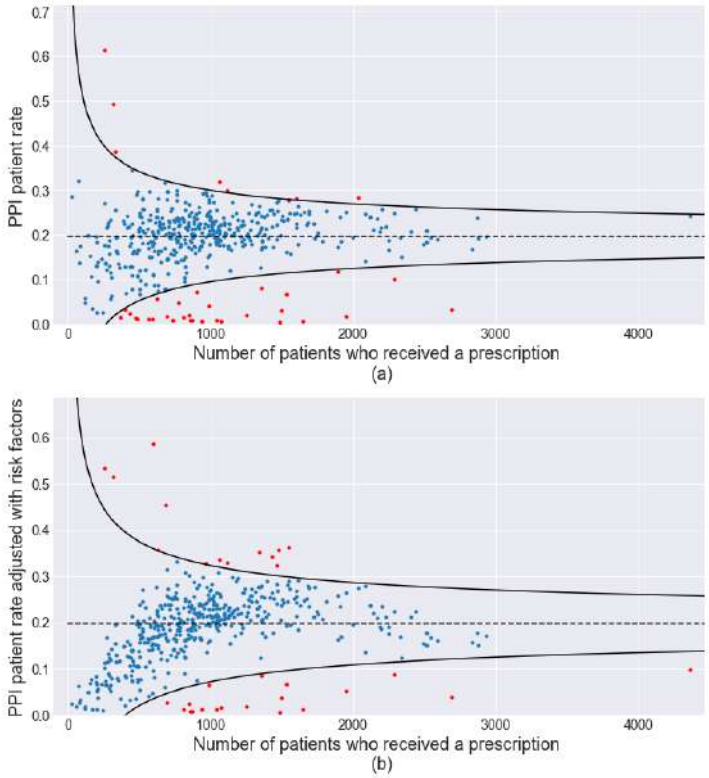


Fig. 1 Funnel plots for Saarland during the 4-th quarter of 2016: (a) funnel plot of raw PPI patient rate; (b) funnel plot of PPI patient rate adjusted with age, gender and cost factors.

4 Conclusions

Funnel plots provide a clear and operative tool for health surveillance and anomaly detection, also in presence of overdispersion. Overdispersion is not an unusual phenomenon in health data, and funnel plot has the ability to easily visualize it. In order to avoid misprescriptions due to laxed standards of providers, the purpose is to provide an easy-to-consult tool which can report in a comprehensible way the possibility that a practice is significantly out of standards. We proposed a joint use

of this graphical instrument and a statistical model as Poisson regression to detect out-performing institutions in terms of rate of patients who receive PPI prescriptions. Comparing suspicious practices in both the outcomes, it is possible to make more robust considerations about the behaviour of physicians.

Acknowledgements We would like to express our gratitude to Central Research Institute of Ambulatory Health Care in Germany (Zi) for data provision.

References

1. Abbas, M. K., Zaidi, A., Robert, C. A., Thiha, S., Malik, B. H.: The Safety of Long-term Daily Usage of a Proton Pump Inhibitor: A Literature Review. Cureus, (2019) doi: 10.7759/cureus.5563
2. Dover, D.C., Schopflocher, D.P.: Using funnel plots in public health surveillance. Popul Health Metr. **9**:58. (2011)
3. Egger, M., Smith, G.D., Schneider, M., Minder, C. Bias in meta-analysis detected by a simple, graphical test. Br Med J, **315**(7109), 629–624 (1997)
4. Gawron, A.J., Feinglass, J., Pandolfino, J.E., Tan, B.K., Bove, M.J., Shintani-Smith, S.: Brand name and generic proton pump inhibitor prescriptions in the United States: insights from the national ambulatory medical care survey (2006-2010). Gastroenterol Res Pract. (2015) doi:10.1155/2015/689531
5. Gemeinsamer Bundesausschuss (Ed.): Richtlinie des Gemeinsamen Bundesausschusses über die Verordnung von Arzneimitteln in der Vertragsärztlichen Versorgung [Guideline of the Common Federal Committee on the Prescription of Drugs in Ambulatory Care]. Gemeinsamer Bundesausschuss: Berlin, Germany. (2016)
6. Halfdanarson, O.O., Pottegard, A., Bjornsson, E.S., Lund, S.H., Ogmundsdottir, M.H., et al.: Proton-pump inhibitors among adults: a nationwide drug-utlization study. Therapeutic Adv Gastroenterol. **11**, 1–11. (2018)
7. Helgadottir, H., Metz, D., Lund, S., Gizurarson, S., Jacobsen, E., Asgeirsdottir, G, Yngvadottir, Y., Bjornsson, E.: Study of Gender Differences in Proton Pump Inhibitor Dose Requirements for GERD: A Double-Blind Randomized Trial. Journal of clinical gastroenterology. (2016) doi: 10.1097/MCG.0000000000000542
8. Hirsch, O., Donner-Banzhoff, N., Schulz, M., Erhart, M.: Detecting and Visualizing Outliers in Provider Profiling Using Funnel Plots and Mixed Effects Models—An Example from Prescription Claims Data. Int J Environ Res Public Health. **15**(9), 2015. (2018)
9. Ieva, F., Paganoni, A.M.: Detecting and visualizing outliers in provider profiling via funnel plots and mixed effect models. Health Care Manag. Sci. **18**, 166–172 (2015)
10. Lazarus, B., Chen, Y., Wilson, F.P., et al.: Proton Pump Inhibitor Use and the Risk of Chronic Kidney Disease. JAMA Intern Med. **176**(2),238–246. (2016)
11. Normand, S.L., Shahian, D.M.: Statistical and clinical aspects of hospital outcomes profiling. Statistical Science. **22**,206–226. (2007)
12. Scarpignato, C., Gatta, L., Zullo, A., Blandizzi, C., SIF-AIGO-FIMMG Group, Italian Society of Pharmacology, the Italian Association of Hospital Gastroenterologists, and the Italian Federation of General Practitioners: Effective and safe proton pump inhibitor therapy in acid-related diseases - A position paper addressing benefits and potential harms of acid suppression. BMC medicine. (2016) doi: 10.1186/s12916-016-0718-z
13. Spiegelhalter, D.J.: Funnel plots for comparing institutional performance. Stat. Med. **24**, 1185–1202. (2005)
14. Vakil, N., Niklasson, A., Denison, H., Rydén, A.: Gender differences in symptoms in partial responders to proton pump inhibitors for gastro-oesophageal reflux disease. United European Gastroenterol J. (2015) doi: 10.1177/2050640614558343.

Selecting optimal thresholds in ROC analysis with clustered data

Scelta di soglie ottimali nell'analisi ROC con dati raggruppati

Duc Khanh To, Gianfranco Adimari and Monica Chiogna

Abstract We tackle estimation of receiver operating characteristic (ROC) surfaces and selection of the optimal pair of thresholds for continuous diagnostic tests given covariates in clustered data when the disease status is described by three ordinal classes. The approach is based on a linear mixed-effect model which accounts for both the clusters and the covariates effects. The asymptotic properties of estimators are studied. Simulation studies are performed to assess the performance of estimators.

Sommario *Per test diagnostici continui, si studia il problema della stima della superficie ROC e della individuazione di valori di soglia ottimali nel caso di dati raggruppati e in presenza di variabili esplicative. Si propone un approccio basato su un modello lineare a effetti misti e si studiano le proprietà degli stimatori individuati, il cui comportamento nel finito è valutato attraverso esperimenti di simulazione.*

Key words: clustered data, ROC surface, linear mixed-effect model

1 Introduction

The receiver operating characteristic (ROC) surface has been popularly used in biomedical studies to evaluate the ability of a diagnostic test (or biomarker) to distinguish among one of three classes of a disease status. Let Y be a diagnostic test result measured on a continuous scale, and let Y_1, Y_2, Y_3 be the test result for subjects in class 1, 2 and 3, respectively. Without loss of generality, we assume that higher

Duc Khanh To, Gianfranco Adimari
Department of Statistical Sciences, University of Padova, Via C. Battisti, 241; I-35121 Padova, Italy. e-mail: duckhanh.to@unipd.it; e-mail: gianfranco.adimari@unipd.it

Monica Chiogna
Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Via Belle Arti, 41; 40126 Bologna, Italy. e-mail: monica.chiogna2@unibo.it

values of test result are associated to higher severity of the disease. Given a pair of thresholds (t_1, t_2), with $t_1 < t_2$ in the range of diagnostic test results, three true class fractions can be defined as $\text{TCF}_1(t_1) = \Pr(Y_1 \leq t_1)$, $\text{TCF}_2(t_1, t_2) = \Pr(t_1 < Y_2 \leq t_2)$ and $\text{TCF}_3(t_2) = \Pr(Y_3 > t_2)$. The ROC surface for the diagnostic test Y is obtained by plotting ($\text{TCF}_1(t_1)$, $\text{TCF}_2(t_1, t_2)$ and $\text{TCF}_3(t_2)$) in a unit cube over all possible values of t_1 and t_2 (Nakas and Yiannoutsos, 2004). For practical uses, choosing optimal thresholds is a relevant issue.

Recently, several methods have been developed to estimate the ROC surface (Nakas and Yiannoutsos, 2004; Xiong et al, 2006) and to select an optimal pair of thresholds (Nakas et al, 2010; Attwood et al, 2014). Most of them consider a setting in which measurements on statistical units can be considered realizations of statistically independent random variables, and the diagnostic test is not influenced by any covariate. In some medical studies, however, statistical units are enrolled in clusters (e.g., families), and the diagnostic test can be affected by some covariates. In this setting, Xiong et al (2018) used a linear mixed-effect model to account for the cluster and the covariates effects, and proposed an approach to estimate the volume under the ROC surface (which is summary measure of the diagnostic test accuracy).

In this paper, we employ the same model as in Xiong et al (2018) to estimate the ROC surface. Then, in order to properly address the problem of selecting an optimal pair of thresholds, we adapt to the clustered-data case three criteria based on the generalized Youden index, the so-called perfection point and the maximum volume of the cuboid under ROC surface.

2 The linear mixed-effect model

Suppose to have p covariates, $X_1, \ldots, X_p$ say, possibly associated with the diagnostic test Y. Let c be the total number of clusters (for instance families), randomly selected from the population. For the cluster k-th, $k = 1, \ldots, c$, let n_{ki} be the total number of subjects belonging to class i-th of disease, $i = 1, 2, 3$. Note that n_{ki} could be equal to 0 for some clusters. In order to account for the clustering effect on Y, as well as for covariates' effects, we consider the following linear mixed-effect model (see also Xiong et al, 2018):

$$\begin{aligned} Y_1 &= \alpha_{k_1} + z_1^\top \beta_1 + \varepsilon_1, \\ Y_2 &= \alpha_{k_2} + z_2^\top \beta_2 + \varepsilon_2, \\ Y_3 &= \alpha_{k_3} + z_3^\top \beta_3 + \varepsilon_3, \end{aligned} \tag{1}$$

where (Y_1, Y_2, Y_3) is a triplet of test scores from three randomly sampled subjects from the three disease classes, (k_1, k_2, k_3), $k_i \in \{1, \ldots, c\}$, are cluster memberships indicating the clusters from which Y_1, Y_2, Y_3 are observed, $z_i = (1, x_{1i}, \ldots, x_{pi})^\top$ are fixed (i.e., not random) covariates values, and $\beta_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})^\top$, $i = 1, 2, 3$, are vectors of parameters representing covariates effects. In model (1), α_k are random effects accounting for the presence of clusters, and ε_i are subject-level random er-

rors. We assume that: (i) the random effects α_k and the subject-level random errors ε_i follow a normal distribution, i.e., $\alpha_k \sim \mathcal{N}(0,\sigma_c^2)$ and $\varepsilon_i \sim \mathcal{N}(0,\sigma_i^2)$ with $i=1,2,3$; (ii) $\alpha_1,\alpha_2,\ldots,\alpha_c$ and $\varepsilon_1,\varepsilon_2,\varepsilon_3$ are all independent. These assumptions are standard in the linear mixed-effect modelling framework (McCulloch and Searle, 2001).

Let $\beta=(\beta_1^\top,\beta_2^\top,\beta_3^\top)^\top$ and $\theta=(\sigma_c^2,\sigma_1^2,\sigma_2^2,\sigma_3^2)^\top$ be the unknown parameters in model (1). An unbiased estimator $\widehat{\gamma}=(\widehat{\beta}^\top,\widehat{\theta}^\top)^\top$ of $\gamma=(\beta^\top,\theta^\top)^\top$, can be obtained resorting on restricted (or residual) maximum likelihood (REML) estimation (McCulloch and Searle, 2001). Under some regularity conditions, the REML estimator $\widehat{\gamma}$ is consistent and asymptotically normal, with mean γ and covariance matrix Λ, i.e., $\widehat{\gamma}\dot{\sim}\mathcal{N}(\gamma,\Lambda)$. The asymptotic covariance matrix Λ can be consistently estimated by using the sandwich formula (Liang and Zeger, 1986).

3 The proposal

3.1 Covariate-specific ROC surface

According to model (1), the marginal distribution of Y_i, $i=1,2,3$, is normal with mean $z^\top\beta_i$ and variance $\sigma_c^2+\sigma_i^2$, i.e., $Y_i\sim\mathcal{N}(z^\top\beta_i,\sigma_c^2+\sigma_i^2)$ with $i=1,2,3$. Given two cut points t_1 and t_2 ($t_1<t_2$) and a vector z of covariates values, the TCFs are:

$$\begin{aligned}
\mathrm{TCF}_1(t_1;z) &= \Phi\left(\frac{t_1-z^\top\beta_1}{\sqrt{\sigma_c^2+\sigma_1^2}}\right),\\
\mathrm{TCF}_2(t_1,t_2;z) &= \Phi\left(\frac{t_2-z^\top\beta_2}{\sqrt{\sigma_c^2+\sigma_2^2}}\right)-\Phi\left(\frac{t_1-z^\top\beta_2}{\sqrt{\sigma_c^2+\sigma_2^2}}\right),\\
\mathrm{TCF}_3(t_2;z) &= 1-\Phi\left(\frac{t_2-z^\top\beta_3}{\sqrt{\sigma_c^2+\sigma_3^2}}\right),
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Plugging the REML estimators $\widehat{\beta}_1,\widehat{\beta}_2,\widehat{\beta}_3,\widehat{\sigma}_c^2,\widehat{\sigma}_1^2,\widehat{\sigma}_2^2,\widehat{\sigma}_3^2$ into these expressions, leads to the estimators $\widehat{\mathrm{TCF}}_1(t_1;z)$, $\widehat{\mathrm{TCF}}_2(t_1,t_2;z)$ and $\widehat{\mathrm{TCF}}_3(t_2;z)$.

3.2 Selection thresholds methods

Under the given model setting, a covariate-specific optimal pair of thresholds (t_1^+,t_2^+) can be obtained by:

(i) maximizing the generalized Youden index $J_3(z)$ (GYI, Nakas et al, 2010), with

$$J_3(z) = \mathrm{TCF}_1(t_1;z) + \mathrm{TCF}_2(t_1,t_2;z) + \mathrm{TCF}_3(t_2;z); \tag{2}$$

(ii) minimizing the Euclidean distance $D_3(z)$ between the perfection point $(1,1,1)$ and $(\mathrm{TCF}_1(t_1;z), \mathrm{TCF}_2(t_1,t_2;z), \mathrm{TCF}_3(t_2;z))$ (CtP, Attwood et al, 2014), with

$$D_3(z) = \sqrt{[1-\mathrm{TCF}_1(t_1;z)]^2 + [1-\mathrm{TCF}_2(t_1,t_2;z)]^2 + [1-\mathrm{TCF}_3(t_2;z)]^2}; \tag{3}$$

(iii) maximizing the volume $V_3(z)$ of the cuboid under the ROC surface (MV, Attwood et al, 2014), with

$$V_3(z) = \mathrm{TCF}_1(t_1;z) \times \mathrm{TCF}_2(t_1,t_2;z) \times \mathrm{TCF}_3(t_2;z). \tag{4}$$

According to (2), (3) and (4), the covariate-specific objective functions $J_3(z)$, $D_3(z)$, $V_3(z)$ and the associated optimal pair of cut points (t_1^+, t_2^+) depend on γ. Plugging the REML estimator $\widehat{\gamma}$ into the expressions of $J_3(z)$, $D_3(z)$ and $V_3(z)$, leads to the estimated versions $\widehat{J}_3(z)$, $\widehat{D}_3(z)$ and $\widehat{V}_3(z)$. The estimators $(\hat{t}^+_{1,\mathrm{GYI}}, \hat{t}^+_{2,\mathrm{GYI}})$, $(\hat{t}^+_{1,\mathrm{CtP}}, \hat{t}^+_{2,\mathrm{CtP}})$ and $(\hat{t}^+_{1,\mathrm{MV}}, \hat{t}^+_{2,\mathrm{MV}})$ are obtained by maximizing $\widehat{J}_3(z)$ and $\widehat{V}_3(z)$, or minimizing $\widehat{D}_3(z)$, under the constraint $t_1 < t_2$. The optimization leads also to the covariate-specific estimated optimality statistics $\widehat{J}_3^+(z)$, $\widehat{D}_3^+(z)$ and $\widehat{V}_3^+(z)$ (e.g., $\widehat{J}_3^+(z)$ is the maximum of $\widehat{J}_3(z)$). Estimators $\widehat{J}_3^+$, $\widehat{D}_3^+$, $\widehat{V}_3^+$, $(\hat{t}^+_{1,\mathrm{GYI}}, \hat{t}^+_{2,\mathrm{GYI}})$, $(\hat{t}^+_{1,\mathrm{CtP}}, \hat{t}^+_{2,\mathrm{CtP}})$ and $(\hat{t}^+_{1,\mathrm{MV}}, \hat{t}^+_{2,\mathrm{MV}})$ are functions of the REML estimators. As a consequence, they are consistent and asymptotically normal. Delta method can be applied to obtain asymptotic variances and covariances (see also Schisterman and Perkins (2007)) as:

$$\mathbb{V}\mathrm{ar}\left(\widehat{H}^+\right) = \left(\frac{\partial H}{\partial \gamma^\top}\right) \Lambda \left(\frac{\partial H}{\partial \gamma^\top}\right)^\top, \tag{5}$$

$$\Sigma_{\hat{t}^+_{1,*}, \hat{t}^+_{2,*}} = \left(\frac{\partial t^+_{1,*}}{\partial \gamma^\top}, \frac{\partial t^+_{2,*}}{\partial \gamma^\top}\right) \Lambda \left(\frac{\partial t^+_{1,*}}{\partial \gamma^\top}, \frac{\partial t^+_{2,*}}{\partial \gamma^\top}\right)^\top, \tag{6}$$

where H stands, in turn, for J_3, D_3 or V_3, and the symbol $*$ stands for the name of the selection method (i.e., GYI, CtP and MV). The involved derivatives are:

$$\frac{\partial t^+_{m,*}}{\partial \beta^\top} = \left(\frac{\partial^2 H}{\partial t^+_{m,*} \partial t^+_{m,*}}\right)^{-1} \left(-\frac{\partial^2 H}{\partial t^+_{m,*} \partial \beta^\top}\right),$$

$$\frac{\partial t^+_{m,*}}{\partial \theta^\top} = \left(\frac{\partial^2 H}{\partial t^+_{m,*} \partial t^+_{m,*}}\right)^{-1} \left(-\frac{\partial^2 H}{\partial t^+_{m,*} \partial \theta^\top}\right),$$

$$\frac{\partial H}{\partial \gamma^\top} = \frac{\partial H}{\partial I}\frac{\partial I}{\partial \gamma^\top} + \frac{\partial H}{\partial t^+_{1,*}}\frac{\partial t^+_{1,*}}{\partial \gamma^\top} + \frac{\partial H}{\partial t^+_{2,*}}\frac{\partial t^+_{2,*}}{\partial \gamma^\top},$$

where I denotes the identity function, with $m = 1,2$. The plug-in method gives consistent estimates of quantities in (5) and (6). Confidence intervals for the entities of interest can be easily constructed by using the normal approximation.

4 Simulation study

We perform several simulation experiments to evaluate the performance of the discussed selection methods. Table 1 gives some results only, referring to the following scenario. The total number of clusters c is taken to be in the set $\{15,30,60\}$. Then, we consider a balanced setting, i.e., an equal size in each cluster, with $n_k \in \{4,10\}$. In the k-th cluster, the disease status of subjects is simulated from a multinomial distribution, $multinomial(n_k,(0.6,0.3,0.1))$. We consider a covariate X simulated from a uniform random variable $\mathscr{U}(-2,2)$. The parameters of model (1) are set to be $\beta_{01}=0.5$, $\beta_{11}=0.5$, $\beta_{02}=2$, $\beta_{12}=0.8$, $\beta_{03}=3.5$ and $\beta_{13}=1.1$. Variances of errors ε_i are set to be $\sigma_1^2=0.3$, $\sigma_2^2=0.8$, $\sigma_3^2=1.3$; σ_c^2, is set to be 0.2 or 1. This choice of variance components allows us to consider two different values for the intra-class correlation coefficient $ICC=\frac{\sigma_c^2}{\sigma_c^2+\sigma_\varepsilon^2}$, with $\sigma_\varepsilon=\frac{\sigma_1+\sigma_2+\sigma_3}{3}$. In particular, ICC equals 0.213 or 0.574, corresponding to the two values of σ_c^2.

Table 1 reports the estimated covariate-specific optimality statistics and the corresponding optimal pairs of thresholds, at $x=1.58$, over 1000 simulations. When the number of clusters and the cluster sizes are small ($c=15,30$ and $n_k=4$), biases are slightly large, but become smaller as the number of clusters and cluster sizes grow larger. The estimated asymptotic standard deviations are comparable with the Monte Carlo ones. Larger ICC values produce larger standard deviations. The empirical coverage probabilities of 95% confidence intervals become close to the nominal level when either the number of clusters or the cluster sizes increase.

References

Attwood, K., Tian, L. and Xiong, C. (2014). Diagnostic thresholds with three ordinal groups. *J. Biopharm. Stat*, **24**(3), 608–633.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Stat. Med*, **23**(22), 3437–3449.

Nakas, C. T., Alonzo, T. A. and Yiannoutsos, C. T. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Stat. Med* **29**(28), 2946–2955.

Schisterman, E. F. and Perkins, N. (2007). Confidence intervals for the youden index and corresponding optimal cut-point. *Commun. Stat. Simulat*, **36**(3), 549–563.

Xiong, C., van Belle, G., Miller, J. P. and Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Stat. Med*, **25**(7), 1251–1273.

Xiong, C., Luo, J., Chen, L., Gao, F., Liu, J., Wang, G., Bateman, R. and Morris, J. C. (2018). Estimating diagnostic accuracy for clustered ordinal diagnostic groups in the three-class case–Application to the early diagnosis of Alzheimer disease. *Stat. Methods. Med. Res*, **27**(3), 701–714.

Table 1 Simulation results for the proposed covariate-specific selection methods (at $x = 1.58$) in cases of equal size within clusters. Bias is the difference between the average of estimated parameters and the true ones. MCSD is the Monte Carlo standard deviation of estimated parameters in 1000 simulation. ASD is the average of estimated standard deviations based on (5) and (6). CP is the empirical coverage probability of confidence intervals, at nominal level 0.95.

		$\sigma_c^2 = 0.2$								$\sigma_c^2 = 1$							
		$n_k = 4$				$n_k = 10$				$n_k = 4$				$n_k = 10$			
		Bias	MCSD	ASD	CP	Bias	MCSD	ASD	CP	Bias	MCSD	ASD	CP	Bias	MCSD	ASD	CP
$c = 15$	$\hat{J}_3^+$	−0.007	0.259	0.154	0.748	−0.010	0.162	0.122	0.836	0.013	0.250	0.154	0.771	0.002	0.170	0.131	0.839
	$\hat{D}_3^+$	0.026	0.173	0.104	0.774	0.013	0.106	0.078	0.841	0.013	0.160	0.095	0.786	0.004	0.102	0.078	0.839
	$\hat{V}_3^+$	−0.004	0.176	0.102	0.750	−0.002	0.106	0.080	0.832	0.008	0.136	0.078	0.780	0.006	0.085	0.066	0.834
	$\hat{I}_{1,GYI}^+$	0.007	0.293	0.232	0.875	0.000	0.235	0.201	0.874	−0.011	0.507	0.426	0.887	0.004	0.476	0.408	0.881
	$\hat{I}_{2,GYI}^+$	0.177	1.106	0.373	0.803	0.028	0.375	0.286	0.866	0.189	1.161	0.527	0.836	0.035	0.566	0.463	0.878
	$\hat{I}_{1,CIP}^+$	0.009	0.284	0.228	0.873	−0.002	0.233	0.198	0.880	−0.012	0.497	0.422	0.890	−0.001	0.472	0.405	0.883
	$\hat{I}_{2,CIP}^+$	0.176	1.142	0.351	0.773	0.022	0.376	0.296	0.848	0.120	1.118	0.503	0.822	0.020	0.557	0.470	0.867
	$\hat{I}_{1,MV}^+$	−0.013	0.335	0.236	0.872	−0.005	0.235	0.201	0.881	−0.038	0.553	0.429	0.884	−0.004	0.476	0.407	0.879
	$\hat{I}_{2,MV}^+$	0.095	1.035	0.344	0.788	0.009	0.352	0.282	0.863	0.050	1.013	0.498	0.828	0.011	0.542	0.456	0.870
$c = 30$	$\hat{J}_3^+$	−0.011	0.172	0.132	0.839	0.002	0.105	0.093	0.894	0.000	0.166	0.131	0.841	0.005	0.113	0.101	0.912
	$\hat{D}_3^+$	0.018	0.117	0.086	0.847	0.002	0.067	0.059	0.895	0.008	0.104	0.080	0.844	−0.001	0.067	0.060	0.909
	$\hat{V}_3^+$	−0.005	0.115	0.087	0.832	0.003	0.069	0.061	0.894	0.003	0.085	0.066	0.845	0.005	0.056	0.050	0.905
	$\hat{I}_{1,GYI}^+$	0.001	0.190	0.173	0.920	0.001	0.158	0.150	0.927	0.006	0.347	0.318	0.925	−0.001	0.320	0.305	0.934
	$\hat{I}_{2,GYI}^+$	0.039	0.379	0.290	0.872	0.013	0.229	0.209	0.909	0.046	0.539	0.404	0.895	0.012	0.366	0.343	0.927
	$\hat{I}_{1,CIP}^+$	−0.003	0.186	0.171	0.908	0.002	0.158	0.149	0.926	0.000	0.344	0.313	0.919	−0.001	0.320	0.304	0.931
	$\hat{I}_{2,CIP}^+$	0.027	0.377	0.289	0.856	0.014	0.238	0.218	0.909	0.017	0.475	0.395	0.894	0.011	0.368	0.346	0.934
	$\hat{I}_{1,MV}^+$	−0.005	0.192	0.174	0.915	0.000	0.159	0.150	0.933	−0.002	0.350	0.317	0.918	−0.003	0.323	0.305	0.930
	$\hat{I}_{2,MV}^+$	0.010	0.351	0.277	0.866	0.008	0.225	0.206	0.916	0.004	0.455	0.385	0.894	0.006	0.359	0.338	0.931
$c = 60$	$\hat{J}_3^+$	−0.005	0.114	0.105	0.917	−0.002	0.071	0.069	0.928	0.003	0.110	0.103	0.924	0.003	0.076	0.074	0.930
	$\hat{D}_3^+$	0.007	0.073	0.067	0.917	0.003	0.045	0.044	0.923	0.002	0.067	0.062	0.922	−0.000	0.045	0.044	0.929
	$\hat{V}_3^+$	−0.001	0.074	0.068	0.913	−0.001	0.047	0.045	0.925	0.003	0.055	0.051	0.923	0.002	0.038	0.037	0.931
	$\hat{I}_{1,GYI}^+$	0.000	0.130	0.124	0.930	0.001	0.111	0.108	0.945	−0.003	0.237	0.227	0.934	0.003	0.228	0.219	0.937
	$\hat{I}_{2,GYI}^+$	0.010	0.237	0.209	0.911	0.010	0.155	0.151	0.942	0.007	0.315	0.290	0.922	0.011	0.252	0.245	0.934
	$\hat{I}_{1,CIP}^+$	−0.002	0.130	0.122	0.931	−0.001	0.109	0.107	0.945	−0.005	0.236	0.225	0.933	0.002	0.226	0.218	0.936
	$\hat{I}_{2,CIP}^+$	0.010	0.246	0.218	0.910	0.007	0.163	0.158	0.937	−0.000	0.317	0.292	0.921	0.006	0.255	0.248	0.934
	$\hat{I}_{1,MV}^+$	−0.002	0.132	0.124	0.930	0.001	0.110	0.108	0.945	−0.004	0.239	0.228	0.933	0.003	0.228	0.219	0.934
	$\hat{I}_{2,MV}^+$	0.004	0.229	0.204	0.905	0.006	0.153	0.149	0.943	−0.004	0.304	0.281	0.919	0.006	0.249	0.243	0.935

Environment, Physics and Engineering

A hidden semi-Markov model for segmenting environmental toroidal data

Un modello a classi latenti semi-markoviane per l'analisi di dati ambientali toroidali

Francesco Lagona and Antonello Maruotti

Abstract Toroidal time series are temporal sequences of bivariate angular observations that often arise in environmental and ecological studies. A hidden semi-Markov model is proposed for segmenting these data according to a finite number of latent classes, associated toroidal densities. The model conveniently integrates circular correlation, multimodality and temporal auto-correlation. A computationally efficient EM algorithm is proposed for parameter estimation. The proposal is illustrated on a time series of wind and sea wave directions.

Abstract *Le serie storiche toroidali sono sequenze di osservazioni circolari bivariate che appaiono di frequente in studi ambientali ed ecologici. Un modello classi latenti semi-markoviane viene proposto per classificare questi dati secondo un numero finito di classi latenti, ognuna associata con una densità toroidale. Il modello è in grado di catturare simultaneamente la correlazione circolare, la multimodalità e l'auto-correlazione temporale dei dati.*

Key words: hidden semi-Markov model, EM algorithm, model-based clustering, toroidal data

1 Introduction

Bivariate sequences of angles are often referred to as toroidal time series, because the pair of two angles can be represented as a point on a torus. These data often arise in environmental and ecological studies. Examples include time series of wind and wave directions [8], time series of wind mean directions and directions of the

F. Lagona
University of Roma Tre, e-mail: francesco.lagona@uniroma3.it

A. Maruotti
LUMSA, e-mail: a.maruotti@lumsa.it

maximum gust observed each day [2] and time series of turning angles in studies of animal movement [10].

The analysis of toroidal time series is complicated by the difficulties in modeling the dependence between angular measurements over time [7]. An additional complication is given by the multimodality of the marginal distribution of the data, because environmental toroidal data are observed under time-varying heterogeneous conditions.

This paper introduces a toroidal hidden semi-Markov model (HSMM) that simultaneously accounts for dependence across circular measurements, temporal autocorrelation, multimodality and latent time-varying heterogeneity. Under this model, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent semi-Markov process. While the toroidal density accommodates dependence between two circular variables, a mixture of toroidal densities allows for multimodality and, finally, a latent semi-Markov process accounts for temporal correlation and, simultaneously, for time-varying heterogeneity.

Our proposal extends previous approaches that are based on toroidal hidden Markov models [9, 1]. Under a toroidal hidden Markov model, the data are approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent, first-order Markov chain with a finite number of states. The sojourn times of each state of a Markov chain are distributed according a geometric distribution. Hence the most likely dwell time for every state of a hidden Markov model with underlying first-order Markov chain is 1. Our proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent semi-Markov model, allowing for sojourn times that are not necessarily geometrically distributed.

2 A hidden semi-Markov model for toroidal data

Let $\mathbf{z} = (x, y)$ be a pair of angles, $x, y \in [0, 2\pi)$. Moreover, let $f(x;\alpha)$ and $f(y;\beta)$ be two circular densities, respectively known up to the parameters α and β. Further, let $F(x;\alpha)$ and $F(y;\beta)$ be the two cumulative distribution functions of x and y, defined with respect to a fixed, although arbitrary, origin. Finally, let $g(u;\gamma), u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then,

$$f_q(\mathbf{z};\theta) = 2\pi g\left(2\pi\left[F(x;\alpha) - qF(y;\beta)\right]\right) f(x;\alpha) f(y;\beta) \quad q = \pm 1 \tag{1}$$

is a parametric toroidal density with support $[0, 2\pi)^2$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x;\alpha)$ and $f(y;\beta)$ [3]. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by decoupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then equation (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the de-

pendence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$).

The proposed hidden semi-Markov model can be described as a dynamic mixture of copula-based toroidal densities. To illustrate, let $\mathbf{z} = (\mathbf{z}_t, t = 1, \ldots, T)$, $\mathbf{z}_t = (x_t, y_t)$, $x_t, y_t \in [0, 2\pi)$, be a toroidal time series. We assume that the distribution of the data is driven by the evolution of an unobserved semi-Markov process with K states, which represents (time-varying) latent classes and can be specified as a sequence $\mathbf{u} = (\mathbf{u}_t, t = 1, \ldots, T)$ of multinomial variables $\mathbf{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes, whose binary components represent class membership at time t. The joint distribution $p(\mathbf{u}; \pi)$ of the chain is fully known up to a parameter π that includes K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \ldots, K, \sum_k \pi_k = 1$, $K^2 - K$ transition probabilities $\pi_{hk} = P(u_{tk} = 1 | u_{t-1,h} = 1), h, k = 1, \ldots, K, \sum_k \pi_{hk} = 1, h \neq k$ (whereas $\pi_{kk} = 0$, $k = 1 \ldots K$), and, finally, p parameters of the dwell time distributions of each state.

The specification of the HSMM is completed by assuming that the observations are conditionally independent, given a realization of the semi-Markov process. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K) = \prod_{t=1}^{T} \prod_{k=1}^{K} f(\mathbf{z}_t; \theta_k)^{u_{tk}}, \tag{2}$$

where $f(\mathbf{z}; \theta_k), k = 1, \ldots, K$, are the K cylindrical densities defined by (1) and known up to a vector of parameters θ_k.

The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation $\mathbf{u}$, namely

$$L(\pi, \theta; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \pi) f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K). \tag{3}$$

By computing the maximum likelihood estimate $\hat{\theta}$, the toroidal time series can be segmented according to the posterior probabilities of class membership

$$\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}; \hat{\theta}), \tag{4}$$

based on $\hat{\theta}$. More precisely, the observation at time t can be allocated to class $k^\star$ if $\hat{\pi}_{tk^\star} \geq \hat{\pi}_{th}$, for each $h = 1 \ldots K$ (maximum a posterior, MAP, allocation).

When the dwell distribution of each latent state is geometric, the model reduces to a hidden Markov model that ignores alternative dwell time distribution. If, additionally, the transition probability matrix of the model has equal rows, the model reduces to a mixture model where observations are clustered by ignoring the information redundancy that is due to temporal correlation.

3 An application to marine data

The proposed methods have been implemented to segment a time series of $T = 1326$ semi-hourly wind and wave directions, taken in wintertime by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. Figure 1 displays the scatter plot of the data. Point coordinates indicate the direction (in radians) from which winds blow and waves travel. For simplicity, these bivariate observations are plotted on the plane, although data points are actually on a torus. The interpretation of these data is not easy. While in the ocean wind and wave directions are strongly correlated, this is not necessarily true in the Adriatic Sea, due to the orography of the basin and the location of the buoy. Coastal winds generate synchronized waves only when the waves travel unobstructed, that is, either northwesterly or southeasterly, along the major axis of the basin. When western and south-western winds blow from the coast, waves are not synchronized with wind and travel along the major axis of the Adriatic basin from SE to NW. This explains the clusters shown in Figure 1 and suggests the occurrence of two latent wind–wave regimes. Accordingly, a HSMM with two states have been estimated from these data.

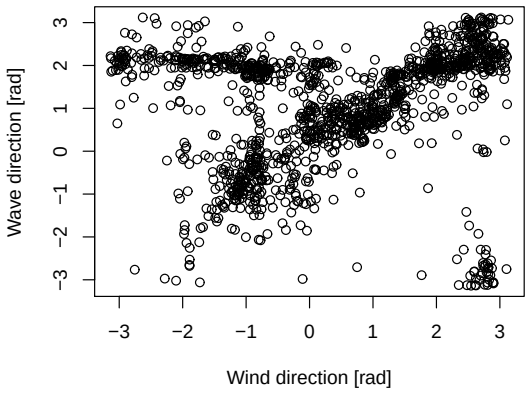


Fig. 1 Wave directions and heights, as observed by the buoy of Ancona in wintertime ($-\pi$, $-pi/2$, 0, $\pi/2$ respectively indicate South, West, North, East). For simplicity, the data are plotted on the plane, although they are points on the torus $[-pi/2, \pi/2)^2$.

The proposed HMM requires a parametric specification of the toroidal density (1), which reduces to the choice of the binding density g and the choice of the marginal densities $f(x;\alpha)$ and $f(y;\beta)$ that respectively model the marginal distribution of the wind and wave direction. However, depending on the choice of the binding density, the density (1) can be multimodal [4]. Using multimodal densities in segmentation and classification problems, such as the one motivating this paper,

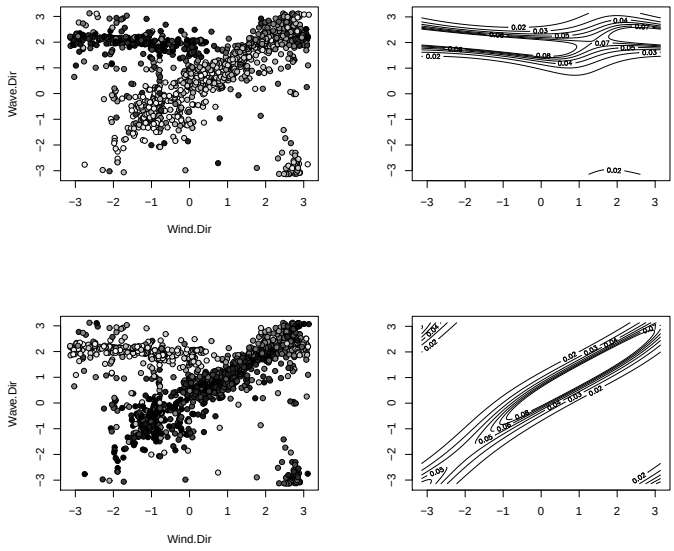


Fig. 2 Segmentation of a time series of wind and wave directions. Left: observations colored with grey levels according to the estimated membership probabilities of each class (black indicates a probability equal to 1). Right: contour plot of state-specific toroidal densities.

may unnecessarily complicate the interpretation of the results. Unimodal densities can however be obtained by using the wrapped Cauchy as a binding density g [4].

Accordingly, for this study, the binding density has been specified as a centered wrapped Cauchy

$$g(u;\gamma) = \frac{1}{2\pi}\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(u)} \quad , \quad u \in [0,2\pi).$$

This circular density depends on a single concentration parameter $\gamma \in [0,1)$ and reduces to the uniform circular density when $\gamma = 0$.

Wrapped Cauchy densities that include additional location parameters α_1 and β_1 have been instead exploited to model the marginal distributions of wind and wave direction, say

$$f(x;\alpha) = \tfrac{1}{2\pi}\frac{1-\alpha_2^2}{1+\alpha_2^2-2\alpha_2\cos(y-\alpha_1)} \quad , \quad x \in [0,2\pi); \tag{5}$$

$$f(y;\beta) = \tfrac{1}{2\pi}\frac{1-\beta_2^2}{1+\beta_2^2-2\beta_2\cos(y-\beta_1)} \quad , \quad y \in [0,2\pi). \tag{6}$$

The proposed toroidal density is therefore obtained by taking a wrapped Cauchy density that binds wrapped Cauchy marginals, a model known as the bivariate wrapped Cauchy model [5].

Figure 2 shows the shapes of the two state-specific toroidal distributions and the segmented observations. The model successfully segment the observations according to two clusters, and offers a clear-cut indication of the distribution of the data under each regime. Under state 1, wind and wave directions are essentially independent, because coastal winds do not generate waves. Under state 2, winds blows along the major axis of the Adriatic basin and their directions are highly correlated with the directions of the wave that they generate.

Overall, the model describes the plasticity of the wind–wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime switching changes not only the modal directions and concentrations around these modes but also, and more interestingly, the correlation structure of the data. As a result, on the one side, the (marginal) weak correlation between wind and wave directions is explained by the presence of coastal winds (State 1). On the other side, the model indicates that the wind direction is an accurate predictor of the wave direction only under a specific regime (State 2). In summary, wind directions should not be used to predict wave directions, without accounting for the latent, environmental heterogeneity of the data under study.

Acknowledgements Francesco Lagona was supported by the 2015 PRIN supported project 'Environmental processes and human activities: capturing their interactions via statistical methods', funded by the Italian Ministry of Education, University and Scientific Research.

References

1. Bulla J, Lagona F, Maruotti A, Picone M (2012) A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series, Journal of Agricultural, Biological, and Environmental Statistics, 17: 544-567
2. Coles S (1998) Inference for circular distributions and processes, Statistics and Computing, 8: 105-113.
3. Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models, Journal of the American Statistical Association 73: 602-606
4. Jones MC, Pewsey A, Kato S (2015). On a class of circulas: copulas for circular distributions, Annals of the Institute of Statistical Mathematics, 67: 843-862
5. Kato S, Pewsey A (2015) A Möbius transformation-induced distribution on the torus, Biometrika, 102: 359-370
6. Lagona F (2019) Copula-based segmentation of cylindrical time series, Statistical and Probability Letters, 144: 16-22
7. Lagona F (2018) Correlated cylindrical data. In: C. Ley and T. Verdebout (Eds) Applied Directional Statistics: Modern Methods and Case Studies, Chapman & Hall/CRC: New York, 45-59
8. Lagona F, Picone M, Maruotti A, Cosoli S (2014) A hidden Markov approach to the analysis of space-time environmental data with linear and circular components, Stochastic Environmental Research and Risk Assessment 29: 397-409
9. Lagona F, Picone M (2013) Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data, Journal of Statistical Computation and Simulation, 83: 1223-1237
10. Mastrantonio G (2018) The joint projected normal and skew-normal: A distribution for poly-cylindrical data, Journal of Multivariate Analysis, 165: 14-26

An experimental analysis on quality and security about green communication

Analisi sperimentale sulla qualità e sicurezza delle comunicazioni biocompatibili

Vito Santarcangelo, Emilio Massa, Davide Scintu, Michele Di Lecce, Massimiliano Giacalone

Abstract The paper introduces a new green approach of data communication through the use of visible light and audio. The study focuses on the quality and security of these two technologies.
Abstract *Il documento introduce un nuovo approccio biocompatibile relativo alla comunicazione dei dati attraverso l'uso di luce visibile e l'audio. Lo studio si concentra sull'analisi della qualità e sicurezza di queste due tecnologie.*

Key words: visible light communication, elusive analysis, data over audio, security analysis, green communication.

1 Introduction

In an innovative context where radio frequencies and 5G technologies are leading actors, our alternative approach wants to focus the reader attention on two experimental data communication infrastructures based on no radio frequencies as visible light and audio. This innovative approach for indoor data transmission can

[1] Vito Santarcangelo, iInformatica Srls – Corso Italia 77; email: vito@iinformatica.it
Emilio Massa, iInformatica Srls – Corso Italia 77; email: emilio@iinformatica.it
Davide Scintu, iInformatica Srls – Corso Italia 77; email: dscintu@iinformatica.it
Michele Di Lecce, iInformatica Srls – Corso Italia 77; email: michele@iinformatica.it
Massimiliano Giacalone, University Federico II of Naples; massiliano.giacalone@unina.it

represent an alternative to the main known BLE (Bluetooth low energy), Bluetooth, RFID (Radio frequency identification) and WIFI communications for short range distances. Scope of our work is very ambitious and aims to represent an interesting challenge path for innovation in data transmission without radio frequency and to analyze and implement relative security approach. A good indoor security level is reached, in fact, these kind of approach allows to border the transmission only to the room instead of radio frequencies (characterized by a better diffusion). Otherwise, these communication approaches can be useful to develop data communication for hostile environments where radio frequency can not be used (e.g. oil and gas environments). Possible applications of these two technologies are for anti-theft, tracking and surveillance purposes for devices, the remote control of mobile objects, the identification of objects using interfaces of internet of things, etc. However, it is important to analyze the possible vulnerabilities of these two technologies considering possible attacks and provide innovative methods for detection and prevention. Also the dynamic combination of these two data transmission mediums considering the environment parameters can be very useful to improve redundancy and reduce errors in data communication.

This paper is structured in four sections. In the second section visible light communication technology and approaches are described considering possible attacks and detection and prevention solutions. The same approach is followed for data over audio technology in the third section of the paper. Experimental results, conclusions and future development application are described in the fourth section.

2 Visible light communication

Visible light communication (VLC) operates, as the name implies, in the portion of the visible spectrum of light.

The visible spectrum is a part of the electromagnetic spectrum (EMI) with a wavelength ranging from ~ 700 (red) to ~ 400 (violet) nanometers with a frequency range between 428THz and 750THz.

The VLC technology has had an exponential growth in recent years, also given by the reduction in costs and the technical evolution of LED such as efficiency and brightness, therefore there is no approved standard defined by the IEEE (Institute of Electrical and Electronic Engineers).

The data are transmitted by modulating the light with a sequence of 0 and 1 through the use of light- emitting diodes (LED) while photodiodes, which are very sensitive to light, are mainly used as the receiver. Unfortunately, at low speeds the human eye will perceive a constant flickering of the light source. However, the LED modulating the light in an ultra- fast way, will be considered as a constant source in time without any variation.

The main data modulation techniques are: OOK (On-Off Keying), VPPM (Variable Pulse Position Modulation), CSK (Color shift keying).

The advantages of using VLC technology over radio frequencies consists in having an unlimited band (400nm - 700nm) without any regulation, it does not generate any electromagnetic interference, it is harmless for the human body and finally it has a higher transmission speed. It has been shown that through the laser diode based on gallium nitride (GaN) it is possible to reach a speed of 15 Gb/s with a distance of 15cm and 197cm respectively. The most commonly used slogan in the security field is "What You See Is What You Send" (WYSIWYS) so if there is an unauthorized device it would be easily recognizable through directionality and visibility. Although there is this assumption, VLC networks are not exempt from potential attacks. “Jamming” is an attack designed to disrupt communication so that the signal-to-noise ratio is reduced. In visible-spectrum communication, the jamming- based attack is easily implemented. To do it, it would be enough to have a higher light source (usually) than the transmitter that would disturb the transmitted data. Unfortunately, with this technique there are no effective solutions, except that of using another transmission medium like DoA in a coupled way. VLC networks can be subject to attacks such as the "man in the middle" (MIM) in which another device is interposed between the sender and the recipient for the purpose of secretly extracting the data and/or altering it by making people believe devices involved to have direct communication. In this case, it is sufficient to insert another device to be able to intercept the flow. The MIM attack can be countered by encrypting the data in both directions. The device that would like to obtain the data will no longer be able to decode them and therefore, it will be removed from the communication.

Arduino-based circuits have been created. The transmitter device is composed of a white LED flanked by a 220 Ohm resistor. On the receiver there are two voltage dividers with two photoresistors, which vary their resistance based on the amount of light that hits it and two 10KOhm resistors. The first photoresistor is directly hit by the LED that deals with transmitting the data, while the second detects the ambient light that will be subtracted from the first. The transmitter adopts an OOK data modulation. The two devices were placed at a distance of about 10cm. By inserting an interfering light source (Jamming) like the torch of the smartphone, it turns out to be very effective to disturb the signal and even to neutralize it. All messages sent were not received correctly.

As regards the MIM connection, another device has been placed in which both a photoresistor and an LED are present, so as to obtain the message from the transmitter and send it to the receiver altering its contents. Even this attack was successful without the devices noticing a third device.

3 Data Over Audio

DoA (Data over audio) is a technology that allows data to be sent between devices via audible or inaudible audio signals to man, maintains the advantages of wired indoor networks allowing the creation of private networks, limited and without

electromagnetic pollution while maintaining convenience of wireless networks based on RF technology.

The information to be transmitted is enriched with information useful for transmission (eg timestamp, sender), followed by encryption to preserve security and subsequently algorithms are used to preserve the loss of information (eg reed-solomon codes). The information is transformed into audio frequencies and then sent. The amount of information that can be sent depends on the frequency range you want to use and on the number of devices you want to communicate simultaneously. Some examples of the communication speed: 7bytes in 3s(ultrasonic, multi-channel), 8bytes in 4s(ultrasonic, single-channel), 32bytes(acoustic, single-channel).
Once the communication is started, algorithms are used to clean up the sound from the background noise and from frequencies not relevant to the communication, finally the sound is saved and transformed back in the original information.

The range of action of the communication depends on the frequencies used, the capacity of the speakers and the background noise. The DDos (Distributed Denial of Service) attack consists of one or more devices within the range of a DoA communication that during this communication send audio signals with the purpose of blocking the transmission of data by interposing its audio signal to that of communication. Algorithms like reed-solomon codes to reduce errors in communication, combined with algorithms to distinguish to noises unrelated to DoA communications, can be useful against DDoS attacks, even though this technology is still weak to such attacks. An interlocutor placed between the DoA-DoA communication (emitter and receiver side) can represent an elusive factor of some systems. The DoA receiver will read the signal through a special microphone and will authenticate. However, if the signal doesn't arrive directly to the receiver but is transmitted remotely, it could be a communication interposed between the emitter and the final DoA receiver (e.g. with the use of RF or VOIP). A method based on a comparison between frequencies constituting an original signal and a radio transmitted signal can be an effective anti-elusive method. An identifying footprint of the real presence of a radio broadcast can be represented by the White Gaussian Noise Additive (AWGN) that is present in every radio broadcast. In the electronic field, a manifestation of these signals is described as thermal noise. The observation of this phenomenon can be physically implemented on the receiving side through a device suitable for monitoring a resistor to which no voltage is applied. Instead with AWGN's presence thermal agitation occurs which generates a noise voltage that varies over time and depending on the temperature. By controlling this resistor is therefore possible to perceive whether the received signal is or is not an attempt to circumvent the DoA system as a function of the voltage at its ends. Thermal agitation is not entirely deterministic, which is why the AWGN is described by a random variable and an average zero value and by its white spectrum as it's flat in a wide range of frequencies. A device equipped with a special anti-elusive system based on scanning and frequency comparison between a reference and the received signal, would make it possible to interpret the incoming signal and perform a check before giving an output. Thus, even managing to evade the first authentication, a second check based on the analysis of the signal spectrum arriving at the receiver would

exclude the possibility of remote elusion; for the prevention of recorded messages, a possible solution concerns the timestamp sent in the encrypted audio message, this added data allows to have an audio sent always different and allows to assign a deadline to the message.

4 Experimental results and conclusions

Real experiments with application of DoA have been conducted for safety in construction site areas (Caldarola patent), in home automation lamps (Thegg Domotica patent) and for coffee vending machines (Coffee Express patent). In the first case study an "ALT" alert is revealed by the fleet in critical areas. The simulation and monitoring of the site areas is made through the use of DoA emitters which allow to have a rigorous control of action areas through augmented reality alerts in the earth moving machinery HMIs. In the second case study the lamp is managed through DoA, in the third case study the vending machine interaction is provided by DoA channel.

Figure 1: DoA experimental Case Study

To provide an extimation of perfomances we have implemented a test scenario with the use of DoA and VLC communication channel. We have used as input the AIN Thesaurus, characterized by over 1200 adjectives and adverbs of Italian language. We have also considered the application of AES cryptography to AIN Theasurus strings. Considering a distance of 0,5 m the DoA performance with a basic microphone (SunFounder Mini Microphone) reaches 97,4%. The transmission rate is 4,52s/32bytes. Some problems of decoding are obtained with the use of AES. In this case the string is characterized by 64 bytes. Different results are obtained with the use of VLC technology, that reaches 99,1% of performance without the use of

cryptography till a distance of 0,5 meters. It represents an emerging challenge to a correct performance estimation.

Table 1: Estimated Performances Results

Input Test Set	Distances	Technology	Estimated Performance
AIN Thesaurus	2 m	DoA	88,3%
AES + AIN Thesaurus	2 m	DoA	56,9%
AIN Thesaurus	0,5 m	DoA	97,4%
AES + AIN Thesaurus	0,5 m	DoA	78,3%
AIN Thesaurus	0,5 m	VLC	99,1%
AES + AIN Thesaurus	0,5 m	VLC	91,5%

These technologies have been the subject of experimental green approaches conducted in some companies of southern Italy. These prototype technologies have represented an important solution in communication field in environments characterized by an explosive atmosphere studied, like in the case of L'Antincendio (Colucci, M., 2019). In this hostile environment there is a mixture of flammable substances in which the radiofrequency can represent a source of ignition. The implementation of an environmental monitoring station that communicates via a VLC-DoA experimental hybrid gateway represented a possible solution for these environments. These technologies represent a considerable commitment in terms of corporate social responsibility (Barile, A., 2019). An other important development is about the communication in a green accommodation area as that of Matera Inerti using a disused aggregate quarry (Antonello Ribba's patent). These case studies represent an evidence of the space available for new challenge in industrial research and experimental development of solutions which require performance estimation.

References

1. Barile, A., (2019), "*Dispositivo e relativo metodo per il monitoraggio degli ambienti produttivi di una azienda di panificazione in ottica di responsabilità sociale*", Patent, UIBM
2. Colucci, M., (2019), "*Dispositivo gateway innovativo biocompatibile e relativo metodo per la trasmissione dati*", Patent, UIBM
3. Ding, L. ,Wang, H., Chen, J. (2009), "*Tracking under additive white Gaussian noise effec*t", 7th Asian Control Conference
4. Kelly, J.J., Joplin, Mo., (Jul. 10, 1990), "*System For The Recognition Of Automated Telephone Answering Devices And Delivery Of Prerecorded Messages To Such Devices*", U.S. Patent
5. Moore, R., Lopes, J., (1999). Paper templates. In TEMPLATE'06, 1st International Conference on Template Production. SCITEPRESS.
6. Nawrocki, W., Berthel, K.H., Doehler, T., Kock, H., (December 1989), "*Measuremnt of thermal noise by d.c. SQUID*", Cryogenics
7. Viola, S., Islim, M.S., Watson, S., Videv, S., Haas, H., Kelly, Anthony E., (2017), "*15 Gb/s OFDM-based VLC using direct modulation of 450 GaN laser diode*," Proc. SPIE 10437, Advanced Free-Space Optical Communication Techniques and Applications III

An improved sensitivity-data based method for probabilistic ecological risk assessment

Un metodo efficiente per la valutazione probabilistica del rischio ecotossicologico basato su dati di sensitivitá

Sonia Migliorati and Gianna Serafina Monti

Abstract Ecological risk assessment procedures are essentially based on Species sensitivity distributions (SSDs), a community-level concentration-response curve. In this contribution we propose a new definition of SSD as a finite mixture of log-logistic distributions, which integrates all the information available in the toxicity dataset, in a conservative perspective. We provide parameter estimation within a Bayesian MCMC approach. Furthermore, we propose the derivation of a more realistic hazardous chemical concentration HCp, the concentration protective of $(1-p)\%$ species in the environment. An application of our methods to real data completes the work.

Abstract *Le distribuzioni di sensibilitá delle specie (SSD) rispetto alle concentrazioni di un certo agente tossico sono sempre piú integrate nelle procedure di valutazione del rischio ecotossicologico. In questo contributo, in un'ottica conservativa nella valutazione del rischio ecotossicologico, viene proposta una nuova definizione di SSD basata su una mistura finita di distribuzioni log-logistiche, in grado di valorizzare il contributo informativo contenuto nei dati, sia a livello di specie testate che di modalitá d'azione delle sostanze tossiche. L'inferenza é condotta secondo un approccio Bayesiano. Viene fornito quindi un modo alternativo per la stima di HCp, ovvero della concentrazione protettiva per l'*$(1-p)\%$ *delle specie nell'ambiente di riferimento, su cui potranno basarsi le decisioni normative. Un caso applicativo mostra l'efficacia del metodo proposto.*

Key words: Species Sensitivity Distribution, MCMC, hierarchical models, log-logistic distribution

Sonia Migliorati
Department of Economics, Management and Statistics, University of Milano Bicocca, e-mail: sonia.migliorati@unimib.it

Gianna Serafina Monti
Department of Economics, Management and Statistics, University of Milano Bicocca, e-mail: gianna.monti@unimib.it

1 Species Sensitivity Distributions in risk assessment

Species Sensitivity Distributions (SSDs) are a widespread method for ecotoxicological risk assessment of chemicals (Posthuma et al., 2002). An SSD model relates the variation in the tolerance of tested species to a specific compound.

Let n denote the number of different species tested with respect to a certain chemical compound $i\,(i = 1,\ldots,k)$, and let $y_{ij}\,(j = 1,\ldots,n)$, denote the effective concentration data, generally log(base 10)-transformed, of the j^{th} species under study. The value y_{ij} refers to a specific toxicity endpoint derived from a single-species toxicity study, for example a median effective/hazardous concentration value (EC50), or a no observed effect concentrations (NOEC). Once the data are arranged in non-decreasing order, let $x_{ij} = i/(n+1)$ indicate the relative rank of each species, also called Weibull plotting positions.

A generic model formulation to fit a probabilistic SSD curve to the observed data can be written as:

$$SSD_i = F_i(y;\boldsymbol{\theta}), \qquad (i = 1,\ldots,k) \tag{1}$$

where $F_i(.;\boldsymbol{\theta})$ is the cumulative distribution function that describes the relationship between the (transformed) species concentration levels, y, and the corresponding plotting positions, whereas $\boldsymbol{\theta}$ is the unknown model parameter vector.

Several distributions were proposed to fit an SSD model. A log-normal distribution is used by the European Commission (Wagner and Lokke, 1991; ECHA, 2008) and it is considered the standard choice for SSD model (Aldenberg and Jaworska, 2000; Posthuma et al., 2002). Though, Newman et al. (2000) pointed out the scarcity of goodness of fit of a log-normal model in several examples. An alternative model is the log-logistic distribution (Aldenberg and Slob, 1993; Wheeler et al., 2002), especially adequate from a conservative perspective, thanks to its extended tails. In this case the parameter vector $\boldsymbol{\theta}$ is equal to (μ, σ), where μ is the location parameter and σ the scale parameter.

SSDs are generally used to assess the degree of hazard posed by a given chemical to species, represented by hazardous chemical concentration (HCp), that is a concentration harmful to the p% of species in a given assemblage. Particular attention is deserved to the estimate of HC5, as it corresponds to an estimate of a predicted no effect concentration (PNEC), frequently used in environmental risk assessment. Several methods are present in literature to obtain HCp estimates. Let's briefly remember the one proposed by Aldenberg and Slob (1993), based on a log-logistic distribution, i.e.

$$\widehat{\text{HC5}}_{AS} = \bar{y} - \kappa_n s \tag{2}$$

where $\bar{y}$ and s are the sample mean and standard deviation of the observed endpoints respectively, and κ_n is an extrapolation factor, tabulated for various sample sizes n.

2 Hierarchical species sensitivity distribution

One of the structural problems when attempting to apply the SSD approach to ecological risk assessment is the scarcity of available data. To overcome this limit, we propose here to incorporate all available information into a hierarchical model for SSDs expressed by

$$\begin{aligned} & y_{ij}|\mu,\delta_i,\gamma_j,\sigma^2 \sim \text{Logistic}(\mu+\delta_i+\gamma_j,\sigma^2), \\ & \mu|\sigma_\mu \sim \text{N}(0,\sigma_\mu^2), \\ & \delta_i|\sigma_\delta \sim \text{N}(0,\sigma_\delta^2), \quad \text{(random substance effect)}, \\ & \gamma_j|\sigma_\gamma \sim \text{N}(0,\sigma_\gamma^2), \quad \text{(random species effect)}, \end{aligned} \tag{3}$$

where y_{ij} is the log-(base 10) toxicity value for compound i, ($i = 1,\ldots,k$), tested on species j ($j = 1,\ldots,n$). Thus, and differently from the standard approach, we let substance and species effects to be represented by random effects with a dedicated variance parameter. Hence, for each substance-species combination (i,j) the residuals are a random sample from a distribution centered about zero with homogeneous variance σ^2 that is independent of experimental conditions, chemical, and species.

In a Bayesian perspective, the (weakly informative) hyper-prior distributions can be assigned as follows:

$$\sigma \sim \text{Uniform}(0,100),\ \ \sigma_\delta \sim \text{Uniform}(0,100),\ \ \sigma_\gamma \sim \text{Uniform}(0,100)\,. \tag{4}$$

furthermore we set σ_u to a high value. The posterior distributions of the parameters of interest are obtained via Markov chain Monte Carlo (MCMC) methods.

By considering substances and species effects as random, we succeed in enriching the model variability structure, usually represented by σ^2 (i.e. the measurement error), by two further components, namely $\sigma_\delta^2, \sigma_\gamma^2$ (Gelman, 2015). The latter account for variability between substances and between species.

Thus, the proposed SSD for substance i, namely $\text{MSSD}_i(y)$, can be expressed in terms of the hierarchical model (3) as a finite mixture of logistic CDFs:

$$\text{MSSD}_i(y) = \frac{1}{n}\sum_{j=1}^{n} F_{ij}\left(y;\mu+\delta_i+\gamma_j,\sigma^2\right). \tag{5}$$

Furthermore, the posterior CDF of the hazardous concentration HCp can be estimated from the posterior of $\text{MSSD}_i(y)$ by observing that, given a value y, the following equality holds

$$P(\text{MSSD}_i(y) \le p) = P(\text{HCp}_M \ge y)\ . \tag{6}$$

Therefore, the CDF of HCp_M at point y can be estimated by the proportion of MCMC sample values of $\text{MSSD}_i(y)$ that are greater than or equal to p. To better

reflect uncertainty in the estimate, essentially due to small sample size, we propose to calculate the lower one-tailed 95% confidence limit.

3 An application to real toxicity data

3.1 Data description

To demonstrate the effectiveness of our procedure we considered data from Versteeg et al. (1999), related to a variety of substances ($k = 11$), including heavy metals, pesticides, surfactants, and general organic and inorganic compounds, tested on several species $n = 56$. This is an unbalanced design, as some species have not been tested on all chemicals. Multiple data for the same species were summarised as geometric means using the NOEC or EC20 concentration.

3.2 Implementation and results

We run three MCMC chains in parallel. For each chain, after a burn-in period of 5,000 samples, to reach stationarity of the chain, 10,000 samples of the random variables were generated with a thinning rate of 50, i.e. we discarded all but every 50-th observation, with the goal of reducing autocorrelation. All analyses were performed using R and OpenBUGS softwares (R Core Team, 2019; Lunn et al., 2000).

In Figure 1 we show the autocorrelation plots and partial time series plots of the parameters σ, σ_γ and σ_δ, which are typical diagnostic tools used to assess convergence.

To appreciate the well-fitting of the $\text{MSSD}_i(y)$, given by (5), let us consider Figure 2, which shows both the MSSD and the standard log-logistic SSD, together with their 90% pointwise confidence bands for three different toxic compounds. The log-logistic SSD confidence bands are obtained using 1000 bootstrap resamples. The proposed model clearly improves the fit of the log-logistic model, as it better captures the variability of sample data gathering information from differences in sensitivity of the selected sample species.

We summarised and compared the results of our analyses by using HC5 point estimates with associated lower 95% confidence intervals (one tailed) (Table 1).

In particular we calculated the sample quantile $\widehat{\text{HC5}}_q$, the HC5 estimated according to Aldenberg and Slob (1993) (see (2)) and its lower left-sided 95% confidence limit (one tailed), i.e. $\widehat{\text{HC5}}_{AS}$ and $\widehat{\text{HC5}}_{ASL}$ respectively, and $\widehat{\text{HC5}}_M$ computed according to (6) along with its lower left-sided 95% confidence limit $\widehat{\text{HC5}}_{ML}$.

As we might expect, the lower left-sided 95% confidence limit of HC5 estimate given by our model gives a reasonable estimate of the PNEC. Indeed, $\widehat{\text{HC5}}_{ML}$ is

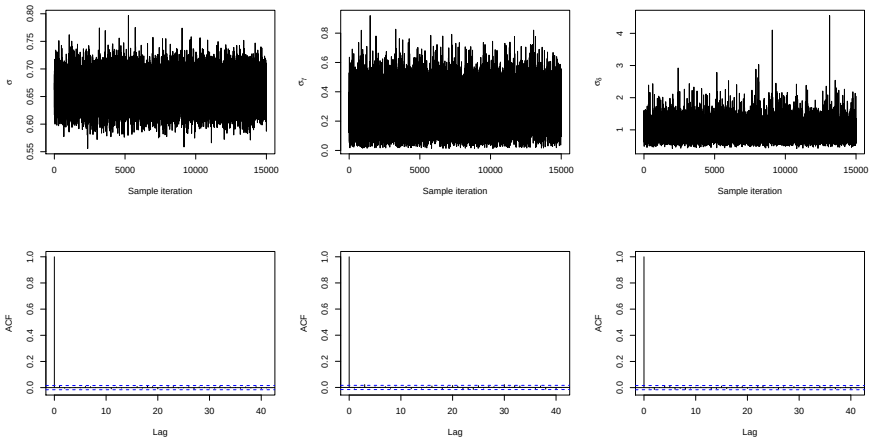


Fig. 1 Diagnostic tools to assess convergence of MCMC samples with respect to the parameters (from left to right) σ, σ_γ and σ_δ. Upper panels: time-series plots (after the burn-in period and thinning regime). Lower panels: autocorrelation functions.

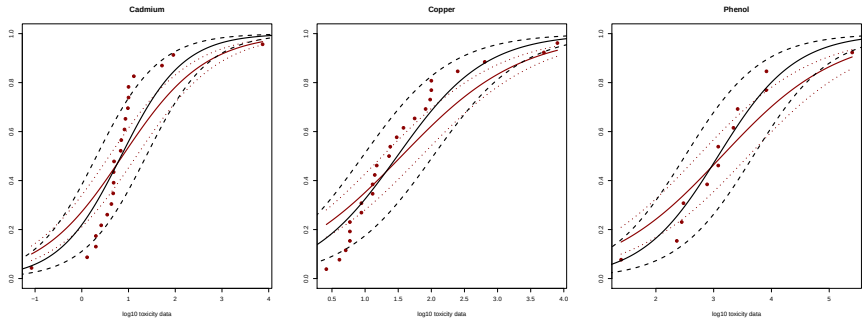


Fig. 2 Fitted SSD curves for three toxic compounds. Log-Logistic SSD with 90% bands (dark red solid and dashed lines), MSSD with 90% pointwise confidence bands (solid and dashed black lines). Points correspond to log-NOEC data of tested species.

always lower than $\widehat{\text{HC5}}_q$, but higher than $\widehat{\text{HC5}}_{ASL}$, thus realizing a good compromise between the two existing proposals.

References

Aldenberg, T. and Jaworska, J. S. (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and Environmental Safety*, 46:1–18.

Table 1 HC5 estimates on original scale for the selected compounds: sample quantile, ($\widehat{\text{HC5}}_q$), HC5 estimated according to Aldenberg and Slob (1993) and its lower left-sided 95% confidence limit (one tailed), and estimates derived from MSSD curve ($\widehat{\text{HC5}}_M$) together with lower left-sided 95% confidence limits.

Compound	$\widehat{\text{HC5}}_q$	$\widehat{\text{HC5}}_{AS}$	$\widehat{\text{HC5}}_{ASL}$	$\widehat{\text{HC5}}_M$	$\widehat{\text{HC5}}_{ML}$
3,4 DCA	2.00	0.09	0.02	3.80	0.91
Ammonia	4.05	0.86	0.19	13.49	3.55
Atrazine	1.22	0.37	0.19	1.58	0.43
C12LAS	206.42	111.39	68.65	125.89	37.15
C12TMAC	44.41	16.28	6.17	18.62	4.79
Cadmium	1.33	0.10	0.06	0.55	0.16
Chlorine	3.44	0.41	0.07	1.82	0.40
Copper	4.28	0.44	0.25	2.40	0.72
Lindane	0.33	0.05	0.01	1.02	0.24
Phenol	84.73	7.56	1.85	93.33	25.12
Zinc	13.76	2.18	0.48	7.94	1.82

Aldenberg, T. and Slob, W. (1993). Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotoxicology and Environmental Safety*, 25(1):48 – 63.

ECHA (2008). Guidance for the implementation of REACH: Guidance on information requirements and chemical safety assessment. *Chapter R.10: Characterisation of Dose [Concentration]-Response for Environment, May 2008. Available at: http://guidance .echa.europa.eu/docs/guidance document/information requirements r10 en.pdf. Accessed on June 2014*.

Gelman, A. (2015). Analysis of variance - why it is more important than ever. *The Annals of Statistics*, 33(1):1–33.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Newman, M. C., Ownby, D. R., Mézin, L. C. A., Powell, D. C., Christensen, T. R. L., Lerberg, S. B., and Anderson, B. (2000). Applying species-sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient numbers of species. *Environmental Toxicology and Chemistry*, 19(2):508–515.

Posthuma, L., Suter, G. W., and Traas, T. P. (2002). *Species Sensitivity Distribution in Ecotoxicology*. Lewis Publishers, Boca Raton, FL.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Versteeg, D. J., Belanger, S. E., and Carr, G. J. (1999). Understanding single-species and model ecosystem sensitivity: Data-based comparison. *Environmental Toxicology and Chemistry*, 18(6):1329–1346.

Wagner, C. and Lokke, H. (1991). Estimation of ecotoxicological protection levels from NOEC toxicity data. *Water Research*, 25(10):1237 – 1242.

Wheeler, J. R., Grist, E. P., Leung, K. M., Morritt, D., and Crane, M. (2002). Species sensitivity distributions: data and model choice. *Marine Pollution Bulletin*, 45(1-12):192–202.

Comparing predictive distributions in EMOS

Distribuzioni predittive per modelli EMOS

Giummolè Federica and Mameli Valentina

Abstract EMOS models are widely used post-processing techniques for obtaining predictive distributions from ensembles for future weather variables. A predictive distribution can be easily obtained by substituting the unknown parameters with suitable estimates in the distribution of the future variable, thus obtaining a so called estimative distribution. Nonetheless, these distributions may perform poorly in terms of coverage probability of the corresponding quantiles. In this work we propose the use of calibrated predictive distributions in the context of EMOS models. The proposed calibrated predictive distribution improves on estimative solutions, producing quantiles with exact coverage level. A simulation study assesses the goodness of the calibrated predictive distribution in terms of coverage probabilities and also logarithmic score and CRPS.

Abstract *I modelli EMOS forniscono un metodo per ottenere distribuzioni predittive a partire da un insieme di previsioni per una variabile meteorologica di interesse. Una distribuzione predittiva si può ottenere facilmente sostituendo i parametri non noti con delle stime opportune nella distribuzione della variabile futura. Questa procedura dà origine alle cosiddette distribuzioni estimative che però spesso risultano inadeguate in quanto la probabilità di copertura associata ai loro quantili differisce da quella nominale. In questo lavoro proponiamo, nel contesto dei modelli EMOS, una distribuzione predittiva calibrata che fornisce quantili la cui probabilità di copertura coincide con quella nominale. Uno studio di simulazione evidenzia la bontà della predittiva proposta, sia in termini di probabilità di copertura che rispetto alle funzioni di perdita logaritmica e CRPS.*

Key words: Coverage probability, CRPS, EMOS, logarithmic score, predictive distribution.

Giummolè Federica
Ca' Foscari University of Venice, Via Torino 155 Mestre, e-mail: giummole@unive.it

Valentina Mameli
University of Udine, Via Tomadini 30 Udine, e-mail: valentina.mameli@uniud.it

1 Introduction

In modern society, weather conditions have wide-ranging economic impacts in fields as diverse as aviation, shipping, tourism and agriculture, just to name a few. All these important applications require accurate forecasts of future weather conditions. Weather forecasts are usually provided as forecast ensembles obtained from multiple numerical models achieved using different initial conditions and different numerical representations of the atmosphere [8]. However, such ensemble forecasts are able to capture only part of the forecast uncertainty exhibiting dispersion errors and systematic biases [7], [2]. For this reason, ensemble forecasts are often statistically post-processed to produce calibrated predictive distributions. Many statistical post-processing methods have been proposed in the literature. The most popular are the ensemble model output statistics (EMOS) that allow for probabilistic forecasts of continuous weather variables ([4]). EMOS is nothing but a linear regression model with heteroschedastic Gaussian errors. The EMOS mean is a linear combination of the ensemble member forecasts, with unknown coefficients that represent the contributions of each member of the ensemble to the interest weather variable. The EMOS variance is a linear function of the ensemble variance that accounts for spread relationship. More precisely, it is assumed that a weather continuos variable Y depends on the ensemble forecasts $X_1,\dots,X_m$ in such a way that its mean is equal to $\beta_0+\beta_1X_1+\ldots+\beta_mX_m$ and its variance is equal to $\gamma+\delta S^2$, where $S^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2$ denotes the ensemble variance and $\beta_0,\dots,\beta_m$, $\gamma>0$ and $\delta>0$ are unknown coefficients. Under normality assumptions, the distribution of Y is $N(\beta_0+\beta_1X_1+\ldots+\beta_mX_m,\gamma+\delta S^2)$. Suitable estimates are then substituted to the unknown parameters, obtaining what is known as an estimative distribution for the future weather quantity Y:

$$N(\hat{\beta}_0+\hat{\beta}_1X_1+\ldots+\hat{\beta}_mX_m,\hat{\gamma}+\hat{\delta}S^2),$$

where $\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_m,\hat{\gamma}$, and $\hat{\delta}$ are suitable estimates of $\beta_0,\beta_1,\dots,\beta_m,\gamma$, and δ, respectively.

Unknown parameters can be estimated using the method of maximum likelihood or of minimum Continuous Ranked Probability Score (CRPS), which respectively optimise the logarithmic score and the CRPS, see [5, 6].

Unfortunately, an estimative distribution can lead to poor prediction statements, since it does not take into account for the uncertainty introduced by substituting estimates to the true parameter values. In particular, the coverage of prediction intervals obtained by the estimative distribution does not achieve the nominal coverage level, see [1, 3].

In this work we recommend for the EMOS model the use of a calibrated predictive distribution based on a bootstrap procedure proposed by [3], which improves on the estimative solution. On a simulation study, we compare the Gaussian estimative distributions obtained with minimum CRPS estimates and maximum likelihood estimates with their calibrated counterparts. We show that the calibrated predictive distributions always improve on the estimative ones in terms of coverage of pre-

diction intervals and of logarithmic score. As regards the CRPS, all the considered distributions perform in a similar way.

2 Calibrated predictive distributions

In this section we briefly review, in the context of EMOS models, the calibrating approach proposed by [3], which provides predictive distributions whose quantiles give well-calibrated coverage probability.

Suppose that $\{Y_i\}_{i\geq 1}$ is a sequence of independent continuous random variables with probability distribution specified by the EMOS model:

$$Y_i \sim N(\beta_0+\beta_1 x_{i1}+\ldots+\beta_m x_{im}, \gamma+\delta s_i^2), \quad i\geq 1,$$

where x_{ij} is the i-th value of the ensemble variable X_j, $j=1,\ldots,m$, s_i^2 is the ensemble variance of $x_{i1},\ldots,x_{im}$, and $\theta=(\beta_0,\beta_1,\ldots,\beta_m,\gamma,\delta)$ is the unknown parameter vector. We assume that $Y=(Y_1,\ldots,Y_n)$, $n>1$, is observable, while $Z=Y_{n+1}$ is a future or not yet available variable. We indicate with $\Phi(z;\theta)$ the Gaussian cumulative distribution function of Z.

Given the observed sample $y=(y_1,\ldots,y_n)$, an α-prediction limit for Z is a function $c_\alpha(y)$ such that, exactly or approximately,

$$P_{Y,Z}\{Z\leq c_\alpha(Y);\theta\}=\alpha, \tag{1}$$

for every $\theta\in\Theta$ and for every fixed $\alpha\in(0,1)$. The above probability is called coverage probability and it is calculated with respect to the joint distribution of (Z,Y).

Consider a suitable asymptotically efficient estimator $\hat{\theta}=\hat{\theta}(Y)$ for θ and the estimative prediction limit $z_\alpha(\hat{\theta})$, which is obtained as the α-quantile of the estimative distribution function $\Phi(\cdot;\hat{\theta})$. The associated coverage probability is

$$P_{Y,Z}\{Z\leq z_\alpha(\hat{\theta}(Y));\theta\}=E_Y[\Phi\{z_\alpha(\hat{\theta}(Y));\theta\};\theta]=C(\alpha,\theta) \tag{2}$$

and, although its explicit expression is rarely available, it is well-known that it does not match the target value α even if, asymptotically, $C(\alpha,\theta)=\alpha+O(n^{-1})$, as $n\to+\infty$, see e.g. [1]. As proved in [3], the function

$$G_c(z;\hat{\theta},\theta)=C\{\Phi(z;\hat{\theta}),\theta\}, \tag{3}$$

which is obtained by substituting α with $\Phi(z;\hat{\theta})$ in $C(\alpha,\theta)$, is a proper predictive distribution function, provided that $C(\cdot,\theta)$ is a sufficiently smooth function. Furthermore, it gives, as quantiles, prediction limits $z_\alpha^c(\hat{\theta},\theta)$ with coverage probability equal to the target nominal value α, for all $\alpha\in(0,1)$.

The calibrated predictive distribution (3) is not useful in practice, since it depends on the unknown parameter θ. However, a suitable parametric bootstrap estimator for

$G_c(z;\hat{\theta},\theta)$ may be readily defined. Let y^b, $b=1,\ldots,B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^b$, $b=1,\ldots,B$, be the corresponding estimates. Since $C(\alpha,\theta)=E_Y[\Phi\{z_\alpha(\hat{\theta}(Y));\theta\};\theta]$, we define the bootstrap calibrated predictive distribution as

$$G_c^{boot}(z;\hat{\theta})=\frac{1}{B}\sum_{b=1}^{B}\Phi\{z_\alpha(\hat{\theta}^b);\hat{\theta}\}|_{\alpha=\Phi(z;\hat{\theta})}. \tag{4}$$

The corresponding α-quantile defines, for each $\alpha\in(0,1)$, a prediction limit having coverage probability equal to the target α, with an error term which depends on the efficiency of the bootstrap simulation procedure.

3 A simulation study

In order to assess and compare the performance of the estimative and the calibrated predictive distributions for the EMOS model we perform several experiments with simulated ensembles. The ensemble members are drawn from a m-variate normal distribution with zero mean and identity covariance matrix, with $m=5,10,15$. The i-th observation is generated from a normal random variable with mean $\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}$ where $\beta_j=(j+1)/\sum_k\beta_k$, $j=0,\ldots,m$, and variance $\gamma+\delta s_i^2$, with $\gamma=0$ and $\delta=1$, $i=1,\ldots,n$ with $n=20$. The bootstrap procedure is based on 500 bootstrap samples. The estimation is based on 1000 Monte Carlo replications. We evaluate the estimative and calibrated predictive distributions in terms of coverage probabilities and also using the logarithmic score and CRPS as loss functions, as commonly used in the literature, see [4, 9]. It should be noted that the calibration procedure is based on asymptotic considerations. Thus the improvement over estimative results is more evident with small sample sizes. Here we have chosen $n=20$, having in mind to use the 20 most recent daily observations of a meteorological variable as the training period for estimating the EMOS model parameters, see [4].

Table 1 provides the results of a simulation study for comparing coverage probabilities of 66.7% and 90% central prediction intervals obtained from the estimative and the calibrated distributions with minimum CRPS and maximum likelihood estimates. For this aim, we consider prediction limits of levels $\alpha=0.05,0.167,0.833$, and 0.95. It can be noted that the coverage probabilities associated to the calibrated quantiles almost equal the nominal values, showing accurate coverage. Average width of prediction intervals, not shown here, demonstrates that the calibrated predictive distributions yield slightly longer prediction intervals with respect to the estimative ones, but this can be explained by the greater coverage of these prediction intervals.

We assess the improvement of the calibrated predictive distributions over the estimative ones by computing also the logarithmic score and the CRPS, averaged over 1000 replicates, as shown in Table 2. The superior performance of the calibrated dis-

tributions is reflected in the values of the logarithmic score. Indeed average values of the logarithmic score for estimative distributions are significantly worse with respect to their calibrated counterparts. In terms of the CRPS, the estimative solutions perform similarly to the calibrated ones.

4 Conclusions

This work proposes a comparison between estimative and calibrated predictive distributions based on a bootstrap resampling procedure. The comparison is carried out on a simulation study, where appropriate verification measures, such as the CRPS, logarithmic score and coverage probabilities, are used for assessing the predictive performance of the considered distributions. From the results one can conclude that the calibrated predictive distributions always improve on the estimative ones, in terms of logarithmic score and coverage. Instead, the considered predictive distributions perform similarly with respect to the CRPS. Future development of the work will explore sliding training periods of constant size in the same way as in the work of [4]. Moreover other verification measures will also be considered.

Acknowledgments
This work was partially supported by the Italian Ministry for University and Research under the PRIN2015 Grant No. 2015EASZFS_003.

References

1. Barndorff-Nielsen, O.E., Cox, D.R.: Prediction and asymptotics. Bernoulli, **2**, 319–340 (1996).
2. Buizza, R., Houtekamer, P.L, Toth, Z., Pellerin, G., Wei, M., Zhu, Y.: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. Monthly Weather Review, **133**, 1076–1097 (2005).
3. Fonseca, G., Giummolè, F., Vidoni, P.: Calibrating predictive distributions. Journal of Statistical Computation and Simulation, **84**, 373–383 (2014).
4. Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman , T.: Calibrated ProbabilisticForecasting Using Ensemble Model Output Statistics and Minimum CPRS Estimation. Monthly Weather Review, **133(5)**, 1098–1118 (2005).
5. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, **102**, 359–378 (2007).
6. Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**, 243–268 (2007).
7. Hamill, T. M., Colucci, S. J.: Verification of Eta-RSM short-range ensemble forecasts. Monthly Weather Review, **125**, 1312–1327 (1997).

$m = 5$				
α	Est log	Cal log	Est crps	Cal crps
0.05	0.125	0.052	0.128	0.050
0.167	0.249	0.181	0.274	0.180
0.833	0.730	0.803	0.722	0.809
0.95	0.865	0.946	0.854	0.935
$m = 10$				
α	Est log	Cal log	Est crps	Cal crps
0.05	0.162	0.050	0.184	0.046
0.167	0.290	0.191	0.304	0.187
0.833	0.719	0.822	0.719	0.823
0.95	0.845	0.939	0.836	0.941
$m = 15$				
α	Est log	Cal log	Est crps	Cal crps
0.05	0.146	0.044	0.151	0.035
0.167	0.265	0.152	0.265	0.134
0.833	0.766	0.864	0.757	0.867
0.95	0.864	0.959	0.852	0.967

Table 1 Coverage probabilities of the four predictive distributions for different nominal levels α. Standard errors are always smaller than 0.015. Est log denotes the estimative distribution with maximum likelihood estimates and Est crps the estimative distribution with CRPS estimates, while Cal log and Cal crps are the respective calibrated counterparts.

$m = 5$			
Est log	Cal log	Est crps	Cal crps
1.672	1.56	1.716	1.582
(0.043)	(0.022)	(0.047)	(0.023)
$m = 10$			
Est log	Cal log	Est crps	Cal crps
2.277	1.825	2.539	1.87
(0.085)	(0.0295)	(0.114)	(0.0368)
$m = 15$			
Est log	Cal log	Est crps	Cal crps
2.493	1.656	3.052	1.687
(0.116)	(0.024)	(0.158)	(0.022)

$m = 5$			
Est log	Cal log	Est crps	Cal crps
0.636	0.636	0.644	0.645
(0.016)	(0.014)	(0.016)	(0.015)
$m = 10$			
Est log	Cal log	Est crps	Cal crps
0.792	0.806	0.809	0.817
(0.020)	(0.0179)	(0.021)	(0.18)
$m = 15$			
Est log	Cal log	Est crps	Cal crps
0.695	0.691	0.716	0.706
(0.018)	(0.015)	(0.019)	(0.015)

Table 2 Logarithmic (left) and CRPS (right) values of the four predictive distributions for different values of m. Est log denotes the estimative distribution with maximum likelihood estimates and Est crps the estimative distribution with CRPS estimates, while Cal log and Cal crps are the respective calibrated counterparts.

8. Palmer, T.N.: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. Quarterly Journal of the Royal Meteorological Society, **128**, 747–774 (2002).
9. Schlosser, L., Hothorn, T., Stauffer, R., Zeileis, A.: Distributional regression forests for probabilistic precipitation forecasting in complex terrain. The Annals of Applied Statistics, **13(3)**, 1564–1589 (2019).

Compositional analysis of fish communities in a fast changing marine ecosystem

Analisi composizionale di comunità ittiche in un ecosistema marino in rapida evoluzione

M. Mingione, P. Alaimo Di Loro, G. Jona Lasinio, S. Martino, F. Colloca

Abstract Concerns about the effects of human activities and climate change on the Mediterranean marine ecosystem are recently demanding new research work on the response of fish communities to environmental pressures to which they are subjected. The main limitation of standard statistical approaches is that they cannot properly account for the complex interactions occurring among sub-populations. In this work, we analyze a community of 129 demersal fish species living in front of the Lazio coast. The goal is to understand if the compositions changed over time and, if so, verify any statistical relationship with changes in water temperature. Results show that this area of Mediterranean Sea is undergoing a slow but progressive meridionalization of the fish inhabiting its waters.

Abstract Le preoccupazioni dovute agli effetti delle attività umane e del cambiamento climatico sull'ecosistema marino del Mediterraneo hanno recentemente richiesto nuovi lavori di ricerca sulla risposta delle comunità ittiche alle pressioni ambientali di cui soffrono. Le limitazioni principali degli approcci statistici standard è che non possono adeguatamente gestire la complessità delle interazioni che si osservano tra sottopopolazioni. In questo lavoro, studiamo una comunità di 129 specie di pesci demersali che abitano la costa del Lazio. Lo scopo principale è quello di capire se le composizioni siano cambiate nel tempo e, nel caso, verificare possibili relazioni statistiche con variazioni della temperatura. I risultati mostrano che questa zona del Mediterraneo sta subendo una lenta ma progressiva meridionalizzazione dei pesci che la abitano.

Key words: Compositional analysis, Evolution of fish community, Climate change

Marco Mingione
University of Rome "La Sapienza", Statistical Science Department, e-mail: marco.mingione@uniroma1.it

Pierfrancesco Alaimo Di Loro
University of Rome "La Sapienza", Statistical Science Department, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

Giovanna Jona Lasinio
University of Rome "La Sapienza", Statistical Science Department, e-mail: giovanna.jonalasinio@uniroma1.it

Sara Martino
Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, e-mail: sara.martino@ntnu.no

Francesco Colloca
Department of Marine Biology - Zoologic station Anton Dohrn, Napoli e-mail: francesco.colloca@szn.it

1 Introduction

A thorough knowledge of fish diversity, distributions and habitat requirements is essential to the management of fisheries and conservation of species. In this respect, several attempts have been made to quantify the abundance of species in specific areas, their distribution and which factors affect the richness and biodiversity of the habitat [4; 3; 8]. In large part of the previous statistical modeling efforts, species have been usually considered individually or biodiversity has been generally quantified in terms of a single summary index. This implicitly neglects the well-established fact that species interact with each other and these interactions influence the assembly of the community. Moreover, analysis in terms of biodiversity indices is not able to detect whether some groups or species replaced others while keeping constant the overall diversity of the community. This means that eventual changes in the structure of the community must be verified considering simultaneously all its components. When dealing with such an high-dimensional outcome, standard multivariate models fail to account properly for the natural relationships among the single components of the response. Furthermore, the information of major interest in this study lies in the relative magnitude between the size of sub-groups, and not in their absolute values. When this is the case case, the outcomes can be expressed as portions of a total and are therefore more adequately modeled through the techniques of compositional analysis [1].

2 Data

We analize data gathered in the context of the MEDITS program [2]. We consider the $n = 40$ fishing stations sampled in front of the Lazio coast (Tyrrhenian Sea) from 1995 to 2015. Locations can be divided into 2 classes according to their depth: the 22 locations at depth $< 200m$ have been classified as *shelf*[1]; the remaining 18 locations at depth $\geq 200m$ have instead been labeled as *slope*. We only consider data for fishes (teleosteans and selaceans) including a total amount of 129 different species. For each of them, the number of sampled individuals and their weight are recorded for each year at each fishing station. The same data have been previously analyzed in [6], which focused on the temporal evolution of the biodiversity in the same area; while this study did not highlight any relevant temporal pattern, we cannot exclude substantial variations in terms of the individual components that constitute the community. In order to reduce the dimensionality of the outcome, species have been grouped together according to the habitat they prefer. This classification identifies 3 different groups: subtropical (S), subtropical-temperate (ST) and temperate (T). T species are the most common in this area of the Mediterranean Sea and it is most represented in the sample with 84 different species ($\approx 64\%$ of the entire catch). S species are the second component in magnitude, counting 38 species ($\approx 32\%$ of the entire catch). ST is a midway class composed only by 7 of the sampled species ($\approx 4\%$), characterised by species with good adaptability to different habitats. Additional information about temperature has been included from MyOcean database [9]. In particular, we added the quarterly average surface temperature recorded each year at each location. Figure 1 shows a progressive increase of the temperature of the four quarters temperatures between 1995 and 2017. In this study, we try to understand if and which groups of species are already (directly or indirectly) affected by the

[1] According to standard nomenclature, this depth class should include both shelf and edge.

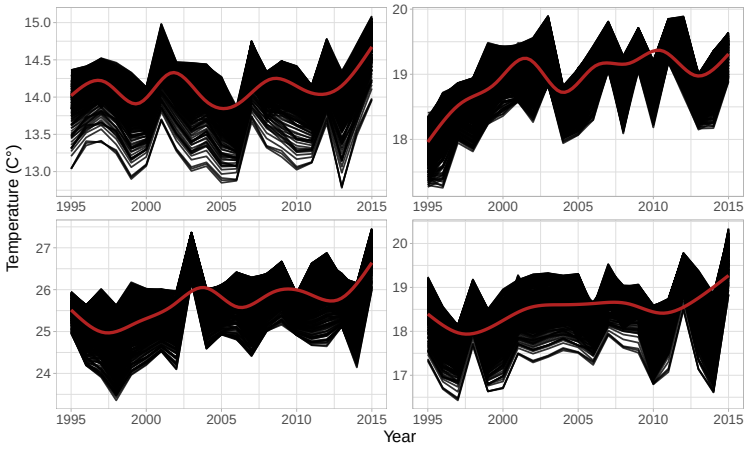


Fig. 1 Time series of the quarterly T1, T2, T3 and T4 average surface temperatures at 40 fishing stations in the study area, from top left panel to right bottom panel respectively.

rapidly evolving climatic condition of the Mediterranean Sea, pushing toward the meridionalization of its habitat and fauna [5].

3 Modeling compositional data

Whenever we are dealing with a collection of variables $\mathbf{s} = (s_1, s_2, \ldots, s_D) \in \mathbb{R}_+^D$ that represent the different parts of a whole, we can treat them as compositions. The compositional approach assumes that the relevant information contained in such data is associated only to the relative magnitude of the parts and not to their absolute value. When this is the case, we can discard all the unimportant information and achieve interpretability in terms of composition by standardizing the collection $\mathbf{s}$ through the closure operator $\mathscr{C}(\cdot)$:

$$\tilde{\mathbf{s}} = \mathscr{C}(\mathbf{s}) = \frac{(s_1, s_2, \ldots, s_D)}{\sum_{i=1}^{D} s_i}.$$

Managing compositions is less straightforward than it seems: the most significant information about compositions is contained in ratios between the components, which are generally difficult to handle; most methods from multivariate statistics developed for real-valued data-sets are misleading or inapplicable to compositions, which belong to the D-dimensional unit simplex $\mathscr{S}^D$. As a matter of fact, the classic algebraic operations used to deal with conventional real vectors are neither subcompositionally coherent nor scaling invariant. Aitchison geometry, introduced by [1], replaces these operators with others that suite the structure of the simplex and allows the development of *compositional analysis*. Let us consider two composition vectors $\mathbf{u}, \mathbf{v} \in \mathscr{S}^D$ and a scalar $\lambda \in \mathbb{R}$.

- *Perturbation*. Corresponds to the sum for compositions and is defined as:

$$\mathbf{z} = \mathbf{u} \oplus \mathbf{v} = \mathscr{C}(u_1 v_1, u_2 v_2, \ldots, u_3 v_3)$$

- *Powering*. Corresponds to the scalar multiplication for compositions and is defined as:

$$\mathbf{z} = \lambda \odot \mathbf{v} = \mathscr{C}\left(v_1^{\lambda}, v_2^{\lambda}, \ldots, v_3^{\lambda}\right)$$

The set of compositions, together with the operations of Aitchison geometry (including also the *compositional scalar product*), builds a $(D-1)$-dimensional Euclidean space structure on the simplex. This means that we can virtually translate any composition to the space of real vectors. One common transformation is the *isometric log-ratio transformation*: $ilr: \mathscr{S}^D \overset{iso}{\to} \mathbb{R}^{D-1}$, which induces an isometric identification of $\mathbb{R}^{D-1}$ and $\mathscr{S}^D$. Perturbation and powering in $\mathscr{S}^D$ are equivalent to sum and scalar product on their real-valued ilr counterparts and this allows to define all the basic probability operations on the simplex by working directly on the ilr space.

This implies that, whenever we define a valid stochastic model on the simplex, we can easily translate it into a standard multivariate regression model in $\mathbb{R}^{D-1}$ through the ilr transformation. In particular, if we consider the *Normal on the simplex* [7] distribution, which is perfectly coherent with the geometric structure induced by Aitchison geometry [1], this would correspond to the standard multivariate Normal in the ilr space. In the domain of this paper, we will focus on how to deal with compositions as dependent variable $\mathbf{y} = \tilde{\mathbf{s}}$ when all the covariates $\{x_k\}_{k=1}^K$ are classic variables. In order to stick with the structure of the data introduced in Section 2, without any loss of generality, we will introduce the regression framework in the context of a fixed effect model.

Let $\mathbf{y}_{it}$ be the $1 \times p$ composition vector and $\mathbf{x}_{it}$ the $K \times 1$ vector of covariates observed on unit i, $i = 1, \ldots, n$ at time t, $t = 1, \ldots, T$. We can express the compositional regression model through the linear Aitchison structure of the simplex as:

$$\mathbf{y}_{it} = \mathbf{a}_i \oplus \bigoplus_{k=1}^{K} \left(x_{it,k} \odot \mathbf{b}_k\right) \oplus \boldsymbol{\varepsilon}_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{1}$$

where $\mathbf{a}_i$ is a unit-specific unknown compositional constant, which is perturbed by each covariate $x_{it,k}$ through the powering of the $1 \times p$ compositional vector of coefficients $\mathbf{b}_k$[2]; $\boldsymbol{\varepsilon}_{it}$ are independent compositional random variables with null compositional expectation $\mathbf{1}_D = (1, \ldots, 1)/D$, constant variance $\Sigma_{\mathscr{S}}$ and *Normal distribution on the simplex*. This model can be translated into a multivariate regression problem in $\mathbb{R}^{D-1}$ by applying the ilr transform at both side of Equation (1):

$$\text{ilr}(\mathbf{y}_{it}) = \text{ilr}(\mathbf{a}_i) + \sum_{k=1}^{K} x_{it,k} \cdot \text{ilr}(\mathbf{b}_k) + \boldsymbol{\varepsilon}_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{\varepsilon}_{it} \overset{iid}{\sim} \mathscr{N}_{D-1}(\mathbf{0}, \Sigma_{\text{ilr}})$. We can then fit the model using standard linear multivariate regression techniques, and then transform back the parameters in their simplicial counterparts in order to interpret their effect in terms of composition. The estimated intercepts $\hat{\mathbf{a}}_i$ can be interpreted as the expected composition for $x_{it,k} = 0$, $\forall k$; the slope $\mathbf{b}_k$ may be interpreted as the perturbation applied to the composition if $x_{it,k}$ increases by 1 unit. It is important to point out that the neutral element of

[2] Each element $b_{k,j}$ of $\mathbf{b}_k$ represents the effect of covariate x_k on the jth component

perturbation is the vector $\mathbf{1}_D$: values of $\mathbf{b}_{k,j}$ greater (lower) than $1/D$ have a positive (negative) effect on component j.

4 Application

In according to the notation introduced in Section 3, let us denote outcome and covariates related to the ith fishing stations at tth time points as $\{(\mathbf{y}_{it}, \mathbf{x_{it}}),\}_{i,t=1}^{n,T}$. The response variable $\mathbf{y}_{it}$ has been transformed into compositions through $\tilde{\mathbf{y}}_{it} = \mathscr{C}(\mathbf{y}_{it})$. We assume that the resulting compositions are the realizations of model (1), where individual fixed effects $\mathbf{a}_i$ have been assigned to each fishing haul. Variables referred to the geographic position of the stations (longitude, latitude and depth) have been excluded from the analysis (individual fixed effects already account for the spatial variability). Six hauls widely change their position along the years and they have been excluded from the analysis because such movements would negatively affect the proper estimation of the respective fixed effects. Furthermore, in order to highlight temporal variation with respect to the first year of observation, all the temperatures have been centered with respect to the ones observed in 1995, which is taken as a baseline. The model can be estimated exploiting its ilr counterpart expressed in Equation (2). The fit produced an $R^2 \approx 0.54$, which is a good fit considering the dimensionality of the outcome and the few predictors considered. Nevertheless, the major interest of this analysis is directed toward the interpretability of the estimated coefficients. Figure 2 shows the estimates of simplicial coefficients for all the statistically significant variables ($p < 0.1$). The values of the fixed effects highlight a large variability between shelf and slope stations (respectively characterized by the dominant presence of S species and of T ones), but also a non-negligible variability within class, that justifies the utilization of individual fixed effects. The bars on the right summarize the effect of time (expressed in years) and of the yearly temperature trend referred to the first quarter. As discussed in Section 3, those values must be interpreted in terms of *perturbation*. From a graphical point of view, the time effect seems to be very close to the neutral element. However, the numeric values of the coefficients are $\mathbf{b}_{year} = (0.3343, 0.3369, 0.3287)$ (respectively for S, ST and T species), which correspond to a slight but general increase of the S and ST component at the expense of the T one. Consider that on the 21st year, through the operation of powering, we get an effect of $21 \odot \mathbf{b}_{year} = (0.3474, 0.4088, 0.2436)$. In particular, the slow transition to the benefit of specifically S species is extremely evident in the years where the temperature of the first quarter was greater than the one of the baseline. Indeed, the estimated coefficient for such co-variate has been estimated to be $\mathbf{b}_{T1} = (0.4108, 0.3084, 0.2807)$.

5 Conclusions and further development

Compositional analysis revealed to be a very useful tool in understanding and modeling the structure of a community by simultaneously taking into account all its parts. While the available time window is still too short for making claims about phenomena that usually evolve at a slow pace and take place along decades, this preliminary regression model already highlighted a slow but progressive meridionalization of the fauna inhabiting the portion of Mediterranean sea in front of the Lazio coast. This may be interpreted as an effect of water temperature increase, which may affect directly or indirectly the indigenous fish community. However, further studies that consider other relevant

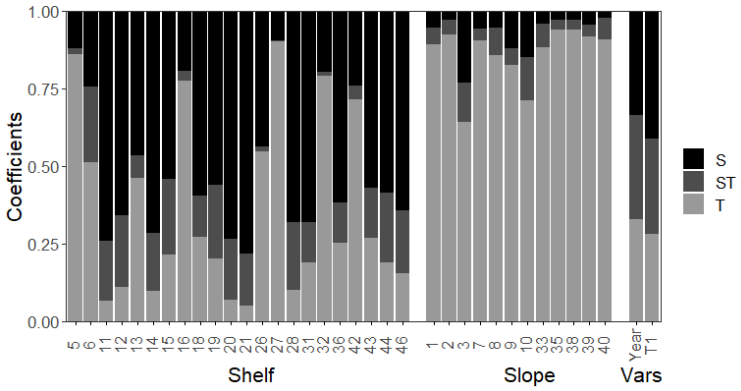


Fig. 2 Barplot of the estimated $\hat{a}_i$'s and $\hat{b}_k$'s arranged from the left to the right: fixed effect (shelf and slope) and variables coefficients.

pressure factors on the fish population (fishing, pollution etc.) are required in order to establish an effective causal effect. A more extensive study on this topic would consider alternative species classifications that may be finer or possibly include exotic individuals.

References

[1] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[2] JA Bertrand, L Gil De Sola, Costas Papaconstantinou, G Relini, and Arnauld Souplet. An international bottom trawl survey in the mediterranean: the medits programme. *Actes de Colloques-IFREMER*, pages 76–96, 2000.

[3] Alan E Gelfand, John A Silander, Shanshan Wu, Andrew Latimer, Paul O Lewis, Anthony G Rebelo, Mark Holder, et al. Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, 1(1):41–92, 2006.

[4] Antoine Guisan and Wilfried Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009, 2005.

[5] Anna M Mannino, Paolo Balistreri, and Alan Deidun. The marine biodiversity of the mediterranean sea in a changing climate: the impact of biological invasions. *Mediterranean Identities-Environment, Society, Culture*, 2017.

[6] M Mingione, G Jona Lasinio, S Martino, and F Colloca. Multivariate analysis and biodeversity partitioning of a demersal fish community: an application to lazio coast. *Book of Short Paper SIS 2019*, pages 985–990, 2019.

[7] Vera Pawlowsky-Glahn et al. Statistical modeling on coordinates. 2003.

[8] S. Shirota, A. Gelfand, and S. Banerjee. Spatial joint species distribution modeling using dirichlet processes. *arXiv preprint arXiv:1711.05646*, 2017.

[9] K. von Schuckmann, P. Le Traon, N. Smith, A. Pascual, P. Brasseur, K. Fennel, S. Djavidnia, S. Aaboe, E. Fanjul, E. Autret, et al. Copernicus marine service ocean state report. *Journal of Operational Oceanography*, 11(sup1):S1–S142, 2018.

FDA dimension reduction techniques and components separation in Fourier-transform infrared spectroscopy

Tecniche di riduzione della dimensionalità e di separazione di componenti nell'analisi di dati funzionali di spettroscopia infrarossa in trasformata di Fourier

Francesca Di Salvo, Elena Piacenza , Delia Francesca Chillura Martino

Abstract FTIR spectroscopy is a measurement technique used to obtain an infrared spectrum of absorption of a solid (or a liquid or a gas), for the characterization of specific chemical components of materials. When repeated measures are taken on samples of materials, the result is a collection of spectra representing a set of samples from continous functions (signals) defined in the domain of the frequencies. An unifying approach to the study of a collection of FTIR spectra is proposed to deal with the presence of random shifts in the peaks, the identification of representative spectra and finally the characterization of the observed differences: in the functional data framework, the performance of a new proposal for alignment of curves and their assignment to clusters of shapes with an intrinsic order is presented and discussed.
Abstract *La spettroscopia FTIR è una tecnica mediante cui si acquisisce uno spettro di assorbimento di una sostanza, contenente informazioni per la caratterizzazione dei materiali e per lo studio dei legami chimici. Effettuando misurazioni ripetute su campioni di materiali, si osserva una collezione di spettri che per la loro struttura rappresentano le realizzazioni campionarie di curve continue definite sul dominio delle frequenze. Si propone un approccio statistico unitario per lo studio di collezioni di spettri che affronti il problema della presenza di traslazioni casuali nei picchi delle curve, dell'identificazione di curve rappresentative e della interpretazione delle dissimilaritá osservate. L'approccio proposto per l'allineamento e la separazione degli spettri in cluster intrinsecamente ordnati é presentato nel contesto dei dati funzionali.*

Francesca Di Salvo
Department of Agricultural Food and Forest Sciences, viale delle Scienze, Università degli studi di Palermo, e-mail: francesca.disalvo@unipa.it

Elena Piacenza , Delia Francesca Chillura Martino
Department of Biological, Chemical and Pharmaceutical Sciences and Technologies, viale delle Scienze, Università degli studi di Palermo, e-mail: delia.chillura@unipa.it, elena.piacenza@community.unipa.it

Key words: Shape analysis, functional data, reduction of dimensionality, FTIR spectroscopy

1 Introduction

The motivation of this contribution is the attempt to give answers to the interpretation of a collection of Fourier Transform InfraRed (FTIR) spectra, for which the core of the study is a chemical analysis, but a better synthesis of the most meaningful information is achieved by statistical methods.
FTIR spectroscopy is a label-free and non-destructive technique based on the absorption of electromagnetic waves in the infrared region of the spectrum [1]; the absorption bands originate from fundamental transitions of molecular vibrations, that are highly characteristic of specific chemical components. A collection of IR spectra represents a set of samples from continous functions defined in the domain of the frequencies: each spectrum can be visualized in a graph of infrared light absorbance on the vertical axis vs frequency or wavelength on the horizontal axis. Typical units of IR frequency are reciprocal centimeters (called wave numbers, with the symbol cm^{-1}). Units of IR wavelength are given in micrometers (microns with symbol μm, which are related to wave numbers in a reciprocal way. The mid-infrared, approximately 4000–400 cm^{-1} (or 2.5–25μm), may be used to study the fundamental vibrations and their associated structure. After having processed the data by chemical analysis, a statistical approach is adopted in order to sinthetize the most meaningful information from the spectral results.

The procedure allows a complementary analysis of the similarities, or differences, among the curves; the presence of random shifts in the peaks, due to variability in measurements and equipment, introduces the warping as a tool to align the observed spectra and reduce their variance [6], [7] and [8]. The alignment problem and assignment of curves to clusters of shapes is handled with principal component technique in order to characterize and correlating the morphological differences, and a natural order among the members of the clusters, can be obtained in terms of their cohesiveness.

2 The methodology

The goal of any absorption spectroscopy is to measure how much light a sample absorbs at each wavelength and to relate this with chemical composition. The raw data are collected rapidly repeating the measurements many times over a short timespan and afterwards, an algorithm takes all this data and works backward to infer what the absorption is at each wavelength. In this process, from recording the raw data to converting into the light absorption, the resulting curves are affected by noise. The couple $(w, Y_i(w))$ denotes respectively the wavelength and the observed inten-

sity for $w \in W = [4000\text{–}400]cm^{-1}$, $W \subset R^+$. In the observed spectrum, the points are separated by equal frequency intervals. The curves are standardized by the linear transformation:

$$Y_i^s(w) = \frac{Y_i(w)}{max_{w \in W}(Y_i(w))}, \tag{1}$$

Because of their structure, tha IR spectra can be represented as a realization of a univariate functional random field:

$$Y_i^s(w) = X_i(w) + \varepsilon_i(w) \tag{2}$$

$X_i(w)$ is the signal and $\varepsilon_i(w)$ is the random component. From a chemical point of view, the focus is on the shape of the curves, independently from their height, as the differences in height are produced by different intensities, while the study aims to compare loci of peaks, independently from their intensities. For this reason, a non linear transforms of the curves is considered, the square root slope functions: this transformation introduces the problem of separation of the phase and amplitude components of variability of the functional data; the phase variability (variability along the x axis) is here interpreted as a negligible inherent variability in the underlying process itself that needs to be separated from the amplitude variability along the vertical axis; the effect of the reduction of phase variability is considered facing the basic questions of interest:

1. estimation of the central tendency
2. estimation of the variability among the curves and the identification of clusters (location and shape types)
3. detection of outlying curves
4. description of the essential modes of variability

Several methods pursue one of these goals, but separately; alternative proposals approaches the basic questions $1-4$ in an unitary framework: a complete overview of the goals, challenges and contributes in literarure is in [5], where methods based on the square root slope functions are presented. Different the approaches performs these tasks, sequentially, [6] and [7] or jointly [9]. For the goal of this analysis, the algorithm presented provides cluster-wise alignment accounting for the phase variability and discovering the principal modes in the square root slope function (SRSF) framework:let $X_i(w)$ be differentiable almost everywhere on W, with derivative denoted by $\dot{X}$, the square root sope function is defined as:

$$q(w) = sign(\dot{X})\sqrt{|\dot{X}|} \tag{3}$$

An agglomerative hierarchical clustering of the SRSF's of the curves provides an optimal configuration, and inside each cluster a dimension-reduction tool to approximate the two components, phase and amplitude (elements of infinite-dimensional spaces), as elements of finite-dimensional spaces is performed: a general model, that takes into account clusters effects and simultaneously represents the i^{th} SRSF of the j^{th} cluster, j from 1 to J, in the finite-dimensional space, is derived by the Karhunen-Loéve expansion:

$$q_{ij}(\gamma_i(w)) = \mu_q(w) + \sum_{h=1}^{\infty} u_{ij_h}\xi_{j_h}(w) \tag{4}$$

where $\gamma_i(w)$ are the warping function, a deformations of the domain W, such that the original functions composed with these warpings are optimally aligned; $\mu_q(w)$ is the Karcher mean, whose amplitude and relative phase denote the sample means of amplitudes and relative phases; u_{ij_h} and $\xi_{j_h}(w)$ are respectively the h^{th} Principal score for the i^{th} curve and the h^{th} eigenfunction inside the j^{th} cluster. The alignment of the coefficients of principal component analysis (FPCA) on the space of the first derivatives

$$(\hat{\gamma}, \hat{\mu_q}, \hat{u}, \hat{\xi}) = argin f_{\gamma,\mu,u,\xi} \left(\sum_{i=1}^{n} |q_{ij} \circ \gamma_i - \mu_q - \sum_{i=1}^{H} u_{ij_h}\xi_{j_h}|^2 \right) \tag{5}$$

A suitable algorithm implements the estimation problem iteratively. On the current estimated curves, the clustering is repetead until the convergence is obtained. Results derived in terms of principal modes of variations inside the clusters, describe the main cluster-effects on the shape of the curves group-wise registered, while the detection of outlying curves with respect to the centre of the cluster is performed by mean of measures based on the modified band depth [4].A previous paper [2] introduced an algorithm based on the notion of Modified Band Depth for improving the clustering of the curves.

3 Data and results

The dataset consists in a collection of 24 absorbance spectra recorded in the mid-infrared range of wavelength $[4000-400]cm^{-1}$. Here results are presented for a smaller range of wavelength, $[3770-2770]cm^{-1}$, characterized the presence of peaks of interest. The original data standardized in the range $[0,1]$ are in Fig.1 (top-left). The focus of the analysis, implementing the techniques described in Section 2, is to determine homogeneous clusters of curves with respect their shape. The optimal number of clusters is determined by hierarchical clustering with agglomerative algorithms and the optimal number of clusters k is selected by maximizing the average silhouette Fig.1 (top-right, bottom-left); as preliminary results, the Fig.1 (bottom-right) reports four well separated clusters. Another goal is finding a representative shape for each cluster. A data depth algorithm, based on the concept of modified band depth [4], measures the centrality of an observation within a cluster and allows the definition of a natural ordering from center outwards, allowing reallocation of the most external curves. In Fig.2 the structure of the four groups is presented, the inner curves (75% of the cluster) and the external curves (the remaining 25%).

4 Conclusion

This study represents a functional data analysis of spectroscopic investigation carried out with the FTIR thecnology. The analytical approach allowed to compare information from a large number of spectra, leading to an in-depth knowledge of the materials. Shapes of the curves are the essential difference between clusters and the comprehensive approach proposed deals with the problem of registering and modeling them in a joint framework working on the spaces of their derivatives. The results obtained are characterized by groups of peculiar shape with peaks located at specific wavelength providing chemical interpretation in terms of composition of the materials.

References

1. Chan K.L.A., Kazarian S.G.:Attenuated total reflection Fourier-transform infrared (ATF-FTIR) imaging of tissue and live cells. Che. Soc. Rev. **45**, 1850–1864.(2015), doi: 10.1039/c5cs00515a.
2. Di Salvo, F., Rotondi, R., Lanzano, G.: Detecting clusters in spatially correlated waveforms, GNGTS conference, Trieste, November 13th - 16th (2017)
3. Lee, S. and S. Jung: Combined analysis of amplitude and phase variations in functional. Stat.Me. **22**, 1–21(2017), data.arXiv:1603.01775
4. Lopez-Pintado, S., Romo, J., : Depth-based inference for functional data, Computational Statistics and Data Analysis 51 (10), 4957-4968, (2007).
5. Srivastava,A., Klassen, E.P.: Functional and Shape Data Analysis. Springer Series in Statistics, Springer-Verlag, New York (2016) DOI 10.1007/978-1-4939-4020-2 8
6. Tucker, D.J., Wu, W. , Srivastava, A., Generative models for functional data using phase and amplitude separation, Computational Statistics and Data Analysis, 61, 50–66, (2013)
7. Tucker J. D., W. Wu, and A. Srivastava, "Phase-Amplitude Separation of Proteomics Data Using Extended Fisher-Rao Metric," Electronic Journal of Statistics, **8**, no. 2, 1724–1733 (2014)
8. Tucker J. D., W. Wu, and A. Srivastava, "Analysis of signals under compositional noise With applications to SONAR data," IEEE Journal of Oceanic Engineering,**29**, no. 2, 318–330 (2014).
9. Tucker J. D., J. R. Lewis, and A. Srivastava, "Elastic Functional Principal Component Regression," Statistical Analysis and Data Mining, **12**, no. 2, pp. 101–115. (2019)
10. Xie, W., S. Kurtek, K. Bharath, and Y. Sun (2016). "A Geometric Approach to Visualization of Variability in Functional Data." Journal of the American Statistical Association **112**, 979–993 (2017)

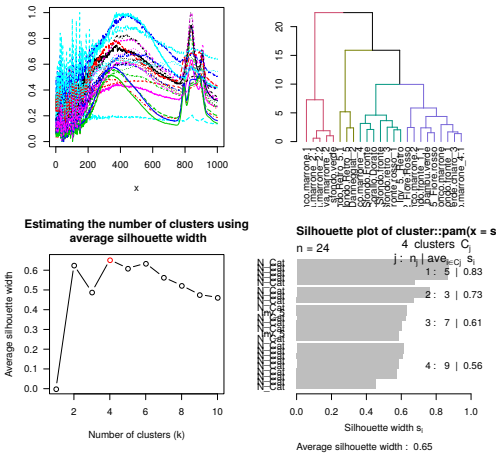


Fig. 1 Observed spectra (top-left). Hierarchical clustering: dendrogram (top-right), optimal number of cluster (bottom-left); silhouette (bottom-right)

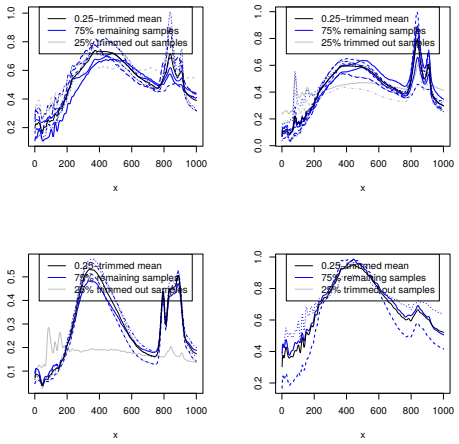


Fig. 2 The final four clusters with indication of the .25-trimmed mean (black line) and 25% most external curves (gray lines)

Functional Data Analysis for Spectroscopy Data

Analisi di Dati Funzionali per Dati di Spettroscopia

Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Diego Perugini, Alessandro Pisello, Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis, Simone Vantini

Abstract We propose a functional data analysis approach for the study of spectroscopy data. The applicative problem concerns the characterization of the spectral response of silicate glasses in order to infer the chemical composition of the materials from spectral data, which can be remotely collected. The analysis performed aims at characterizing the phase variability and the amplitude variability of the data in order to extract meaningful information. The technique used is k-mean alignment.
Abstract *Proponiamo un approccio di analisi di dati funzionali per lo studio di dati di spettroscopia. Il problema applicativo riguarda la caratterizzazione della risposta spettrale di vetri silicati al fine di dedurre la composizione chimica dei materiali a partire da dati spettrali, che possono essere raccolti in remoto. Lo scopo dell'analisi è la caratterizzazione della variabilità di fase e della variabilità di ampiezza dei dati al fine di estrarre informazioni significative. La tecnica utilizzata è il k-mean alignment.*

Key words: functional data, spectroscopy data, functional registration, k-mean alignment

Mara S. Bernardi
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: marasabina.bernardi@polimi.it

Matteo Fontana, Alessandra Menafoglio, Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy

Diego Perugini, Alessandro Pisello
Department of Physics and Geology, University of Perugia, Italy

Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis
Institute for Space Astrophysics and Planetology, INAF-IAPS, Rome, Italy

1 Introduction

Silicates are the main constituents of terrains on terrestrial planets in the solar system. Silicate glasses represent the amorphous phase of silicate crystals, and they are widely present in natural volcanic rocks, in which amorphous phases can be present as a small fraction or as the only constituent. Since it is known that terrestrial planets are largely covered by volcanic products, it is of great importance to study the spectral response of glasses to better interpret available and future remotely sensed spectra from past and future missions [5, 3, 2]. Spectral characterization of crystalline materials is usually made through pointing peaks position, intensity, and FWHM (full width at half maximum). Since amorphous material spectral response does not represent a clear set of peaks, but a series of "bulges" and "valleys" whose shape is blurred, a new approach has to be found. To tackle this problem, we propose to apply advanced statistical techniques in the framework of Functional Data Analysis. The advantage of such techniques is the possibility to consider in the analysis the whole spectrum, while classical approaches based on multivariate statistic rely on a dimensionality reduction step (such as the extraction of features characterizing the peaks), thus losing information from the original data.

Functional data analysis [6] is the branch of statistical analysis that considers functional data (such as curves, surfaces, spectra, etc.) as statistical units. In recent years, many functional data analysis techniques have been proposed and applied to many different fields of application. Functional data analysis is suitable for the analysis of the data considered in this work as the underlying phenomenon is continuous in nature (continuous curves on the frequency domain). Resorting to a functional representation of the data allows to automatically extract characterizing features and to properly describe the shape of the curves.

The dataset analyzed in this work, described in Section 2, concerns the spectral response of 4 different magmatic series and natural rocks. The statistical analysis of the data is described in Section 3. Section 4 draws the conclusions of the work and outlines possible directions of future research in this context.

2 Dataset

In order to build up an exhaustive database of amorphous, four different series of silicate glasses were produced, with the aim of reproducing the widest possible chemical variability, according to what we find on Earth. Each series of glasses was produced by crushing, melting and mixing two natural endmembers from different areas, whose volcanic products have different chemical variability: Vulcano (Aeolian islands, Italy) Snake River Plain (USA) Etna (Sicily, Italy) and Pantelleria (Mediterranean Sea, Italy). The resulting series of glasses present a range of chemical compositions that is covering the majority of rocks on planet Earth. Of peculiar interest are the variations in SiO2 content (linked to evolution of magma) and alkaline content (linked to geodynamic setting). Samples were analyzed using

electron probe microanalyses, and spectroscopically characterized in reflectance at the INAF-IAPS in Rome in the mid- infrared range (2-14 micron). Figure 1 shows the data.

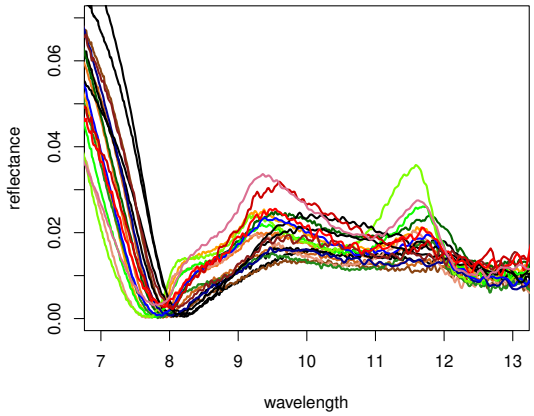


Fig. 1 Plot of the original data.

3 Statistical analysis

The available data are considered as a discrete noisy sampling of an underlying smooth function. Therefore, as a preprocessing step, spline smoothing is applied to the rough data represented in Figure 1 to obtain a smooth functional representation of the data and to remove sampling error and spurious oscillations that are not of interest for the analysis of the phenomenon. The smoothed curves are represented in Figure 2. Spline smoothing is performed using the R package fda [7].

As can be seen in Figure 2, the variability among the data is characterized by an horizontal displacement and a difference in the shapes (peaks more or less pronounced). To properly account for these features of the data, we resort to functional alignment (also called functional registration), a technique to decompose phase variability and amplitude variability. This technique is particularly suitable for the analysis of these data as the horizontal displacement and the shape of the curves are meaningful for inferring the chemical composition of the material.

The technique applied is k-mean alignment [8, 9, 10], an algorithm that simultaneously aligns and clusters curves through an iterative procedure. The use of this technique is motivated by the practical interest in grouping samples of materials with similar spectral characteristics, as this could imply similar chemical compositions.

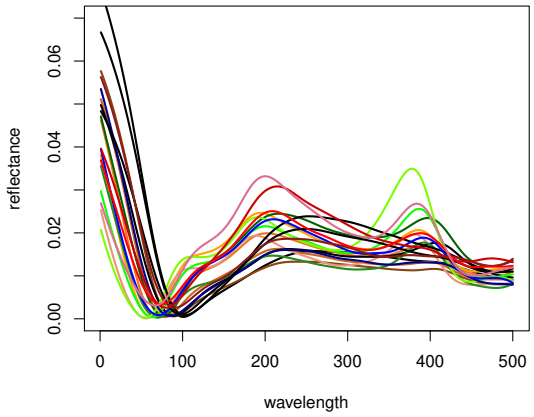


Fig. 2 Plot of the smoothed data.

Functional alignment requires a proper choice of similarity index and class of warping functions. Indeed, aligning a curve $f(s)$ to a reference consists in finding a warping function of the abscissa $h(s)$ such that the curve $f(h(s))$ is the most similar to the reference.

Regarding the choice of the similarity index, in this work we use the following similarity index, which measures the cosine of the angle between the derivatives of two functions g and g:

$$\rho(f,g) = \frac{< f', g' >_{L^2}}{\|f'\|_{L^2}\|g'\|_{L^2}} = \frac{\int_{\mathbb{R}} f'(s)\, g'(s)\, ds}{\sqrt{\int_{\mathbb{R}} f'(s)^2\, ds}\sqrt{\int_{\mathbb{R}} g'(s)^2\, ds}}.$$

This choice reflects the need to analyze the shape of the curves, neglecting their magnitudes. Indeed, the similarity index chosen considers as identical two curves that differ for vertical shifts and vertical dilations:

$$\rho(f,g) = 1 \iff \exists a \in \mathbb{R}^+, \exists b \in \mathbb{R} : f = ag + b.$$

Regarding the choice of the class of warping functions, we consider in our analysis the following three classes:

$$W_{shift} = \{h : h(s) = s + q \text{ with } q \in \mathbb{R}\}.$$
$$W_{dilation} = \{h : h(s) = ms \text{ with } m \in \mathbb{R}^+\}.$$
$$W_{affine} = \{h : h(s) = ms + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}.$$

The best results in terms of similarity were obtained with the classes W_{shift} and W_{affine}, with no significant difference between these two classes, meaning that the actual phase variability is well captured by a simple translation.

The results of the analysis highlight a strong correlation between the values of the shifts and the silica content of the materials. Moreover, the clustering obtained by k-mean alignment provides meaningful information about the total alkali-silica composition of the materials. These results demonstrate the importance of the analysis of both phase and amplitude variability as they are both informative for the goals of the analysis.

Further details about this analysis can be found in [1]. K-mean alignment is performed using the R package fda [4].

4 Conclusions and future work

The analysis performed in this work shows the usefulness of functional data analysis techniques for the study of spectral data and the relevance of the information extracted from the data through k-mean alignment in retrieving the chemical composition of the material.

Directions of future research concerns the development of a functional regression model to predict the chemical composition using the spectral data as covariates. In this model, the response is a compositional data, while the covariate is a bivariate functional data composed by the warping function (describing the phase variability) and the aligned data (describing the amplitude variability).

5 Acknowledgements

This work was supported by ACCORDO Quadro ASI-POLIMI “Attività di Ricerca e Innovazione” n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano and by the ASI-UniPG agreement 2019-2-HH.0.

References

1. Bernardi, M.S., Fontana, M., Menafoglio, A., Perugini, D., Pisello, A., Ferrari, M., De Angelis, S., De Sanctis, M.C., Vantini, S.: Silicates Glasses Spectroscopy: A Functional Data Analysis Perspective. Manuscript
2. De Sanctis, M. C., Altieri, F., Ammannito, E., Biondi, D., De Angelis, S., Meini, M., Pirrotta, S., Vago, J.L., Mugnuolo, R.: Ma_MISS on ExoMars: mineralogical characterization of the martian subsurface. Astrobiology, 17(6-7), 612-620 (2017)
3. Maturilli, A., Helbert, J., Moroz, L.: The Berlin emissivity database (BED). Planetary and Space Science, 56(3-4), 420-425 (2008)
4. Parodi, A., Patriarca, M., Sangalli, L. M., Secchi, P., Vantini, S., Vitelli, V.: fdakma: Functional Data Analysis: K-Mean Alignment. R package version 1.2.1. (2015) https://CRAN.R-project.org/package=fdakma

5. Pisello, A., Vetere, F. P., Bisolfati, M., Maturilli, A., Morgavi, D., Pauselli, C., Iezzi, G, Lustrino, M., Perugini, D.: Retrieving magma composition from TIR spectra: implications for terrestrial planets investigations. Scientific reports, 9(1), 1-13 (2019)
6. Ramsay, J. O., Silverman, B. W.: Functional Data Analysis. 2nd edn Springer. New York (2005)
7. Ramsay, J. O., Wickham, H., Graves, S., Hooker, G.: fda: Functional Data Analysis. R package version 2.4.8.1. (2018)
8. Sangalli, L. M., Secchi, P., Vantini, S., Vitelli, V.: Classification of Functional Data: Unsupervised Curve Clustering When Curves are Misaligned. In Joint Statistical Meetings pp. 4034-4047 (2010)
9. Sangalli, L. M., Secchi, P., Vantini, S., Vitelli, V.: K-mean alignment for curve clustering. Computational Statistics & Data Analysis, 54(5), 1219-1233 (2010)
10. Vitelli, V., Sangalli, L. M., Secchi, P., Vantini, S.: Functional clustering and alignment methods with applications. Communications in Applied and Industrial Mathematics 1(1): 205-224 (2010)

Functional graphical model for spectrometric data analysis

Modello grafico funzionale per l'analisi di dati spettroscopici

Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini

Abstract Motivated by the analysis of spectrographic data, we introduce a functional graphical model for learning the conditional independence structure of spectra. Absorbance spectra are modeled as continuous functional data through a cubic B-spline basis expansion. A Gaussian graphical model is assumed for basis expansion coefficients, where a sparse structure is induced for the precision matrix. Bayesian inference is carried out, providing an estimate of the precision matrix of the coefficients, which translates into an estimate of the conditional independence structure between frequency bands of the spectrum. The proposed model is applied to the analysis of the infrared absorbance spectra of strawberry purees.

Abstract Motivati dall'analisi di dati spettroscopici, introduciamo un modello grafico funzionale per l'apprendimento della struttura di indipendenza condizionale degli spettri. Gli spettri di assorbimento sono modellati come dati funzionali continui attraverso una espansione in base B-spline cubica. Un modello grafico gaussiano è utilizzato per i coefficienti dell'espansione di base, attraverso il quale viene indotta una struttura sparsa per la matrice di precisione dei coefficienti. L'inferenza bayesiana del modello permette di ottenere una stima della struttura di indipendenza condizionale tra le bande di frequenza degli spettri. Il modello proposto è applicato all'analisi degli spettri di assorbimento infrarosso di puree di fragole.

Key words: Bayesian inference, functional data analysis, graphical model selection

Laura Codazzi, Alessandro Colombi and Matteo Gianella
Politecnico di Milano, e-mail: laura.codazzi@mail.polimi.it, alessandro3.colombi@mail.polimi.it, matteo1.gianella@mail.polimi.it

Raffaele Argiento, Lucia Paci and Alessia Pini
Università Cattolica del Sacro Cuore e-mail: raffaele.argiento@unicatt.it, lucia.paci@unicatt.it, alessia.pini@unicatt.it

1 Introduction

Spectrography is a technique that is used to summarize all the molecular components of a substance. Using this technique, for each substance it is possible to acquire an absorbance spectrum, that is a signal containing detailed information about the substance composition. From a mathematical point of view, the spectrum is a continuous function of the wavelength. To make this signal more meaningful it is important to understand which wavelength bands are related to the different components. The dependency structure between the signal at different wavelenghts is particularly informative in this sense: if two different bands of the spectrum are dependent, we can conclude that they refer to the same components, or to two closely related components. Our goal is then to investigate the dependence among different portions of an absorbance spectrum.

Let $y_i(s)$ be the absorbance at wavelength $s \in [l,u]$ for unit i, $i = 1,\dots,n$, where $[l,u] \subset \mathbb{R}^+$ is the common domain of all curves. The whole curve $y_i(s)$, $s \in T$ is the absorbance spectra of unit i. Using standard smoothing techniques of functional data analysis, the model that we assume for the i-th spectra is:

$$y_i(s) = \sum_{j=1}^{p} \beta_{ij}\varphi_j(s) + \varepsilon_i(s), \quad s \in [l,u]$$

where $\varphi_1,\dots,\varphi_p$ are suitable B-spline basis, $\beta_i = (\beta_{i1},\dots,\beta_{ip})$ is the individual-specific spline coefficients' vector and $\varepsilon_i(s) \overset{\text{iid}}{\sim} \mathrm{N}(0,\tau_\varepsilon^2)$. Within the Bayesian setting a customary prior for the β_i coefficients is:

$$\beta_1,\dots,\beta_n \mid \mu,\Sigma \overset{\text{iid}}{\sim} N_p(\mu,\Sigma), \tag{1}$$

where μ is the prior mean vector and Σ is the prior covariance matrix.

In this framework, the dependence between two bands of the spectrum is reflected in the relationship among the corresponding smoothing regression parameters. Hence, our goal boils down to the study of the dependence structure of the β's, that is on matrix Σ, or on its inverse Σ^{-1}, that is called precision or concentration matrix.

2 Functional graphical model

We employ a graphical modeling approach for learning conditional independence structures among the coefficients. Let $G = (V,E)$ be the graph defined by the node set $V = \{1,\dots,p\}$ representing the coefficients and by the edge set $E \subset V \times V$. In practice, the graph is unknown and must be estimated. For Gaussian graphical models, the problem translates in assuming that the precision matrix Σ^{-1} in (1) is Markov with respect to G. In other words, coefficients β_{ij} and β_{ik} are conditionally

independent given all the remaining variable $\beta_{\setminus\{j,k\}}$ whenever $\{j,k\} \notin E$, if and only if the corresponding entry in matrix Σ^{-1} is zero, i.e., $\beta_{ij} \perp \beta_{ik}|\beta_{\setminus\{j,k\}} \Leftrightarrow \Sigma^{-1}_{j,k} = 0$; see Figure 1 for an example.

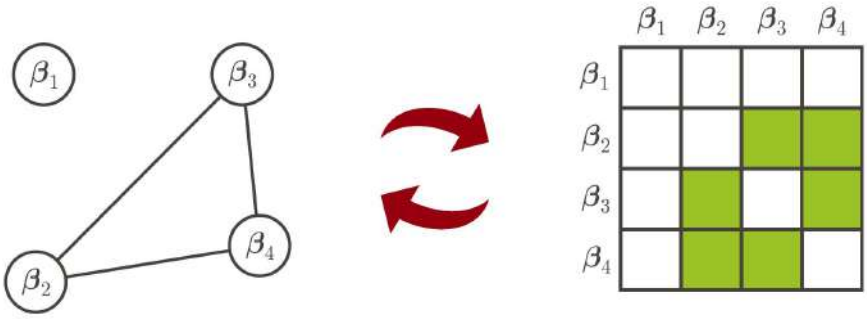


Fig. 1 Example of a Gaussian graphical model with 4 nodes. On the left, the graph with 4 nodes and, on the right, the corresponding precision matrix where green cells represent nonzero entries.

Hence, the Bayesian hierarchical model becomes

$$\begin{aligned} y_i \mid \beta_i, \tau^2_\varepsilon &\overset{\text{iid}}{\sim} \mathrm{N}_r\left(\Phi\beta_i, \tau^2_\varepsilon \mathbf{I}_r\right) \\ \beta_1, \ldots, \beta_n \mid \mu, \Sigma &\overset{\text{iid}}{\sim} \mathrm{N}_p\left(\mu, \Sigma\right) \\ \mu &\sim \mathrm{N}_p\left(\mathbf{0}, \sigma^2_0 \mathbf{I}_p\right) \\ \Sigma^{-1} \mid G &\sim \text{G-Wishart}\left(d_0, \Sigma_0\right) \\ G &\sim \pi\left(G\right) \\ \tau^2_\varepsilon &\sim \mathrm{IG}\left(a, b\right) \end{aligned} \tag{2}$$

where Φ is the matrix collecting the B-spline basis. The G-Wishart distribution has density

$$p\left(\Sigma^{-1} \mid G, d_0, \Sigma_0\right) = I_G\left(d_0, \Sigma_0\right)^{-1} |\Sigma^{-1}|^{\frac{d_0-2}{d_0}} \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1}\Sigma_0\right)\right), \quad \Sigma^{-1} \in P_G$$

where $d_0 > 2$ are the degree of freedom of the parameter, Σ_0 is a $p \times p$ positive definite symmetric matrix, I_G is the normalizing constant and P_G is the set of all the $p \times p$ positive definite symmetric matrices constrained by G. A noninformative prior is placed on graph G that is, the prior probability of any edge ξ to be present in the graph is $\pi(\xi) = 0.5$.

A Monte Carlo Markov Chain (MCMC) algorithm is employed to approximate the full joint posterior probability of model (2). In particular, we adopt a Gibbs sampling strategy to sample from the posterior distribution of the β's, τ^2_ε and μ, while the covariance matrix Σ and the graph G are sampled using a birth-death step as implemented in the `R package BDgraph` [2]. As a result, the vector of

all the visited graph is returned along with the associated weight that provide an approximation of the posterior inclusion probability of each edge ξ.

Given the MCMC output, a variety of summaries can be adopted to estimate the graph. Here, we employ a Bayesian version of the (approximate) expected false discovery rate (FDR; [3]), i.e., we estimate the graph considering those edges whose posterior probability of inclusion is greater than$1-r$, where r is determined so that the FDR is at most 5%.

3 Analysis of fruit purees data

We analyze the spectrum of absorbance of 351 fruit purees prepared using exclusively fresh whole strawberries (without the addition of adulterants) measured on an equally-spaced grid of 235 wavelengths levels and then normalized with respect to the area under the curve; see the left panel of Figure 2. A description of the data can be found in [1]. The goal of the analysis is to provide useful insights about which wavelength bands are related to the different components of the purees, and more specifically about the dependency structure between wavelength bands.

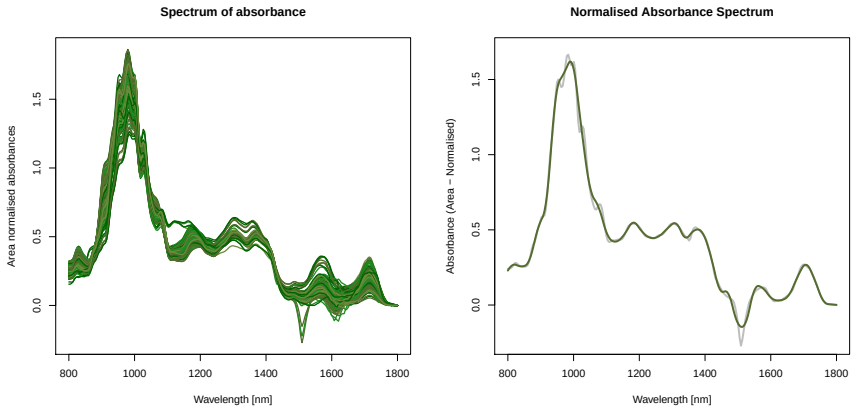


Fig. 2 Plot of 351 spectra of absorbance measured in 235 different wavelengths in left panel and an example of a smoothed curve in the right panel.

We fit model (2) by assuming $p = 40$ B-spline cubic basis functions. The right panel of Figure 2 presents an example of a smoothed curve (green) compared to the original one (gray). Observe that the smoothed curve follows precisely the shape of the original one, smoothing away some pointwise variability.

The left panel of Figure 3 shows the posterior mean of the coefficients while the right panel displays the estimated graph. As expected, the underlying graph is quite sparse and characterized by a block structure. The identified blocks are located

mainly close to the diagonal, with some extra-diagonal exceptions. The prevalence of blocks close to the diagonal is expected, since close wavelength bands are likely to be associated to similar components. The blocks far from the diagonal are also of great interest, since they suggest that components associated to very different wavelengths are strongly related according to their presence in the strawberry purees.

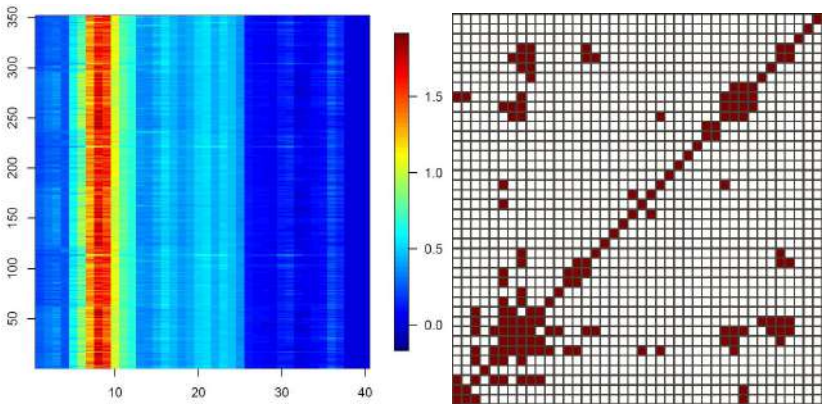


Fig. 3 Posterior mean of all β coefficients (left panel) and the estimated graph (right panel).

References

[1] J. K. Holland, E. K. Kemsley, and R. H. Wilson. Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. *Journal of the Science of Food and Agriculture*, 76(2):263–269, 1998.

[2] A. Mohammadi and E. C. Wit. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

[3] C. Peterson, M. Vannucci, and F. Stingo. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2014.

Local LGCP estimation for spatial seismic processes

Stima di LGCP locali per processi sismici spaziali

Nicoletta D'Angelo, Marianna Siino, Antonino D'Alessandro, Giada Adelfio

Abstract Using recent results for local composite likelihood for spatial point processes, we show the performance of advanced and flexible statistical models to describe the spatial displacement of earthquake data. Local models described by [1] allow for the possibility of describing both seismic catalogs and sequences. When analysing seismic sequences, the analysis of the small scale variation is the main issue. The interaction among points is taken into account by Log-Gaussian Cox Processes models through the estimation of the parameters of the covariance of the Gaussian Random Field. In their local version these parameters are allowed to vary spatially, and this is a crucial aspect for describing and characterizing the study area through a multiple underlying process.

Abstract *Sulla base di recenti risultati per la verosimigianza locale composita per i processi di punto spaziali, in questo lavoro si analizza la performance di modelli statistici avanzati al fine di descrivere la dislocazione spaziale di dati sismici. I modelli locali descritti da [1] risultano adatti per la caratterizzazione di cataloghi e sequenze sismiche. Per quanto riguarda le sequenze, l'analisi della variazione a piccola scala è cruciale. L'interazione fra i punti è considerata dai modelli Log-Gaussian Cox Processes attraverso la stima dei parametri della covarianza del Gaussian Random Field. Nella loro versione locale questi parametri sono assunti variabili nello spazio, il che rappresenta un aspetto chiave nella descrizione e rapprensentazione della zona analizzata, caratterizzata da processi multiscala sottostanti.*

Key words: earthquakes; Gibbs process; local composite likelihood; Cox process; point process

Nicoletta D'Angelo, Marianna Siino, Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Univeristà degli Studi di Palermo, Palermo, e-mail: nicoletta.dangelo@unipa.it; marianna.siino01@unipa.it; giada.adelfio@unipa.it

Antonino D'Alessandro
Istituto Nazionale di Geofisica e Vulcanologia, Palermo, e-mail: antonino.dalessandro@ingv.it

1 Local composite likelihood for Poisson Models

When dealing with the analysis of spatial point processes, spatial log-linear model are often used. Although, for these for models, parameters are usually assumed to be constant across the entire study region, this assumption may be too simplistic in contexts that are characterised by multiscale and fractal features, like the seismic one. The relationship between the intensity of earthquakes and other possible characteristics of the area where events occur, can be described by a log-linear model with spatial varying coefficient. That is:

$$\lambda(\mathbf{u}) = \lambda(\mathbf{u};\theta(\mathbf{u})) = \exp(\theta(\mathbf{u})^{\top} Z(\mathbf{u}) + B(\mathbf{u})) \tag{1}$$

where $\theta(\cdot)$ is a function of the spatial location $\mathbf{u} \in D$. In the local context, the template model is the one in Equation (1) and the local log-likelihood associated with location $\mathbf{u}$ is $\log \mathrm{L}(\mathbf{u};\theta) = \sum_{i=1}^{n} w_h(\mathbf{u}-u_i) \log \lambda(u_i;\theta) - \int_D \lambda(v;\theta) w_h(\mathbf{v}-\mathbf{u}) dv$ where $w_h(\mathbf{u}) = h^{-d} w(\mathbf{u}/h)$ is a weight nonparametric function, and $h > 0$ is a smoothing bandwidth [5]. Maximising the local likelihood provides spatial varying parameter estimates, confidence intervals, hypothesis tests and other standard tools [1].

1.1 Spatial Log-Gaussian Cox Processes

Cox processes are models for point phenomena that are environmentally driven and have a clustered structure. Let N be a point pattern defined on a spatial region D. N is said to be a Cox process driven by Λ, if its conditional distribution, given a realisation $\Lambda(\mathbf{u}) = \lambda(\mathbf{u})$, is a Poisson process on D with intensity function $\lambda(\mathbf{u})$. Following the inhomogeneous specification in [7], the Log-Gaussian Cox Process (LGCP) for a generic point in space has the following intensity $\Lambda(\mathbf{u}) = \lambda(\mathbf{u}) \exp\{S(\mathbf{u})\}$ where S is a Gaussian Random Field (GRF) with $E(S(\mathbf{u})) = \mu = -\frac{\sigma^2}{2}$ and so $E(\exp\{S(\mathbf{u})\}) = 1$ and with variance-covariance matrix $C(S(\mathbf{u}), S(\mathbf{v})) = C(||\mathbf{u}-\mathbf{v}||) = \sigma^2 R(r)$ under the stationary assumption, where $R(\cdot)$ is the correlation function of the GRF, completely specified by its first and second moments. A LGCP is completely determined by $(\mu, \sigma^2, R(\cdot))$ of the GRF of the point process. It is isotropic if and only if the underlying Gaussian process is isotropic. Several authors have discussed on the issue of separability of LGCPs in space and time [6]. We only refer to models where C is isotropic, i.e. only depends on the distance $r = \|u-v\|$ between locations u and v, with exponential isotropic structure for the covariance function as in [4]. The default template used in this paper is $R(r) = \exp(-r)$ which rises to the exponential covariance function

$$C(r) = \sigma^2 \exp\left(-\frac{r}{\alpha}\right) \tag{2}$$

depending only on α and σ^2 that are the scale parameter for the spatial distance and the variance respectively. The effect of increasing σ^2 is to generate higher peaks in the surface intensity which leads to clusters of points. Increasing the spatial scale parameter α, the underlying GRF presents a strong spatial correlation and it corresponds to a diffuse aggregation of points of the LGCP [9]. Estimation of local LGCP models che be performed by the adaptation of the Palm likelihood to a general non-stationary point process, because of its formal similarity to the Poisson likelihood, though minimum contrast estimation is also possible [1].

2 Application to seismic data

The used catalogue data (1105 events) concern earthquakes occurred in Greece between 2005 and 2014. Only seismic events with a magnitude larger than 4 are considered in this study, and the analyses are marginal with respect to time, focusing on the spatial dependence of events. Starting from the information of the area under study, we are interested in assessing if local characteristics are present in the observed area and if the presence of seismic sources can affect the intensity of the process, that is if the effect of the available covariates can improve the fitting of the model. Therefore, the considered explanatory variables are the Distance from the faults D_f, Distance from the plate boundary D_{pb} and the Distance from volcanoes D_v, computed as the Euclidean distances from the spatial location u of events and the map of the available geological information [3]. Earthquakes, together with plate boundary (in red) and faults (in blue), are displayed in Figure 1 (a). Several models are fitted, with different specifications of the linear predictor. Diagnostics and model selection are carried out through the inspection of the smoothed raw residuals and the inhomogeneous K-functions, but these are not reported here for the sake of brevity. The chosen local Poisson model is

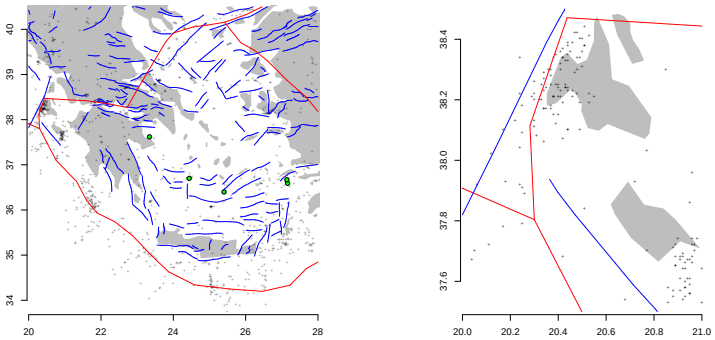


Fig. 1 Left panel: Earthquakes occurred in Greece between 2005 and 2014 (a). Right panel: Earthquakes occurred in Ithaki, Kefalonia and Zakynthos between 2005 and 2014 (b). Volcanoes, in green, faults, in blue, and plate boundary, in red.

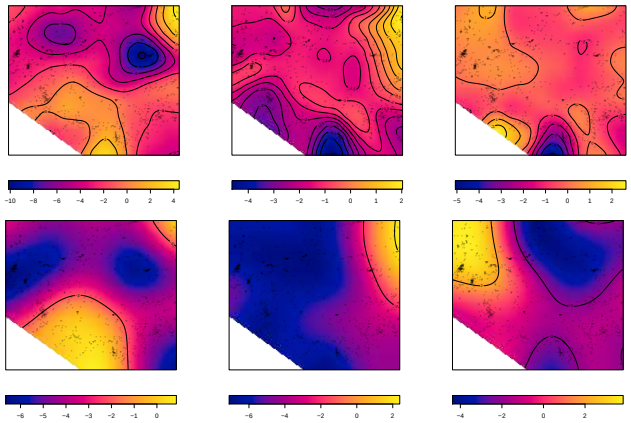


Fig. 2 Top three panels: Spatial varying estimated coefficients $\hat{\beta}_1(\mathbf{u})$, $\hat{\beta}_2(\mathbf{u})$ and $\hat{\beta}_3(\mathbf{u})$ of the local Poisson model in Equation (3). Bottom three panels: the corresponding T-tests.

Table 1 Summary of the spatial varying estimated coefficients of the local Poisson model in Equation (3).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Intercept	-0.8465	3.7821	4.6536	4.7818	5.4126	13.7942
D_f	-10.3862	-4.0474	-2.3888	-2.4529	-0.6133	4.6906
D_{pb}	-4.735	-1.925	-1.413	-1.448	-1.04	2.17
D_v	-5.3552	-0.6595	-0.2799	-0.2987	0.2351	2.9546

$$\lambda(\mathbf{u}) = \exp(\beta_0(\mathbf{u}) + \beta_1(\mathbf{u})D_f(\mathbf{u}) + \beta_2(\mathbf{u})D_{pb}(\mathbf{u}) + \beta_3(\mathbf{u})D_v(\mathbf{u})) \tag{3}$$

The maps of the varying estimated coefficients $\hat{\beta}_1(\mathbf{u})$, $\hat{\beta}_2(\mathbf{u})$ and $\hat{\beta}_3(\mathbf{u})$ are shown on the top three panels in Figure 2. Their summary statistics are reported in Table 1. The corresponding T-tests are reported on the bottom three panels, where the level curves correspond to the ± 1.96 threshold, associated with a 0.95 confidence level. The blue regions are those in which the null hypothesis does not hold, and so where the coefficients are significantly different from zero. For more details see [1]. As expected, the estimated coefficients are negative quite in all the area, since usually, the intensity decreases as the distance from the seismic source increases, but it is important to notice how the coefficients take quite different values along the whole areas. The lowest values are found exactly in correspondence of earthquakes occurred along the seismic source. These previous analyses of the entire catalogue highlighted the need to focus the analysis on a smaller region. Therefore we move to the analysis of a seismic sequence. The chosen area, that displays an inhomogeneous behaviour, is situated in the Ionian Sea, and it comprehends three of 'the Seven Islands', namely Ithaki, Kefalonia and Zakynthos, from North to South. In this area, earthquakes appear to be clustered in the Western area of Kefalonia island and South to the Zakynthos island. For this analysis, the variables D_f and D_{pb} are taken into account, in order to get an interpretation of the two effects. In Figure 1 (b)

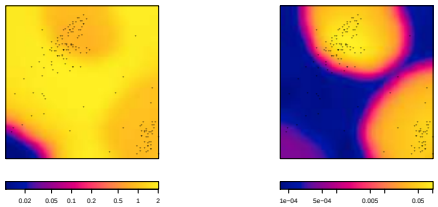


Fig. 3 Spatial varying estimated coefficients $\hat{\sigma}(\mathbf{u})$ and $\hat{\alpha}(\mathbf{u})$ of the local LGCP model in Equation (4).

Table 2 Summary of the spatial varying estimated coefficients of the local LGCP model in Equation (4).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
sigma	0.000105	0.802612	2.451644	2.192861	3.627219	4.157830
alpha	0.00001	0.00009	0.00012	0.01988	0.03764	0.11281

earthquakes, together with plate boundary (in red) and faults (in blue), are displayed. In this application we first consider a local Poisson model with:

$$\lambda(\mathbf{u}) = \exp(\beta_0 + \beta_1 D_f(\mathbf{u}) + \beta_2 D_{pb}(\mathbf{u}) + \beta_3 D_f(\mathbf{u}) D_{pb}(\mathbf{u})) \tag{4}$$

and then it considered as the deterministic part of a local LGCP model, trying to analyse interaction features among points of the selected sequence. The local version of the LGCP model estimates the spatial varying additional parameters $\sigma^2(\mathbf{u})$ and $\alpha(\mathbf{u})$, relaxing the usual assumption that in LGCP models the underlying process is unique and therefore one could obtain only a single value of the covariance of the process describing the correlation among points of the process. With the local version of the LGCP model it is now possible to spot those regions in which, combining together the two additional parameters, the covariance is higher, and therefore it is possible to identify regions in which different processes generate the observed events. Recalling the formula of the covariance of the process from Equation (2), we can introduce the modification of the latter $C(r,\mathbf{u}) = \sigma^2(\mathbf{u}) \exp\left(\frac{-r}{\alpha(\mathbf{u})}\right)$ that is, now the covariance depends also on the location and it is not constant. Basically, while usually it is assumed that the process generating the events is unique, the local approach allows for the assumption that the events may be generated from multiple processes. This is particularly useful since we are able to describe areas where points show different correlation patterns, fitting a unique model. The estimates $\hat{\sigma}^2(\mathbf{u})$ and $\hat{\alpha}(\mathbf{u})$ are displayed in Figure 3 and their summary statistics are reported in Table 2. The results are coherent with the estimated coefficients $\hat{\sigma}^2$ and $\hat{\alpha}$ of the global LGCP model that are respectively equal to 1.92 and 0.2. From Equation (2) we know that a unit increase of $\alpha(\mathbf{u})$ has a grater effect on the computation of the covariance, if compared to a unit increase of $\sigma^2(\mathbf{u})$. Therefore, from Figure 3, we know that the difference in the correlation among points is evident. In particular, this is higher in those areas in which cluster of events occurred, i.e. on the top of the region and on the bottom-right.

3 Conclusions

In this article the local models described by [1] are applied for the description of the seismic events occurred in Greece. This is an area of high seismic hazard that has been characterised by many destructive earthquakes in the last century. In [8], Hybrid of Gibbs models [2] are used to describe the same Hellenic catalogue. In this paper, interaction among points, crucial in the context of the analysis of seismic events, has been taken into account by fitting the local version of LGCP models. This approach makes possible to account both for the effect of covariates and also for the interaction among points, as function of the spatial location. From our application, local models provide good inferential results. It seems that fitting parameters that vary locally leads to more parsimonious models, explaining the spatial underlying variation of the phenomenon. One limit of the provided study is that the local composite likelihood works well if kernel smoothing well performs. Dealing with earthquake data, this is not always true, since their well known features of inhomogeneity and clustering in space. Finally, another important limit of this work is the neglected time component. Indeed, when analysing seismic data, often the focus is on the time sequence data analysis, since aftershocks are the proof of the dependence of earthquakes from the past. One possible future development could take into account both the spatial component, treated through local methods, such as the locally weighted regression, and the time component, moving in the context of space-time point processes.

References

1. Baddeley, A. (2017). Local composite likelihood for spatial point processes. Spatial Statistics, 22, 261-295.
2. Baddeley, A., Turner, R., Mateu, J., Bevan, A. (2013). Hybrids of Gibbs point process models and their implementation.
3. Baddeley, A., Rubak, E., Turner, R. (2015). Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC.
4. Brix, A., Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(4), 823-841.
5. Daley, D. J., Vere-Jones, D. (2008). Spatial point processes. An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure, 457-536.
6. Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G., Gooding, S. (2007). Second-order analysis of inhomogeneous spatial point processes using case–control data. Biometrics, 63(2), 550-557.
7. Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Statistical Science, 28(4), 542-563.
8. Siino, M., Adelfio, G., Mateu, J., Chiodi, M., D'alessandro, A. (2017). Spatial pattern analysis using hybrid models: an application to the Hellenic seismicity. Stochastic environmental research and risk assessment, 31(7), 1633-1648.
9. Siino, M., Adelfio, G., Mateu, J. (2018). Joint second-order parameter estimation for spatio-temporal log-Gaussian Cox processes. Stochastic environmental research and risk assessment, 32(12), 3525-3539.

Observation-driven models for storm counts

Modelli observation-driven per il conteggio delle tempeste

Mirko Armillotta, Alessandra Luati and Monia Lupparelli

Abstract New technologies have made available rich datasets of counts and binary data, focusing attention on time series models for discrete-valued processes. From the recent establishment of their theory, not much has been said about the practical usefulness of these data. Here, to partially address this task, an illustrative example on number of storm detected in the North Atlantic is presented.

Abstract *Nuove tecnologie hanno reso disponibile un ricco ammontare di dati binari e di conteggio, concentrando l'attenzione sui modelli di serie storiche per processi discreti. Dal recente sviluppo della loro teoria, non molto è stato detto riguardo l'utilità pratica di questi dati. Qui, per svolgere parzialmente questo compito, è presentato un esempio illustrativo sul numero di tempeste rilevate nel Nord Atlantico.*

Key words: Discrete data, QMLE, Generalized ARMA, Poisson autoregression, Calibration, Sharpness

1 A general framework for discrete-valued observation-driven models

Let us consider $\{Y_t\}_{t\in\mathbb{Z}}$ as a discrete-valued time series process. We define a class of observation-driven models such that:

$$Y_t \,|\, \mathscr{F}_{t-1} \sim q(\cdot;\mu_t)\,, \tag{1}$$

Mirko Armillotta, Alessandra Luati
Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126, Bologna, e-mail: mirko.armillotta2@unibo.it, e-mail: alessandra.luati@unibo.it

Monia Lupparelli
Department of Statistics, Computer Science, Applications, University of Florence, Viale Morgagni 59, 50134, Florence, e-mail: monia.lupparelli@unifi.it

$$g(\mu_t) = \alpha + \sum_{j=1}^{k} \gamma_j g(\mu_{t-j}) + \sum_{j=1}^{p} \phi_j h(Y_{t-j}) + \sum_{j=1}^{q} \theta_j \left[\frac{h(Y_{t-j}) - \bar{g}(\mu_{t-j})}{s_{t-j}} \right] \quad (2)$$

where $\mathscr{F}_{t-1} = \sigma(Y_{t-s}, s > 0)$, with. It is assumed that $q(\cdot; \mu_t)$ is a density (or mass) function whose dynamic is captured by the parameter μ_t. We define $g(\cdot)$, $h(\cdot)$ and $\bar{g}(\cdot)$ some one-to-one monotonic functions with range $\mathbb{R}$. The process s_t is some scaling sequence, usually $s_t = \sqrt{\mathrm{V}[h(Y_t)|\mathscr{F}_{t-1}]}$. The parameters ϕ_j measure an autoregressive effect of the observations; γ_j state the long memory dependence of the process (since μ_{t-j} depends on past observations Y_r, $r < t-j$); θ_j represents the analogous of a moving average part. In general, all the functions involved are not constrained to assume the same shape. This leads to a quite general and flexible framework which encompasses the most frequently used models on discrete-valued observation processes. Ergodicity conditions for model (1, 2) has been found in [1]. In [2] the same authors proved consistency and asymptotic normality of the quasi maximum likelihood estimator (QMLE), as defined in [7]. Once inference theory is established, we present an application.

2 Named storms in the North Atlantic Hurricane Basin

Here, by using (1, 2), we analyse the annual number of named storms in the North Atlantic Hurricane Basin (1851-2018) . In the label "named storm" are counted tropical storms, hurricanes and subtropical storms, from HURDAT database. The time series, with $n = 168$, is plotted in Figure 1 along with its autocorrelation function. There is a temporal correlation which spread over several lags in the past; this suggest the use a long memory feedback mechanism. We decided to fit models coming from two different distributions; the Poisson distribution:

$$q(Y_t, \mu_t) = \mathrm{P}(Y_t = y|\mathscr{F}_{t-1}) = \frac{\exp(-\mu_t)\mu_t^y}{y!}, \quad (3)$$

and the Negative binomial distribution (NB, henceforth):

$$q(Y_t, \mu_t) = \mathrm{P}(Y_t = y|\mathscr{F}_{t-1}) = \frac{\Gamma(\nu + y)}{\Gamma(y+1)\Gamma(\nu)} \left(\frac{\nu}{\nu + \mu_t} \right)^{\nu} \left(\frac{\mu_t}{\nu + \mu_t} \right)^{y}, \quad (4)$$

where $\nu > 0$ is the dispersion parameter and μ_t is the conditional expectation; the latter is the same for both distributions; in fact (4) is defined in terms of mean instead of probability parameter $p_t = \frac{\nu}{\nu + \mu_t}$ and it accounts for overdispersion in the data, since in the Poisson (3) mean and variance are the same, whereas in (4), $\mathrm{V}(Y_t|\mathscr{F}_{t-1}) = \mu_t(1 + \mu_t/\nu) \geq \mu_t$. We define a model selection procedure by estimating the following sub-models of (2). For simplicity, all the models include only the first lag. The log-linear autoregression (log-AR [8]):

$$\log(\mu_t) = \alpha + \phi \log(y_{t-1} + 1) + \gamma \log(\mu_{t-1}). \quad (5)$$

is obtained from (2) when the index $q = 0$, $g(\mu_t) = \log(\mu_t)$ and $h(y_t) = \log(y_t + 1)$. The Generalized ARMA model (GARMA, [3]):

$$\log(\mu_t) = \alpha + \phi \log(y^{\star}_{t-1}) + \theta \left[\log(y^{\star}_{t-1}) - \log(\mu_{t-1})\right] \tag{6}$$

is defined here from (2) with $k = 0$, $g \equiv \bar{g} \equiv h = \log(\cdot)$, where $y^{\star}_{t-1} = \max\{y_t, c\}$ and $c = 0.1$. The Generalized linear ARMA model (GLARMA, [6]) can be derived by (2) with $p = 0$, $g(\mu_t) = \log(\mu_t)$, $h(y_t) = y_t$ and $\bar{g}(\mu_t) = \mathrm{E}(Y_t|\mathscr{F}_{t-1})$:

$$\log(\mu_t) = \alpha + \gamma \log(\mu_{t-1}) + \theta \left(\frac{y_{t-1} - \mu_{t-1}}{s_{t-1}}\right) \tag{7}$$

where $s_t = \sqrt{\mu_t}$ for the Poisson distribution and $s_t = \sqrt{\mu_t\,(1 + \mu_t/\nu)}$ for the NB. A more general model is also introduced:

$$\log(\mu_t) = \alpha + \gamma \log(\mu_{t-1}) + \phi \log(y_{t-1} + 1) + \theta \left[\frac{\log(y_{t-1} + 1) - \bar{g}(\mu_{t-1})}{s_{t-1}}\right] \tag{8}$$

where $\bar{g}(\mu_t) = \mathrm{E}\left[\log(Y_t + 1)|\mathscr{F}_{t-1}\right]$ and $s^2_t = \mathrm{V}\left[\log(Y_t + 1)|\mathscr{F}_{t-1}\right]$. These conditional moments are computed by using the second order Taylor expansion. The score functions written in terms of predictor $x_t = \log \mu_t$ are:

$$\chi_n(\rho) = \frac{1}{n} \sum_{t=1}^{n} \Big(y_t - \exp x_t(\rho)\Big) \frac{\partial x_t(\rho)}{\partial \rho} \tag{9}$$

$$\chi_n(\rho) = \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \frac{(y_t + \nu)\exp x_t(\rho)}{\exp x_t(\rho) + \nu}\right) \frac{\partial x_t(\rho)}{\partial \rho} \tag{10}$$

The solution of the system of non-linear equations $\chi_n(\rho) = 0$, if it exists, provides the QMLE of ρ (denoted by $\hat{\rho}$). In NB models, the estimation of ν is required. We used the moment estimator, as in [4]:

$$\hat{\nu}_1 = \left\{1/n \sum_{t=1}^{n} \left[(y_t - \hat{\mu}_t)^2 - \hat{\mu}_t\right] / \hat{\mu}^2_t\right\}^{-1} \tag{11}$$

where $\hat{\mu}_t = \mu_t(\hat{\rho})$ from the Poisson model. Then, with $\nu = \hat{\nu}_1$ we estimate the NB model and obtain the new estimates for $\hat{\mu}_t$, plug them into (11), obtain a new value for $\hat{\nu}_1$, and repeat the procedure until desired convergence.

Clearly, we replaced each quantity with the sample counterparts computed at $\hat{\rho}$. The results of the analysis are summarized in Table 2. The intercept is not significant, at a 5% level, for the General models and the NB log-AR and GARMA models. The same can be said for the parameter $\hat{\gamma}$ in the NB General model. All the other coefficients are significant. The parameter $\hat{\nu}$ is generally around 5. Both AIC and BIC select the Pois GLARMA model as the best, in goodness-of-fit sense.

We then assess the adequacy of fit. We check the behaviour of the standardized Pearson residuals $e_t = [Y_t - \mathrm{E}(Y_t|\mathscr{F}_{t-1})] / \sqrt{\mathrm{V}(Y_t|\mathscr{F}_{t-1})}$ which is done by taking the

empirical version $\hat{e}_t$ from the estimated quantities. If the model is correctly specified, the residuals should be white noise sequence with constant variance. The ACF in our case appear quite uncorrelated (not presented for space constraints).

Another check comes form the probability and marginal calibrations (mc), as defined in [9]. In particular [5] introduced a Probability Integral Transform (PIT) for discrete data. It can be build by the following conditional distribution function

$$F(u|y_t) = \begin{cases} 0, & u \leq P_t(y_t - 1) \\ \frac{u - P_t(y_t-1)}{P_t(y_t) - P_t(y_t-1)}, & P_t(y_t) \leq u \leq P_t(y_t - 1) \\ 1, & u \geq P_t(y_t) \end{cases} \tag{12}$$

where $P_t(\cdot)$ is the cumulative distribution function (CDF) at time t (in our case Poisson or NB). If the model is correct, $u \sim Uniform(0,1)$ and the PIT (12) will appear to be the cumulative distribution function of a Uniform(0,1). the PIT (12) is computed for each realisation of the time series $y_t, t = 1\ldots,n$ and for values $u = j/J, j = 1,\ldots,J$, where J is the number of bins (usually equal to 10 or 20); then its mean $\bar{F}(j/J) = 1/n\sum_{t=1}^{n} F(j/J|y_t)$ is taken. The outcomes are probability mass functions, which are obtained in terms of differences $\bar{F}(\frac{j}{J}) - \bar{F}(\frac{j-1}{J})$ plotted in Figure 2. The Poisson PIT's appear to be closer to Uniform(0,1).

The marginal calibration is defined as in [9] and [4]. Practically, it compares the average of CDF selected, $\bar{P}(x) = 1/n\sum_{t=1}^{n} P_t(x)$, against the average of the empirical CDF, $\bar{G}(x) = 1/n\sum_{t=1}^{n} \mathbf{1}(y_t \leq x)$; the difference $\bar{P}(x) - \bar{G}(x)$ is plotted in Figure 1 for GLARMA model and the model (8). In the other models the results are similar to those of GLARMA. Clearly, the nearer the difference to zero axis, the better. A far better concordance with empirical distribution for the Poisson case is found in the mc plot in Figure 1.

In order to assess the power of prediction we refer to the concept of sharpness of the predictive distribution defined in [9]. It can be measured by some average quantities related to the predictive distribution, which take the form $1/n\sum_{t=1}^{n} d(P_t(y_t))$, and $d(\cdot)$ is some function called scoring rule. We used some of the usual scoring rules employed in the literature: the logarithmic score (logs) $-\log p_t(y_t)$, where $p_t(\cdot)$ is the probability mass at the time t; the quadratic score (qs) $-2p_t(y_t) + \|p\|^2$, where $\|p\|^2 = \sum_{k=0}^{\infty} p_t^2(k)$; the spherical score (sphs) $-p_t(y_t)/\|p\|$ and the ranked probability score (rps) $\sum_{k=0}^{\infty}[P_t(k) - \mathbf{1}(y_t \leq k)]$, for different models and distributions. The results are summarized in Table 1. The Poisson General model have the best predictive performance for all the scoring rules. This shows the usefulness of our framework for the model selection in the existing literature as well as different new models which lead to improved performances.

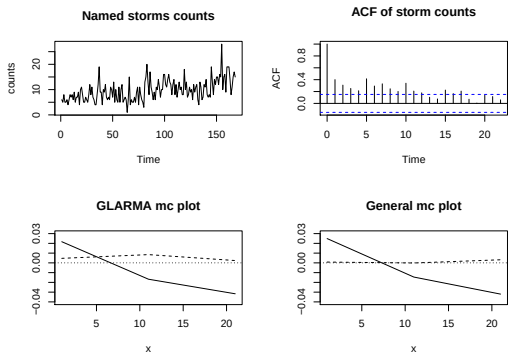


Fig. 1 Top-left: named storms counts. Top-right: ACF. Bottom-left: mc plot GLARMA model. Bottom-right: mc plot General model. Dashed line Poisson. Black line NB. Dotted line zero axis.

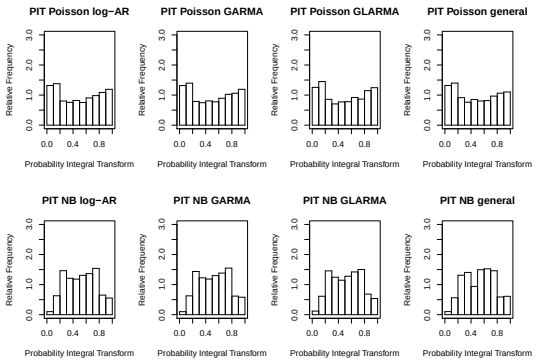


Fig. 2 Top: PIT's for the Poisson model. Bottom: PIT's for the NB model.

Table 1 Predictive performance for named storms.

Models	Distribution	logs	qs	sphs	rps
log-AR	Poisson	2.7257	-0.0775	-0.2808	2.0320
	NB	2.8018	-0.0727	-0.2723	2.1235
GARMA	Poisson	2.7293	-0.0774	-0.2807	2.0342
	NB	2.8059	-0.0724	-0.2718	2.1285
GLARMA	Poisson	2.7247	-0.0768	-0.2796	2.0384
	NB	2.7927	-0.0735	-0.2736	2.1073
General	Poisson	**2.6989**	**-0.0812**	**-0.2859**	**1.9925**
	NB	2.7987	-0.0732	-0.2733	2.0905

Table 2 MLE results storms.

Models	$\hat{\alpha}$	$\hat{\phi}$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\nu}$	*AIC*	*BIC*
Pois log-AR	0.212	0.231	0.673	-	-	11.361	20.733
	(0.082)	(0.058)	(0.089)	-			
Pois GARMA	0.289	0.882	-	-0.684	-	11.368	20.740
	(0.092)	(0.039)	-	(0.083)			
Pois GLARMA	0.314	-	0.864	0.071	-	**11.359**	**20.731**
	(0.103)	-	(0.046)	(0.018)			
Pois General	-0.037	0.790	0.210	-0.178	-	13.307	25.803
	(0.029)	(0.115)	(0.109)	(0.003)			
NB log-AR	0.390	0.286	0.540	-	5.262	11.528	20.900
	(0.310)	(0.114)	(0.246)	-			
NB GARMA	0.483	0.797	-	-0.556	5.190	11.536	20.908
	(0.354)	(0.154)	-	(0.248)			
NB GLARMA	0.376	-	0.836	0.139	5.402	11.510	20.881
	(0.194)	-	(0.086)	(0.041)			
NB General	0.016	0.916	0.084	-0.365	5.157	13.505	26.001
	(0.184)	(0.190)	(0.117)	(0.124)			

References

1. Armillotta, M., Luati, A., Lupparelli, M.: Stationarity of a general class of observation driven models for discrete valued processes. Book of short Papers SIS 2019, Pearson, 2019, 31–39 (2019)
2. Armillotta, M., Luati, A., Lupparelli, M.: Observation driven models for discrete-valued time series. Forthcoming.
3. Benjamin, M.A., Rigby, R.A., Stasinopoulos, D.M.: Generalized autoregressive moving average models. J. Amer. Stat. Assoc. **98**, (461), 214–223 (2003)
4. Christou, V., Fokianos, K.: On count time series prediction. J. Stat. Comput. and Sim. **85**, (2), 357–373 (2015)
5. Czado, C., Gneiting, T., Held, L.: Predictive model assessment for count data. Biometrics **65**,(4), 1254–1261 (2009)
6. Davis, R., Dunsmuir, W., Streett, S.: Observation-driven models for Poisson counts. Biometrika **90**, (4), 777–790 (2003)
7. Douc, R., Fokianos, K., Moulines,E.: Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. Electron. J. Stat. **11**, (2), 2707–2740 (2017)
8. Fokianos, K., Tjøstheim, D.: Log-linear Poisson autoregression. J. Multivar. Anal. **102**, (3), 563–578 (2011)
9. Gneiting, T., Balabdaoui, F., Raftery, A. E.: Probabilistic forecasts, calibration and sharpness. J. Royal Stat. Soc.:B **69**, (2), 243–268 (2007)

Statistical control of complex geometries, with application to Additive Manufacturing

Controllo statistico di geometrie complesse, con applicazioni all'Additive Manufacturing

Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio, Piercesare Secchi

Abstract New production processes, as Additive Manufacturing (AM), allow the production of objects and shapes characterized by a growing complexity, especially when compared with those normally manufactured in traditional production processes. It is thus necessary to develop methods which can be applied to the study of the variability of a dataset whose elements are manufactured realizations of the same nominal model, which can carry a very high degree of complexity. In the present work, we propose a method allowing modeling a wide variety of possibly local geometric deviations of the manufactured object with respect to the nominal model. A way of identifying and monitoring these kind of deviations, based on Principal Component Analysis in Hilbert spaces, is proposed as well. The proposed method is tested on a real datased of items produced via AM.

Abstract Lo sviluppo di nuovi metodi di produzione, quali l'Additive Manufacturing (AM), rende realizzabili forme geometriche sempre più articolate e molto più complesse di quelle normalmente realizzate in processi produttivi più tradizionali. Si rende dunque necessario lo sviluppo di metodi statistici applicabili all'esplorazione della variabilità di dataset in cui ogni osservazione è una realizzazione effettiva-

Riccardo Scimone
MOX, Dipatimento di Matematica, Politecnico di Milano
Center for Analysis, Decision and Society, Human Technopole

Tommaso Taormina
Dipartimento di Meccanica, Politecnico di Milano

Bianca Maria Colosimo
Dipartimento di Meccanica, Politecnico di Milano

Marco Grasso
Dipartimento di Meccanica, Politecnico di Milano

Alessandra Menafoglio
MOX, Dipartimento di Matematica, Politecnico di Milano

Piercesare Secchi
MOX, Dipartimento di Matematica, Politecnico di Milano

mente prodotta di un modello nominale, che può presentare un elevato grado di complessità. Nel presente lavoro viene proposto un metodo che consente, in linea di principio, di modellare un insieme molto ampio di deviazioni geometriche, anche locali, degli oggetti prodotti rispetto al modello nominale, e di rilevare successivamente tali deviazioni via Analisi delle Componenti Principali in spazi di Hilbert. Tale metodo viene altresì testato su un dataset di oggetti reali prodotti via AM.

Key words: Statistical Process Control, Object Oriented Statistics, Compositional Data Analysis, Functional Data

1 Introduction

Additive Manufacturing processes are becoming more and more important, since they are facing a continuous technological improvement, and they allow the realization of a wide variety of shapes and geometries. The production of complex shapes (e.g, objects with a complete asimmetry, or with an internal lattice structure) constitutes an interesting challenge for Statistical Process Control. The necessity of adequate methods, suitable to the statistical control of this kind of objects, is moreover driven by the fact that the industrial sectors in which AM presents the greatest potentialities are the aerospace and the biomedical sectors ([10]), both areas in which the identification of production errors is foundamental.
After the manufacturing phase, data usually come as reconstructed shapes (e.g., via tomography), in the form of a point cloud, possibly associated to a triangulated mesh. Point clouds are a class of data whose statistical process monitoring has been studied by several authors ([6], [7], [11], [12]). In all cited works, points in the reconstructed point cloud were associated with their deviation from the nominal geometry (usually a CAD model), hence producing a *deviation map*. However, the deviation map generated by the distances of the points in the reconstructed point cloud is usually different from the deviation map generated by the distances of the points in the nominal point cloud from the reconstructed object. Different information are, in general, carried by the corresponding deviation maps, as follows from the formal definition of Hausdorff distance between subsets of a metric space, which was firstly introduced, in a completely different context, by Hausdorff in [1]. In [15] we propose to summarize all the information content about the differences between an object and its corresponding model using two *deviation maps*. These maps are then used in a monitoring method based on a functional Principal Component Analysis for compositional data (SFPCA, [8]), allowing to exploit this information content completely. In this communication, performances of the method shall be showcased on a dataset of manufactured objects, produced via Additive Manufacturing at the Department of Mechanical Engineering at Politecnico di Milano. These objects will be briefly introduced in the next section.

2 Test Dataset: a motivating example

The dataset on which all the analyses are carried out consists of a sample of 16 meshes, resulting from the tomography of plastic objects produced via Additive Manufacturing at the Department of Mechanical Engineering of Politecnico di Milano. These objects are trabecular egg-shaped shells, and they constitute a quite realistic example of the geometric complexity that can be achieved by Additive Manufacturing processes. We show the nominal model in Fig. 1.

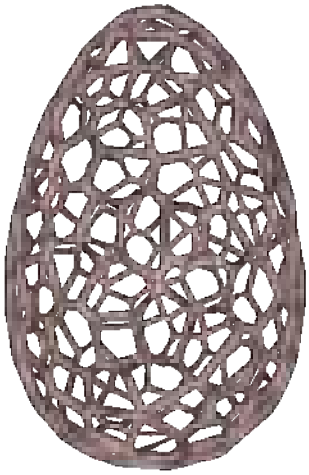


Fig. 1 Nominal CAD model for the produced trabecular structures.

Among the sixteen produced items, we have objects with no evident defects (In Control), as well as items affected by different kinds of geometrical deviations (Out of Control). In Fig. 2 we show two realizations of defective elements, on the left an item affected by irregularities on the geometrical structure, on the right a case in which a struct is missing.

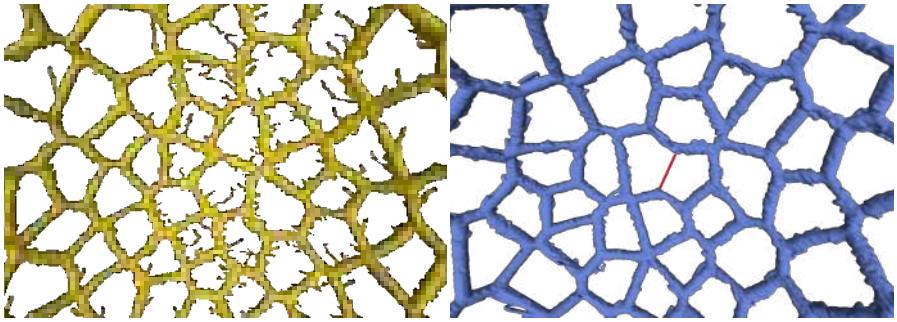


Fig. 2 Different kind of geometric deviations in the manufactured objects

It is worth noting that these geometrical deviations are an example of the complementariety of the two deviation maps mentioned in Section 1. Indeed, while the local irregularities can be detected if one measures the distance of points of the manufactured object from the nominal model, a missing struct is visible only when considering distance of points of the nominal model from the manufactured object.

3 Methodology: a density representation of Hausdorff maps

The general problem of analyzing the differences between two meshes or point clouds can be efficiently stated by referring to the definition of Hausdorff distance, given below (see [2] for a deeper insight).

Definition 1. Let X, Y be two closed, bounded, non-empty subsets of a metric space (U, d). Their Hausdorff distance is defined as

$$d_H(X,Y) := \max \left\{ \sup_{x\in X} \inf_{y\in Y} d(x,y), \sup_{y\in Y} \inf_{x\in X} d(x,y) \right\} \tag{1}$$

This definition implicitly introduces two maps, which are

$$d_X : Y \to \mathbb{R}^+, d_X(y) := \inf_{x\in X} d(x,y) \tag{2}$$

and

$$d_Y : X \to \mathbb{R}^+, d_Y(x) := \inf_{y\in Y} d(x,y) \tag{3}$$

which will be called, as already done, *deviation maps*. A key idea underlying the method we propose in [15] is that both deviation maps should be analyzed to re-

trieve all possible differences between a reconstructed geometry and the nominal one. In the case of a dataset of point clouds, these maps are easily computed, since they are represented by finite, discrete sets. From each map we estimate a density obtained by dropping the spatial information carried by the data. These are f_X and f_Y, representing the PDFs of the values of d_X and d_Y respectively (or of a monotone transformation of these). The dataset of PDFs can be analyzed on the common support of the densities through SFPCA ([8]), which extends Functional Principal Component Analysis ([3]) to probability density functions. The natural space in which PCA of density functions can be coherently performed has been defined by ([4]), and laterly studied in other works ([5], [9]). It is the space $B^2(a,b)$, defined as the set (of equivalence classes) of positive functions with square-integrable logarithms

$$B^2(a,b) := \{f > 0 \text{ s.t. } \log f \in L^2(a,b)\} \tag{4}$$

where the equivalence relation is defined as

$$f_1 = f_2 \iff \exists c > 0 \text{ s.t. } f_1 = c f_2 \text{ a.e..} \tag{5}$$

In $B^2(a,b)$, operations and inner product are defined as

$$f_1 \oplus f_2 = f_1 f_2, \tag{6}$$

$$\alpha \odot f_1 = f_1^{\alpha}, \alpha \in \mathbb{R} \tag{7}$$

$$\langle f_1, f_2 \rangle_{B^2} = \frac{1}{2(b-a)} \iint \log \frac{f_1(x)}{f_1(y)} \log \frac{f_2(x)}{f_2(y)} \, dx \, dy. \tag{8}$$

This space is a useful extension of the Aitchison geometry (see [13] for a complete reference), in which PCA can be applied in the form of SFPCA ([8]), and it can be practily applied relying on the centered log-ratio transformation, as in [14]. Dimensionality reduction via SFPCA allows reducing the monitoring problem to a multivariate one based on the scores along the first K principal components. In [15], we build a control chart scheme based on the scores vectors, which allows detecting out of control conditions for the produced objects.

4 Results and conclusion

SFPCA, based on a geometric structure coherent with the features of probability density functions, provides interpretation tools which are quite powerful when coupled with the geometric iterpretation of the introduced deviation maps. This dimensional reduction technique provides a very good compromise between spatial simplification (all spatial information is dropped) and loss of information (the functional approach allow to performs data analysis directly on density functions). The control chart scheme based on SFPCA proves to be very effective in detecting out-of-control conditions deriving from either widespread or localized defects. Indeed,

the method performs well also in presence of local defects affecting just a small part of the whole point cloud, exhibiting then good detection power.

5 Acknowledgemetns

The present work has been supported by ACCORDO Quadro ASI-POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano.

References

1. Hausdorff, F.: Grundzüge der mengenlehre, Leizpig, Von Veit, (1914)
2. Henrikson, J.: Completeness and Total Boundedness of the Hausdorff Metric (1999)
3. Ramsay, J., Silverman, B. W.: Functional Data Analysis. Springer (2005)
4. Egozcue, J.J., Díaz–Barrero, J.L., Pawlowsky-Glahn, V.: Hilbert Space of Probability Density Functions Based on Aitchison Geometry. Acta Mathematica Sinica, English Series 22, 1175–1182 (2006)
5. Gerard van den Boogaart, K., Egozcue, J.J., Pawlowsky-Glahn, V.: Bayes Linear Spaces. Statistics and Operations Research Transactions 34 (2010)
6. Megahed, F., Wells, L., Camelio, J.: A Spatiotemporal Method for the Monitoring of Image Data. Quality and Reliability Engineering International 28, 967–980 (2012)
7. Wells, L.J., Megahed, Fadel M., Niziolek, C.B., Camelio, J., Woodall, W.H.: Statistical process monitoring approach for high-density point clouds Journal of Intelligent Manufacturing 24, 1267–1279 (2013)
8. Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics & Data Analysis 94, (2014)
9. Pawlowsky-Glahn, V., Egozcue, J.J., Van den Boogaart, K.: Bayes Hilbert Spaces. Australian and New Zealand Journal of Statistics 56, 171–194 (2014)
10. Grasso, M., Colosimo, B.: Statistical Process Monitoring of Additive Manufacturing via In-Situ Sensing. QPRC 2017, The 34th Quality and Productivity Research Conference Quality and Statistics (2017)
11. He, K., Zhang, M., Zuo, L., Alhwiti, T., Megahed, F.: Enhancing the monitoring of 3D scanned manufactured parts through projections and spatiotemporal control charts. Journal of Intelligent Manufacturing 28, 899–911 (2017)
12. Stankus, S., Castillo-Villar, K.: An Improved multivariate generalised likelihood ratio control chart for the monitoring of point clouds from 3D laser scannersInternational Journal of Production Research 57, 1–12 (2018)
13. Filzmoser, P., Hron, K., Templ, M.: Applied Compositional Data Analysis. Springer (2018)
14. Menafoglio, A., Grasso, M., Secchi, P., Colosimo, B.: Profile Monitoring of Probability Density Functions via Simplicial Functional PCA With Application to Image Data, Technometrics 60:4, 497–510 (2018)
15. Scimone, R., Taormina, T., Colosimo, B.M., Grasso, M., Menafoglio, A., Secchi, P.: Statistical modelling and monitoring of geometrical deviations in complex shapes with application to Additive Manufacturing. Manuscript. (2020)

Tree attributes map by 3P sampling in a design-based framework

Mappa degli attributi forestali attraverso il campionamento 3P in un approccio basato sul disegno

Lorenzo Fattorini and Sara Franceschi

Abstract As mapping is becoming a burning issue in environmental and forest surveys, the estimation of marks for a finite population of points (e.g. trees) scattered onto a survey region when a 3P sampling is adopted is here considered. The estimation is approached in a design-based setting, consequently avoiding the massive modelling usually involved in model-based framework to handle with marked point processes. Since 3P sampling involves marks prediction for each population unit, prediction errors are here interpolated instead of marks in order to better meet design-based consistency requirements. The effectiveness of the proposed strategy is checked theoretically and by means of a simulation study performed on both artificial and real populations.

Abstract *La mappatura di attributi ecologici costituisce un argomento di notevole interesse nelle indagini di carattere ambientale e forestale. In questo lavoro è stato affrontato quindi il problema della stima degli attributi di una popolazione finita di punti (ad esempio alberi) collocati in un'area di studio qualora venga adottato un campionamento 3P. La procedura di stima è stata considerata in un approccio basato sul disegno, evitando dunque la modellistica solitamente utilizzata qualora si lavori in un contesto da modello. Al fine di soddisfare più facilmente le condizioni necessarie alla coerenza, sono stati interpolati gli errori di previsione piuttosto che l'attributo di interesse. Al fine di verificare empiricamente le proprietà teoriche, è stato effettuato uno studio di simulazione sia su popolazioni reali che artificiali.*

Key words: 3P sampling, marked point populations

[1] Lorenzo Fattorini, Università di Siena; email: Lorenzo.fattorini@unisi.it
Sara Franceschi, Università di Siena; email: franceschi2@unisi.it

1 Introduction

Wall-to-wall mapping of spatial pattern of ecological attributes is becoming a burning issue in forest surveys and inventories, since it represents a crucial information for environmental managing.

Maps of interest attributes are usually obtained in a model-based approach (see e.g. [5], [7], [8]) supposing that units are outcomes of a marked point process in the plane. However, in the last years the issue has been addressed in a design-based framework ([1], [3], [4]) where, contrary, marks are treated as fixed characteristics and properties of the resulting maps are determined by the probabilistic sampling scheme adopted to select units on which marks are recorded. The authors exploited the very simple Tobler's first law of geography as assisting model and accordingly, they adopted the so-called inverse distance weighting (IDW) interpolator in which the marks at un-sampled units are estimated by a weighted sum of the sampled marks with weights inversely decreasing with distances to the unit under estimation.

Nevertheless, as to concern natural finite populations of marked points such as trees or shrubs, maps are rarely possible owing to the prohibitive acquisition of units' locations. In most cases, natural populations are sampled without knowing their lists, by means of plots or transects located in the study area according to a suitable probabilistic design.

Probably, the unique relevant case in which the mapping of natural populations becomes possible is under the so called 3P sampling. Indeed, under 3P sampling all the units of the population are visited by a crew of experts in such a way that population list together with predictions of the survey variable for each population unit can be obtained. Then, units are independently selected with probabilities proportional to prediction (3P) as in the Poisson sampling. In this context, for the sampled units, both predictions and actual values are known, in such a way that also the prediction errors are known for them. Therefore, prediction errors can be interpolated by IDW, in such a way that the interpolated values are given by the predictions plus the interpolated errors. Fattorini et al. [8] pointed out that this strategy is likely to provide considerable improvement with respect to the direct interpolation of marks.

The goal is now to pursue this strategy and to check its effectiveness.

2 Notations and theoretical results

Let $\mathcal{A}$ be a survey region and suppose a discrete population U of N points $p_1,\ldots,p_N$ scattered onto $\mathcal{A}$. For brevity denote points by their indices $1,2,\ldots,N$. For each point $j\in\mathsf{U}$ let y_j be the amount of the survey variable Y, usually referred to as the mark of unit j. The interest is to reconstruct the marked population of points on the basis of the values recorded in a sample S of n points selected from U by means of 3P

sampling. Exploiting the Tobler law of geography in a model-assisted framework, Bruno et al. [1] propose to estimate the y_j s by means of the IDW interpolator

$$\hat{y}_j = Z_j y_j + (1 - Z_j) \sum_{i \in U} w_{ij} y_i$$

where Z_j is the random variable equal to 1 if $j \in S$ and 0, otherwise

$$w_{ij} = \frac{Z_i \phi(d_{ij})}{\sum_{l \in U} Z_l \phi(d_{lj})}$$

is the weight attached to the mark of unit *i* to estimate the mark of unit *j*, d_{ij} represents the Euclidean distance between points *j* and *h* and $\phi : [0, \infty) \to \Re^+$ is a not increasing function on $]0, \infty)$, with $\phi(0) = 0$ and $\lim_{d \to 0^+} \phi(d) = \infty$.

Fattorini et al. [8] derived conditions under which 3P sampling ensures design-based asymptotic unbiasedness and consistency of the IDW interpolator. Furthermore, they propose to interpolate predictions errors rather than marks. Indeed, for the sampled trees, both predictions, here indicated by y_j^0, and actual values y_j are known, in such a way that also the prediction errors $e_j = y_j - y_j^0$ are known for each sampled point. Therefore, the IDW interpolator can be used for interpolating the prediction errors for any population unit by means of

$$\hat{e}_j = Z_j e_j + (1 - Z_j) \sum_{i \in U} w_{ij} e_j$$

As to the mean squared error estimation, following [8], a simple estimator is

$$\hat{V}_j = (\hat{e}_j - e_{near(j)})^2$$

where *near*(*j*) is the label of the sampled unit nearest to point *j*.

In this way the interpolated marks are given by the predicted values plus the interpolated errors, that is

$$\hat{y}_j = y_j^0 + \hat{e}_j, \quad p_j \in U$$

It can be proven that smoothness assumptions for marks, necessary for ensuring design consistency, are more realistic when considering prediction errors. Indeed, jumps and irregularities of marks throughout $\mathcal{A}$ are presumably absorbed by similar jumps and irregularities in the corresponding predictions. As consequence, if predictions are good, prediction errors turn out more smoothed than marks.

Following [2] and [6], it is reasonable to assume that there exists a maximum error rate of predictions $r \in (0,1)$ that occurs at the extremes of marks, in such a way that small values near the lower bound *l* are over-evaluated and large values near the upper bound *L* are under-evaluated. If for simplicity we further assume that the predictions increase linearly with marks, then

$$y_j^0 = a + b y_j, \; p_j \in U$$

where

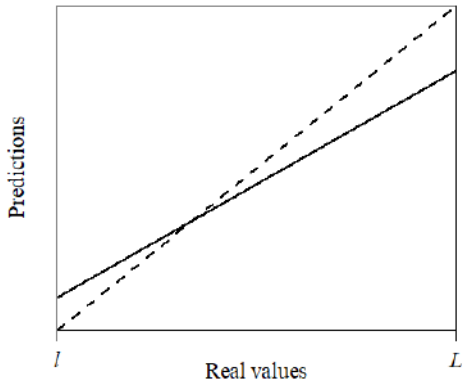
$$b = 1 - r\frac{L+l}{L-l}, \quad a = (1+r)l - bl$$

with $r < (L-l)/(L+l)$ in order to ensure a slope $b \in (0,1)$ in such a way that the prediction line increases less slowly than the line of perfect predictions (see Figure 1).

In such a situation it has been proven that, under some conditions, if prediction errors are interpolated instead of marks, the gain in precision is expected to be about

$$(1-b)^{-1} = \frac{L+l}{(L-l)r}.$$

Figure 1: The continuous line shows predictions as linear function of the true values with maximum error rates occurring at the extremes. The dotted line is the line of perfect predictions



3 Simulation study

A simulation study is performed on a set of artificial nested populations of points located in a unit square in accordance with different spatial patterns in order to check the gain provided by the error interpolation with respect to the direct interpolations of marks under predictions of different precision. The same comparison is performed on a real population of about 1400 trees within a stand of 9*ha* located in Monte Cimino (South Italy).

For any considered population, 10000 samples were selected according to 3P sampling and absolute relative bias, relative root mean squared error and absolute bias of the relative root mean squared error estimator were computed from the resulting Monte Carlo distributions of the density estimates and of the relative root mean squared error estimates.

Results, not reported here for sake of brevity, completely confirm the theoretical finding of section 2.

4 Conclusion

Theoretical and simulation results highlight the failure of IDW technique in the direct interpolation of marks under 3P sampling. This problem tends to disappear when prediction errors are interpolated instead of marks. In this case, under good predictions, prediction errors are much more smoothed than marks. Therefore, the interpolation gains a satisfactory level of accuracy and precision notwithstanding the unbalanced selection provided by 3P sampling.

In conclusion, in accordance with the results of this study, when the purpose is simply to produce a map of a mature population of trees without giving explanation about the process that dislocated trees and generated volumes, the complex task of modelling the marked point process and estimating its parameters can be avoided and we can opt to produce the map in a design-based framework adopting 3P sampling and the IDW interpolation of prediction errors: under reliable predictions, the resulting map has the potential of being a reliable picture of the actual map.

References

1. Bruno, F, Cocchi, D. and Vagheggini, A.: Finite population properties of individual predictors based on spatial pattern. Environ. Ecol. Stat. (2013) doi: 10.1007/s10651-012-0229-9
2. Corona P.M.: Metodi di inventariazione delle masse e degli incrementi legnosi in assestamento forestale, Rome: Aracne Editrice (2007)
3. Fattorini, L., Marcheselli, M. and Pratelli, L.: Design-based maps for finite populations of spatial units. J. AM. Stat. Assoc. (2018) doi: 10.1080/01621459.2016.1278174
4. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based mapping for finite populations of marked points. Electron. J. Stat. (2019) doi: 10.1214/19-EJS1572
5. Karr, A.F.: Inference for stationary random fields given Poisson samples (1986). Adv. Appl. Probab. doi: 10.2307/1427306
6. Kinnunen, J., Maltamo, M., Paivinen, R.: Standing volume estimates of forest in Russia: how accurate is the published data?. Forestry (2007) doi: 10.1093/forestry/cpl042
7. Lai Ping Ho and Stoyan, D.: Modelling marked point patterns by intensity-marked Cox processes. Stat. Probabil. Lett. (2008) doi: 10.1016/j.spl.2007.11.013
8. Mase, S.: The threshold method for estimating total rainfall. Ann. I. Stat. Math (1996) doi: 10.1007/BF00054785

Unsupervised classification of texture images by gray-level spatial dependence matrices and genetic algorithms

Classificazione di immagini attraverso matrici di dipendenza spaziale dei livelli di grigio e algoritmi genetici

Roberto Baragona[1] and Laura Bocci[2]

Abstract Recognition of objects and regions of interest in digital image processing often relies on texture classification. The source image is divided according to a rectangular grid to form textured regions each of which is characterized by some numerical significant measure called feature. A new approach is introduced that uses the gray-level spatial dependence matrices and the genetic clustering with unknown *K* algorithms to locate sets of homogeneous regions and enhance the discrimination amongst them. There is no need to select and compute complicated features transforms as the procedure is based on the optimal weighting of the simple basic features. A simulation experiment is performed using the well-known Brodatz textures to demonstrate that the new procedure is able to define well separated clusters according to the principle of strong internal cohesion and high inter-clusters separation.

Abstract *Nell'elaborazione di immagini digitali il riconoscimento di oggetti e di regioni di interesse si basa spesso sulla classificazione delle trame. L'immagine viene divisa secondo una griglia rettangolare individuando regioni caratterizzate da una specifica caratteristica. Allo scopo di individuare nell'immagine insiemi di regioni omogenee si presenta un approccio basato sulle matrici di dipendenza spaziale dei livelli di grigio e sugli algoritmi genetici di clustering con numero di gruppi K incognito. Secondo questo approccio non è necessario selezionare e calcolare complicate trasformazioni di funzioni poiché la procedura si basa sulla ponderazione ottimale delle semplici funzioni di base. Un esperimento di simulazione che utilizza le note trame di Brodatz dimostra che la nuova procedura è in grado di definire gruppi ben definiti secondo il principio di forte coesione interna ed elevata separazione tra i gruppi.*

Key words: texture classification, gray-level spatial dependence matrix, genetic clustering algorithms.

---

[1] Roberto Baragona, Department of Communication and Social Research, Sapienza University of Rome; email: roberto.baragona@uniroma1.it

[2] Laura Bocci, Department of Communication and Social Research, Sapienza University of Rome; email: laura.bocci@uniroma1.it

1 Introduction

A large amount of digital imagery data is available from uncountable sources within continuous processes. A thorough examination of even a single image by human visual inspection is often practically impossible and therefore discovering some characteristics of interest hidden in a noisy framework is challenging. Examples may be found in aerial photograph, photo micrograph, satellite and/or medical images.

In this paper we are concerned with the task of texture classification in a given 2-dimensional image, which consists in identifying and portraying, using a gray-scale (or a colormap), the features occurring in the image in terms of the object or type of land cover these features actually represent on the ground. Many problems in image processing may be handled conveniently by texture analysis. Recognition of objects and regions of interest in digital texture processing mainly relies on either supervised or unsupervised texture classification. Uniform textured regions are usually identified by defining and computing numerical significant measures of certain characteristics of the image. Relevant applications may be found in surface recognition and medical imagery framework. Dealing with textures often means that a 2-dimensional image is available that contains important information about the structural arrangement of surfaces and their relationship to the surrounding environment. From the visual image of a texture special characteristics of interest have to be extracted to draft a kind of 3-dimensional picture of reality. Methods for describing image textures (see e.g. Depeursinge et al., 2017) may be approximately divided in geometrical, model based, signal processing and statistical.

In this paper we concentrate on the statistical method based on Gray-Level Spatial Dependence Matrix (GLSDM) (Haralick et al., 1973). Spatial statistics are commonly used within this approach. Such texture features summarize the spatial gray level co-occurrence in an image. Actually the image is an array of gray levels, one gray level for each and every pixel. Each entry of the associated GLSDM is the frequency of the pair of gray levels that distance apart and oriented according to that angle. Several image texture properties may be estimated from spatial dependence matrices and homogeneous regions may be identified and classified accordingly.

Our contribution for improvement mainly resides in using two devices: 1) the discrimination enhancement problem as stated by Zizzari (2003) is solved by using genetic algorithms instead of numerical optimization; specially for large equation systems numerical algorithms may not converge to the optimal solution; 2) the genetic clustering for unknown K (Bandyopadhyay and Maulik, 2002, GCUK) is used to overcome the limitation to only 3 clusters in the early formulation of the discrimination enhancement problem.

The plan of the paper is as follows. In the next Section 2 the feature extraction from GLSDM procedure is summarized and the discrimination enhancement problem is stated formally. In Section 3 the genetic algorithms solution to the discrimination enhancement problem is presented. A Monte Carlo simulation experiment is described in Section 4 which aims at demonstrating the capability of the genetic algorithm-based procedure in the recovering of textures in an image.

2 The discrimination enhancement problem

An image in its digital form is usually recorded as an $L_x \times L_y$ matrix $\mathbf{R}$ of resolution cells, or pixels. Any of the elements r_{pq} (p = 1,…, N_x, q=1,…, N_y) of the matrix $\mathbf{R}$ is a pixel with gray tone $g \in G$, $G = \{1, \dots, N_g\}$.

The Gray-Level Spatial Dependence Matrix (GLSDM) $\mathbf{P}_{(d,\theta)}$ related to the matrix $\mathbf{R}$ and for displacement (d, θ) as defined in Haralick et al. (1973) is a square symmetric matrix, with row and column dimensions equal to the number of discrete gray levels (intensities) N_g in the image being examined, for each distance d and orientation θ. The generic element of $\mathbf{P}_{(d,\theta)}$ in row i (gray tone i) and column j (gray tone j) is the frequency of pairs (r_{pq}, r_{st}) of pixels distance d far apart along direction θ and such that $r_{pq} = i$ and $r_{st} = j$. The dependence spatial matrix reveals certain properties about the spatial distribution of the gray levels in the texture.

Considering that all texture information is contained in $\mathbf{P}_{(d,\theta)}$ related to the image $\mathbf{R}$, a set of textural features can be extracted from GLSDM. These features may be defined as possibly non-linear real functions of the elements $p(d, \theta)_{ij}$. Though any feature contains information about the textural characteristics of the image, the interpretation may not be as easy to perform. One may be often left in doubt what characteristic or property it is exactly concerned with. Moreover, the result is strongly affected by the ability of the feature functions to map the underlying structures of an image in order to identify objects or regions of interest in the image itself. This task is strictly dependent on the discriminatory power of the chosen feature functions, in the sense that the functions have to assign very different values to different types of image blocks.

Let $\mathbf{W} = [w_{ij}]$ $(i, j = 1, \dots, N_g)$ be the $N_g \times N_g$ matrix of weights associated to the GLSDM $\mathbf{P}_{(d,\theta)}$, we introduce the features functional form as follows

$$f(\mathbf{W}) = \mathbf{1}'_{N_g}\left(\mathbf{W} \odot \mathbf{P}_{(d,\theta)}\right)\mathbf{1}_{N_g}\,, \tag{1}$$

where $\mathbf{1}_{N_g}$ denotes the column vector with N_g ones and $\odot$ is the Hadamard product (i.e. the entry-wise product). The discrimination power of the feature f in (1) will depend directly on how well each of the individual elements w_{ij} $(i, j = 1, \dots, N_g)$ discriminate between spatial gray tone related to clusters of pixels. The purpose is to find the weights w_{ij} associated to each element of $\mathbf{P}_{(d,\theta)}$ that are large for those elements which produce good discrimination, while are small (ideally zero) for the elements giving poor discrimination.

Often a real problem consists in distinguishing textures within the same textural image. Let $\mathbf{R}$ be divided in n sub-matrices (blocks) $\{\mathbf{R}_1, \dots, \mathbf{R}_h, \dots, \mathbf{R}_n\}$ each of dimension $L_x^h \times L_y^h$ (h = 1, . . . , n). For each and every block $\mathbf{R}_h$ (h = 1,…, n) a gray-level spatial dependence matrix $\mathbf{P}_h$ for given (d, θ) may be computed and features f_h evaluated as in (1)

$$f_h(\mathbf{W}) = \mathbf{1}'_{N_g}(\mathbf{W} \odot \mathbf{P}_h)\mathbf{1}_{N_g}\,, \qquad (h = 1,\dots, n). \tag{2}$$

Each one of the n blocks $\mathbf{R}_h$ may be assigned to one of K clusters $C = \{C_1, \dots, C_k, \dots, C_K\}$ using the block feature. Let T_k be the cardinality of the discrete set C_k, the $n = \sum_{k=1}^{K} T_k$.

The objective is the improvement of such a discriminatory power of textural features by identifying an optimal set of weights **W** which solves the problem of maximizing this level of separation among clusters of pixels. The criterion is to enhance the separation amongst feature values that may be associated to different classes of pixels and simultaneously improve internal cohesion. The optimization problem that allows the set of weights **W** to be computed is called *discrimination enhancement problem* (Zizzari, 2003). This way the functional form itself is determined by the data and is not imposed arbitrarily.

The requirement of greatest separation amongst clusters C_k (k = 1,…, K) may be conveniently achieved by maximizing the total *inter - cluster separation measure*

$$TISM(\mathbf{W}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{v=1}^{K}\left(\bar{f}(\mathbf{W})^{(k)} - \bar{f}(\mathbf{W})^{(v)}\right)^2$$

where $\bar{f}(\mathbf{W})^{(k)} = \frac{1}{T_k}\sum_{h\in C_k} f_h(\mathbf{W})$ (k = 1,…, K), f_h defined by Equation (2).

For maximizing internal cohesion the total *within - cluster scatter measure*

$$TWSM(\mathbf{W}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{h\in C_k}\left(f_h(\mathbf{W}) - \bar{f}(\mathbf{W})^{(k)}\right)^2$$

The discrimination enhancement problem consists in searching for optimal weighting parameters **W** so that $(TWSM - TISM)$ is minimized. Note that the features themselves depend on the weights and their values are involved as well within the optimization procedure, so usual optimization may not be used directly.

The number of clusters K is either known or may be assumed known from some preliminary analysis, but it seems preferable that the number of clusters be included as an additional parameter within the optimization procedure. A measure of cluster validity which is in close agreement with the principles of the discrimination enhancement problem is the Davies - Bouldin index (Bezdek and Pal, 1998).

3 A Genetic algorithm for texture analysis

The optimization problem described in Section 2 may be decomposed in two steps. The first step is concerned with the estimation of the weighting parameters. The second step regards the allocation of the blocks to clusters. The genetic clustering with unknown K (GCUK) algorithm (Bandyopadhyay and Maulik, 2002) is adapted to implement both steps sequentially in each iteration. The GCUK algorithm is a natural choice, as it minimizes both the chromosome length and the execution time, and uses as optimality criterion a transform of the Davies-Bouldin index.

A GCUK algorithm is a population based meta-heuristic that starts from a set of tentative solutions, called chromosomes, to the optimization problem and proceed for a pre-specified number of iterations in each of which the set of tentative solutions in the current population are updated to produce a new set of improved tentative solutions that replace the current one. Each iteration is composed of three steps usually called selection, crossover and mutation.

A crucial point in the definition of a genetic algorithm is the choice of the encoding of a tentative solution. In this framework, any tentative solution to the discrimination enhancement problem is coded as a chromosome of length $N_g^2 + K_{\max}$. The chromosome has two parts: 1) a sequence of N_g^2 real numbers that are assumed as the weights $\{w(i,j)\}$ $(i,j = 1,\dots,N_g)$; 2) a sequence of length $K_{\max}$ which may include either real numbers or # symbols, the former are a centre of a non-empty cluster, the latter represents an empty cluster; the number of non-empty clusters K is required to take value in the interval $[K_{\min}, K_{\max}]$.

A disadvantage in operating with the whole chromosome resides in that the chromosome itself may be too large for the genetic algorithms to be able to output a reliable solution in a reasonable time. In addition, the two parts of the chromosome may happen to be unbalanced, i.e. one part may be larger than the other in such a way that the genetic algorithm works essentially on the longer part and neglects the shorter. Then we may run the procedure in two steps: 1) weights estimation: the second part of the chromosomes is assumed constant, and chromosomes are evolved by improving their fitness function only using the first part of the chromosome; 2) allocation of the blocks to clusters: changes are allowed only in the second part of the chromosome. In this second step the optimal allocation of blocks to clusters takes place by using the weights obtained in the first step.

Should the number of weights be excessively large we could resort to maximum likelihood estimation of the weights assuming normality and using Monte Carlo methods. In this case only step 2 would be executed achieving a significant reduction of computations.

4 A Monte Carlo experiment

To check the effectiveness of the genetic algorithm-based procedure for unsupervised classification of textures some Monte Carlo simulations have been performed. We downloaded all 111 Brodatz textures from D1.gif to D112.gif (texture D14.gif is missing). These are 75×75 pixels images. In the present simulation experiment 49 textures selected at random were arranged in a 7×7 grid to assemble a 525×525 pixels mosaic. The random selection has been replicated 1000 times to yield 1000 texture mosaics. Three levels of noise have been considered, 1%, 5% and 10%, that is pixels in each image have been altered at random according to the given percentage. For each texture mosaic the genetic algorithm-based clustering procedure has been used to recover the right classification of the original textures.

For the computation of the gray-level spatial dependence matrices and the features extraction the Matlab Image Processing Toolbox has been used, available in the version R2017b or higher.

Results from the complete Monte Carlo experiment are reported in Table 1 where the well-known corrected Rand index (CRI) and number of cluster bias (difference between actual and estimated number of clusters) are averaged across 1000 replications. The standard errors are enclosed in parentheses. The genetic algorithm parameters have been: population size 10; number of generations 20; crossover probability 0.7; mutation probability 0.01; interval where the number of clusters has to be searched ($K_{\min} = 2$, $K_{\max} = 10$). Ng = 16 gray levels have been used (bit depth = 4 bits) and features have been computed averaged across angles $\theta = (0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3}{4}\pi)$ for each of three distances $d = (1, 2, 3)$. As a matter of fact, angular dependencies should be invariant under rotations. This condition may be fulfilled by using some suited functions of the angularly dependent features (Haralick et al., 1973, p. 615), for instance their average.

Table 1: Assessment of the two steps genetic algorithm-based procedure for texture analysis: Corrected Rand index and number of cluster bias, averaged across 1000 replications of Monte Carlo simulation experiment. The standard errors are enclosed in parentheses.

Level of noise	*Corrected Rand index*	*Number of cluster bias*
1%	0.8451 (0.1480)	- 0.5420 (2.0298)
5%	0.8247 (0.1495)	- 0.2950 (2.0600)
10%	0.8099 (0.1556)	- 0.2090 (2.0453)

According to the CRI values shown in Table 1 the genetic algorithm-based procedure seems able to recognize the different textures. It has to be expected that, increasing the noise level, the agreement between the actual and estimated partitions deteriorates while the variability of the estimates increases. Nevertheless, the figures reported in Table 1 suggest that the loss of precision is rather small. The bias of the estimates of the number of clusters is always negative, that is the procedure estimates in general more clusters than the actual ones. However, the standard errors of the estimates are large as maybe often a cluster is split into two similar clusters.

References

1. Bandyopadhyay, S., Maulik, U.: Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recognition, 35, 1197 – 1208 (2002)
2. Bezdek, J.C., Pal, N.R.: Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, 28, 301 – 315 (1998)
3. Depeursinge, A., Al-Kadi, O.S., Mitchell, J.R.: Biomedical Texture Analysis. Academic Press, London (2017)
4. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3, 610 – 621 (1973)
5. Zizzari, A.: Methods on Tumor Recognition and Planning Target Prediction for the Radiotherapy of Cancer. Otto von Guericke University Magdeburg, PhD dissertation (2003)

Finance, business and official statistics

A discrete choice approach to analyze contractual attributes in the durum wheat sector in Italy

Un approccio a scelte discrete per l'analisi degli attributi contrattuali nel settore del grano duro in Italia

Stefano Ciliberti[1] and Simone Del Sarto[1] and Giulia Pastorelli[1] and Angelo Frascarelli[1] and Gaetano Martino[1]

Abstract This study investigates the preferences of durum wheat farmers as regards the characteristics of a contract, considering this tool as a mechanism for transferring decisions and property rights. To this aim, a sample of durum wheat Italian farmers is interviewed for expressing their preferences about some contractual features, such as price, production and quality rules and sustainable environmental techniques. Using a discrete choice analysis through a multinomial logistic model, results shows how contracts, in addition to price stabilization, are relevant for their impact on technology, quality, and the environment.

Abstract *Questo lavoro indaga le preferenze degli agricoltori di grano duro circa i termini contrattuali di un contratto, considerato come meccanismo di trasferimento delle decisioni e dei diritti di proprietà. A tale scopo, è stato intervistato un campione di agricoltori italiani di grano duro, i quali hanno espresso le loro preferenze inerentemente al prezzo, regole di produzione e di qualità e tecniche ambientali sostenibili. I risultati ottenuti mediante modelli a scelta discreta mostrano come i contratti, oltre alla stabilizzazione dei prezzi, siano rilevanti per il loro impatto sulla tecnologia, sulla qualità e sull'ambiente.*

Key words: Discrete choice models, contracts, contractual terms, uncertainty

1 Introduction

Contracts between farmers and agri-food processors have received increasing attention in different theories and by policy-makers. Recent developments in Transaction Cost Economics have shed light on the possibility of analyzing the contracts in terms of the allocation of rights between transacting parties [2]. This new perspective allows for a model of contract analysis to be developed which, on one hand, connects

[1] Department of Agricultural, Food and Environmental Sciences, University of Perugia

the contract with the organization and provides a detailed perspective and, on the other, highlights the transfer of rights between the parties and the role this plays in organizing the transaction. The three basic components (the transaction, the property and decision rights, and the contract) are analytically interconnected in the golden triangle and provide key "tools to investigate the institutional dimensions framing the economies" [3]. This approach conceives the contract as the way to organize the transfer of rights between the parties to a transaction [2].

Moreover, the theory states that the choice of governance structure is made by aligning the attributes of the transaction (asset specificity, uncertainty, and frequency) with the attributes of the governance structures [10]. Transaction attributes determine the costs associated with each organizational form and agents seek to choose an efficient organization to minimize these costs. According to empirical studies on contract farming [1, 5], uncertainty plays the major role in this background of contractual choice. From a general standpoint, uncertainty incurs higher governance costs when coordination is decentralized. As a consequence, an agent would exclude a contractual term/condition when faced with a high level of uncertainty. By keeping the decision rights in his/her own hands and out of the contract, the costs of adaptation are reduced compared to the case of a low level of uncertainty in which the decision rights are allocated in the contract. In this regard, the concept of uncertainty can be split into three subcategories: market uncertainty ("unpredictability of demand and inability to accurately forecast and schedule production" [4]), technological uncertainty (acceleration of technological changes that increase the rate of obsolescence in assets [7]) and behavior uncertainty ("the difficulty of observing and measuring the adherence of the transacting parties to the contractual arrangements and the difficulty of measuring the performance of these parties" [6]).

Against this backdrop, the present paper aims at testing the main hypothesis that the transaction attribute of uncertainty affects the allocation of decision rights within a contract. In more details, the contractual attributes related to the three types of uncertainty are: price (market uncertainty), rules of production, quality levels and sustainable cultivation techniques (technological uncertainty), and forms of payment and technical assistance (behavioral uncertainty).

The durum wheat sector in Italy certainly represents an interesting case study for testing the above hypothesis. In fact, durum wheat is usually seen as a commodity market and this context enables us to put less emphasis on asset specificity and more focus on the uncertainty attribute. Moreover, despite contracts have not been traditional largely used in this sector as a means of decision coordination, they have gained momentum even in order to address the different type of uncertainties that have negatively affected the stability of exchanges between farmers and processing companies.

To this aim, we analyze data coming from questionnaires administrated to farmers. They have to choose among several contracts, specified in terms of various attribute levels (contractual terms) that reflect the three uncertainties just mentioned. In particular, a choice-based conjoint analysis is performed by means of an extended multinomial logit model [9].

This article is organized as follows. Section 2 describes the way we obtained data and the statistical method employed for our analysis. Results are shown in Section 3, while some concluding remarks are given in Section 4.

2 Material and methods

We use data from questionnaires administrated during the period October 2019-February 2020 to farmers from several regions that represent the main areal of production of durum wheat in Italy. In these questionnaires, different potential contracts are proposed, characterized by six attributes (each one with three levels), reflecting different types of uncertainties (Table 1).

For investigating the farmers' preferences as regards contractual terms, a discrete choice experiment is carried out. A full factorial design is obtained by combining the levels of all the attributes. Hence, $3^6 = 729$ different contracts may arise, but this amount is too hard to handle in a single experiment. For this reason, a fractional factorial design is created and contracts are randomly distributed into 18 choice sets, each of them with three possible contracts (alternatives) plus the no-choice option, that is, "none of the previous contracts". Ultimately, the farmer randomly faces three out of these 18 choice sets and, for each one, the choice is among four alternatives: either one of the three proposed contracts or the no-choice option.

Let us suppose that decision-maker $i = 1,\dots,n$ has to choose among J alternatives (i.e., contract profiles in our case). According to the random utility modeling framework, he/she will choose the alternative that maximizes his/her utility, U_{ij}, with $j = 1,\dots,J$. Then, the probability of choosing alternative l is

$$P_{il} = Prob(U_{il} > U_{ij}, j \neq l). \tag{1}$$

Decision-maker's utility can be decomposed as $U_{ij} = V_{ij} + \varepsilon_{ij}$, where V_{ij} is function of known covariates (alternative and/or decision-maker specific), while ε_{ij} represents the impact of all the unobserved factors (not included in V_{ij}) on the utility of selecting a particular alternative. In this work, we assume that V_{ij} contains only alternative specific variables (i.e., alternative attributes) and can be specified linearly, i.e., $V_{ij} = \boldsymbol{x}_j^\top \boldsymbol{\beta}$, where $\boldsymbol{x}_j$ is a vector of p attributes of alternative j and $\boldsymbol{\beta}$ is the p-dimensional parameter vector. Moreover, we suppose that the random component is independently, identically extreme value distributed. Hence, we can derive the multinomial logit choice probabilities as follows [8]:

$$P_{il} = \frac{\exp(\boldsymbol{x}_l^\top \boldsymbol{\beta})}{\sum_{j=1}^{J} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}. \tag{2}$$

Model parameters $\boldsymbol{\beta}$ are estimated through maximum likelihood methods (see [8] for further details). Finally, in this work we consider the extended no-choice multi-

Table 1 Contractual attributes, levels and choice relative frequency (%). Note: DSS stands for decision support system.

Uncertainty	Attribute	Level	Label	%
Market	Price	Guaranteed minimum price	P_100_GMP	36.3
		100% open price (based on the average of the commodity exchange)	P_100_open	17.5
		Mixed (50% open - 50% fixed) price based on production costs	P_50_50	25.8
Technological	Technique	Freely decided by the producers	T_free	29.1
		Agreed with industry	T_coll	29.8
		Imposed by industry	T_imp	20.6
	Quality threshold	Proteins content $> 12.5\%$	Th_12.5	29.8
		Proteins content $> 13.5\%$	Th_13.5	27.2
		Proteins content $> 14.5\%$	Th_14.5	22.5
	Sustainable cultivation techniques	Optimized nitrogen application	SCT_Frac_N	26.5
		Conservation Agriculture	SCT_Ca	25.7
		Both	SCT_both	27.4
Behavioral	Time of payment	100% in September	TP_100	28.3
		50% in September, 50% in March	TP_50_50	26.9
		Monthly payments	TP_mon	24.4
	Technical assistance	No	TA_no	19.5
		Yes	TA_yes	30.0
		Yes, through a DSS	TA_dss	30.0

nomial logit (ENCMNL) [9], where an extra dummy variable is added to $\boldsymbol{x}_j$ for taking into account the no-choice option.

3 Results

In this section we report the results obtained on a final sample of 191 questionnaires, related to farmers who completely fulfilled the questionnaire (i.e., without missing values in the choices). As each farmer has to express three choices (three choice sets), a total of 573 choices are available, of which 117 (20.4%) consist in the no-choice option. The remaining choices are summarized in Table 1, in which the raw choice relative frequency is shown for each attribute level. As we can see, most of farmers chooses contracts with a guaranteed minimum price (36.3%), more than twice with respect to those selecting contracts with 100% open price. Moreover, as far as the techniques of sustainable cultivation are concerned, the choices are equally distributed over the three possible techniques.

Table 2 shows the results of the ENCMNL model estimated on the data at issue. Estimates for price attribute reveal that the presence of a 100% opened price in a

Table 2 Parameter estimates of the extended no-choice multinomial logit: no significant estimates are reported in *Italic* style.

Parameter	**Estimate**	***p*-value**
P_100_open vs. P_100_GMP	-0.690	< 0.001
P_50_50 vs. P_100_GMP	-0.430	< 0.001
T_coll vs. T_free	*0.089*	*0.433*
T_imp vs. T_free	-0.367	0.004
Th_13.5 vs. Th_12.5	*-0.189*	*0.123*
Th_14.5 vs. Th_12.5	-0.400	0.001
SCT_Ca vs. SCT_Frac_N	*-0.011*	*0.928*
SCT_both vs. SCT_Frac_N	*-0.043*	*0.719*
TP_50_50 vs. TP_100	*0.066*	*0.573*
TP_mon vs. TP_100	*-0.171*	*0.188*
TA_yes vs. TA_no	0.376	0.003
TA_dss vs. TA_no	0.379	0.003
No choice constant	-0.638	0.002

contract negatively affects (-0.690 on the logit scale) the probability of choosing a contract with respect to a contract having the guaranteed minimum price, other things being equal. The same also applies for a mixed price, even if the magnitude is lower. This result shows that farmers rely on contracts in order to reduce market uncertainty, thanks to a guaranteed minimum price that – compared to open price and mixed price – is not influenced by market price volatility.

As regards the technique attribute, results show that the presence of technical requirements imposed by the industry significantly hinders the probability of choosing a contract, with reference to contracts where farmers are free to decide. Therefore, what emerges is that durum wheat producers prefer to maintain decisional rights on technique in their hands, in order to autonomously address technological uncertainty.

Furthermore, the quality threshold attribute significantly influences the probability of choosing a contract only as regards the highest content of protein ($>$14.5%): in particular, a higher quality requested by the contract discourages farmers from a potential choice.

Lastly, the presence of service of technical assistance impacts on the probability of adopting a contract with reference to the no technical assistance option. Such a result confirms that farmers are willing to share their decision rights with the counterpart to share technological uncertainty, with a direct positive impact on behavioral uncertainty as well.

4 Conclusions

The main purpose of the present paper is to investigate the hypothesis that the transaction attribute of uncertainty affects the allocation of decision rights within a contract. To this aim, a choice-based conjoint analysis is performed through a multinomial logistic model in order to investigate how several contractual terms (considered as proxies of different types of uncertainty) influence the potential contract choice by durum wheat farmers. Main findings reveal that a guaranteed minimum price is preferred by farmers so as to decrease market uncertainty. Moreover, farmers are discouraged from a higher product quality requested by the contract, but are potentially interested in a contract with technical assistance, in order to share technological uncertainty with the counterpart.

References

1. Martino, G., Polinori, P.: An analysis of the farmers contractual preferences in process innovation implementation. Br. Food J. **121**, 426–440 (2019)
2. Ménard, C.: Organization and governance in the agrifood sector: How can we capture their variety?. Agribusiness **34**, 142–160 (2018a)
3. Ménard, C.: Research frontiers of new institutional economics. RAUSP Manage. J. **53**, 3–10 (2018b)
4. Parmigiani, A.: Why do firms both make and buy? An investigation of concurrent sourcing of complementary components. Strateg. Manage. J. **34**, 1145–1161 (2007)
5. Qin, P., Carlsson, F., Xu, J.: Forest tenure reform in China: a choice experiment on farmers' property rights preferences. Land Econ., **87**, 473–487 (2011)
6. Robertson, T.S., Gatignon, H.: Technology development mode: A transaction cost conceptualization. Strateg. Manage. J. **19**, 515–531 (1998)
7. Schnaider, P.S.B., Ménard, C, Saes, M.S.M.: Heterogeneity of plural forms: A revised transaction cost approach. Manage. Decis. Econ. **39**, 652–663 (2018)
8. Train, K.E.: Discrete choice methods with simulation. Cambridge University Press, Cambridge (2009)
9. Vermeulen, B., Goos, P., Vandebroek, M.: Models and optimal designs for conjoint choice experiments including a no-choice option. Int. J. Res. Mark. **25**, 94–103 (2008)
10. Williamson, O.E.: The economic institutions of capitalism. The Free Press, New York (1985)

A fuzzy approach to the measurement of the employment rate

Un approccio sfocato alla misura del tasso di occupazione

Bruno Cheli, Alessandra Coli, Andrea Regoli

Abstract The computation of employment and unemployment rates is conditioned by the rigid classification of labour force into the mutually exclusive groups of people in employment and people not in employment. This implies a relevant loss of statistical information and wipes out all the nuances that exist between fully employed people and those ones who work only occasionally, while needing to work more. In our view employment should not be considered as a simple attribute that is present or absent but rather as a matter of degree. The aim of this paper is to define and compute fuzzy measures of employment by considering both the number of hours actually worked and wished.

Abstract *Il calcolo dei tassi di occupazione e di disoccupazione è condizionato dalla rigida ripartizione delle forze di lavoro nei due gruppi mutuamente esclusivi degli occupati e dei non occupati. Ciò conduce ad una rilevante perdita di informazione statistica e non consente di cogliere le sfumature che esistono tra la condizione di occupato a tempo pieno e quella di chi, pur risultando occupato, lavora meno ore rispetto a quanto desiderato. In realtà, l'occupazione non dovrebbe essere considerata un semplice attributo di cui rilevare la presenza o l'assenza ma piuttosto una grandezza presente con livelli di intensità diversi. Lo scopo di questo articolo è quello di definire e calcolare misure sfocate di occupazione, tenendo conto sia del numero di ore effettivamente lavorate che della quantità desiderata di ore di lavoro.*

Bruno Cheli
Department of Economics and Management, University of Pisa,Via Cosimo Ridolfi 10, 56124 Pisa, Italy e-mail: bruno.cheli@unipi.it

Alessandra Coli
Department of Economics and Management, University of Pisa, Via Cosimo Ridolfi 10, 56124 Pisa, Italy e-mail: alessandra.coli1@unipi.it

Andrea Regoli
Department of Management and Quantitative Studies, University of Naples Parthenope, Via Generale Parisi 13, 80132 Naples, Italy e-mail: andrea.regoli@uniparthenope.it

Key words: employment rate, fuzzy approach, labour force survey

1 Introduction

The analysis of the labor force is traditionally based on the clear contrast between the two complementary sets of employed and unemployed. This classification, however, appears too rigid since it neglects all the nuances that exist between those who work full-time and those who, on the other hand, work only occasionally even if they wish to work more. Furthermore, this way of proceeding involves a significant loss of statistical information which could instead be used for a more accurate representation and measurement of the phenomenon considered. For this purpose, it is necessary to stop treating employment and unemployment as simple binary attributes that manifest themselves in terms of presence or absence and instead treat them as quantitative variables that can assume different degrees or intensities. On a logical level, this implies the passage from a boolean conception to a "fuzzy" one. The fuzzy set theory of (Zadeh, 1965) has been widely applied in the most varied research fields. One of the areas in which it has been most successful is that of poverty analysis (Cerioli and Zani, 1990; Cheli and Lemmi, 1995; Betti et al., 2006). Apparently, however, no application has yet been registered for the analysis of the labor force. The aim of this paper is to define fuzzy measures of employment using available information on the number of hours weekly worked and about the fact that part-time workers may not be satisfied of their condition and wish to work more. In a future development of this research, we intend to extend the same approach to the measurement of unemployment.

2 Employment and unemployment rates in official statistics

According to the International Labour Organization directives (ILO 1982), working age population (persons aged 15 and older) can be split into three mutually exclusive groups: employed, unemployed and inactive population. The way in which these groups are identified affects the value assumed by unemployment and employment rates. In fact, the unemployed rate is defined as the proportion of unemployed persons in the total labour force population whereas the employment rate expresses the number of employed persons as a percentage of the working age population. Employed, unemployed and inactive population groups are identified on the basis of information collected by labour force surveys (e.g. the European Union Labour Force Survey) and following a hierarchical process, which first identifies the employed persons, then the unemployed ones and eventually the inactive population.

In line with the ILO guidelines, an employed person is a person aged 15 and over who meets one of the following conditions: i) during the reference week he/she performed some work - even if for just one hour - for pay, profit or family gain; ii)

he/she was not at work but had a job or business from which he/she was temporarily absent.

The unemployed population comprises all the persons in working age who: i) were without work during the reference period, ii) are currently available for work, and have taken concret actions in a specified recent period to seek paid employment or self-employment.

Finally, the inactive population includes persons in working age who are neither employed, nor unemployed.

3 The fuzzy measure of employment

The fuzzy measure of employment intends to overcome the sharp distinction between persons in employment and persons not in employment. The application of the fuzzy set theory (Zadeh, 1965) to the concept of employment rejects the definition of employment as a binary status identified by the one-hour criterion. This concept would include among the employed every individual who performed some work during the reference week, regardless of the number of hours worked. Instead, we define a membership function in the fuzzy subset of the employed, which is measured on a scale from 0 to 1, whereby 1 means full membership to the set of the employed persons and 0 full non-membership.

A fuzzy measure of employment is able to reflect the unmet need for working additional hours among the employed, thus accounting for the labour under-utilization. The concept of labour under-utilization encompasses both time-related underemployment and involuntary part-time. These groups include persons who share some characteristics with the unemployed though they are officially included among the employed. In the fuzzy approach that we propose, we treat them as employed to a certain degree: the lesser the time they work, the lower their degree of employment.

4 Empirical application on Italian data

The fuzzy approach for measuring employment has been implemented on the 2018 yearly Italian data of the EU Labour Force Survey (Eurostat, 2019).

People who are not employed (i.e. unemployed and inactive population) are assigned a membership function equal to 0.

Employed people (according to ILO working status), are assigned a membership value that ranges from 0 (excluded) to 1. Fig. 1 sketches the assumptions underlying the membership function for persons in employment and shows some results.

Membership function is set equal to 1 in case of: i) full-time workers working no less than a specified threshold (defined as we explain below) who represent the 67.6% of the employed population, ii) full-time workers who do not wish to work more, even if employed for less than the specified threshold (13.4%), iii) volun-

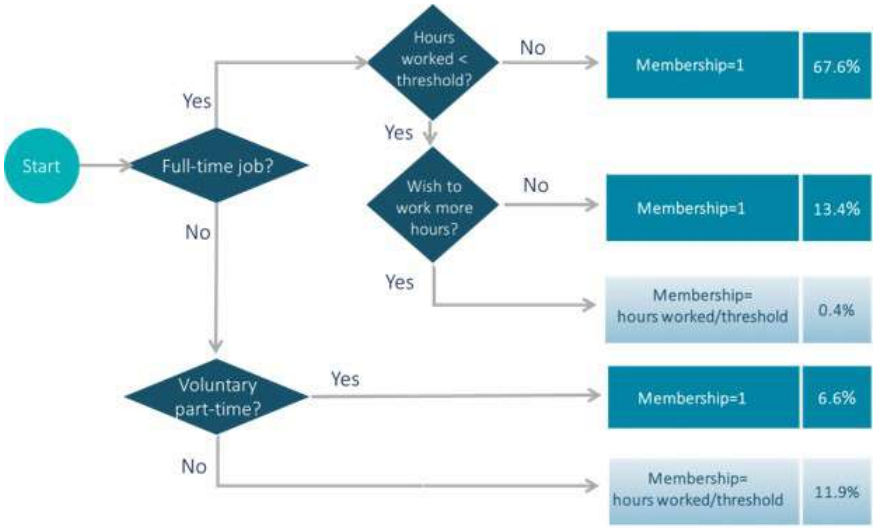


Fig. 1 Membership to the employment set

tary part-time workers (6.6%). The threshold acts as a limit set by statutory or collectively agreed standards and corresponds approximately to the number of hours worked on average by full-time workers. In particular it has been set equal to 40 weekly hours for workers of the private sector (employees, self-employed and family workers) and to 36 weekly hours for employees of the public sector. We have verified that these figures represent the modes of the distribution of the hours worked for the two groups. For both of them, the mode is equal to the median value, which supports the choice of the mode as a suitable measure of central tendency.

In the middle, with membership values greater than 0 and lower than 1, we find the underemployed workers. The largest part of them (11.9%) is composed by involuntary part-time workers, i.e. workers who could not find a full-time job, because of aggregate demand problems (linked with the lack of full-time jobs) or supply side problems (due for example to the lack of qualifications for the available full-time jobs). The remaining part consists of full-time workers who have worked less than the specified threshold and are willing to work additional hours. Their membership function has been defined as the ratio of the hours actually worked to the threshold value.

Finally, the fuzzy measure of employment results from the weighted mean of the individual membership values. As such, it can be compared with the traditional employment-to-population ratio (Table 1). The results show that, with reference to the total population between 15 and 64 years, the fuzzy rate is 55.2%, smaller by 5.6% compared to the official rate. The downward correction is more marked among females (9.3%) compared to males (2.8%), which proves that the unmet need for working more hours is by far stronger among the female workers compared to the male ones. Young adults (aged less than 35 years) and workers with a low educa-

tion level emerge as the groups of workers who are the most severely penalized by labour under-utilization. With reference to the geographical area of residence, the downward correction of the fuzzy measure is slightly above the average in the Southern regions and in the Islands and slightly below the average in the Northern regions.

Table 1 Employment-to-population ratio (15-64 years)

	Traditional measure (T)	Fuzzy measure (F)	F/T(%)
Total	58.5	55.2	94.4
Male	67.6	65.7	97.2
Female	49.5	44.9	90.7
15-24 years	17.7	15.8	89.3
25-34 years	61.7	57.3	92.9
35-44 years	73.4	69.4	94.6
45-54 years	72.3	68.6	94.9
55-64 years	53.7	51.4	95.7
Lower secondary education	43.8	40.7	89.3
Upper secondary education	64.3	60.9	92.9
Third secondary education	78.7	75.4	95.8
North-West	66.8	63.5	95.1
North-East	68.1	65.0	95.4
Centre	63.2	59.0	87.0
South	44.9	41.9	93.3
Islands	43.7	40.3	92.2

References

1. Betti G., Cheli B., Lemmi A., Verma V. (2006) On the construction of fuzzy measures for the analysis of poverty and social exclusion, Statistica & Applicazioni, Vol. IV, numero speciale 1, pp. 77-97.
2. Cerioli A., Zani S. (1990) A Fuzzy Approach to the Measurement of Poverty. In: Dagum C., Zenga M. (eds) Income and Wealth Distribution, Inequality and Poverty, Studies in Contemporary Economics, Springer-Verlag, Berlin, pp. 272-284.
3. Cheli B., Lemmi A. (1995) A "Totally" Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. Economic Notes 24: 115-134.
4. Eurostat (2019) EU Labour Force Survey Database User Guide, Version: November 2019, Luxembourg.
5. ILO (1982) Statistics of Labour Force, Employment, Unemployment and Underemployment, Report II of the 13th International Conference of Labour Statisticians, Geneva.
6. Zadeh L.A. (1965) Fuzzy sets. Information and Control 8: 338-353.

A proposal to model credit risk contagion using network count-based models

Un approccio network basato su dati di conteggio per modellizzare il rischio di contagio nel credito

Arianna Agosto and Daniel Felix Ahelegbey

Abstract Interconnectedness between institutions and between economic sectors is known to be the main source of systemic risk, and was recognised as a trigger of the great financial crisis in 2008-2009. In this paper we study contagion effects between corporate sectors using a financial network model where the significant links are identified through conditional independence testing.
While the existing financial network literature is mostly focused on Gaussian processes, our approach is based on discrete data. We indeed assess pairwise dependence of default counts in different economic sectors, finding several contagion channels in both the conditional mean and the shocks of the corporate credit risk dynamics.
Abstract *Le interconnessioni tra imprese e tra settori economici sono note come la principale causa del rischio sistemico e hanno contribuito in modo decisivo alla diffusione della crisi finanziaria globale degli anni 2008-2009. Nel presente lavoro studiamo gli effetti di contagio tra settori economici con una metodologia network, in cui i link significativi sono individuati mediante test di indipendenza condizionale. Mentre la letteratura dei network finanziari utilizza prevalentemente processi gaussiani, il nostro approccio è basato su dati discreti. Testando le interdipendenze nella dinamica dei conteggi di default di diversi settori di impresa, troviamo possibili canali di contagio significativi sia nella media condizionata sia nella componente di shock della dinamica del rischio di credito.*

Key words: Conditional Granger causality; Contagion risk; Financial networks; PC-algorithm; Poisson autoregressive models; Vector autoregressive models

Arianna Agosto
Department of Economics and Management, University of Pavia, Via San Felice 5 e-mail: arianna.agosto@unipv.it

Daniel Felix Ahelegbey
Department of Economics and Management, University of Pavia, Via San Felice 5 e-mail: danielfelix.ahelegbey@unipv.it

1 Introduction

The increase in credit risk arising from interconnectedness between institutions and individuals, commonly referred to as *contagion*, is known as one of the main sources of systemic risk. Contagion channels originate from the inter-linkages between companies, sectors, and countries. Corporate default dependency can be conditional on the business cycle, affecting all companies in a financial system, or arise from trading and legal relationships. Studying interconnections and their changes is of major interest for both policy makers and lending institutions. Correlation network models have been recently proposed in the econometric literature to study systemic risk. In particular, [1] and [3] derived contagion measures based on Granger causality tests and variance decompositions. More recently, [2] and [6] extended the methodology introducing stochastic correlation networks. So far, stochastic correlation network models have been based on Gaussian processes, a natural approach when data are market prices. Several works have instead applied discrete data models to analyse possible dependence between default events. For example, [5] tested conditional independence of default times based on the Poisson assumption, while [4] applied Poisson autoregressive models with exogenous covariates to the US default count time series.

In this work we combine the financial network approach with the default count modelling. We indeed estimate a network representing pairwise conditional dependencies between corporate sectors based on their default count dynamics, modelled through log-linear Poisson autoregressive processes.

2 Data

We focus on the default dynamics of Italian non-financial companies.
In particular, as a proxy of default counts we consider the quarterly bad loans flow provided in Bank of Italy statistical database[1] for the period 1996-2018. Using the classification provided in the dataset, we divide corporates into the following sectors: Commerce, Energy, Manufacturing, Primary, Real Estate (including both constructions and real estate corporates), Transports, Information and Communication Technologies, Services. Figure 1 plots the default count time series, showing an increase in level and variability starting from 2008, with the mean returning close to the pre-crisis level after 2015. All the eight series confirm the empirical evidence that defaults cluster in time and show peaks during economic downturns.

[1] http://www.bancaditalia.it/statistiche.

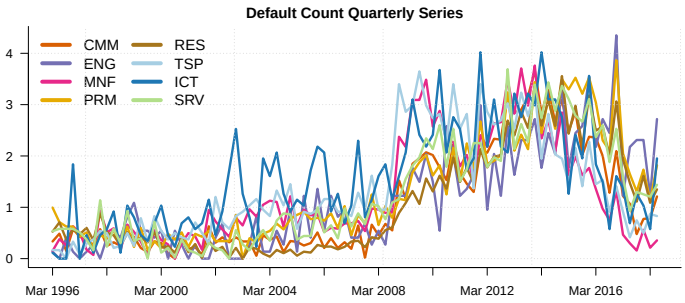


Fig. 1 Normalised quarterly default counts in the Italian credit system (March 1996 - June 2018) by economic sector (CMM = Commercial, ENG = Energy, MNF = Manufacturing, PRM = Primary, RES = Real Estate, TSP = Transports, ICT = Information and Communication Technologies, SRV = Services).

3 Model

3.1 Log-linear Poisson autoregressive model

We assume the conditional distribution of the default counts in economic sector i at time t to be Poisson distributed with a log-linear autoregressive intensity:

$$\begin{aligned} Y_{it}|\mathscr{F}_{t-1} &\sim Poisson(\lambda_{it}) \\ \log(\lambda_{it}) &= \omega_i + \sum_{j=1}^{p} \alpha_{ij} \log(1+y_{it-j}) + \sum_{j=1}^{q} \beta_{ij} \log(\lambda_{it-j}) \end{aligned} \tag{1}$$

where
$\omega_i \in \mathbb{R}$, $y_{ij} \in \mathbb{N}$, $\alpha_i \in \mathbb{R}^p$, $\beta_i \in \mathbb{R}^q$, $\forall i = 1,...,k$ with $k = 8$ sectors.

In Model 1, whose properties and inference theory were studied by [8], both the α_i and the β_i coefficients express dependence of the expected number of defaults on past default counts.
For each of the considered sector, we estimate Model 1 by maximum likelihood, choosing the number of p and q lags between 1 and 4 which minimises the Akaike Information Criteria (AIC). The estimated default intensities are the expected number of defaults in a given quarter.
In this framework, the error component is instead the difference between the observed counts and the estimated intensity, scaled to take into account the time-varying variance: the so-called Pearson residuals $e_{it} = \frac{y_{it}-\hat{\lambda}_{it}}{\sqrt{\hat{\lambda}_{it}}}$.

In the following sections, we estimate inter-sector linkages through testing dependencies in both the intensity ($\hat{\lambda}_{it}$) and the shock (e_{it}) processes.

3.2 Network model

We use the estimates obtained from application of the log-linear Poisson model to build two networks, where the nodes are the considered sectors and the directed edges represent significant dependencies between the default counts of different sectors.
The first network is made up of the links between the estimated default intensities and represents lagged effects between sectors, while the second one is based on the shock component of the default count processes.

3.2.1 Inter-sector network based on lagged effects

We use the default intensities estimated in Section 3.1 to investigate interdependencies in the conditional mean (and variance) of different economic sectors by testing conditional Granger causality [7].

To perform the Granger causality test, we consider the following 8-variate Vector AutoRegressive (VAR) specification:

$$\Delta\lambda_t = \Phi_0 + \sum_{k=1}^{p} \Phi_k \Delta\lambda_{t-k} + u_t, \qquad u_t \sim \mathcal{N}(0,\Sigma) \tag{2}$$

where the lag order p is chosen through model selection based on the Bayesian Information Criteria (BIC). In this setting, the hypothesis that a change of default intensity in sector l does not Granger cause a change of default intensity in sector i can be verified by testing the restriction $\phi_{il,j} = 0$ for $j = 1,...,p$.

Once the test is performed for each of the ϕ_{il} coefficients, we are able to build the A_λ adjacency matrix with binary entries such that $A_{\lambda,il} = 0$ if $\Delta\lambda_{lt-j}$ does not Granger cause $\Delta\lambda_{it}$, 1 otherwise. This defines a directed network where an edge from node l to node i is drawn if changes in the expected number of defaults in sector i are conditionally dependent on changes in the expected number of defaults in sector l.

Figure 2 plots the network structure in two sub-periods: 1996-2007 (*pre-crisis*) and 2008-2018 (*post-crisis*). To aid interpretation, we distinguish the links using the signs of estimated coefficients: positive dependencies are depicted in green and negative in red. It can be seen from Figure 2 that the networks based on lagged effects in the conditional mean are sparse. In the first subsample (1996-2007) the only positive edge is directed from the commerce sector to ICT, while a negative link is present between Energy and the commerce sector. The period during and following the global financial crisis (2008-2018) shows instead three positive connections,

all involving the energy sector, which turns out to affect Primary, Real Estate and Commerce.

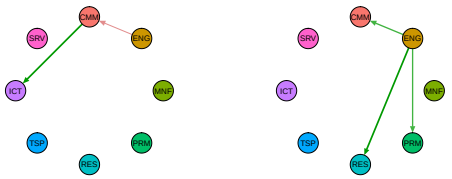


Fig. 2 Inter-sector network based on the quarterly expected number of defaults (left: March 1996 - December 2007; right: January 2008 - June 2018; CMM = Commercial, ENG = Energy, MNF = Manufacturing, PRM = Primary, RES = Real Estate, TSP = Transports, ICT = Information and Communication Technologies, SRV = Services).

3.2.2 Inter-sector network based on contemporaneous effects

Credit risk contagion can also spread through shocks in the default dynamics, generating contemporaneous systemic effects. We thus consider conditional inter-sector dependencies between the residuals of the log-linear models, corresponding to the random (Poisson) part of the default count dynamics.

In order to estimate directed links between sectors based on the estimated shocks, we employ the PC algorithm, which is a constrained-based network inference developed by [10] for learning partially directed networks. We thus build an adjacency matrix called A_e with binary entries such that $A_{e,il} = 0$ if e_{it} does not depend on e_{lt}, 1 otherwise. This way we draw a directed network where an edge from node l to node i means that shocks in the number of defaults in sector i are conditionally dependent on shocks in the number of defaults in sector l.

The networks of connections in default count shocks are shown in Figure 3. Many connections - all with positive sign - can be found in both sub-periods. This indicates that shocks in the default dynamics of different sectors are highly and positively correlated, acting as possible contagion channels. It can also be noticed that the period including the financial crisis is characterised by a larger number of inter-sector links and, as expected, a higher centrality of the real estate sector, which is strongly infected.

4 Conclusions

In this paper we have considered inter-sector contagion effects in corporate default counts, combining Poisson dynamic models with the financial network approach. We estimated directed links between corporate sectors based on their default dy-

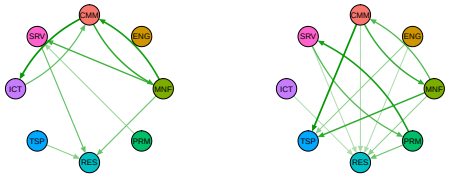


Fig. 3 Inter-sector network based on shocks to the quarterly number of defaults (left: March 1996 - December 2007; right: January 2008 - June 2018; CMM=Commercial, ENG=Energy,MNF=Manufacturing, PRM=Primary, RES=Real Estate, TSP=transports, ICT=Information and Communication Technologies, SRV=Services).

namics, considering both lagged and contemporaneous effects.
Focusing on Italian corporate default counts in eight economic sectors from 1996 to 2018 we found evidence of a high inter-sector vulnerability especially in the global financial crisis period and in the following years. Several significant links between corporate sectors were found in both the mean and the shock component of the default count dynamics.

References

1. Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon: Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors. Journal of Financial Economics 104 (3), 535-559 (2012)
2. Ahelegbey, D. F., M. Billio, and R. Casarin: Bayesian Graphical Models for Structural Vector Autoregressive Processes. Journal of Applied Econometrics 31 (2), 357-386 (2016)
3. Diebold, F. and K. Yilmaz: On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms. Journal of Econometrics 182 (1), 119-134 (2014)
4. Agosto, A., Cavaliere, G., Kristensen, D., Rahbek, A.: Modeling Corporate Defaults: Poisson Autoregressions with Exogenous Covariates (PARX). Journal of Empirical Finance **38**(B), 640-663 (2016)
5. , Lando, D. and M. Nielsen: Correlation in Corporate Defaults: Contagion or Conditional Independence? Journal of Financial Intermediation 19, 35-62 (2010)
6. Giudici, P. and A. Spelta: Graphical Network Models for International Financial Flows. Journal of Business and Economic Statistics 34 (1), 128-138 (2016)
7. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods: Econometrica 37 (3), 424-438 (1969)
8. Fokianos, K. and Tjøstheim, D. (2011): Log-linear Poisson autoregression, Journal of Multivariate Analysis, 102, 563-578 (2011)
9. Lando, D., and Nielsen, M.: Correlation in corporate defaults: Contagion or conditional independence? Journal of Financial Intermediation **19**, 355–372 (2010)
10. Spirtes, P., C. Glymour, and R. Scheines: Causation, Prediction, and Search. MIT Press, London (2000)

A similarity matrix approach to empower ESCO interfaces for testing, debugging and in support of users' experience

Le matrici di similarità a sostegno della fruizione delle interfacce di ESCO per il test, il debug e il miglioramento delle esperienze di utilizzo

Adham Kahlawi, Cristina Martelli, Lucia Buzzigoli, Laura Grassini[1]

Abstract This paper intends to apply the Latent Semantix Indexing approach to ESCO, the European language for labour description and coding, overcoming the limits attributable to the current querying instruments. This objective is strategic to bridge the gap existing between ESCO repository, whose richness and granularity suggests to adopt it as a current operative language for labour processes management, and the users' current language, not always able to select the proper keywords to satisfy their information needs. The paper also discusses an application of the proposed approach to testing and debugging of ESCO linguistic objects, intending to identify redundancies and anomalies in the overall knowledge base.

Abstract *Questo lavoro applica l'approccio Latent Semantic Indexing a ESCO, il linguaggio europeo del lavoro, per rafforzarlo e migliorarne la fruizione, superando alcuni limiti degli strumenti di interrogazione della piattaforma. Si tratta di un obiettivo strategico per colmare il divario esistente tra la grande granularità e ricchezza espressiva del repository di ESCO, che ne suggeriscono l'utilizzo nei processi di gestione del lavoro, e le capacità linguistiche degli utenti, non sempre in grado di individuare le parole chiave più adeguate. Questo lavoro discute, inoltre, un'applicazione del metodo al problema del collaudo e del debug del repository di ESCO, per individuarne eventuali ridondanze ed anomalie.*

Keywords: ESCO, labor market, natural language interfaces, similarity measures.

[1] Adham Kahlawi, Università di Firenze, adham.kahlawi@unifi.it
Cristina Martelli, Università di Firenze, cristina.martelli@unifi.it
Lucia Buzzigoli, Università di Firenze, lucia.buzzigoli@unifi.it
Laura Grassini, Università di Firenze, laura.grassini@unifi.it

1 Introduction

The dynamism of the current labour market and the rapid changes to which it is subjected require specific attention for the role of skills (competences and knowledge) in the professional placement / re-placement of workers, and for the preparation of training offers consistent with the market demands.

For years, numerous international organizations have reported the problem with the identification of megatrends and critical issues (WEF, 2018; ILO, 2019; OECD, 2019). The European Union, in particular, has stressed the role of skills in the program of the European Pillar of Social Rights (European Commission, 2016b, European Commission 2017) and in 2016 adopted the New Skills Agenda for Europe (European Commission, 2016a), which outlines the strategic role of skills in supporting employment, growth and competitiveness.

In this framework, which is part of Europe 2020 strategy (European Commission, 2010), a specific project was aimed at the implementation of an operational tool that establishes a European common language for the classification of Skills, Competences, Qualifications and Occupations, named ESCO (European Commission, 2019).

The main scope is to organize the knowledge on the European labour market and also on the sector of education and training, to improve both the matching between qualifications and labour market needs, and the matching between jobseekers and employers.

ESCO combines the assets of (i) a rigorous official coding standard and (ii) of a specialized language. Thanks to its high granularity, it can be used, in fact, also to provide a clear description of job profiles for job seeking or CV writing. Therefore, according to the most accredited information systems approaches, ESCO can be used for generating administrative data sets, easily suitable to be used as statistical registers.

This work intends to empower the ESCO functionalities in labour market governance and management, proposing a new approach to improve its capability of crossing users' natural language; the proposed approach also intends to support in ESCO components testing and debugging.

In particular, we intend to apply the Latent Semantic Indexing (LSI) approach to measure the similarity between ESCO objects descriptions; the aim is to allow explorative pathways for queries optimization and for identifying relations and redundancies among ESCO descriptive elements.

2 The current version of ESCO

The current full version, ESCO v1, can be freely consulted on the dedicated portal since 2017 and is published as Linked Open data. The classification is available in 27 languages and is organized in three interlinked ‘pillars’: occupations (2942 entries), skills, competences and knowledge (13485 entries) and qualifications (9457 entries).

The first two pillars have a rich and articulated content, while the qualifications pillar is not as well populated (European Commission, 2015). This is also due to the different methods of development: the pillars of occupations and skills are developed by ESCO contributors (sectoral reference groups and online consultations), managed and coordinated by the Commission, while for the qualifications pillar only external sources are used (typically national qualifications databases of the member states).

In our study we will focus on the first two pillars.

The pillar of occupations is organized hierarchically: the top four levels coincide with those of the international standard classification of occupations ISCO08 (ILO, 2012); the lower levels are specific to ESCO (ESCO occupations). For each occupation, ESCO provides a description, a set of alternative labels, the position in the hierarchy, and – only for the fifth level or lower, organized by means of broader/narrower relation – the essential or optional skills, competences and knowledge required for the occupation. Consequently, a connection is created between the pillars.

The pillar of skills, competences and knowledge, is usually called simply the skills pillar because in ESCO the concept of Skill is a general class that includes skill, competences and knowledge, that are different concepts. Nevertheless, it does not distinguish between skills and competences.

ESCO provides various criteria that the user can apply to classify what is generally called Skill. A first simple criterion is the distinction between skills (denoted by a verb) and knowledge (denoted by a noun), called *type*. Another criterion is the reusability level, which in the intentions of the ESCO drafters should be "crucial for supporting occupational mobility" (European Commission, 2019): each Skill is labelled as transversal, cross-sector, sector-specific and occupation-specific. Other distinctions arise on the basis of the "Skills to occupations" and "Skills to Skills" relationships. In both cases, each Skill can be defined 'essential' or 'optional' for other specific Skills or for specific occupations (another way to generate the interlinking among pillars).

The organization of the skills pillar is therefore not hierarchical, but a sort of hierarchy among Skills is made indirectly by the broader/narrower relation. In brief, for each Skill ESCO provides a description, a set of alternative labels, the Skill type, the reusability level, the occupations for which the Skill is essential/optional, the relationship with other Skills (essential/optional, broader/narrower).

All this information produces a rich database which lists occupations, skills, knowledge and qualifications that are relevant for the European labor market showing the relationships among them, that can be used in other European platforms for labour: for instance, it is the standard language expected for exchange of information in EURES, the network of European employment services (European Commission, 2016c).

Any ESCO object, both skill, competence or occupation is structured in a label and a description (Figure 1).

Up to now, ESCO is queryable only using keywords applied to the object labels: the only way to get an answer to users' need is to use the exact keyword that individuates the required object. This fact represents a major limitation to the usage of ESCO for administrative and management purposes: in practice, it is impossible to

perform evolutive queries that, starting from a fuzzier request, allow aiming to the needed answer.

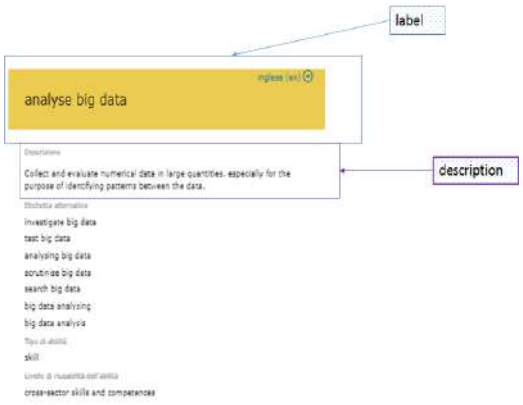


Figure 1: ESCO objects structure

Another aspect that hampers the current usage of ESCO in labour market management and governance is attributable to the redundancy of descriptions' contents. It is a sort of side effect of the impressive richness of ESCO knowledge base: sometimes it is not an easy task to discriminate between different elements, which appear quite similar even if allocated in very different areas of ESCO base.

3 The LSI approach to ESCO interface empowerment

In the application, we analyse similarity between every couple of skills by two different approaches.

Firstly, the similarity between the ESCO descriptions of two different skills is measured using LSI, and this measure is applied to all possible pairs of skills. LSI is an indexing and retrieval technique known for years in the literature on information retrieval, that is able to project queries and documents into a space with latent semantic dimensions (Rosario, 2000) exploring the co-occurrence of each word with every single word. In synthesis, LSI is a text mining processing that produces a corpus-based similarity measure for each pair of descriptions. The measure ranges from 0 to 1 (the greater the more similar).

Secondly, we calculate a use/connection-similarity index. This index is based on the connection between occupations and Skills.

Each ESCO occupation is described by a set of skills. The similarity indicator measures the extent to which a pair of skills are used together. In the specific:
N_x indicates the number of occupations requiring Skill x
N_y indicates the number of occupations requiring Skill y
N_{xy} indicates the number of occupations requiring both Skill x and Skill y
$Sim_{x|y}$ and $Sim_{y|x}$ are defined as follows:

$$Sim_{x|y} = \frac{N_{xy}}{N_y} \qquad Sim_{y|x} = \frac{N_{xy}}{N_x}$$

Note that an index similar to $Sim_{y|x}$ has been used in Opik, et. al. (2018) to measure, in their case, the similarity between two occupations.

The analysis of the different indicators (LSI, $Sim_{x|y}$ and $Sim_{y|x}$) shows a number of interesting cases, some of which are presented in table 1.

Table 1. Some preliminary results

Case	Skill *x* (label)	Skill *y* (label)	$Sim_{x|y}$	$Sim_{y|x}$	LSI
Case 1	IBM Informix	DB2	1.0	1.0	0.9999
Case 2	liaise with union officials	perform internal investigations	0.0	0.0	1.0000
Case 3	provide health psychological concepts	Health psychology	1.0	0.5	1.0000
Case 4	assess chiropractic intervention	prescribe treatments related to surgical procedures	1.0	1.0	0.2181

Case 1 refers to skills that are always required together to describe an occupation; on the other hand, they have two almost identical descriptions. This observation puts the issue whether these skills have an own separate identity.

Case 2 refers to a pair of skills that are never required together to describe an occupation, even if they have identical descriptions. This situation encourages to check whether errors or inaccuracies may have occurred while describing the Skills.

Case 3 refers to a pair of skills that have identical descriptions; on the other hand, when a specific occupation requires Skill *x*, also Skill *y* is required but the opposite is not true. As a result, this is an indication of a possible hierarchy among skills.

Case 4 refers to skills that are always required together to describe an occupation but, as opposed to Case 1, they have two different descriptions. ESCO querying approach can take advantage of this occurrence: at present, it is impossible to perform incremental querying protocols, and the success of an interrogation relies only on the keywords' exact match. Embedding these indicators in ESCO interface it will be possible to move from a skill to the closer ones.

4 Conclusion

This paper intends to expand the information potentialities of ESCO repository, using also the description field as a query target, overcoming the limitations actually in place which limit to the labels the target of the queries.

The proposed approach relies on the LSI methodology which allows evaluating the similarity among textual objects: in this case, ESCO skills descriptions. These

preliminary results encourage to pursue the strategy of embedding the proposed indicators in ESCO interfaces, to empower queries protocols and to test and debug the repository contents on automatic ground.

References

1. European Commission: EUROPE 2020. A strategy for smart, sustainable and inclusive growth, COM (2010).
2. European Commission: The qualifications pillar in ESCO – building the bridge between education and training and employment with ESCO, ESCO SEC 046 DRAFT (2015).
3. European Commission: New Skill Agenda for Europe. Working together to strengthen human capital, employability and competitiveness, COM(2016) 381 final. (2016a).
4. European Commission: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM(2016) 127. (2016b) https://eur-lex.europa.eu/resource.html?uri=cellar:bc4bab37-e5f2-11e5-8a50-01aa75ed71a1.0004.02/DOC_1&format=PDF
5. European Commission: Regulation (EU) 2016/589 of the European Parliament and of the Council (2016c) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0589&from=EN
6. European Commission: European Pillar of Social Rights. (2017). https://ec.europa.eu/commission/sites/beta-political/files/social-summit-european-pillar-social-rights-booklet_en.pdf .
7. European Commission: ESCO Handbook (2019).
8. ILO: International Standard Classification of Occupations (ISCO-08), Geneva (2012).
9. ILO: Work for a brighter future – Global Commission on the Future of Work, International Labour Office – Geneva (2019).
10. OECD: OECD Employment Outlook 2019: The Future of Work, OECD Publishing, Paris (2019). https://doi.org/10.1787/9ee00155-en.
11. Opik, R., Kirt, T. and Liivinnar, I.: Megatrend and Intervention Impact Analyzer for Jobs: A Visualization Method for Labor Market Intelligence. Journal of Official Statistics, 34(4), 961–979 (2018).
12. Rosario B.: Latent Sematic Indexing: An Overview, INFOSYS 240, Spring 2000. https://pdfs.semanticscholar.org/9598/1f057cb76a24329fcf2b572f75d8c2b1613e.pdf (2018)
13. WEF: The Future of Jobs. Report 2018, World Economic Forum, Geneva (2018).

Adding MIDAS terms to Linear ARCH models in a Quantile Regression framework

Regressione quantilica con l'aggiunta di termini MIDAS per modelli Linear ARCH

Vincenzo Candila, Lea Petrella

Abstract Recent financial crises have placed an increased accent on methods dealing with risk management. Despite some critiques, the Value-at-Risk (VaR) still plays today a leading role among the risk measures. For this reason, the financial econometrics literature has been involved in proposing as much as possible accurate VaR models. Recently, the quantile regression (QR) approach has been used to directly forecast the VaR measures. Within such a QR framework, we add a (MI(xed)-DA(ta) Sampling) term to the well known Linear ARCH (LARCH) model. The MIDAS term allows the inclusion of macroeconomic variables usually observed at low frequencies (monthly, quarterly, and so forth) in contexts where the dependent variable is generally observed at higher frequencies (mainly, daily). The resulting model, named Quantile LARCH-MIDAS (Q–LARCH–MIDAS), is the first model incorporating the MIDAS approach within the QR framework.

Abstract *Le recenti crisi finanziarie hanno portato un enorme interesse verso i metodi per la gestione del rischio. Nonostante alcune critiche, il Value-at-Risk (VaR) ha ancora oggi un ruolo primario tra le misure di rischio. Per questa ragione, la letteratura econometrica-finanziaria ha posto l'attenzione sui modelli per la stima del VaR. Recentemente, la regressione quantilica (QR) è stata usata per calcolare direttamente il VaR. In questo contesto di QR, un termine MIDAS (MI(xed)-DA(ta) Sampling) è aggiunto al noto modello Linear ARCH (LARCH). Il termine MIDAS permette l'inclusione di variabli macro, solitamente osservate a frequenza mensile o quadrimestrale, in contesti dove, di solito, la variabile dipendente è osservata a cadenza giornaliera. Il modello risultante, chiamato Quantile LARCH-MIDAS (Q–LARCH–MIDAS), è il primo modello che incorpora l'approccio MIDAS all'interno di un contesto di QR.*

Key words: Value-at-Risk, Quantile Regression, MIDAS term.

Vincenzo Candila
MEMOTEF Department, Sapienza University of Rome, Italy, e-mail: vincenzo.candila@uniroma1.it e-mail: lea.petrella@uniroma1.it

1 Introduction

During the last decades, financial econometrics literature has paid particular attention to the methods for risk management. Among the different risk measures proposed in the literature, the Value-at-Risk (VaR) plays still today a leading role. Despite some criticisms (Artzner et al., 1999), according to the Basel frameworks, the VaR measures are fundamental in order to set aside risk capital adequately (Burchi and Martelli, 2016). The methodology used to obtain the VaR measures can be broadly divided into three main categories: parametric, non-parametric, and semi-parametric (Jorion, 1997). The parametric approach requires the estimation of the volatility of the asset under investigation as a primer step. Typically, the GARCH (Bollerslev, 1986) class of models is used. Secondly, the VaR measures are indirectly obtained by considering these volatility estimates and the quantile at a fixed level of the presumed distribution of the asset. Usually, the Normal distribution is taken into account. Contrary to the parametric approach, the non-parametric technique does not make any distribution assumption concerning the daily returns. The semi-parametric technique specifies the updating dynamics of the model but does not require any distributional assumptions. Contributions based on the quantile regression (QR) (Koenker and Bassett, 1978; Engle and Manganelli, 2004) framework belong to the class of semi-parametric methods for the estimation of the VaR measures. Recent works employing the QR methods are Laporta et al. (2018); Bernardi et al. (2015), among others. Within the QR context, this paper aims to investigate the profitability of including a MIDAS (MI(xed)-DA(ta) Sampling) (Ghysels et al., 2007) component in the well known Linear ARCH (LARCH) (Taylor, 1986) model. The MIDAS component allows to filter the information coming from variables observed at lower frequencies (say, monthly, quarterly) in contexts where the dependent variable is usually observed daily. In this respect, many contributions (Amendola et al. (2019) and Conrad and Loch (2015), among others) highlight that the macroeconomic variables, which are typically observed at a monthly or quarterly frequencies, are driving forces of (daily) assets' variability. To the best of our knowledge, this is the first time that the MIDAS approach is incorporated within the QR framework. Therefore, the advantage of using the proposed Quantile LARCH–MIDAS (Q–LARCH–MIDAS) model to estimate the VaR measures is that these latter values are directly obtained as conditional quantiles of the daily return process, which in turn may depend on some exogenous variables observed at lower frequencies.

The rest of the paper is organized as follows. Section 2 introduces the Q–LARCH–MIDAS model and Section 3 is devoted to the empirical application.

2 Q–LARCH–MIDAS model

Let r_i be a (log-) return of an asset observed at time i, with $i = 1, \cdots, N$. Usually, i represents a day, but sometimes it can refer to periods observed at lower frequencies. In the general heteroskedastic framework, it is assumed that:

$$r_i = \sigma_i z_i, \tag{1}$$

where σ_i denotes the standard deviation, conditionally to the information set $\mathscr{F}_{i-1}$, and z_i is an *iid* random variable with $E(z_i) = 0$ and $Var(z_i) = 1$.

The conditional (one-step-ahead) VaR for day i, at τ level, is defined as VaR_i, and represents the quantity such that

$$Pr(r_i < VaR_i | \mathscr{F}_{i-1}) = \tau. \tag{2}$$

Therefore, by definition, the VaR at time i is the $\tau-$th conditional quantile of the series r_i, given $\mathscr{F}_{t-1}$. For this reason, the VaR can be also expressed as $Q_{r_i}(\tau|\mathscr{F}_{i-1})$.

Within the parametric context, VaR_i can be (indirectly) obtained once got an estimate of the (one-step-ahead) conditional standard deviation, $\hat{\sigma}_i$ and once a distribution function $F(\cdot)$ for z_i is assumed. That is:

$$VaR_i = F(z_i;\tau)\hat{\sigma}_i, \tag{3}$$

where $F(z_i;\tau)$ denotes the quantile at $\tau\%$ of z_i. Many options are available for the conditional standard deviation of r_i. For instance, it can be estimated by means of a specification belonging to the GARCH class of models. The density function of z_i is usually assumed Normal, as in the seminal work of Engle (1982). Instead of following a parametric approach to obtain the VaR measures, the goal of this paper is to directly estimate the VaR by means of the quantile regressions, in a semi-parametric context.

In the linear regression model, the relationship between a dependent variable y_i, at time i, and a set of covariates $\boldsymbol{x}_i$ is represented by the following equation:

$$y_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + u_i, \tag{4}$$

where the vector $\boldsymbol{x}_i$ includes an intercept and $k-1$ covariates, while the zero mean *iid* error term u_i, with quantile function $Q_u(\tau)$, is left with an unspecified distribution. As demonstrated by Koenker and Bassett (1978), the τ-th quantile of y_i, conditional to $\boldsymbol{x}_i$, is:

$$Q_{y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}(\tau), \tag{5}$$

where, in line with Xiao et al. (2015), the $k \times 1$ vector $\boldsymbol{\beta}(\tau) = (\beta_1 + Q_u(\tau), \beta_2, \cdots, \beta_{k-1})^{'}$ is obtained minimizing the following loss function:

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}\in\mathscr{R}^k}{\arg\min} \left[\sum_{i\in\left\{i:y_i\geq \boldsymbol{x}_i^{'}\boldsymbol{\beta}\right\}} \tau|y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}| + \sum_{i\in\left\{i:y_i< \boldsymbol{x}_i^{'}\boldsymbol{\beta}\right\}} (1-\tau)|y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}| \right]. \tag{6}$$

The asymptotic properties of the regression quantile estimator in (6) are discussed in Bassett and Koenker (1978).

The work of Koenker and Zhao (1996) is the first contribution where the VaR (or any other quantile of interest) is estimated within the QR framework. In particular,

the authors consider the LARCH(q) model of Taylor (1986), defined by:

$$r_i = (\beta_0 + \beta_1|r_{i-1}| + \cdots + \beta_q|r_{i-q}|)z_i, \quad \text{with } i = 1, \cdots, N, \tag{7}$$

where $z_i \overset{iid}{\sim} (0,1)$. The vector $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \cdots, \beta_q(\tau))'$ is obtained minimizing the function in (6), replacing y_i with r_i and $\boldsymbol{x}_i = (1, |r_{i-1}|, \cdots, |r_{i-q}|)'$. Koenker and Zhao (1996) illustrate the asymptotic properties of such estimator. Under this notation, the τ-th conditional quantile of r_i, that is the VaR at τ%, is:

$$\hat{Q}_{r_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}(\tau). \tag{8}$$

The innovation of this work is to enlarge the set of covariates of Eq. (7). More in detail, we add a MIDAS term, allowing the inclusion of variable(s) observed at different frequencies with respect to that of the dependent variable. More in detail, let t be the period of observation for the MIDAS variable. This period may be a week, a month, a quarter, and so forth. The LARCH(q) of Eq. (7) changes to the proposed Q–LARCH–MIDAS specification, that is:

$$r_{i,t} = (\beta_0 + \beta_1|r_{i-1,t}| + \cdots + \beta_q|r_{i-q,t}| + \theta \sum_{j=1}^{K} \delta_k(\omega)|X_{t-j}|)z_{i,t}, \tag{9}$$

where $r_{i,t}$ is the log-returns of day i within the period t, the coefficient θ signals the impact of the weighted summation of the K realizations of the additional variable X_t, observed each period t. The variable X_t could be a macro-economic variable driving the log-returns of $r_{i,t}$ or a proxy of volatility at lower frequency (for instance: weekly or monthly aggregated realized volatility). The only condition that the Q–LARCH–MIDAS requires is the (weak) stationarity of X_t. Globally, there are N_t days for the period t and there are T different "lower frequency" periods. In total, there are N days, obtained from $N = \sum_{t=1}^{T} N_t$. In order to take benefit of the information coming from variable(s) observed at lower frequencies, the MIDAS component is a one-sided filter of the K lagged realizations of a given variable X_t, through the weighting function $\delta_k(\omega)$, calculated for $k = 1, \cdots, K$. As in the related literature, we use the Beta function, which is:

$$\delta_k(\omega) = \frac{(k/K)^{\omega_1 - 1}(1 - k/K)^{\omega_2 - 1}}{\sum_{j=1}^{K}(j/K)^{\omega_1 - 1}(1 - j/K)^{\omega_2 - 1}}. \tag{10}$$

Given that we are only interested in the cases where the most recent observations have a larger weight, we set $\omega_1 = 1$ and $\omega_2 \geq 1$. This will allow only for a monotonic decreasing system of weights.

The parameter space of the Q–LARCH–MIDAS model consists of the following vector $\Theta = \{\beta_0, \beta_1, \cdots, \beta_q, \theta, \omega_2\}$. The estimation of $\hat{\boldsymbol{\beta}}(\tau)$ is obtained by minimizing the loss function in (6), where as above y_i is replaced by $r_{i,t}$ and $\boldsymbol{x}_i = (1, |r_{i-1,t}|, \cdots, |r_{i-q,t}|, WS_{i-1,t})'$, with $WS_{i-1,t} = \sum_{j=1}^{K} \delta_k(\omega)|X_{t-j}|$.

3 Empirical Analysis

The main application of this work focuses on the estimate of the VaR measures for the S&P 500 Index. Most of the data have been collected from the "realised library" of the Oxford-Man Institute. The returns of interest are the open-to-close daily log-returns, for the period 3 July 2000 – 12 November 2019 (4861 daily observations). The additional MIDAS component in the Q–LARCH–MIDAS model is the realized volatility, after a weekly aggregation (with $K = 4$ and $K = 8$). Moreover, we also use the Q–LARCH–MIDAS with a monthly MIDAS term: the (first difference of the) U.S. Industrial Production (IP), collected from the Federal Reserve Economic Data (FRED) archive, with $K = 6$ and $K = 12$ as number of lagged realizations. The competing models of the proposed Q–LARCH–MIDAS specification are: Q–LARCH, GARCH (G) and GARCH with Student's t-distribution (G–t), RiskMetrics (RM), standard GARCH–MIDAS (G–M) Engle et al. (2013), with IP as MIDAS component and $K = 12$, Symmetric Absolute Value (SAV), Asymmetric Slope (AS) and Indirect GARCH (IG) specifications for the CAViaR (Engle and Manganelli, 2004) model.

We evaluate the performance of the proposed model through the Model Confidence Set (MCS) (Hansen et al., 2011), employed with the VaR loss function proposed by González-Rivera et al. (2004). The full sample period has been further divided into three sub-periods: the first two correspond to the same periods used in Laurent et al. (2012) and the third period represents the Great Recession (according to the NBER dates). Overall, the proposed model behaves very well. In fact, for $\tau = 0.05$, the Q–LARCH–MIDAS with weekly MIDAS component model always enters the MCS. Moreover, looking at the full sample period, only the models based with the MIDAS component belong to the set of superior models.

Table 1 S&P 500: Model Confidence Set for 5% VaR measures

	Q–L–M	Q–L–M	Q–L–M	Q–L–M	Q–L	G	G–t	RM	G–M	SAV	AS	IG
X_t	RV	RV	IP	IP					IP			
Freq.	W	W	M	M					M			
K	4	8	6	12					12			
Period 1	0.1405	0.1405	0.1438	0.1432	0.1436	0.1406	0.1404	0.1406	0.1373	0.1414	0.1349	0.1386
Period 2	0.0786	0.0786	0.0796	0.0796	0.0807	0.0780	0.0781	0.0785	0.0774	0.0785	0.0794	0.0781
Period 3	0.2193	0.2199	0.2073	0.2088	0.2158	0.2351	0.2345	0.2320	0.2283	0.2410	0.2356	0.2325
Full Period	0.1109	0.1110	0.1109	0.1109	0.1118	0.1119	0.1121	0.1118	0.1100	0.1131	0.1116	0.1116

Notes: The table reports the averages for the loss function proposed by González-Rivera et al. (2004), under four different periods. Shades of gray denote inclusion in the MCS at significance level $\alpha = 0.25$. Models' labels are in the text. X_t indicates the MIDAS component, "Freq." its frequency (Weekly or Monthly) and K the number of lagged realizations of the MIDAS component included in the model. Period 1: July 2000 to March 2003 (681 obs.); Period 2: April 2003 to July 2007 (1088 obs.); Period 3: December 2007 to June 2009 (397 obs.); Full Period: July 2000 to November 2019 (4861 obs.).

References

Amendola, A., V. Candila, and G. M. Gallo (2019). On the asymmetric impact of macro–variables on volatility. *Economic Modelling 76*, 135–152.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). Coherent measures of risk. *Mathematical finance 9*(3), 203–228.

Bassett, G. and R. Koenker (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association 73*(363), 618–622.

Bernardi, M., G. Gayraud, and L. Petrella (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis 10*(3), 553–603.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Burchi, A. and D. Martelli (2016). Measuring market risk in the light of Basel III: New evidence from frontier markets. In P. Andrikopoulos, G. Gregoriou, and V. Kallinterakis (Eds.), *Handbook of Frontier Markets*, pp. 99–122. Elsevier.

Conrad, C. and K. Loch (2015). Anticipating long-term stock market volatility. *Journal of Applied Econometrics 30*(7), 1090–1114.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*, 987–1007.

Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*(3), 776–797.

Engle, R. F. and S. Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics 22*(4), 367–381.

Ghysels, E., A. Sinko, and R. Valkanov (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews 26*(1), 53–90.

González-Rivera, G., T.-H. Lee, and S. Mishra (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting 20*(4), 629–645.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Jorion, P. (1997). *Value at Risk*. Chicago: Irwin.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R. and Q. Zhao (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory 12*(5), 793–813.

Laporta, A. G., L. Merlo, and L. Petrella (2018). Selection of value at risk models for energy commodities. *Energy Economics 74*, 628–643.

Laurent, S., J. V. Rombouts, and F. Violante (2012). On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics 27*(6), 934–955.

Taylor, S. (1986). *Modelling financial time series*. New York: Wiley.

Xiao, Z., H. Guo, and M. S. Lam (2015). Quantile regression and value at risk. In C.-F. Lee and J. C. Lee (Eds.), *Handbook of Financial Econometrics and Statistics*, pp. 1143–1167. Springer.

Company requirements in Italian tourism sector: an analysis for profiles

Requisiti aziendali nel settore turistico in Italia: una lettura per profili

Paolo Mariani, Andrea Marletta, Lucio Masserini, Mariangela Zenga

Abstract The recruitment process represents a way to evaluate the skills that a candidate needs to have in the workplace. This paper aims to evaluate the requirements for new hires in the tourism sector. In particular, we analysed the profiles of 1.526 workers recruited in 2017 by The Adecco Group in Italy. A multinomial logistic regression is carried out to obtain more in-depth knowledge regarding the most selected (or preferred) profiles by employers, among those evaluated.

Abstract *Il processo di reclutamento rappresenta un modo per valutare le skills necessarie ai candidati per entrare nel mercato del lavoro. Lo scopo di questo lavoro è valutare i requisiti necessari per le nuove assunzioni nel settore turistico. In particolare, sono stati analizzati i profili di* 1.526 *lavoratori assunti nel 2017 da The Adecco Group in Italia. Una regressione logistica multinomiale è stata effettuata per ottenere informazioni riguardo ai profili più selezionati (o preferiti) tra quelli valutati.*

Key words: Job market, Multinomial logit, Tourism sector

Paolo Mariani, University of Milano-Bicocca, Department of Economics, Management and Statistics
e-mail: paolo.mariani@unimib.it

Andrea Marletta, University of Milano-Bicocca, Department of Economics, Management and Statistics
e-mail: andrea.marletta@unimib.it

Lucio Masserini, University of Pisa, Department of Economics and Management
e-mail: lucio.masserini@unipi.it

Mariangela Zenga, University of Milano-Bicocca, Department of Statistics and Quantitative Methods
e-mail: mariangela.zenga@unimib.it

1 Introduction

In the job market, the recruitment process represents a primary utility for companies: in fact, finding the right person for the right job at the right time is essential for the sake of company's performance. On the contrary, a wrong choice in the recruitment process could have bad consequences, both from a monetary point of view and in terms of loss of time.

The successful way to select the best candidate is to measure her/his attributes and competencies that could be divided into 3 categories: knowledges, abilities and attitudes. Knowledges are defined as a set of structured principles and theories useful for the correct implementation of the profession. Abilities involve procedures and processes that define the capabilities to accomplish the professional tasks. Attitudes are cognitive features affecting the professional development and execution of job activities.

In this paper the attributes and competencies for employers (travel consultants and clerks) in the tourism sector are analyzed. According to Isfol (2017) [3], skills more required for these professional figures are customer and personal service and foreign language. Customer and personal service are related to the knowledge of principles and processes for providing customer and personal services (i.e. customer needs assessment, meeting quality standards for services, and evaluation of customer satisfaction). Foreign language regards knowledge of the structure and content of a foreign language including the meaning and spelling of words, rules of composition and grammar and pronunciation.

2 Data and methods

In this paper, the data is sourced from the 2017 Adecco Group data-base, where the statistical unit is represented by a candidate receiving a job offer and the explanatory variables are the mandatory requirements needed to pass the recruitment process. The job positions are made comparable using the ESCO international classification. Information about job offers generates knowledge about the criteria used for the selection of the best candidate. In 2017, there were more than 120.000 job positions divided into the following 9 industries: Information Technology and digital, engineering, medical, finance, tourism, Human Resource, commercial, food services and production. In this paper, the work positions analysed are in the tourism sector, so the sample size is $n = 1.526$. The job profiles included are hotel concierge, airport baggage handler and travel consultant. The results are presented only for the subsample related to the travel consultant professional role. The number of the analysed job offers is $n = 626$.

Using these data, a multinomial logistic regression approach has been proposed to identify some relevant profiles. Given an unordered response variable y with $h = 1, 2, \ldots, J$ categories, the equation that expresses a multinomial logit model directly

in terms of the response probabilities, indicated with π_j, conditional to a vector of explanatory variables (x), can be written as it follows [1]:

$$\pi_j(x) = P(y = j|x) = \frac{exp(\alpha_j + \beta_j' x)}{1 + \sum_{h=1}^{J-1} exp(\alpha_h + \beta_h' x)}$$

For identification purpose, $\alpha_j = 0$ and $\beta_j = 0$, where J is the reference category. Furthermore, the numerator for various j sum to the denominator, so $\sum_j^J \pi_j(x) = 1$. Parameter estimation is performed via ML by using the Newton-Raphson method.

3 Results

A multinomial logistic regression [2] was carried out in order to obtain a more in-depth knowledge of the most selected (or preferred) profiles by employers, among those evaluated. In particular, four profiles of candidates have been obtained on the basis of some pre-requirements: the base outcome is a candidate without any specific characteristics served as the reference group against which all the other profiles were compared; candidates with a degree or a high school diploma; candidates with previous experience; and candidates with both a degree or a high school diploma and a previous experience. Furthermore, a set of covariates was used to explain the choice of the profile. Some variables refer to the characteristics of the employers as follows: two dummy variables for identifying the enterprise's sub-sector of economic activity: Fashion, Show and Events (not = 0; yes = 1); and Tourism, Tour Operator and Travel Agencies (not = 0; yes = 1); two binary variables for identifying the macro region of Italy where the enterprise is located: North (not = 0; yes = 1) and Centre (not = 0; yes = 1); two binary variables for identifying the size of the enterprise: Small (not = 0; yes = 1) and Medium (not = 0; yes = 1); a quantitative variable regarding the length of the employment contract (expressed in months). Moreover, two dummy variables that refer to features of the candidates have been included in the model: motivation (not = 0; yes = 1) and adaptability (not = 0; yes = 1).

Table 1 shows the maximum likelihood estimates of the multinomial logistic regression model that allows to identify the variables that affect the choice of candidates' profile by employers. The model's results can also be interpreted in terms of relative-risk ratios by raising the regression coefficients to an exponent. The relative risk ratio describes how much the probability of being in a certain category of the response variable relative to that of the referent group is expected to change for a unit change in the predictor variable, given that the other variables in the model are held constant. After estimation, the small p-value from the Likelihood Ratio (LR) test ($p < 0.001$) leads us to conclude that at least one of the regression coefficients in the model is not equal to zero.

For candidates with a degree or a high school diploma, the only variable that influence the choice is the size of the enterprise. In particular, medium-sized en-

terprises are those more interested in selecting graduates; for these enterprises the relative risk of choosing a candidate with a degree or a high school diploma over candidates without any specific characteristics (Profile 1) is 2.14 relative to that of other enterprises.

Candidates with a previous experience showed two influential variables: the size of the enterprise and the macro region where the enterprise is located. As for candidates with a degree or a high school diploma, medium-sized enterprises prefer this kind of candidates; the relative risk of 2.27 means that the probability of choosing candidates with a previous experience over candidates with the baseline profile is more than twice that of the enterprises of other size. For enterprises located in the north of Italy, the probability of selecting candidates with a previous experience over candidates with the baseline profile is 4.97 times higher than that of enterprises located in other macro regions.

Table 1 Maximum likelihood estimates of multinomial logistic regression model and estimates ranking

Base outcome	**Adaptability & Medium size**	**Motivation & North**	**Travel Agencies & Duration**
Intercept	-0.733	-2.707	-0.280
Fashion, Show and Events	-0.477	-0.623	-0.452
Tourism, Tour Operator and Travel Agencies	0.720	1.232	1.840
North	-0.688	1.603	-0.797
Center	0.184	0.670	-0.860
Small	-0.287	-0.327	-0.074
Medium	0.762	0.818	1.132
Duration of the employment contract	-0.001	-0.001	0.001
Motivation	0.038	0.439	-0.987
Adaptability	0.947	-0.807	0.960
Fashion, Show and Events	8	8	6
Tourism, Tour Operator and Travel Agencies	3	2	1
North	9	1	7
Center	4	4	8
Small	7	7	5
Medium	2	3	2
Duration of the employment contract	6	6	4
Motivation	5	5	9
Adaptability	1	9	3

Source: elaboration on AdeccoGroup data

Candidates who have both a degree or a high school diploma and a previous experience presented six variables affecting the choice of employers. Again, medium-sized enterprises prefer this kind of a candidates, with a relative risk of 3.10. However, the macro region also shows relevant differences; enterprises located in the north and centre of Italy have a lower probability of choosing these candidates, with a relative risk of 0.451 and 0.423, respectively. Moreover, enterprises operating in the sectors of tourism, tour operator and travel agencies are particularly interested in recruiting these candidates, with a probability of selection over candidates with

the base outcome that is 6.30 times higher than that of enterprises in other sectors of economic activity. Two further variables influence the choice, adaptability with a relative risk of 2.96 and the length of the employment contract, since for this profile selected candidates tend to have, on average, contracts with a longer duration.

In the bottom part of the table 1 rankings are presented comparing the estimates of the covariates in descending order for the three profiles. There is a big difference among the three rankings. About the sub-sector, coefficients of Tourism, Tour Operator and Travel Agencies is in the top three for all profiles representing the highest estimate for candidates who have both a degree and a previous experience. On the other hand, the coefficient for Fashion, Show and Events stands in the last position, this means that is more related to the base outcome profile. The estimate for the North macro-region is the lowest for candidates with a degree (no experience) and the highest for those with previous experience (no degree or high school diploma). For candidates with a degree and no experience the highest estimate is for adaptability. For all three profiles, the coefficient for medium companies is very high representing one of the top three positions in the rank.

4 Conclusions

In this paper an analysis of the matching of labour supply and demand was performed based on the job offer data of an important multinational company operating in the field of personnel selection. More specifically, this study aims at evaluating how knowledges, skills and attitudes of the candidates can influence the recruitment phase and choice. The analysis was carried out using a multinomial logistic regression model by taking the most selected (or preferred) profiles by employers as categories of the dependent variable. In particular, these profiles are differentiated by the hard skills, that is, knowledge of the English language, education level and previous experience. A set of the company's covariates such as size, the specific sub-sector of activity within the tourism sector, the geographic area of location were also considered to characterise the employers. Furthermore, candidates' skills such as motivation and adaptability were considered as well as the length of the contract proposed by the company to the candidate. Results show that candidates' profiles with hard skills are preferred by medium-sized companies, whereas companies located in the north of Italy seem to prefer candidates with a previous experience.

References

1. Agresti, A., Categorical Data Analysis. 2nd Edition, John Wiley & Sons, Inc., New York (2002)
2. Greene, W.H., Econometric Analysis. 7th ed. Upper Saddle River. NJ: Prentice Hall (2012)
3. Isfol, Ministero del lavoro - classificatore delle professioni. URL http: //fabbisogni.isfol.it (2017)

Determinants of Firms' Default Risk after the 2008 and 2011 Economic Crises: a Latent Growth Models Approach

Rischio di bancarotta dell'ultimo decennio per le imprese manifatturiere italiane: un modello di curva a crescita latente

Lucio Masserini, Matilde Bini and Alessandro Zeli

Abstract Between 2008 and 2009, European countries became mired in the Great Recession, derived from the US subprime crisis, with financial contraction and bank failures spreading out across the Atlantic and causing damage to financial markets and the real economy. Then, starting in 2011, the crisis affected the euro zone and the sovereign debts of European countries until 2014. For Italy, the crisis period led to a deep negative conjuncture until 2009 and then to a slight recovery until the first half of 2011 and an intense recession from 2011 to 2015. The aims of this work are to identify some determinants of the defaults of Italian firms during the years 2008–2012 and to understand the effects of the Great Recession on the default probability. To perform this analysis, a Latent Growth Curve Model is proposed, using an important Italian private database of Italian firms.

Abstract *In Italia, il periodo di crisi economica è stato caratterizzato da una profonda congiuntura negativa fino al 2009 e da una leggera ripresa fino alla prima metà del 2011 e poi da un'intensa recessione dal 2011 al 2014. Lo scopo di questo lavoro è di studiare gli effetti della grande crisi sul rischio di fallimento delle imprese manifatturiere nel periodo post crisi (2017) e di misurare l'impatto di alcuni fattori su tale rischio. Viene proposto un modello di Curva di crescita latente, sui dati contabili delle società di capitale italiane.*

[1] Lucio Masserini, Department of Economics and Management, University of Pisa; email: lucio.masserini@unipi.it

Matilde Bini, Department of Human Sciences, European University of Rome; email: matilde.bini@unier.it

Alessandro Zeli, Division for Data Analysis and Economic, Social and Environmental Research, Italian National Statistical Institute; email: zeli@istat.it

Key words: Bankruptcy, Firms' default, Interest coverage ratio, Latent Growth Curve Model, Panel data

1 Introduction

In Italy, the crisis period from 2007 to 2014 can be divided into two sub-periods: the first, from 2008 to 2011, led to a deep negative conjuncture until 2009 with a subsequent slight recovery until the first half of 2011, while the second led to an intense recession, begun in 2011 and enduring until 2014. The aim of this paper is to analyse firms' defaults, using information on financial ratios from a large panel of Italian manufacturing firms. Since in the majority of financial literature major attention is devoted to financial conditions [1,2,3,6,7], this study intends to take into consideration more factors characterising a firm that, together with financial and economic conditions, can determine a firm's success or failure. Hence, the main motivation of this study is to assess whether classification variables such as a firm's size, industrial sector, and territorial location can better explain its default risk. An important feature of this study is the large number of enterprises included in the sample, covering a large range of sizes, economic activities, and geographic locations. The analysis is carried out using a statistical approach based on a Latent Growth Curve Model (LGCM) with an Italian private database containing the book-value data of the joint-stock company Italian firms. Unlike traditional longitudinal data analysis techniques, a LGCM allows inferences to be made about individual level effects as well as group effects [4]. Results we obtained allow us to answer the main research questions, such as whether the economic cycle influenced the default risk, how the default risk differs throughout different industries, and whether cumulative indicators have an effect on default probability.

2 Italian manufacturing database

We used a private database called *Analisi Informatizzata delle Aziende Italiane* (AIDA) that contains the book-value data of 16,383 joint-stock Italian firms over the period 2008–2017. First, we checked and controlled the data in order to avoid any inconsistencies in the items of the financial statement and strong outliers among all variables and indicators of interest. We included firms that underwent insolvency proceedings (e.g. failure, liquidation, and extraordinary administration). The target population is represented by Italian manufacturing firms (traced from 2008 to 2017). As a result, we ended up with a balanced panel of 7,689 companies that were used in the statistical analysis, repeated for five years.

3 Latent Growth Curve Model

The analysis was carried out using a Latent Growth Curve Model [5], which assumes the existence of latent trajectories for each firm, measured by the repeated values of the response variable (y_i) over time. In this analysis, the response variable is Interest Coverage, whose values are measured on a continuous scale and observed from 2008 to 2012. This approach was chosen for its flexibility in longitudinal data modelling: it allows for the estimation of different functional forms, the incorporation of both time-varying covariates (TVCs) and time-invariant covariates (TICs), as well as the assessment of the goodness of fit through indices (see, for more details, [9]) and the possibility of dealing with missing data [8]. In general, the LGCM can be expressed in terms of a trajectory equation and a structural model. In particular, given $\mathbf{y}_i$, a $T \times 1$ vector of repeated observed measures for firm *i* at time points $t = 1,2...T$, the trajectory equation, is defined as it follows:

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

$\boldsymbol{\eta}_i$ is the $m \times 1$ vector of the underlying latent factors which identifying the latent growth factors (e.g. initial status and trend factors); $\mathbf{\Lambda}$ is a $T \times m$ matrix of factor loadings for T time points, whose elements are fixed to specify the functional form for the individual trajectories; $\boldsymbol{\varepsilon}_i$ is a random vector of time-specific error terms. On the other hand, the structural model can be formulated as follows:

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_\eta + \mathbf{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i .$$

$\boldsymbol{\mu}_\eta$ is an $m \times 1$ vector of factor means; $\mathbf{\Gamma}$ is an $m \times k$ matrix of regression coefficients between the latent factors and the observed covariates; $\boldsymbol{\zeta}_i$ is an $m \times 1$ vector of error terms, which represents the random components of the model. Finally, a further response variable u_i (Bankruptcy 2017) is added as a distal outcome, whose values can be predicted by the latent growth factors, as follows:

$$u_i = \tau + \boldsymbol{\beta}\boldsymbol{\eta}_i + \boldsymbol{\upsilon}_i$$

Here, the vector of regression coefficients $\boldsymbol{\beta}$ represents the effects of the growth factors on the distal outcome, τ is the intercept and $\boldsymbol{\upsilon}_i$ are the error terms. The final model is shown in Fig. 1, which summarizes graphically the whole system of equations. The rectangular boxes in this figure represent the observed variables

(repeated outcome variables, as well as TICs and TVC), while the circles represent the latent variables (growth factors):

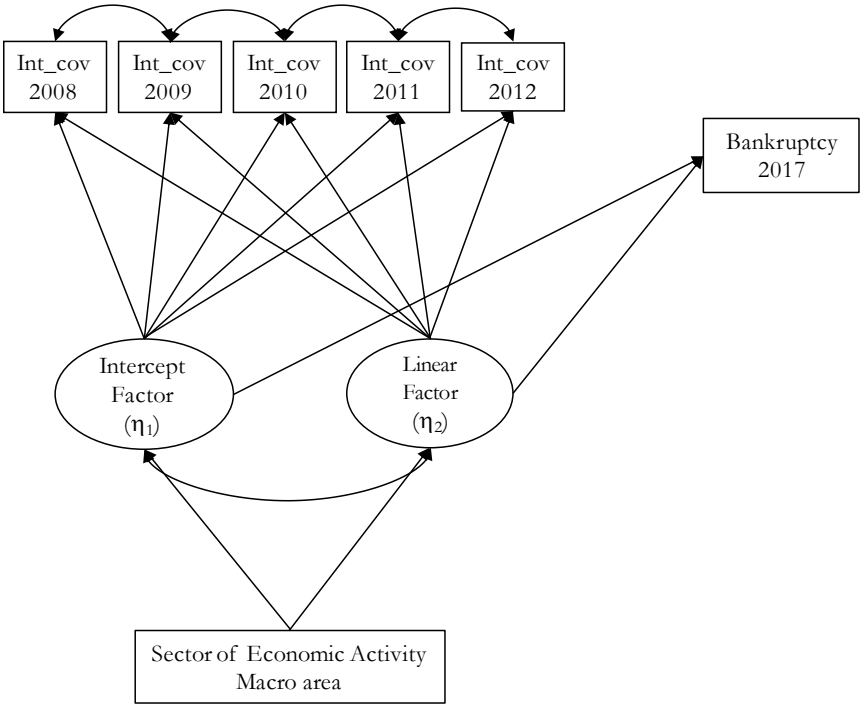


Figure 1: Path diagram of the LGCM

4 Results

Below, the results from the analysis performed by fitting a LGCM are shown. First, a set of alternative unconditional LGCMs with correlated measurement errors between adjacent time points were estimated (such as linear, quadratic, piecewise linear with two knots, and latent basis). These estimates allowed for the identification of a more suitable functional form for the individual latent trajectories. The model parameters were estimated by using the WLSMV (Weighted Least Square Mean and Variance adjusted) estimator with robust standard errors [10]. The model's goodness of fit was evaluated by RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), TLI (Tucker-Lewis Index) and SRMR (Standardized Root Mean Squared Residual) [9]. Based on the results, the linear form LGCM was preferred given the satisfactory values of the indices

(RMSEA = 0.041, CFI = 0.999, TLI = 0.997, SRMR = 0.023). The corresponding results show an adequate fit of the estimated model.

Table 1: LGCMs model parameter estimates for the response variable Interest coverage

Parameter	***Estimate***	***P-value***
Intercept Factor Mean (μ_1)	15.794	0.000
Intercept Factor Variance (ψ_{11})	1609.466	0.000
Linear Factor Mean (μ_2)	0.987	0.012
Linear Factor Variance (ψ_{22})	84.385	0.000
Correlation Intercept Factor vs. Linear Factor (ψ_{12})	0.084	0.429

Results in Table 1 show that the starting point of the response variable Interest Coverage is 15.794 with a significant and remarkable variability (1609.466). Moreover, Interest Coverage shows a linear growth trend (0.987) which also exhibits a significant variability. Differences in growth parameters can be analysed by sector of economic activity and a firm's geographical location in Table 2, where the asterisks show significant differences with respect to the average at the 0.05 significance level.

Table 2: Differences of growth curve parameters relative to baseline level for various sectors of economic activity and geographic location

Parameter	***Intercept***	***Slope***
Food and Tobacco	1.825	0.852
Textile and Leather	-6.775	1.886
Wood, Publishing, and Paper	6.171	-2.094
Refining	1.159	1.723
Chemistry and Rubber	-0.234	0.258
Metallurgy and Steel Industry	2.578*	-1.035
Electric Machines and Mechanical	7.628	0.486
Transport	4.762	-3.008*
Other Industries and Maintenance (Reference Category)	5.431	1.418
North	7.823*	0.197
Centre	6.515	0.174

Firms in the metallurgy and steel industry sector and those located in the North of Italy show a higher level of interest coverage at the beginning of the observed period. On the other hand, a significant and lower rate of growth is found in those firms that belong to the transport sector.

Table 3 shows the parameter estimates where Interest Coverage is used to predict a firm's situation in terms of risk of failure after five years, as measured by Bankruptcy 2017, introduced as a distal outcome. In particular, this approach allows for the evaluation of the Interest Coverage trend during the crisis period (2008–2012) and the impact of this trend on the risk of failure after five years.

Table 3: Effects of Interest coverage on Bankruptcy 2017

Parameter	***Estimate***	***P-value***
Intercept (τ)	1.360	0.000
Intercept Factor (β_1)	-0.004	0.002
Linear Factor (β_2)	-0.064	0.000

The Latent Growth Curve Model successfully detects a firm's outcome in terms of bankruptcy five years later. Indeed, the bankruptcy risk is well predicted by the linear function, and the model estimates show that in the considered period, industries in which Italian manufacturing is stronger in terms of interest coverage have a lower risk of bankruptcy.

References

1. Alaka, H.A., Oyedele, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O., Bilal, M.: Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94, 164–184 (2018)
2. Becchetti, L., Sierra, J.: Bankruptcy risk and productive efficiency in manufacturing firms. *Journal of Banking and Finance*, 27: 2099-2120 (2003)
3. Bellovary, J.L., Giacomino, D.E., Akers, M.D.: A review of bankruptcy prediction studies: 1930 to Present. *Journal of Financial Education*, 33, 1–42 (2007)
4. Bini, M., Masserini, L., Zeli, A.: A longitudinal analysis of riskiness indicators after the 2008 and 2011 economic crises: the case of Italian manufacturing. *Social Indicators Research*, accepted (2020).
5. Bollen, K. A., Curran, P. J.: *Latent Curve Models*. New York: Wiley (2006)
6. Bottazzi, G., Grazzi, M., Secchi, A., Tamagni, F.: Financial and economic determinants of firm default. *Journal of Evolutionary Economics*, vol. 21, issue 3, 373–406 (2011)
7. Dimitras, A. I., Slowinski, R., Susmaga R., Zopounidis, C.: Business failure prediction using rough sets. *European Journal of Operational Research*, vol. 114, issue 2, 263–280 (1999)
8. Enders, C.: A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modelling*. 8(1), 128–141 (2001)
9. Hu, L., Bentler, P. M.. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55 (1999)
10. Muthén, B., Satorra, A.: Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, 60(4), 489–503 (1995)

Double Asymmetric GARCH–MIDAS model - new insights and results

Modello Double Asymmetric GARCH–MIDAS: nuovi approfondimenti e risultati

Alessandra Amendola, Vincenzo Candila, Giampiero M. Gallo

Abstract The recently proposed Double Asymmetric GARCH-MIDAS (DAGM) model aims at separating the positive and negative macro variable variations within the long-run term and adds an asymmetric effect in the short-run component. In this work, the intent is to further extend the model in two main directions. A realized measure is included as a daily lagged variable in the short-run component (the so-called "–X" term) and a multi-step-ahead forecasting procedure is implemented for the class of GARCH–MIDAS (GM) models with the additional "–X" term. The extended DAGM-X model, which nests the DAGM and GM, is extensively evaluated under alternative configurations concerning the S&P 500 Index.

Abstract *Il presente lavoro illustra una estensione del modello Double Asymmetric GARCH–MIDAS (DAGM), recentemente proposto. Nella modellizzazione, oltre agli effetti asimmetrici nelle componenti di lungo e di breve periodo, è stata introdotta una misura di volatilità realizzata giornaliera come variabile addizionale per la componente di breve periodo (la cosiddetta parte "–X"). Inoltre, è stata sviluppata una procedura per le previsioni multi-step-ahead, valida per tutti i modelli GARCH–MIDAS (GM), anche con un termine aggiuntivo "–X". La performance del DAGM–X, che generalizza il modello DAGM e il modello GM, è stata valutata in riferimento all'indice S&P 500.*

Key words: Volatility, Asymmetry, GARCH–MIDAS, Forecasting.

Alessandra Amendola
Department of Economics and Statistics, University of Salerno, Italy, e-mail: alamendola@unisa.it

Vincenzo Candila
MEMOTEF Depart., Sapienza University of Rome, Italy, e-mail: vincenzo.candila@uniroma1.it

Giampiero M. Gallo
Italian Court of Audits (Corte dei conti – disclaimer) and NYU in Florence, Italy, e-mail: giampiero.gallo@nyu.edu

1 Introduction

The connection between the volatility of financial assets and the macroeconomic variables (MVs) has a long history. During the last decade, the GARCH–MIDAS model has been successfully used to include such a MVs into the so-called long-run volatility component, which varies with the same frequency as the MVs and around which the daily short-run component fluctuates as well. The Double Asymmetric GARCH–MIDAS (DAGM) model, recently proposed in Amendola et al. (2019), separates the positive and negative MV variations within the long-run term and adds an asymmetric effect in the short-run part. In this paper, we push one step further the DAGM model, including a realized measure as a daily lagged variable in the short-run component (the so-called "–X" term). Moreover, we introduce a multi-step-ahead forecasting procedure suitable for the class of GARCH–MIDAS (GM) models (Engle et al., 2013) with the "–X" term. The proposed DAGM–X model, which nests the standard DAGM and GM, is extensively evaluated under several alternative configurations, focusing on the S&P 500 Index. The daily S&P 500 5-minute realized volatility is included as a daily lagged variable in the short-run component while the rate of change of the monthly US Industrial Production, Housing Starts, and New Orders are used as the driver of market volatility. The idea is that a succession of negative values (over a somewhat long past, and suitably weighted) transmits a different type of increasing impulse to volatility than its positive counterpart. The empirical results give evidence that the DAGM–X model can outperform the standard DAGM, GM, and GM–X models, independently of the forecasting horizon considered. Furthermore, the proposed model is very often superior to the well known HAR model of Corsi (2009), mainly if longer forecasting horizons are taken into account.

2 DAGM–X model

Letting $r_{i,t}$ represent the log–return, that is, the first difference of the log–closing prices for day i of the period (week or month) t (with $i = 1, \cdots, N_t$, where N_t is the number of days for period t), our GM framework defines:

$$r_{i,t} = \sqrt{\tau_t \times g_{i,t}}\varepsilon_{i,t}. \tag{1}$$

In this expression, $\varepsilon_{i,t}$ is the innovation term, $g_{i,t}$ follows a unit–mean reverting GJR-GARCH(1,1) process (short–run component), and τ_t provides the slow–moving local level of volatility (long–run component).

The short–run component is then given by:

$$g_{i,t} = (1 - \alpha - \beta - \gamma/2) + \left(\alpha + \gamma \cdot \mathbb{1}_{\left(r_{i-1,t}<0\right)}\right) \frac{\left(r_{i-1,t}\right)^2}{\tau_t} + \beta g_{i-1,t}, \tag{2}$$

where $\mathbb{1}_{(.)}$ is an indicator function.

The long–run component is defined as:

$$\tau_t = exp\left(m + \theta^+ \sum_{k=1}^{K} \delta_k(\omega)^+ MV_{t-k}\mathbb{1}_{(MV_{t-k}\geq 0)} + \theta^- \sum_{k=1}^{K} \delta_k(\omega)^- MV_{t-k}\mathbb{1}_{(MV_{t-k}<0)}\right), \tag{3}$$

where m plays the role of an intercept, θ^+ and θ^- represent the asymmetric responses to the one–sided filter, and $\delta_k(\omega)^+$ and $\delta_k(\omega)^-$ are suitable functions weighing the past K realizations of the additional stationary predetermined variable labelled MV_t as the *MIDAS variable*. Throughout this work, the Beta function will be used as weighting function of all the GM models, that is:

$$\delta_k(\omega)^+ = \frac{(k/K)^{\omega_1^+ - 1}(1-k/K)^{\omega_2^+ - 1}}{\sum_{j=1}^{K}(j/K)^{\omega_1^+ - 1}(1-j/K)^{\omega_2^+ - 1}}, \tag{4}$$

$$\delta_k(\omega)^- = \frac{(k/K)^{\omega_1^- - 1}(1-k/K)^{\omega_2^- - 1}}{\sum_{j=1}^{K}(j/K)^{\omega_1^- - 1}(1-j/K)^{\omega_2^- - 1}}, \tag{5}$$

with the restriction $\omega_1^+ = \omega_1^- = 1$, which gives a higher weight to the most recent observations (monotonically decreasing weighting scheme). This is in line with what suggested in Ghysels and Qian (2019). Note that the Beta functions assure that $\sum_{k=1}^{K}\delta_k(\omega_2^+) = 1$ and $\sum_{k=1}^{K}\delta_k(\omega_2^-) = 1$.

The extension we suggest is to extend the short-run equation to some additional volatility determinants, observed at the same frequency of $r_{i,t}$. This allows us to move out of the classical GM framework with just MVs. In particular, the short-run component will change to:

$$g_{i,t} = (1-\alpha-\beta-\gamma/2) + \left(\alpha + \gamma\cdot\mathbb{1}_{\left(r_{i-1,t}<0\right)}\right)\frac{\left(r_{i-1,t}\right)^2}{\tau_t} + \beta g_{i-1,t} + \tilde{z}\cdot X_{i-1,t}, \tag{6}$$

where the variable $X_{i-1,t}$ is observed at the same frequency as $r_{i,t}$.[1] We define the model with the short-run component in (6) as DAGM–X, which nests the DAGM if $z = 0$. In order to estimate the DAGM–X (and, hence, also the GM–X), the following assumptions are made:

Assumption 1 *The innovation $\varepsilon_{i,t}$ in (1) is iid, with $E\left[\varepsilon_{i,t}\right] = 0$ and $E\left[\varepsilon_{i,t}^2\right] = 1$.*

Assumption 2 *The short-run parameters are subject to: $\alpha > 0$; $\beta \geq 0$; $\alpha + \beta + \gamma/2 < 1$;*

Assumption 3 *$\tilde{z} \geq 0$; $X_{i,t} \geq 0$, $\forall i$ and t; $X_{i,t}$ is stationary and ergodic.*

[1] Note that, because of this extension,

$$E(g_{i,t}) = 1 + \frac{\tilde{z}}{(1-\alpha-\beta-\gamma/2)}\bar{X} \equiv 1 + z\bar{X}.$$

Assumptions 1–2 are standard in the GARCH literature. Together with Assumption 3 (Han and Kristensen, 2014), the positiveness of $g_{i,t}$, $\forall i$ and t is guaranteed. The parameter space of the DAGM–X- model[2] is $\Theta = \left\{\alpha, \beta, \gamma, \tilde{z}, m, \theta^+, \omega_2^+, \theta^-, \omega_2^-\right\}$. The maximum likelihood (ML) estimates for Θ are obtained once a distribution is chosen for the innovation term in (1). If $\varepsilon_{i,t}$, conditional on the information set up to day $i-1$ of period t, denoted by $\Phi_{i-1,t}$, is assumed to be standard normally distributed, then the following log-likelihood can be maximized:

$$\mathscr{L}(\Theta) = -\frac{1}{2}\sum_{t=t_s}^{T}\sum_{i=1}^{N_t}\left[\log(2\pi) + \log(g_{i,t}\tau_t) + \frac{r_{i,t}^2}{g_{i,t}\tau_t}\right], \tag{7}$$

where t_s is taken sufficiently large in order to assure that different models, with different choices of K, will have the same information set. This is crucial when several models, with alternative K lags, are evaluated in terms of information criteria.

If it is straightforward to derive the one-step-ahead volatility, the multi-step-ahead predictions require some preliminary assumptions, mainly if the model at hand includes the "–X" part. Firstly, let $\sigma_{i,t} = \sqrt{\tau_t \times g_{i,t}}$ be the standard deviation in a GM framework, conditionally to $\Phi_{i-1,t}$. Moreover, let $\Phi_{N_T,T}$ denote the information set available the last day (N_T) of the last period (T). Let $\tilde{\sigma}_{f,T+1}$ be the conditional standard deviation forecast for the day f of the period $T+1$, conditional on the information set $\Phi_{N_T,T}$, that is: $\tilde{\sigma}_{f,T+1} = \left(\sqrt{g_{f,T+1} \times \tau_{T+1}} | \Phi_{N_T,T}\right)$. The main difference between $\tilde{\sigma}_{f,T+1}$ and $\sigma_{i,t}$ is that the former is based on a fixed information set, while the latter on an information set which updates daily. The one-step-ahead predictions depend on τ_t, which in turn is based on the MIDAS variable observed up to period $t-1$. Therefore, in the multi-step-ahead context, τ_{T+1} is a function of historical values of the MIDAS variable observed up to period T, which are included in $\Phi_{N_T,T}$ and is kept constant for all days in $T+1$. The multi-step-ahead predictions require an estimate for the short-run component $g_{f,T+1}$: extending, e.g. Conrad and Kleen (2019), we have:

$$E(g_{f,T+1}|\Phi_{N_T,T}) = 1 + z\cdot\bar{X} + (\alpha+\beta+\gamma/2)^{f-1}(g_{1,T+1} - 1 - z\cdot\bar{X}), \tag{8}$$

where $\bar{X}$ is the average of the X variable calculated over the period up to $\Phi_{N_T,T}$. Note that, as usual, the short-run prediction depends just on $g_{1,T+1}$ (a function of the information set $\Phi_{N_T,T}$) and reduces to the customary $1 + (\alpha+\beta+\gamma/2)^{f-1}(g_{1,T+1}-1)$ when the short-run unconditional expectation is equal to one ($z=0$). In either case, as f increases, the short-run prediction will approach its unconditional mean, and that will combine with τ_{T+1} to produce the overall forecast.

[2] Recall the DAGM–X nests the DAGM and the GM. Therefore, the following discussion can be naturally applied to all the other models, once that the parameter space has been reduced.

3 Empirical application

The empirical application focuses on the volatility of S&P 500 Index, while the MVs are the US Industrial Production (*IP*), Housing Starts (*HS*), and New Orders (*NewOr*). Data on S&P 500 Index and its realized volatility (calculated aggregating at 5-minutes the squared intradaily returns) have been collected from the realised library of the Oxford-Man Institute. Data on the MVs come from the Federal Reserve (St.Louis) Economic Data archive. We use the returns as daily close–to–close log–differences of the S&P 500 Index, on a sample period between 2 January 2002 and 31 December 2019 (4500 daily observations). All the MVs have been considered in terms of month-to-month growth rate. The S&P 500 realized volatility, labelled as $RVol_{i,t}$, entering the "–X" part of the short-run component includes the overnight volatility. Some summary statistics (minimum, maximum, mean, standard deviation, skewness and kurtosis) for all variables considered are in Table 1. The last column of Table 1 reports the estimated coefficient of the predictive regression as recently proposed by Conrad and Schienle (2018), to which we refer for the details on the procedure. Overall, these variables appear to be good predictors of the long-run component.

Table 1 Summary statistics

	Obs.	Min.	Max.	Mean	SD	Skew.	Kurt.	π_1
Daily data								
$r_{i,t}$	4500	-9.688	10.642	0.023	1.158	-0.231	9.523	
$RVol_{i,t}$	4500	0.128	8.958	0.835	0.605	3.392	20.363	
Monthly data								
ΔIP	215	-4.337	1.517	0.088	0.678	-1.984	9.569	-0.125^{***}
ΔHS	215	-18.681	24.647	0.326	8.431	0.176	-0.278	-0.01^{***}
$\Delta NewOr$	215	-9.680	10.363	0.250	2.250	-0.427	3.617	-0.042^{***}

Notes: The table presents the summary statistics for the close-to-close S&P 500 log-returns ($r_{i,t}$), S&P 500 realized volatility with overnight returns ($RVol_{i,t}$), US Industrial Production, Housing Starts and New Orders growth rates (ΔIP, ΔHS and $\Delta NewOr$, respectively). Sample period: 2 February 2002 - 31 December 2019. Percentage scale. The table reports the number of observations (Obs.), the minimum (Min.) and maximum (Max.), the mean, standard deviation (SD), Skewness (Skew.) and Kurtosis (Kurt.). The symbol Δ denotes the first difference. π_1 represents the estimated coefficient of the predictive regression as proposed by Conrad and Schienle (2018). *, ** and *** represent the significance at levels $10\%, 5\%, 1\%$, respectively, associated to HAC robust standard errors, for the null of $\pi_1 = 0$.

In what follows, we use six different specifications for the two models at hand, the DAGM(–X) and GM(–X), labelled as $M_1, \cdots, M_6$. Even–numbered models include the $RVol_{i,t}$ in the "–X" part. Moreover, we also use the HAR model of Corsi (2009). The volatility proxy used is the 5-minute realized volatility. The results of the comparison of all the models, carried out through the Model Confidence Set (MCS, Hansen et al. (2011)), are reported in Table 2 .

Table 2 MSE losses and MCS composition for the out-of-sample period

		MIDAS VAR.	*–X spec.*	1-day	5-days	10-day	1-month	2-months	3-months
DAGM(–X)	M_1	ΔIP		0.122	0.114	0.147	0.197	0.237	0.253
	M_2	ΔIP	$RVol_{i-1,t}$	0.106	0.104	0.133	0.165	0.187	0.204
	M_3	ΔHS		0.114	0.107	0.14	0.189	0.229	0.246
	M_4	ΔHS	$RVol_{i-1,t}$	0.097	0.093	0.122	0.152	0.166	0.177
	M_5	$\Delta NewOr$		0.142	0.128	0.163	0.221	0.276	0.316
	M_6	$\Delta NewOr$	$RVol_{i-1,t}$	0.12	0.116	0.147	0.191	0.24	0.25
GM(–X)	M_1	ΔIP		0.125	0.114	0.148	0.2	0.252	0.278
	M_2	ΔIP	$RVol_{i-1,t}$	0.104	0.099	0.127	0.159	0.184	0.194
	M_3	ΔHS		0.117	0.108	0.141	0.188	0.224	0.242
	M_4	ΔHS	$RVol_{i-1,t}$	0.103	0.098	0.126	0.162	0.188	0.189
	M_5	$\Delta NewOr$		0.135	0.122	0.156	0.212	0.268	0.304
	M_6	$\Delta NewOr$	$RVol_{i-1,t}$	0.111	0.104	0.131	0.161	0.182	0.194
	HAR			0.097	0.113	0.127	0.163	0.207	0.199

Notes: The table reports the average MSE losses for the full out-of-sample period, starting on 2 January 2014 and ending on 31 December 2019 (1506 daily obs.), according to the 1-day to 3-months volatility forecasts. Shades of gray denote inclusion in the MCS at significance level $\alpha = 0.25$.

We find that the extended models with the realized volatility as the "–X" term always belong to the MCS and are often able to outperform the HAR model, independently of the forecasting horizon considered.

References

Amendola, A., V. Candila, and G. M. Gallo (2019). On the asymmetric impact of macro–variables on volatility. *Economic Modelling 76*, 135–152.

Conrad, C. and O. Kleen (2019). Two are better than one: Volatility forecasting using multiplicative component GARCH models. *Jour. of Applied Econometrics*.

Conrad, C. and M. Schienle (2018). Testing for an omitted multiplicative long-term component in GARCH models. *Jour. of Business & Economic Statistics*, 1–14.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*, 174–196.

Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*(3), 776–797.

Ghysels, E. and H. Qian (2019). Estimating MIDAS regressions via OLS with polynomial parameter profiling. *Econometrics and Statistics 9*, 1–16.

Han, H. and D. Kristensen (2014). Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates. *Journal of business & economic statistics 32*(3), 416–429.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The Model Confidence Set. *Econometrica 79*(2), 453–497.

European SMEs and Circular Economy Activities: Evaluating the Advantage on Firm Performance through the Estimation of Average Treatment Effects.

PMI europee e attività di economia circolare: una valutazione del vantaggio sulla performance delle imprese attraverso la stima degli effetti medi del trattamento.

Luca Secondi

Abstract This paper aims at analysing the effects of carrying out Circular Economy activities within European SMEs by means of the micro-data from the Flash Eurobarometer 441 survey and the estimation of endogenous Average Treatment Effects. Valuable findings about the economic advantages of carrying out Circular Economy sustainable practices within the corporate management systems were obtained.

Abstract *Questo studio ha lo scopo di analizzare gli effetti delle attività di economia circolare implementate all'interno delle PMI europee utilizzando i micro-dati dell'indagine Flash Eurobarometer n.441 e attraverso la stima degli effetti medi del trattamento. Risultati interessanti emergono circa i vantaggi economici derivanti dall'esecuzione di pratiche sostenibili di economia circolare all'interno dei sistemi di gestione aziendale*.

Key words: circular economy, firm performance, ATE, ATET, endogeneity

1 Introduction

The transition towards a circular model of economies has been officially recognized by the European Commission (EC) in 2015 (December, 2nd 2015) with the adoption

[1] Luca Secondi, University of Tuscia; email: secondi@unitus.it

of the Circular Economy Action Plan (CEAP), which included measures that can help to stimulate Europe's transition towards a Circular Economy (CE), enhance global competitiveness, promote sustainable economic growth, create new jobs and increase firm efficiency in using resources (Özbuğday et al, 2020).

The European Union (EU) CEAP establishes a concrete and ambitious programme of action and can represent an accelerator of the 2030 Agenda for Sustainable Development Goals foreseen by United Nations (Principato et al. 2019). In a circularity perspective, the value of products and materials is kept as long as possible so that waste and resources are minimized similarly to when a product reaches the end of its life and is reused to create further value. The proposed actions aim at enlarging product lifecycles by means of a greater level of product recycling and re-use bringing benefits to both the environment and the overall economy. However, the transition from the current linear economic model to a circular model requires the rethinking of market strategies towards a new integrated model of production, distribution, consumption as well as the introduction of new business models in order to gain economic benefits, contribute to innovation, growth and social inclusion (Ruggieri et al, 2016; Zamfir et al, 2017).

The EC has constantly monitored the level of EU-country transitions towards CE and according to the most recent published report (European Commission, 2019) the implementation of the CEAP has accelerated the evolution towards CE in Europe. As an example, it was found that in 2016, the sectors considered relevant to the CE employed more than 4 million workers, a 6% increase compared to 2012. The circularity approach has also opened new business opportunities and in 2016, circular activities such as repair, reuse or recycling generated almost 147 billion Euro value added while accounting for around 17.5 billion Euros value of investments (European Commission, 2019).

The topic of CE has also stimulated much interest both among academics and practitioners (Kirchherr et al, 2017) regarding the analyses at both micro and macro levels of firms' knowledge, awareness and implementation of CE practices and, ultimately, of the level of transition of economies towards CE.

By focusing on the number of scientific papers published in international journals within Scopus, the locution "circular economy" is present in approximately 5,500 papers (abstract, title and keywords) in the period 2016-2020, covering above all environmental science, engineering, energy, business management and accounting subject macro-areas. On the other hand, official statistics at European level has also implemented measures for monitoring CE with the introduction of 10 key indicators by Eurostat covering each phase of the product lifecycle as well as competitiveness related aspects.

Among the initiatives carried out at European level for monitoring the implementation of CE activities at single-firm level, the EC has commissioned a sample survey carried out in 2016 addressing the specific topic of CE within Small and Medium Enterprises (SMEs) in Europe - i.e. the Flash Eurobarometer (FE) 441 survey – with the main focus of exploring CE activities carried out within SMEs (European Commission, 2016).

This paper aims at analysing the effect of implementing CE practices within European SMEs on the company's economic performance. By considering the micro

data from the FE 441 survey, we firstly explore the level of implementation of the different activities that contributes to the CE. Secondly, we estimate the effects and the potential advantages (if any) of carrying out at least one of the CE related activities. To achieve the aim of the study we considered the firm's own decision of implementing a CE activity as a treatment effect which can affect its economic performance (in terms of total turnover and the related growth) and we estimated the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATET), by controlling for the endogeneity of the treatment assignment (Wooldridge 2010; 2014). The treatment effects were estimated by considering the observational nature of the data and estimating endogenous treatment effects which enabled us to create an analysis that resembles what would occur had the treatment been randomly assigned (D'Agostino, 2007).

2 Materials and Method

2.1 The Flash Eurobarometer (FE) 441 survey

The microdata used in this study were collected within the FE 441 survey carried out by the TNS political and social network in the EU28 member states from 18th to 27th of April 2016 as a business to business survey which involved 10,406 SMEs employing 1 or more persons (the upper limit considered for defining a SME was 250 employees). The concerned NACE sectors were: Manufacturing (NACE category C), Retail (NACE category G), Services (NACE categories H/I/J/K/M/N) and Industry (NACE categories B/D/E/F).

The FE-441 survey investigated the following aspects: i) the proportion of SMEs undertaking activities related to the CE, and the kind of activities being taken; ii) the SME's awareness of the kinds of finance available for activities related to the CE as well as the most common sources of finance used; iii) the quality of information available to help companies access finance, iv) actual and perceived issues with regard to undertaking activities related to the CE and iv) the awareness of government programmes supporting activities related to the CE.

Specifically, the questionnaire examined if and to what extent the following activities has been undertaken within each company in the last three years: i) re-plan of the way water is used to minimise usage and maximise re-usage; ii) use of renewable energy; iii) re-plan energy usage to minimise consumption; iv) minimise waste by recycling or reusing waste or selling it to another company; v) re-design products and services to minimise the use of materials or use recycled materials).

2.2 The estimation of Endogenous Treatment Effects

The main identification challenge for estimation purposes is to know what would have happened with the firms implementing at least one activity related to CE if they had not implemented the action/s. In order to achieve this purpose it would be essential to have experimental data with a control group of SMEs similar to those implementing the activities (the treatment group), since firms implementing CE activities can differ analytically from firms not implementing any CE activities. Indeed, SMEs self-select into the decision of carrying out at least one of the activities related to CE and it may be more likely according to the firms' sector, dimension, localization and other firm structural characteristics. Moreover, some variables used to measure the impact on firm total turnover of carrying out the CE activity might also endogenously determine the firm's likelihood of implementing a CE activity.

Bearing the above considerations in mind, we therefore specified a linear model for modelling the outcome variable defined as the SMEs' level of total turnover growth in the last 3 years and a probit model for modelling the SMEs' decisions of implementing CE activities. We verified the existence of endogeneity between the two processes and we adopted a control-function approach which explicitly includes residuals from the treatment models for the potential outcomes.

More specifically, the selected Endogenous Treatment Effect estimation (Wooldridge, 2010; 2014) enabled us to extract experimental-style causal effects from observational data thus looking like as the treatment would have been randomly assigned (D'Agostino, 2007) and to estimate the Average Treatment Effect (ATE), the Average Treatment Effect on the Treated (ATET) and the Potential Outcome Means (POMs), by also considering treatment assignment as correlated with the potential outcome. The treatment-effect models we considered for estimating the effects of CE activities on firms' turnover follow this specification:

$$y_{i0} = E(y_{i0}|\mathbf{x_i}) + \epsilon_{i0} \tag{1}$$

$$y_{i1} = E(y_{i1}|\mathbf{x_i}) + \epsilon_{i1} \tag{2}$$

$$t_i = E(t_i|\mathbf{z_i}) + v_i \tag{3}$$

$$y_i = t_i y_{i1} + (1 - t_i) y_{i0} \tag{4}$$

$$E(\epsilon_{ij}|\mathbf{x_i}, \mathbf{z_i}) = E(\epsilon_{ij}|\mathbf{z_i}) = E(\epsilon_{ij}|\mathbf{x_i}) = 0 \quad for \;\; j \in \{0,1\} \tag{5}$$

$$E(\epsilon_{ij}|\mathrm{t}) \neq 0 \quad for \;\; j \in \{0,1\} \tag{6}$$

where for the ith firm: y_{i0} is the potential outcome when the treatment is not received, y_{i1} identifies the potential outcome when receiving the treatment (in our study of implementing at least one CE activity), t_i identifies the observed binary treatment (equal to 1 if at least one CE activity is carried out and 0 if none of the CE activities is implemented within the firm) and y_i is the observed outcome, therefore representing the percentage points by which the company's turnover has increased or decreased over the last three years. Each one of the potential outcomes is determined by its expected value conditional of a regressor vector $\mathbf{x}_i$ and an unobserved random component ϵ_{ij}, for j $\in$ {0,1}. The same specification was adopted for the treatment equation (3) which is given by its expectation conditional on a set of regressors $\mathbf{z}_i$ –

which, according to a methodological perspective are not required to differ from $\mathbf{x}_i$ – and an unobserved component v_i. The equation (6) explicitly takes into consideration endogeneity within the model specification.

We included as covariates in the models: NACE sector (Outcome Model, OM, and Treatment Model, TM), company establishment year (OM), country where the company is located (TM), the percentage of the company's turnover invested in Research and Development (TM), types of activities carried out (OM), size of the company as expressed in terms of number of employees (TM) as well the unemployment rate (year 2016) and the related compound annual average rate of change (years 2013-2015) at national level (OM) as contextual territorial variables.

3 Preliminary Results and Conclusion

As a general descriptive result, it emerged that approximately three quarters of companies (73%) carried out at least one CE related activity among the activities listed in the questionnaire prepared by the EC, with almost all companies located in Malta (approximately 95%) which has implemented at least one of these CE activities, followed by 89% in Ireland and 85% in Luxembourg and Spain, compared to 44% in Bulgaria and Estonia and 47% in Lithuania (European Commission, 2016).

By focusing on the specific activities implemented, it was found that the majority of companies (55%) mentioned minimising waste by recycling, reusing or selling it to another company. Follows the activity of re-planning energy usage to minimise consumption implemented by approximately 38% of the companies. Moreover, about a third of companies (34%) have redesigned products and services to minimise the use of materials or use recycled materials. Further, almost 1 out of 10 companies is planning to redesign products and services for these reasons while approximately 1 out of 5 SMEs has re-planned the way water is used to minimise usage and maximise re-usage. Lastly, among of all the activities investigated, companies are least likely to be using renewable energy (16%).

Table 1 reports the estimated ATEs and ATETs. The existence of endogeneity between the two processes was confirmed by the Wald (endogeneity) test which lead us to reject the null hypothesis that treatment and outcome unobservable are uncorrelated (Chi-square(2) =35.20, p-value=0.0000).

Table 1: Estimation results: Endogenous treatment effect estimation

Endogenous treatment effect estimation	***Coef.***	***Robust SE***	***z***	***P > |z|***
ATE - Average Treatment effect				
Treatment (1=implemented CE activities *vs* 0=not implemented)	14.897	3.855	3.86	0.000
PO-mean - Treatment (0=not implemented CE activities)	-7.663	3.740	-2.05	0.040
ATET – Average Treatment Effect on the Treated				
ATET-Treatment	13.974	4.958	2.82	0.005

Endogenous treatment effect estimation	*Coef.*	*Robust SE*	*z*	*P > |z|*
(1=implemented CE activities *vs* 0=not implemented)				
Potential Outcome Mean - Treatment (0=not implemented CE activities)	-10.544	4.943	-2.13	0.033

Notes: outcome dependent variable: turnover increase/decrease in the last three years expressed in percentage points; treatment binary variable: decision of implementing at least one CE activities.

As an overall result, it was found that the average effect of the treatment in the population – $ATE = E(y_1 - y_0)$ – would be an average increase of the total turnover by approximately 14.9 percentage points, thus enforcing the importance of implementing CE initiatives, activities and practices within company management systems. On the other hand, the average treatment effect among those that received the treatment – i.e. the ATET estimated as the mean of the difference y_1 and y_0 among the subjects that actually receive the treatment – revealed that the companies' turnover would lead to a decrease (observed considering the last three years) of approximately 10.5 percentage points if none of these SMEs carried out any CE activities. Additionally, for those companies which carried out at least one CE activity, the average increase in the company's total turnover is approximately 14 percentage points greater than if none of these companies carried out any CE activities.

These preliminary results focused on the estimates concerning the overall decisions of implementing at least one CE activity. Further estimations focusing on specific CE single-activities foreseen in the FE-441 survey, as well as separate models distinguishing for turnover thresholds in order to explore equal access of companies to circularity practices are ongoing, with the aim of effectively identifying those factors which can support the importance of a sustainable management system as priority in the firms' administration and growth.

References

1. D'Agostino, R. B. (2007). Estimating treatment effects using observational data. Jama, 297(3), 314-316.
2. European Commission (2016) European SMEs and the Circular Economy – FE441. Final report.
3. European Commission (2019) Report from the Commission to the European Parliament (…) on the implementation of the Circular Economy Action Plan, Brussels, 4.3.2019 COM(2019) .
4. Kirchherr, J., Reike, D., & Hekkert, M. (2017). Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, conservation and recycling*, 127, 221-232.
5. Özbuğday, F. C., Fındık, D., Özcan, K. M., & Başçı, S. (2020). Resource efficiency investments and firm performance: Evidence from European SMEs. *Journal of Cleaner Production*, 252, 119824.
6. Principato, L., Ruini, L., Guidi, M., & Secondi, L. (2019). Adopting the circular economy approach on food loss and waste: The case of Italian pasta production. *Resources, Conservation and Recycling*, *144*, 82-89.
7. Ruggieri, A., Braccini, A. M., Poponi, S., & Mosconi, E. M. (2016). A meta-model of inter-organisational cooperation for the transition to a circular economy. Sustainability, 8(11), 1153.
8. Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.
9. Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. Journal of Econometrics, 182(1), 226-234.
10. Zamfir, A. M., Mocanu, C., & Grigorescu, A. (2017). Circular economy and decision models among European SMEs. Sustainability, 9(9), 1507.

Financial Spillover Measures to Assess the Stability of Basket-based Stablecoins

Misure di Spillover Finanziari per Valutare la Stabilità di Basket-based Stablecoins

Paolo Pagnottoni

Abstract The paper aims to assess, from an empirical viewpoint, the advantages of a stablecoin whose value is derived from a basket of underlying currencies, against a stablecoin which is pegged to the value of one major currency, such as the dollar. To this aim, we first find the optimal weights of the currencies that can comprise our basket. We then employ variance decomposition methods to understand which foreign currency mostly drives the others. We then look at how the stability of either stablecoin is affected by currency shocks. Our empirical findings show that our basket based stablecoin is less volatile than all single currencies. This results is fundamental for policy making, and especially for emerging markets with a high level of remittances: a librae (basket-based stablecoin) can preserve its value during turbolent times better than a libra (single currency based stablecoin).

Abstract *Questo studio ha l'obiettivo di individuare, da un punto di vista empirico, i vantaggi di uno stablecoin derivato da un paniere di valute, confrontato con uno stablecoin legato al valore di una valuta di punta, come il dollaro. A questo proposito, troviamo prima i pesi ottimali delle valute che compongono il nostro paniere. Impieghiamo poi metodi di decomposizione della varianza per capire quale delle valute traina maggiormente l'andamento delle altre. I nostri risultati empirici mostrano che il nostro stablecoin fondato su un paniere di valute è meno volatile di tutte le singole valute. Questi risultati sono fondamentali per il policy making, e specialmente per i mercati emergenti con un alto livello di rimesse: un "librae" (stablecoin basato su un paniere di valute) può preservare il suo valore durante periodi turbolenti meglio di un "libra" (stablecoin basato su una singola valuta).*

Key words: Bitcoin; Forecast error variance decomposition; Market linkages; Market risk; Spillovers; Vector autoregression; Vector error correction

---

Paolo Pagnottoni
University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: paolo.pagnottoni01@universitadipavia.it

1 Introduction

Many researchers posed the question of whether a Synthetic Hegemonic Currency (SHC) would be best provided by the public sector. The rationale would be that a global currency, underpinned by a basket of reserve assets, could better support global outcomes. For example, an SHC could dampen the dominating influence of the US dollar on global trade, it could alleviate spillovers to exchange rates from shocks to the US economy, and trade across countries could become less dependent on the dollar.

The revival of discussions concerning an SHC, have somewhat been sparked by the discourse surrounding central bank digital currency (CBDC) and stablecoins. In particular, Facebook announced plans for its own privately issued stablecoin that could emulate some of the characteristics of an SHC. The proposition is to construct a stablecoin that can circulate globally with a value that is derived from an underlying basket of assets comprised of the major currencies. Whilst the exact composition of the underlying basket of assets is yet unspecified, the objective is to devise a digital currency whose exchange rate fluctuations are minimised against several currencies. These plans have been met with resistance from regulators and Facebook itself has repeatedly stated that the Libra stablecoin could be backed by a single currency (the dollar).

Against this background, we investigate the consequences of a global SHC ("Librae", in a literal sense), regardless of whether issued by a private company such as Facebook, or by a central bank. In particular, we first look at the optimal design of an SHC that is backed by a basket of underlying reference currencies, such as those included in the International Monetary fund Special Drawings Rights (SDRs). We then study the currencies which mostly determine price change spillovers among exchange rates, using the framework of (3). For the optimal construction of a basket of currencies, we follow (1) to compute a minimum variance currency basket using major currencies. We construct a reference basket that contains the Dollar (USD), the Euro (EUR), the Yen (JPY), the Renminbi (CNY) and the Pound Sterling (GBP). By construction, our basket based currency should be the least varying in comparison to those contained in the basket, and our results confirm this.

Our price change spillover decomposition shows that the dollar is the currency that has the largest impact on the others, especially in terms of exporting contagion. As a consequence, a negative shock on the dollar causes a shock on all currencies and, through high order contagion, on the dollar itself, leading to a new lower equilibrium. Differently, a shock in the value of the SHC, caused by a shock of a currency in the basket, is offset by the diversification effect and, therefore, the starting equlibrium is maintained. This implies that remittances converted in basket based stablecoin better maintain their value, with respect to those converted in dollars (or dollar based stable coins).

2 Methodology

We aim to build a basket of predetermined (reference) currencies with optimal weights, namely, weights which minimize the variability of a basket based stablecoin. This translates into an optimal control problem which minimizes the variance of the basket constructed with the above mentioned currencies. We follow (1) to build the reduced normalized values (RNVALs) associated to the currencies as well as the above mentioned basket.

We evaluate spillovers through the methodology by (2). As in their seminal paper, we start from estimating a Vector AutoRegressive (VAR) model which we convert into its corresponding vector moving average (VMA) representation, that is

$$x_t = \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \Psi_2 \varepsilon_{t-2} + \cdots \tag{1}$$

where x_t being the $(n \times 1)$ vector of first differences in RNVALs at time t, ε_t a zero-mean white noise process having variance-covariance matrix Σ_ε, $\Psi_1, \Psi_2, ...$ the $(n \times n)$ are the matrices of VMA coefficients. The VMA coefficients are recursively computed as $\Psi_i = \Phi_1 \Psi_{i-1} + \Phi_2 \Psi_{i-2} + ... + \Phi_i \Psi_1$, having $\Psi_i = 0 \; \forall i < 0$ and $\Psi_1 = I_n$.

(2) build their methodology on the KPPS H-step ahead forecast error variance decomposition. Considering two generic variables x_i and x_j, they define the own variance shares as the proportion of the H-step ahead error variance in predicting x_i due to shocks in x_i itself, $\forall i = 1, ..., n$. On the other hand, the cross variance shares (spillovers) are defined as the H-step ahead error variance in forecasting x_i due to shocks in x_j, $\forall i = 1, \cdots, n$ with $j \neq i$. In other words, denoting as $\theta^g_{ij}(H)$ the KPPS H-step forecast error variance decompositions, with $h = 1, \cdots, H$, we have:

$$\theta^g_{ij}(H) = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} (e_i' \Psi_h \Sigma e_j)^2}{\sum_{h=0}^{H-1} (e_i' \Psi_h \Sigma \Psi_h' e_i)} \tag{2}$$

with σ_{jj} being the standard deviation of the innovation for equation j and e_i the selection vector, i.e. a vector having one as i^{th} element and zeros elsewhere. Intuitively, the own variance shares and cross variance shares (spillovers) measure the contribution of each variable to the forecast error variance of itself and the other variables in the system, respectively, thus giving a measure of the importance of each variable in predicting the others.

Note that the row sum of the generalized variance decomposition is not equal to 1, meaning $\sum_{h=0}^{H-1} \theta^g_{ij}(H) \neq 1$. (2) circumvent this problem by normalizing each entry of the variance decomposition matrix by its own row sum, i.e.

$$\tilde{\theta}^g_{ij}(H) = \frac{\theta^g_{ij}(H)}{\sum_{j=1}^{n} \theta^g_{ij}(H)} \tag{3}$$

We make use of directional spillovers indexes (DSI) to measure the spillover from exchange i to all other exchanges J (cfr. Eq. 4) and the spillover from all exchanges J to exchange i (cfr. Eq. 5) as:

$$DSI_{J\leftarrow i}(H) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} \tilde{\theta}^{g}_{ji}(H)}{\sum_{j,i=1}^{n} \tilde{\theta}^{g}_{ij}(H)} \cdot 100 \tag{4}$$

$$DSI_{i\leftarrow J}(H) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} \tilde{\theta}^{g}_{ij}(H)}{\sum_{j,i=1}^{n} \tilde{\theta}^{g}_{ij}(H)} \cdot 100 \tag{5}$$

From the definitions of directional spillover indexes, it is natural to build a net contribution measure, impounded in the net spillover index (NSI) from market i to all other markets J, namely:

$$NSI_i(H) = DSI_{J\leftarrow i}(H) - DSI_{i\leftarrow J}(H) \tag{6}$$

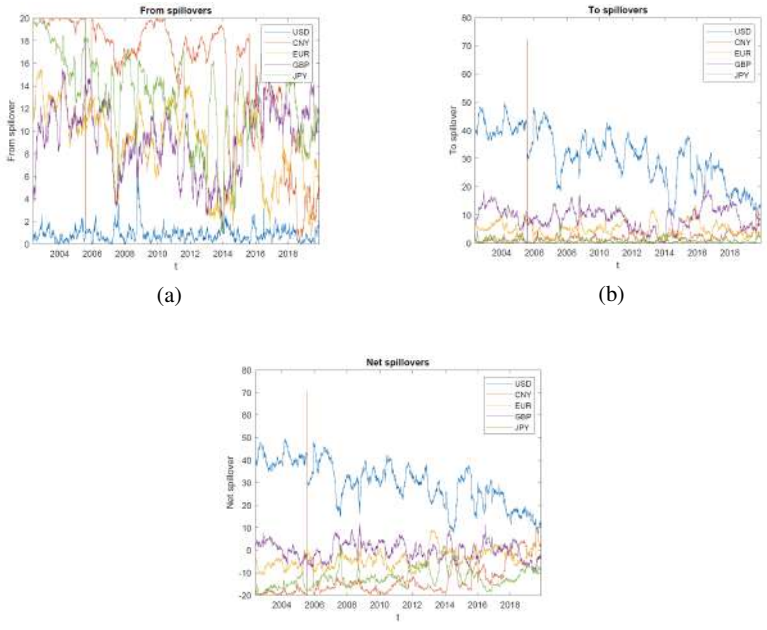
All the metrics discussed above are able to yield insights regarding the mechanisms of market exchange spillovers both from a system-wide and a net pairwise point of view. Furthermore, performing the analyses on rolling windows we are able to study the dynamics of spillover indexes over time.

3 Data and Empirical Findings

To test our proposal, we make use of historical data, according to a retrospective analysis. In particular, we use daily foreign exchange rate data over the period 1 January 2002 - 30 November 2019. To build our optimal basket of currencies, we collect data relative to the foreign exchange pairs between the currencies that are included in the IMF's Special Drawings Rights: the US dollar, the Chinese Renmimbi, the Euro, the British pound and the Japanese Yen.

We now consider spillovers between exchanges, to evaluate the price change connectedness of the currencies that compose the basket, and to understand which is the relative importance of each of the currencies in transmitting shocks. In this way, we are also able to determine which currencies potentially cause strong (or weak) price changes in our proposed stablecoin value. Dynamic directional spillovers can shed light on which of the currencies transmit price change spillovers to others and which of them receive price change spillovers from others. We plot from, to, net and pairwise spillovers in Figure 1.

We can highlight that that USD is the most influential currencies in terms of return spillovers. Indeed, the magnitude of spillovers received from others is weak



(a) (b)

(c)

Fig. 1: From, to and net spillovers

compared to that transmitted to others. Moreover, the net spillover dynamics summarizes the dominant position of the USD, being it always positive and taking relatively high values over the sample period. However, the magnitude of spillovers transmitted by USD follows a negative trend over time, meaning the currency is gradually losing its potentiality to contribute to the evolution of the others, perhaps due to the affirmation of emerging economies in the latter period, especially after the 2009 crisis. Despite that, the latter considerations are in line with the full sample results obtained above, which point to the dominance of USD as a spillover transmitting currency.

The dynamic analysis shows that CNY is not such a leading currency in transmitting price change shocks. Indeed, the full sample result is arguably driven by a noticeable spike which occurred on 21 July 2005. Indeed, during that day the Chinese Central Bank officially announced the abandonment of the eleven-year-old peg to the dollar and pegged the CNY to a basket of currencies whose composition was not disclosed. This caused a prompt revaluation to CNY 8.11 per USD, as well as to 10.07 CNY per Euro. However, the peg to the dollar was reinstituted as the financial crisis strengthened in July 2008. These results indicate that CNY does not particularly contribute to the price change evolution of the other currencies in the basket, although it can exert shocks through sudden policy decisions.

4 Conclusion

In the paper we present a methodology to build a basket based stable coin whose weights can maximise stability over a long time period. The proposed stable coin (Librae) appears to be less volatile than single currencies and, therefore, with respect to single currency stable coins (Libra). It can thus constitute a valuable proposal especially for workers who live abroad and make remittances to their own country, a market segment with a high potential of being attracted by payments in stablecoins.

We have also proposed a variance decomposition technique based on a VAR model, aimed at showing which currencies mostly impact the Foreign Exchange market and whether a single currency or a basket based stablecoin is more resilient to currency shocks. Our results show that the dollar is the currency which mostly impact the market, and that a basket based coin is better than a dollar based one, from a stability and value maintenance viewpoint.

With a basket based stablecoin it is possible to offset the risk of currencies shocks. This is of relevance for different policy purposes and, in particular, for emerging markets and countries having high remittances. Indeed, by holding stablecoins rather than single currencies the risks associated to currency shocks are mitigated and stablecoins holder can count on a currency whose value is less volatile than traditional fiat currencies and, thereby, more reliable. The latter fact has also positive consequences on cross-border payments side, provided that the stability of the stablecoin mitigates the foreign exchange risk, thus contributing to the fact that buyers and sellers give or receive an amount of money whose value is less sensitive to variations over time.

References

[1] Hovanov, N., Kolari, J., Sokolov, M.: Computing currency invariant indices with an application to minimum variance currency baskets. Journal of Economic Dynamics and Control, 28(8), 1481-1504 (2004)

[2] Diebold, F., and K. Yilmaz: Better to give than to receive: Predictive directional measurement of volatility spillovers. International Journal of Forecasting, 28(1), 57-66 (2012)

[3] Diebold, F. and K. Yilmaz: On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms. Journal of Econometrics 182 (1), 119-134 (2014)

Forecasting Banknote Flows in BdI Branches: Speed-up with Machine Learning

Forecast dei Flussi di Banconote nelle Filiali della Banca d'Italia con metodi di Machine Learning

Brandi Marco, Fusaro Monica, Laureti Tiziana and Rocco Giorgia

Abstract Every day banknotes are withdrawn from and deposited to Banca d'Italia branches by banks. Up to 455 time series need to be predicted simultaneously. In order to find a model with the best trade-off between forecasting performance and time consumption, a comparison between different machine learning approaches is presented.

Abstract *Le banche prelevano e depositano quotidianamente banconote presso le filiali della Banca d'Italia. É necessario prevedere contemporaneamente fino a 455 serie storiche. Viene presentato un confronto tra diversi modelli di machine learning in termini di previsione e di tempo di esecuzione, con lo scopo di trovare un modello con il miglior compromesso tra i due aspetti.*

Key words: Forecast, time series, seasonality, machine learning, bagging, random forest, neural network

---

Brandi Marco
Currency Circulation Management Directorate, Banca d'Italia, Rome e-mail: marco.brandi@bancaditalia.it

Fusaro Monica
Currency Circulation Management Directorate, Banca d'Italia, Rome e-mail: monica.fusaro@bancaditalia.it

Laureti Tiziana
Department of Economics, Engineering, Society and Business Organization, University of Tuscia, Viterbo e-mail: laureti@unitus.it

Rocco Giorgia
Currency Circulation Management Directorate, Banca d'Italia, Rome e-mail: giorgia.rocco@bancaditalia.it

The views expressed in this article are those of the authors and do not involve the responsibility of Banca d'Italia.

1 Introduction

In Italy, the provision of banknotes relies on a complex supply chain (cash cycle) [16]. Commercial banks and the postal system provide cash to the public through their branches and ATMs. Every day, these stakeholders request banknotes from the branches of Banca d'Italia, which must guarantee their availability to fully satisfy demand; at the same time, they deposit unfit banknotes and excess stocks that have to be processed by the branches to retire unfit banknotes from circulation. Therefore, branches must be supplied with new banknotes in time to satisfy cash handlers' requirements and at the same time they should be streamlined in case of excessive stocks.
In this framework, Banca d'Italia needs to plan the transfers from the central hub to the branches and back beforehand, to guarantee security standards. This requires a forecast of the deposited and withdrawn banknotes for each branch and denomination. Furthermore, the forecast needs to be updated every week.
Daily withdrawals (W) and deposits (D) data for each branch and denomination are aggregated, due to the weekly nature of banknote transfers. Weekly time series refer to 35 branches (distributed in all Italian regions), 7 denominations for deposits (€5 - €500), and 6 for withdrawals (€5 - €200).[1] Then, a total of 455 time series should be analyzed. Data from January 2009 to June 2019 for a total of 546 observations are used for each time series. Considering the data-intensive nature of the problem, the use of classical forecasting approaches implicates a high degree of computational complexity, as well as the need to verify many assumptions for the model. Machine learning approaches for forecasting time series are explored to overcome these issues since they relax assumptions, such as the distribution form, and link the dependent variable to the covariates using an unknown distribution form [5]. Moreover, machine learning is a data-driven approach oriented to predict and forecast, which is able to pick up non-linear relationships. It also facilitates the implementation of ensemble methods, which generally outperform single methods [15, 8]. The main disadvantage of machine learning methods resides in the fact that their output is often hard to be interpreted. Given the focus in forecasting of the study, this drawback is not a concern. Time series problems are more difficult to handle with machine learning algorithms, due to the time dependency between the response variables. The forecasting procedure has been adapted in this respect.

2 Methodology

As discussed above, machine learning algorithms are not easily implemented to treat time series due to the time dependency of the data, but there are several examples of models that can be used for time series forecasting [2, 9, 10]. In this work, a

[1] The Eurosystem' national central banks stopped the issuance of €500, with the exception of Germany and Austria which stop issuing the €500 at the end of April 2019.

forecast procedure is provided starting from the work of Laurinec [14], who uses regression trees in order to forecast electricity consumption with double seasonality modeling. Each of the 455 time series is treated individually, whereas the two flows of banknotes W and D are modeled equally. Let Y_t be an observation at time t (with $t = 1,\ldots,546$) for a single time series referred to one of the two flows for a denomination in a branch.
First, Y_t has been decomposed in trend, seasonal and remainder components, using nonparametric regression LOESS (STL) introduced in [7] as follows:

$$Y_t = T_t + S_t + R_t \ , \tag{1}$$

Then, the forecast is obtained by estimating separately the two components:

$$\hat{Y}_{t+h} = \hat{T}_{t+h} + \hat{\tilde{Y}}_{t+h} \ , \tag{2}$$

1. The trend component $\hat{T}_{t+h}$ is estimated using simple ARIMA models with automatic parameter selection as proposed by [13];
2. The detrended component $\hat{\tilde{Y}}_{t+h}$ is obatained applying machine learning algorithms to the detrended time series $\tilde{Y}_t = S_t + R_t$.

In order to handle time series with machine learning, an extraction of time-based features is necessary. The following explanatory variables are therefore considered:

- Seasonal variable S_t: estimated within the decomposition of Y_t, it is related to the week of the year and it is constant between years;
- Lag variables to $t-4$: $\tilde{Y}_{t-1}, \tilde{Y}_{t-2}, \tilde{Y}_{t-3}, \tilde{Y}_{t-4}$;
- Fourier coefficients: $a_i \cos(\omega_i t)$, $b_i \sin(\omega_i t)$ for $i = 1,\ldots,K$ as suggested in [18] in order to model seasonality, where ω_i is the Fourier basis;
- Dummy variable indicating Easter week (variable across years) E_t.

The relationship between the response variable and the explanatory variables can be synthesized by the following formula:

$$\tilde{Y}_t = f\left(S_t, \tilde{Y}_{t-1}, \tilde{Y}_{t-2}, \ldots, a_1, b_1, \ldots, E_t\right) \ , \tag{3}$$

where the link function f can be estimated using machine learning algorithms.
The final forecast is obtained using a MIMO (*Multi Input Multi Output*) strategy, which can potentially improve the performance of the forecast [1]. Furthermore, this strategy is able to preserve the stochastic dependency among the predicted values characterizing the time series. However, preserving the stochastic dependencies constrains all the horizons to be forecast with the same model structure, reducing the flexibility of the forecast approach. With this approach the computational time is lower than other existing strategies.

2.1 Machine learning algorithms

Machine learning methods used in this work mainly include tree-based algorithms as decision trees (*CART*) [6] and conditional inference trees (*CTREE*) [11].
In order to improve the forecast, ensemble methods are considered, since they combine several decision trees to produce better predictive performance. The underlying idea is that a group of weak learners gather to produce a strong learner, hence increasing the accuracy of the model although losing interpretability. In this study, bagging and random forest algorithms are considered. The bagging algorithm (bootstrap aggregating: *BCART* and *BCTREE*) [3] takes the average of the predictions obtained by several decision trees (or conditional inference trees) applied to different bootstrap samples of the original one. Random forests (RF) are a combination of decision trees where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [4]. The splits in each tree allow evaluating the variable importance in predicting the values.
Additionally, the most commonly used artificial neural networks (ANN) [17] is performed. The multilayer perceptrons consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function and the backpropagation is used for training the algorithm.
The seasonal ARIMA (SARIMA) model with automatic parameter selection [13] is considered as a benchmark model to compare the results of machine learning with a classical approach.

2.2 Evaluating forecasting performance

In order to evaluate the forecasting performance of the different algorithms, each time series is preliminarily divided into train and several test sets. Specifically, to evaluate the forecast performance at different time horizons, the test set is considered adding one week each time, so that the dimensions of the different test sets are equal to $h = 1,\ldots,52$ up to an entire year of observations. Therefore, the forecast performances are evaluated using the MASE (*Mean Absolute Scaled Error*), which is scale-independent, proposed by [12] to compare different methods across different series. Let $e_t = Y_t - \hat{Y}_t$ be the forecast error, a scaled error is defined as follows:

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}. \tag{4}$$

Mean absolute scaled error is then defined as $MASE = \mathbb{E}(|q_t|)$. Furthermore, the time consumption is evaluated considering the time of execution of each single algorithm on all the time series.[2]

[2] All the algorithms were executed on a PC with processor INTEL(R) Core(TM) i5-4300U 1.90GHz and 8,00 GB RAM. R version 3.6.2

3 Results

The results in terms of MASE and time execution are analyzed among the different models. (Fig. 1). In terms of forecast performance, machine learning ensemble methods outperform single methods. The distribution of MASE for random forest and bagging methods is more concentrated than for SARIMA with RF presenting the lowest variability. RF has a MASE lower than the SARIMA one in above 60% of time series forecast and only 5% of cases have a relevant greater MASE than SARIMA.

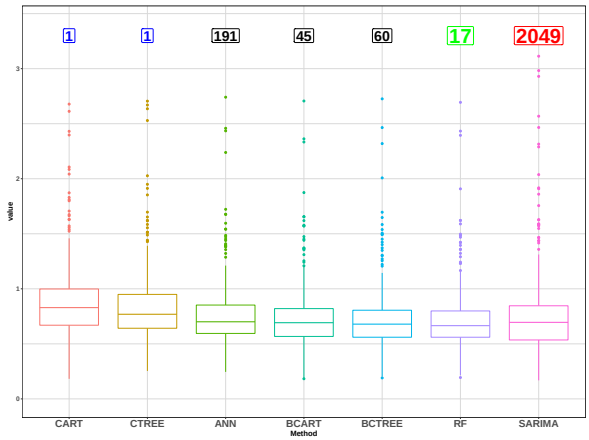


Fig. 1 Distribution of MASE for different methods. Execution time in minutes is reported in the upper part of the figure.

Time of execution is also considered (top of Fig. 1). Although the single decision tree models have the lowest execution time (1 minute), RF looks very promising with 17 minutes, especially compared to SARIMA, which takes up to 34 hours to run for all the time series.
The random forest algorithm is the model with the best trade-off between accuracy and time of execution.

4 Conclusions

This study considers the problem of forecasting a high number of time series. The forecast for each single time series is investigated and machine learning approaches are exploited. Even considering auto-selection of the parameters, classical approaches, with many assumptions that need to be verified, are not the most

suitable solution for forecasting such a high number of time series. Moreover, the predictive performance of machine learning methods generally outperforms, or at least matches, that of classical methods, especially in the case of random forests algorithms.
Machine learning methods are very time-saving with respect to the classical approach. Lower time of execution is the main advantage, considering the applied and operative nature of the problem.
Machine learning algorithms also allow to include more variables and the models could be improved moving from Easter dummy variable to a special dummy variable in order to include local holidays or events at each of the Banca d'Italia branches. A formal selection of variable lags could also be included in the training of machine learning algorithms.

References

1. S. Ben Taieb, G. Bontempi, A. F. Atiya, and A. Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert Syst. Appl.*, 39(8):7067–7083, June 2012.
2. G. Bontempi, S. Ben Taieb, and Y. Le Borgne. *Machine Learning Strategies for Time Series Forecasting*, pages 62–77. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
3. L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
4. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.
5. L. Breiman. Statistical modeling: the two cultures. *Statist. Sci.*, 16(3):199–231, 2001. With comments and a rejoinder by the author.
6. L. Breiman et al. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.
7. R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73, 1990.
8. T. G. Dietterich. Ensemble methods in machine learning. In *MULTIPLE CLASSIFIER SYSTEMS, LBCS-1857*, pages 1–15. Springer, 2000.
9. T. G. Dietterich. Machine learning for sequential data: A review. pages 227–246. 2002.
10. T. Fischer, C. Krauss, and A. Treichel. Machine learning for time series forecasting - a simulation study. FAU Discussion Papers in Economics 02/2018, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics, 2018.
11. T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.
12. R. Hyndman. Another look at forecast accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, 4:43–46, 01 2006.
13. R. Hyndman and Khandakar. Y. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, Articles*, 27(3):1–22, 2008.
14. P. Laurinec. *Using Regression Trees for Forecasting Double-Seasonal Time Series with Trend in R.*, 2017. https://petolau.github.io/Regression-trees-for-forecasting-time-series-in-R/.
15. R. Maclin and D. W. Opitz. Popular ensemble methods: An empirical study. *CoRR*, abs/1106.0257, 2011.
16. G. Rocco, R. Rinaldi, M. Calderini, A. Longo, and F. Sasso. The cash cycle in Italy: trends, players and strategies. *Bancaria*, 12:22–29, 2019.
17. W. Sarle. Neural networks and statistical models, 1994.
18. P. C. Young, D. J. Pedregal, and W. Tych. Dynamic harmonic regression. *Journal of Forecasting*, 18(6):369–394, 1999.

Fully reconciled GDP forecasts from Income and Expenditure sides

Previsioni riconciliate del PIL dal lato del reddito e della spesa

Luisa Bisaglia, Tommaso Di Fonzo and Daniele Girolimetto

Abstract We propose a complete reconciliation procedure, resulting in a 'one number forecast' of the *GDP* figure, coherent with both Income and Expenditure sides' forecasted series, and evaluate its performance on the Australian quarterly *GDP* series, as compared to the original proposal by Athanasopoulos *et al.* (2019).
Abstract *In questo lavoro viene proposta una procedura di riconciliazione delle previsioni del PIL e delle sue componenti tanto dal lato del Reddito quanto da quello della Spesa, volta a produrre previsioni coerenti rispetto ad entrambi i lati. Tale procedura, applicata alle serie trimestrali del PIL australiano, viene posta a confronto con la proposta originale di Athanasopoulos et al. (2019).*

Key words: forecast reconciliation, cross-sectional (contemporaneous) hierarchies, GDP, Income, Expenditure

1 Introduction and summary

In a recent paper, Athanasopoulos *et al.* (2019, p. 690) propose "the application of state-of-the-art forecast reconciliation methods to macroeconomic forecasting" in order to perform aligned decision making and to improve forecast accuracy. In their empirical study they consider 95 Australian Quarterly National Accounts time series, describing the Gross Domestic Product (*GDP*) at current prices from Income and Expenditure sides, interpreted as two distinct hierarchical structures. In the former case (Income), *GDP* is on the top of 15 lower level aggregates (figure 1), while in the latter (Expenditure), *GDP* is the top level aggregate of a hierarchy of 79 time series (see figures 21.5-21.7 in Athanasopoulos *et al.*, 2019, pp. 703-705).

In this paper we re-consider the results of Athanasopoulos *et al.* (2019), where the forecasts of the Australian quarterly *GDP* aggregates are separately reconciled from Income ($\widetilde{GDP}^{I}$) and Expenditure ($\widetilde{GDP}^{E}$) sides. This means that $\widetilde{GDP}^{I}$ and

L. Bisaglia
Dept. Statistical Sciences, University of Padova, e-mail: luisa.bisaglia@unipd.it

T. Di Fonzo
Dept. Statistical Sciences, University of Padova, e-mail: tommaso.difonzo@unipd.it

D. Girolimetto
Dept. Statistical Sciences, University of Padova, e-mail: daniele.girolimetto@studenti.unipd.it

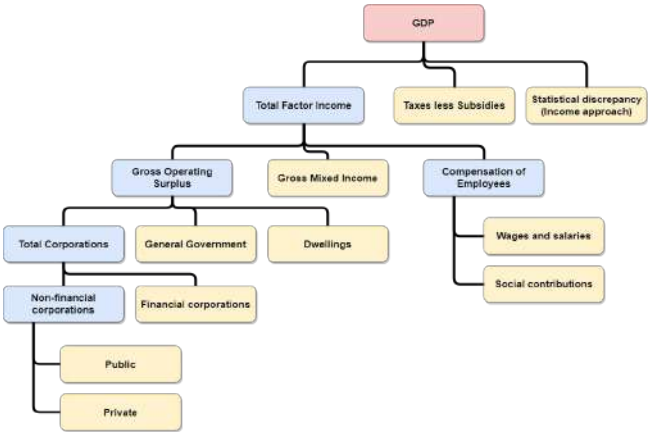


Fig. 1 Hierarchical structure of the income approach for Australian GDP. The pink cell contains the most aggregate series. The blue cell contain intermediate-level series and the yellow cells correspond to the most disaggregate bottom-level series. Source: Athanasopoulos *et al.*, 2019, p. 702.

$\widetilde{GDP}^{E}$ are each coherent within its own pertaining side with the other forecasted values, but in general $\widetilde{GDP}^{I} \neq \widetilde{GDP}^{E}$ at any forecast horizon. This circumstance could confuse and annoy the user, mostly when the discrepancy is not negligible (see Figure 2), and calls for a complete reconciliation strategy, able to produce a 'one number forecast' of the *GDP* figure, which is the main target of the paper.

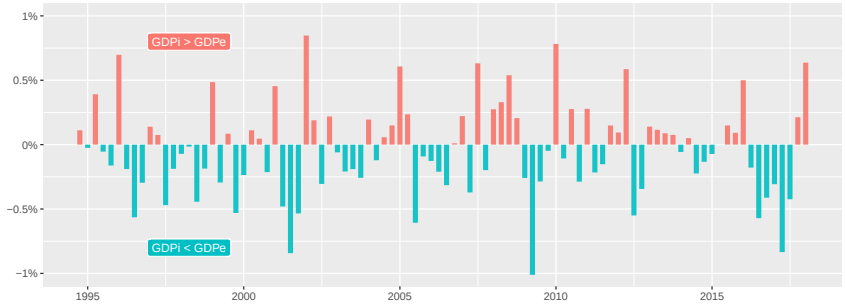


Fig. 2 Discrepancies in the reconciled 1-step-ahead *GDP* forecasts from Income and Expenditure sides. ARIMA base forecasts reconciled according to MinT-shr procedure (see section 3). Source data: Athanasopoulos *et al.* (2019).

We show that fully reconciled forecasts of *GDP*, coherent with all the reconciled forecasts from both Expenditure and Income sides, can be obtained through the classical least squares adjustment procedure proposed by Stone *et al.* (1942). It should be noted that the proposed solution has been considered by van Erven and Cugliari (2015) and Wickramasuriya *et al.* (2019) as an alternative formulation, equivalent to the regression approach by Hyndman *et al.* (2011). As far as we know, however, it has never been applied so far to distinct hierarchies sharing only the top level series. The procedure can be seen as a forecast combination (Bates and Granger, 1969) -

working on different series rather than on the output of multiple models - which makes additional use of external constraints valid for the series and their forecasts.

2 From single side to complete aggregation constraints

Denoting with x_t the actual *GDP* at time t, the relationships linking the series of, respectively, the Income and Expenditure sides hierarchies can be expressed as

$$\mathbf{y}_t^I = \mathbf{S}^I \mathbf{b}_t^I, \qquad \mathbf{y}_t^E = \mathbf{S}^E \mathbf{b}_t^E, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{y}_t^I = \begin{bmatrix} x_t & \mathbf{a}_t^{I\prime} & \mathbf{b}_t^{I\prime} \end{bmatrix}'$, $\mathbf{y}_t^E = \begin{bmatrix} x_t & \mathbf{a}_t^{E\prime} & \mathbf{b}_t^{E\prime} \end{bmatrix}'$, $\mathbf{b}_t^I$ and $\mathbf{b}_t^E$ are (10×1) and (53×1), respectively, vectors of bottom level (disaggregated) series, $\mathbf{a}_t^I$ and $\mathbf{a}_t^E$ are (5×1) and (26×1), respectively, vectors of higher levels (aggregated) series, and

$$\mathbf{S}^I = \begin{bmatrix} \mathbf{1}_{10}' \\ \mathbf{C}^I \\ \mathbf{I}_{10} \end{bmatrix}, \qquad \mathbf{S}^E = \begin{bmatrix} \mathbf{1}_{53}' \\ \mathbf{C}^E \\ \mathbf{I}_{53} \end{bmatrix}$$

are contemporaneous (cross-sectional) summing matrix mapping the bottom level series into the higher-levels variables in each hierarchy, where $\mathbf{1}_k$ denotes a $(k\times 1)$ vector of ones, $\mathbf{I}_k$ denotes the identity matrix of order k, and $\mathbf{C}^I$ and $\mathbf{C}^E$ are the (5×10) and (26×53), respectively, matrices of 0's and 1's describing the aggregation relationships between the bottom level series and the higher level series (apart *GDP*) for Income ($\mathbf{C}^I$) and Expenditure ($\mathbf{C}^E$) sides. The relationships (1) can be equivalently written as

$$\mathbf{U}^{I\prime}\mathbf{y}_t^I = \mathbf{0}, \qquad \mathbf{U}^{E\prime}\mathbf{y}_t^E = \mathbf{0}, \quad t = 1, \ldots, T, \tag{2}$$

where $\mathbf{U}^I = \begin{bmatrix} \mathbf{I}_6 \\ -\mathbf{1}_{10} \; -\mathbf{C}^{I\prime} \end{bmatrix}$, and $\mathbf{U}^E = \begin{bmatrix} \mathbf{I}_{27} \\ -\mathbf{1}_{53} \; -\mathbf{C}^{E\prime} \end{bmatrix}$ are (16×6) and (80×27) matrices, respectively. The only variable subject to linear constraints on both the Income and Expenditure sides in expressions (1) and (2) being x_t (i.e., *GDP*), we can express the aggregation relationships linking the 95 'unique' variables as

$$\mathbf{U}'\mathbf{y}_t = \mathbf{0}, \quad t = 1, \ldots, T, \tag{3}$$

where $\mathbf{y}_t = \begin{bmatrix} x_t & \mathbf{a}_t^{I\prime} & \mathbf{b}_t^{I\prime} & \mathbf{a}_t^{E\prime} & \mathbf{b}_t^{E\prime} \end{bmatrix}'$ is a (95×1) vector , $\mathbf{0}$ is a (33×1) null vector, and $\mathbf{U}'$ is the following (33×95) matrix:

$$\mathbf{U}' = \begin{bmatrix} 1 & \mathbf{0}_5' & -\mathbf{1}_{10}' & \mathbf{0}_{26}' & \mathbf{0}_{53}' \\ 1 & \mathbf{0}_5' & \mathbf{0}_{10}' & \mathbf{0}_{26}' & -\mathbf{1}_{53}' \\ \mathbf{0}_5 & \mathbf{I}_5 & -\mathbf{C}^I & \mathbf{0}_{5\times 26} & \mathbf{0}_{5\times 53} \\ \mathbf{0}_{26} & \mathbf{0}_{26\times 5} & \mathbf{0}_{26\times 10} & \mathbf{I}_{26} & -\mathbf{C}^E \end{bmatrix}. \tag{4}$$

3 Optimal point forecast reconciliation

Forecast reconciliation is a post-forecasting process aimed at improving the quality of the *base* forecasts for a system of hierarchical/grouped, and more generally linearly constrained, time series (Hyndman *et al.*, 2011, Panagiotelis *et al.*, 2019) by exploiting the constraints that the series in the system must fulfill, whereas in general the base forecasts don't. In this framework, as base forecasts we mean the $(n \times 1)$ vector $\hat{\mathbf{y}}_{T+h} \equiv \hat{\mathbf{y}}_h$ of unbiased point forecasts, with forecast horizon $h > 0$, for the $n > 1$ variables of the system.

Following Stone *et al.* (1942), we consider the classical measurement model

$$\hat{\mathbf{y}}_h = \mathbf{y}_h + \boldsymbol{\varepsilon}_h, \quad E(\boldsymbol{\varepsilon}_h) = \mathbf{0}, \quad E\left(\boldsymbol{\varepsilon}_h \boldsymbol{\varepsilon}_h'\right) = \mathbf{W}_h, \tag{5}$$

where $\hat{\mathbf{y}}_h$ is the available measurement, $\mathbf{y}_h$ is the target forecast vector, and $\boldsymbol{\varepsilon}_h$ is a zero-mean measurement error, with covariance $\mathbf{W}_h$, which is a $(n \times n)$ p.d. matrix, for the moment assumed known. Given a $(n \times K)$ matrix of constant values $\mathbf{U}$, summarizing the K linear constraints valid for the n series of the system ($n > K$), in general it is $\mathbf{U}'\hat{\mathbf{y}}_h \neq \mathbf{0}$, and we look for reconciled forecasts $\tilde{\mathbf{y}}_h$ such that $\mathbf{U}'\tilde{\mathbf{y}}_h = \mathbf{0}$.

The reconciled forecasts $\tilde{\mathbf{y}}_h$ can be found as the solution to the linearly constrained quadratic minimization problem:

$$\tilde{\mathbf{y}}_h = \underset{\mathbf{y_h}}{\arg\min} \left(\hat{\mathbf{y}}_h - \mathbf{y}_h\right)' \mathbf{W}_h^{-1} \left(\hat{\mathbf{y}}_h - \mathbf{y}_h\right), \quad \text{s.t. } \mathbf{U}'\mathbf{y}_h = \mathbf{0},$$

which is given by

$$\tilde{\mathbf{y}}_h = \left[\mathbf{I}_n - \mathbf{W}_h\mathbf{U}\left(\mathbf{U}'\mathbf{W}_h\mathbf{U}\right)^{-1}\mathbf{U}'\right]\hat{\mathbf{y}}_h. \tag{6}$$

The key item in expression (6) is matrix $\mathbf{W}_h$, which is generally unknown and must be either assumed known or estimated. In agreement with Athanasopoulos *et al.* (2019), denoting with $\widehat{\mathbf{W}}_1$ the $(n \times n)$ covariance matrix of the in-sample one-step-ahead base forecasts errors of the n series in the system, we consider 3 cases:

- OLS: $\mathbf{W}_h = \sigma^2\mathbf{I}_n$
- WLS: $\mathbf{W}_h = \widehat{\mathbf{W}}_D = \text{diag}\{\hat{w}_{11}, \ldots, \hat{w}_{nn}\}$
- MinT-shr: $\mathbf{W}_h = \widehat{\mathbf{W}}_{shr} = \lambda\widehat{\mathbf{W}}_D + (1-\lambda)\widehat{\mathbf{W}}_1$

where $\widehat{\mathbf{W}}_{shr}$ is the shrinked version of $\widehat{\mathbf{W}}_1$, with diagonal target and shrinkage intensity parameter λ proposed by Schäfer and Strimmer (2005) (more details can be found in Wickramasuriya *et al.*, 2019).

4 The accuracy of the reconciled forecasts of the Australian GDP

According to the notation of the previous section, for the complete Australian *GDP* accounts from both Income and Expenditure sides, it is $n = 95$, $K = 33$, and matrix $\mathbf{U}'$ is given by (4). In addition, the available time series span over the period 1984:Q1 - 2018:Q4.

Base forecasts for the $n = 95$ separate time series have been obtained by Athanasopoulos *et al.* (2019) through simple univariate ARIMA models[1], selected using the `auto.arima` function of the R-package `forecast`. We did not change this first, crucial step in the forecast reconciliation workflow, since the focus is on the potential of forecast reconciliation.[2]

Our reconciliation proposal is applied within the same forecasting experiment designed by Athanasopoulos *et al.* (2019). They consider forecasts from $h = 1$ quarter ahead up to $h = 4$ quarters ahead using an *expanding* window, where the first training sample is set from 1984:Q4 to 1994:Q3 and forecasts are produced for 1994:Q4 to 1995:Q3. The base forecasts are reconciled using OLS, WLS and MinT-shr procedures, and the accuracy is measured by the Mean Squared Error (MSE).

Figure 3 shows the *skill scores* using MSE, that is the percentage changes in MSE registered by each reconciliation procedure, relative to base forecasts, computed such that positive values signal an improvement in forecasting accuracy over the base forecasts. The left and the central columns of the figure refer to the results for the Income and Expenditure sides variables separately considered, while the right column shows the results of the procedure proposed in this paper.

The results confirm also for the enlarged system the findings of Athanasopoulos *et al.* (2019, p. 709):

- reconciliation methods improve forecast accuracy relative to base forecasts;
- negative skill scores are registered only for OLS-reconciled forecasts of bottom level series ($h = 2,3,4$);
- MinT-shr is the best reconciliation procedure in most cases.

In addition, looking at the second row of figure 3, we see that for any forecast horizon the improvements in the unique *GDP* reconciled forecasts are always larger than those registered for $\widetilde{GDP}^E$. The same happens with $\widetilde{GDP}^I$, $h = 3,4$, while for $h = 1,2$ the skill scores are very close.

References

1. Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R.J., Affan, M. (2019), Hierarchical Forecasting, in Fuleky, P. (ed.), *Macroeconomic Forecasting in the Era of Big Data*, Cham, Springer, 689–719.
2. Bates, J.M., Granger, G.W.J. (1969), The combination of forecasts, *Operational Research Quarterly*, 20, 4, 461–468.
3. Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L. (2011), Optimal combination forecasts for hierarchical time series, *Computational Statistics & Data Analysis*, 55, 9, 2579–2589.
4. Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R.J. (2019), *Forecast reconciliation: A geometric view with new insights on bias correction*, Department of Econometrics and Business Statistics, Monash University, Working Paper 18/19.
5. Schäfer, J.L., Strimmer, K. (2005), A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, *Statistical Applications in Genetics and Molecular Biology*, 4, 1.

[1] The R scripts, the data and the results of the paper by Athanasopoulos *et al.* (2019) are available in the github repository located at https://github.com/PuwasalaG/Hierarchical-Book-Chapter.

[2] Athanasopoulos *et al.* (2019) point out that this fast and flexible approach performs well in forecasting Australian GDP aggregates, even compared to other more complex methods.

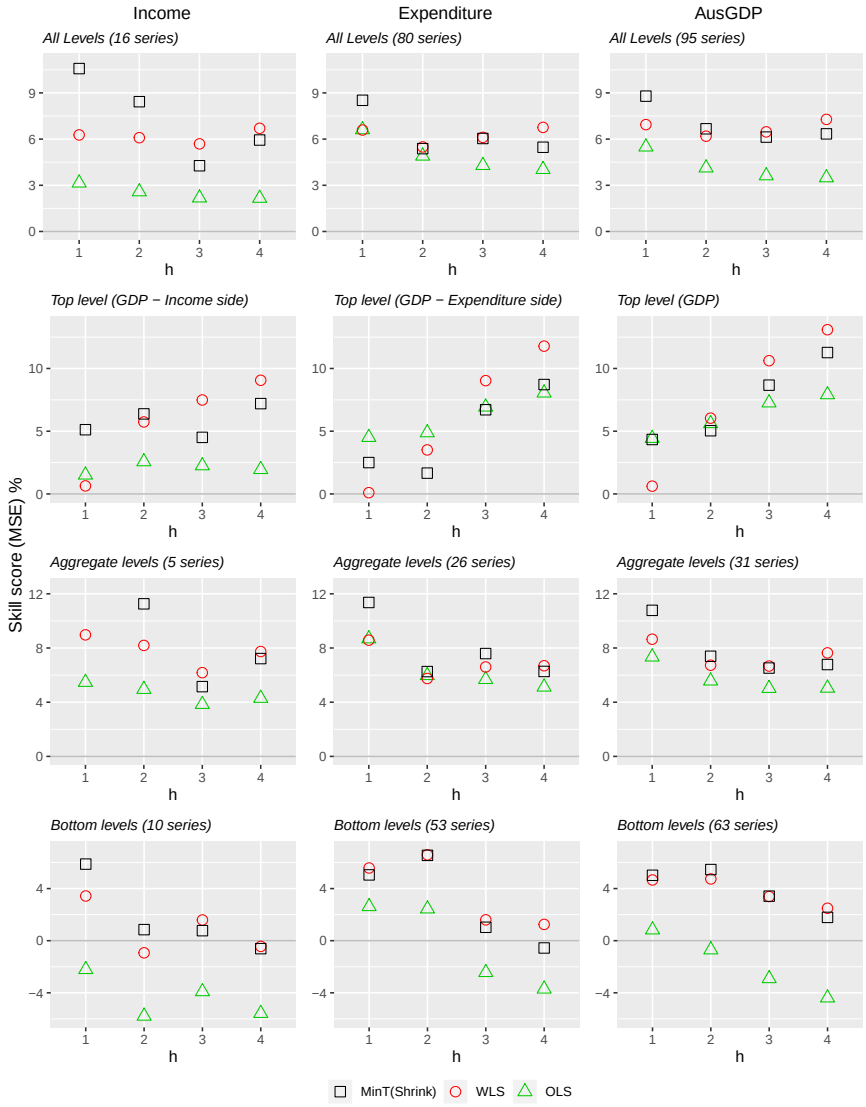


Fig. 3 Skill scores for reconciled point forecasts from alternative methods (with reference to base forecasts) using MSE.

6. Stone, R., Champernowne, D.G., Meade, J.E. (1942), The precision of national income estimates, *The Review of Economic Studies*, 9, 2, 111–125.
7. Van Erven, T., Cugliari, J. (2015), Game-theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts, in Antoniadis, A., Poggi, J.M., Brossat, X., (eds.) *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Berlin, Springer, 297–317.
8. Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J. (2019), Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization, *Journal of the American Statistical Association*, 114, 526, 804–819.

GLASSO Estimation of Commodity Risks
Stima GLASSO dei Rischi delle Commodities

Beatrice Foroni and Saverio Mazza and Giacomo Morelli and Lea Petrella

Abstract In this paper we apply the Graphical LASSO (GLASSO) procedure to estimate the network of twenty-four commodities divided in energy, agricultural and metal sector. We follow a risk management perspective. We use GARCH and Markov-Switching GARCH classes of models with different specifications for the error terms, and we select those that best estimate Value-at-Risk for each commodity. We achieve GLASSO estimation exploring the precision matrix of the multivariate Gaussian distribution obtained from a Gaussian Copula, with marginals given by the residuals of the models, selected via backtesting procedure. The analysis of interdependences in the resulting network is carried out by using the eigenvector centrality metric.

Abstract *In questo articolo applichiamo la procedura GLASSO per stimare il network di ventiquatto commodities divise nei settori dell'energia, agricoltura e metalli. Seguiamo un approccio di valutazione del rischio. Usiamo modelli GARCH e Markov-Switching GARCH con differenti specificazioni per il termine di errore, scegliendo il modello che meglio stima il Value-at-Risk. Effettuiamo la stima GLASSO analizzando la matrice di precisione di una distribuzione Normale ottenuta a partire da una Copula Gaussiana, le cui marginali sono date dai residui*

Beatrice Foroni
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano 9, e-mail: beatrice.foroni@uniroma1.it

Saverio Mazza
MEMOTEF Department, Sapienza Univerity of Rome, Via del Castro Laurenziano 9, e-mail: saverio3107@gmail.com

Giacomo Morelli
DSS Department, Piazzale Aldo Moro 5, Sapienza University of Rome, e-mail: giacomo.morelli@uniroma1.it

Lea Petrella
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano 9, e-mail: lea.petrella@uniroma1.it

dei modelli selezionati attraverso la procedura di backtesting. L'analisi delle interdipendenze nel network viene effettuata usando la metrica di centralità dell'autovettore.

Key words: Value at Risk, GARCH, Energy Commodities, Gaussian Copula

1 Introduction

Over the last two decades, financial markets have recorded an increase of investments in commodities, a feature known as *financialization of commodities*. In particular, investments in commodity futures have tripled making their prices more vulnerable to fluctuations originated in, e.g., volatility. During stress periods, such price swings together with the strong interconnections of commodity markets create alerts for contagion effects. In this paper, we investigate interdependences among and within twenty-four time series of commodity returns representative of the energy, agriculture and metals sectors. We use GARCH and Markov-Switching GARCH (MSGARCH) models with different error term distributions to forecast the Value-at-Risk (VaR) of each commodity. For a given time horizon t and a confidence level p, the VaR is the loss in market value that is exceeded with probability 1-p thus, evaluating the quality of its estimates is of utmost importance. To do this in the context of risk management, backtesting is the most recognized test procedure. We use three different backtesting methodologies that allow us to quantify the quality of the forecasts from a risk perspective. These are the Unconditional Coverage (UC) test of [15], the Conditional Coverage (CC) test of [6] and the Dynamic Quantile (DQ) test of [9]. The residuals of the models, selected through VaR backtesting procedures, are then used as marginals in a Gaussian copula. The Gaussianity achieved allows to exploit Graphical LASSO (GLASSO) estimation of the Gaussian Graphical model in order to identify the underneath co-dependence structure in the network of commodities, where vertices represent the commodities and their dependence is visualized by edges. We carry out the analysis of the network using the eigenvector centrality metric that provides indications on which are the most geographically central and important nodes in the graph. A recent study of within and between interdependences among commodities is in [4]. We differentiate from their analysis in that we operate in a risk management perspective and obtain i) model selection of GARCH and MSGARCH according to backtesting procedures of VaR forecasts and ii) a sparse Gaussian Graphical representation of the commodities achieved via GLASSO estimation. Co-movements in commodity markets are addressed in [7] to examine the relation between Gold and Silver in the metal sector. The seminal paper of [17] detects inter-sectorial co-movements between oil price and metals finding short and long-run equilibrium relationships whereas dependence between oil and agriculture are studied in e.g.,[16].Gold (GC1)[1], Silver (SI1), Palladium (PA1), Copper (HG1), and Zinc (LX1) are representative of the metals sector; WTI Crude

[1] Bloomberg tickers

Oil (CL1), Heating Oil (HO1), Gasoline (XB1), Low Sulfur Gasoline (QS1), Natural Gas (NG1), Natural Gas UK (FN1), and Ethanol (DL1) are representative of the energy sector; Corn (C1), Oats (O1), Rough Rice (RR1), Soybeans (S1), Wheat (W1), Cocoa (CC1), Cotton (CT1), Coffee (KC1), Sugar (SB1), Soybean Oil (BO1), Soybean Meal (SM1), and Orange Juice (JO1) are representative of the agricultural sector. The results show a prevalence of the two-regime models over the single regime ones and Soybean Oil (BO1) for the agriculture sector, Natural Gas (NG1) for energy and Copper (HG1) and Palladium (PA1) for the metal sector seem to be the most central node in the estimated graph. The rest of the paper is organized as follows. In Section 2 we describe the approach and the models used whereas in Section 3 we discuss the main results regarding the model selection, the dependence structure in the Graphical model, the network metrics and conclude.

2 Model Specification

We define y_t the log-return of a financial asset at time t. We follow [14] for the general specification of the MSGARCH model:

$$y_t|(s_t = k, \mathscr{I}_{t-1}) \sim \mathscr{D}(0, h_{k,t}, \xi_k) \tag{1}$$

$$h_{k,t} \equiv \alpha_{k,0} + \alpha_{k,1}\varepsilon_{t-1}^2 + \beta_k h_{k,t-1} \tag{2}$$

where $\mathscr{D}(0, h_{k,t}, \xi_k)$ is a continuous distribution with zero mean, time-varying variance equal to $h_{k,t}$ and additional shape parameters gathered in the vector ξ_k and $\mathscr{F}_{t-1}$ the information set up to time $t-1$. We assume that the latent variable s_t, defined on a discrete space $\{1,\ldots,K\}$, evolves according to an unobserved first order ergodic homogeneous Markov chain with transition probability $\mathbf{P} \equiv \left\{p_{ij}\right\}_{i,j=1}^{K}$, where $p_{ij} = P[s_t = j|s_{t-1} = i]$. When $k = 1$ we obtain the GARCH(1,1) specification. To ensure positivity of $h_{k,t}$ it is required that $\alpha_{k,0} > 0$, $\alpha_{k,1} > 0$, and $\beta_k \geq 0$. Covariance-stationarity in each regime is obtained by requiring that $\alpha_{k,1} + \beta_k < 1$. For each commodity, we choose the model that best predicts its VaR accurately and replicates the well known stylized facts using backtesting procedures. We use the residuals of the models selected via backtesting procedure as marginal of a Gaussian copula function. This way, the Graphical LASSO (GLASSO) procedure introduced in [11] that relies on the gaussianity assumption can be applied to estimate the sparse inverse covariance matrix $\Omega = K^{-1}$. GLASSO builds on a linear regression model to shrink to zero some coefficient, that results in a maximum likelihood problem using a L_1-norm penalty term. It solves the following problem:

$$\max_{\Omega} \log\det\Omega - tr(\Sigma\Omega) - \rho\left\|\Omega\right\|_1 \tag{3}$$

where tr indicates the trace of the matrix and $\|\Omega\|_1$ the L_1-norm that can be calculated as the sum of the absolute values of the elements of Ω. The parameter ρ controls the size of the penalty and it determines the number of zeros in the sparse precision matrix Ω: a higher (lower) value is responsible for a more (less) sparse matrix. Like most of the shrinking methodologies, to get a reliable selection it is fundamental the right choice of the penalization parameter ρ. We follow [12], in selecting $\rho = \hat{\rho}$ such that $\hat{\rho}$ minimizes the cross-validation estimator of the risk. We use the K-fold cross-validation with $K = 10$, which allows the tuning parameter in the GLASSO to remain persistent. In the following Section we emphasize the fundamental results obtained with the techniques described above.

3 Main Results and Conclusion

We study the returns of three different commodity sectors: agriculture, energy and metals. We estimate the models described in Section 2 for $K = 1$ (i.e. GARCH(1,1) models) and for $K = 2$ (i.e. MSGARCH(1,1) models) where we assume several distributions for the error term in (1) to account for well known stylized facts. In particular we use the standard Normal ("norm"), the Student-t ("std") and the Generalized Normal ("ged") distributions, as well as their skewed version (see e.g [3] and [10]) labeled "snorm", "sstd", and "sged", respectively. In Table 1 we show all the models considered for the application. It is worth noting that for the MS-GARCH models we fix $K = 2$ apriori where the two states idendify a low and a high volatility regime. The estimation procedure has been conducted using the maximum likelihood estimation as described in [3]. For each time series, we perform model selection by testing the forecast performance of two specific estimated quantiles, the 95-th and the 99-th, thus offering a risk management perspective of the analysis. Among the models available after the backtesting, we select those that succeeds in at least two out of three tests. We compute network metrics that quantify the position of the nodes in the Gaussian Graphical model obtained. In particular, we focus on the *eigenvector centrality*, which provides indications on which are the most geographically central and important nodes [13]. In a risk management framework, eigenvector centrality is a measure used to capture the capacity of a node (a single commodity) to cause systemic risk, that is, a contagion of risks on other nodes [5]. Table 1 shows the outcome of the backtesting procedure. On the one hand, we observe that there is a considerable prevalence of the Markov-Switching specification in the selected models: 18 selected models present a Markov-Switching specification. This confirms what is pointed out in [2] that is, the two-regime specification better captures the jumps in the dynamic of the volatility in financial returns. On the other hand, we find that there is no prevalence for the asymmetric distributions over the symmetric ones. In Figure 1 we show the graph representation of the adjacency matrix obtained from the estimation of the sparse covariance matrix through the GLASSO algorithm. The dimension of the nodes is proportional to the eigenvector centrality score. According to such metric, we observe the most important nodes in

Commodity sector					
Energy		Metals		Agricolture	
Commodity	Model	Commodity	Model	Commodity	Model
Heating Oil	GARCHged	**Gold**	MSGARCHstd	**Oats**	GARCHsstd
Gasoline	GARCHsged	**Silver**	MSGARCHged	**Wheat**	MSGARCHsged
Low Sulfur Gasolio	GARCHsged	**Copper**	MSGARCHsnorm	**Soybeans**	MSGARCHsnorm
Natural Gas	MSGARCHged	**Palladium**	MSGARCHsstd	**Coffee**	MSGARCHnorm
Ethanol	MSGARCHsstd	**Zinc**	MSGARCHged	**Cocoa**	GARCHsstd
WTI Crude Oil	MSGARCHsstd			**Cotton**	MSGARCHnorm
Natural Gas UK	GARCHsged			**Corn**	MSGARCHged
				Rough Rice	GARCHged
				Sugar	MSGARCHsged
				Soybean Oil	MSGARCHnorm
				Soybean Meal	MSGARCHstd
				Orange Juice	MSGARCHnorm

Table 1 Best models selected

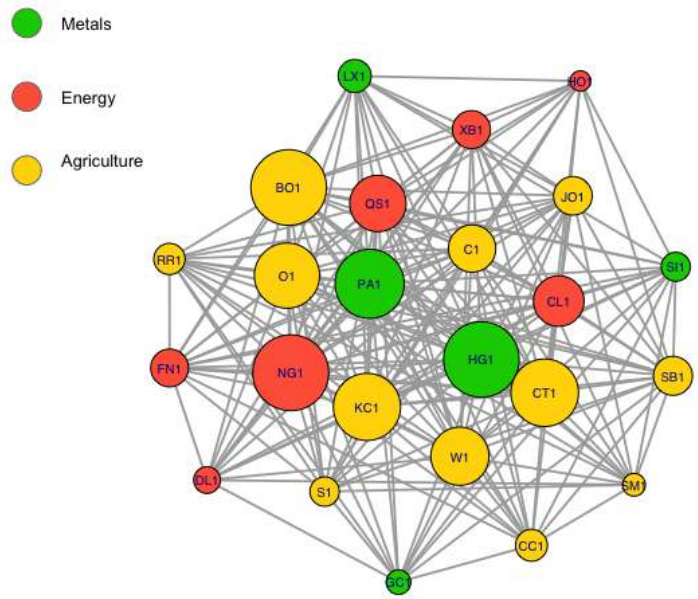


Fig. 1 Graphical model of the commodity futures obtained with GLASSO. The size of the nodes is proportional to their centrality.

the graph: Soybean Oil (BO1) for the agriculture sector, Natural Gas (NG1) for the energy sector and Copper (HG1) and Palladium (PA1) for the metal sector.

Regarding Soybean Oil (B01), the fact that it presents the maximum connection rate with the energy sector can also be found in and [8]; indeed it is known that Biodiesel production in the USA is based predominantly on Soybean Oil, precisely for the 82% [1]. Among the least central commodities we find the Gold (GC1) with a position in the dependence structure that highlights its nature of refuge commodity: showing a marginal role this commodity can be a good investment in anticipation

of high volatility periods. Concluding, the application of VaR backtesting procedure for the set of commodity returns studied has allowed to extend the literature on the model selection with GARCH and MSGARCH dynamics, in a risk management framework. The structure of interdependences in the commodity network has been detected in a Gaussian Graphical model. The choice of exploiting network metrics in a risk oriented framework could bring relevant informations both in terms of effects of risk contagion and for improvements of policy formulation in developing and developed countries.

References

1. Ajanovic, A.: Biofuels versus food production: Does biofuels production increase food prices? Energy **36**(4), 2070–2076 (2011)
2. Ardia, D., Bluteau, K., Boudt, K., Catania, L.: Forecasting risk with markov-switching garch models: A large-scale performance study. International Journal of Forecasting **34**(4), 733–747 (2018)
3. Ardia, D., Bluteau, K., Boudt, K., Catania, L., Trottier, D.A.: Markov-switching garch models in r: The msgarch package. Journal of Statistical Software **91**(4) (2019)
4. Balli, F., Naeem, M.A., Shahzad, S.J.H., de Bruin, A.: Spillover network of commodity uncertainties. Energy Economics **81**, 914–927 (2019)
5. Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L.: Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of financial economics **104**(3), 535–559 (2012)
6. Christoffersen, P.F.: Evaluating interval forecasts. International Economic Review **39**(4), 841–862 (1998)
7. Ciner, C.: On the long run relationship between gold and silver prices a note. Global Finance Journal **12**(2), 299–303 (2001)
8. Dunis, C.L., Laws, J., Evans, B.: Modelling and trading the soybean-oil crush spread with recurrent and higher order networks: A comparative analysis. In: Artificial Higher Order Neural Networks for Economics and Business, pp. 348–366. IGI Global (2009)
9. Engle, R.F., Manganelli, S.: CAViaR: conditional autoregressive value at risk by regression quantiles. Journal of Business & Economic Statistics **22**(4), 367–381 (2004)
10. Fernandez, C., Steel, M.: On bayesian modeling of fat tails and skewness. Journal of The American Statistical Association - J AMER STATIST ASSN **93**, 359–371 (1998). DOI 10.1080/01621459.1998.10474117
11. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England) **9**, 432–41 (2008). DOI 10.1093/biostatistics/kxm045
12. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**(1), 1 (2010)
13. Gould, P.R.: On the geographical interpretation of eigenvalues. Transactions of the Institute of British Geographers pp. 53–86 (1967)
14. Haas, M., Mittnik, S., Paolella, M.S.: A new approach to markov-switching garch models. Journal of Financial Econometrics **2**(4), 493–530 (2004)
15. Kupiec, P.H.: Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives **3**(2), 73–84 (1995)
16. Nazlioglu, S.: World oil and agricultural commodity prices: Evidence from nonlinear causality. Energy policy **39**(5), 2935–2943 (2011)
17. Sari, R., Hammoudeh, S., Soytas, U.: Dynamics of oil price, precious metal prices, and exchange rate. Energy Economics **32**(2), 351–362 (2010)

Measuring the Effect of Unconventional Policies on Stock Market Volatility

Gli Effetti delle Politiche Monetarie Non Convenzionali sulla Volatilità del Mercato Azionario

Giampiero M. Gallo, Demetrio Lacava and Edoardo Otranto

Abstract Our research analyzes the impact of the ECB's unconventional monetary policy on stock market volatility in four Eurozone countries (France, Germany, Italy and Spain). We estimate Multiplicative Error Models to quantify the part of volatility depending directly on unconventional policies: we distinguish between the announcement effect (through a dummy variable) and the implementation effect (measured by a proxy for securities held for monetary policy purpose). While we observe an increase in volatility on announcement days, we find a negative implementation effect, which provides a remarkable dampening effect in the long term.

Abstract *Il presente studio rappresenta un'analisi quantitativa dell'effetto delle politiche monetarie non convenzionali, stabilite dalla BCE, sulla volatilità del mercato azionario nei principali paesi dell'Eurozona. Lo studio è basato sulla stima di Multiplicative Error Models, i quali permettono di distinguere la parte di volatilità derivante direttamente dalle politiche monetarie non convenzionali. Il principale risultato è rappresentato da una riduzione della volatilità dovuta all'implementazione delle politiche non convenzionali (e misurata attravero una proxy degli asset presenti nel bilancio della BCE per scopi di politiche monetarie non convenzionali), nonostante un immediato aumento della volatilità stessa, causato dagli annunci di politica monetaria stabilita dalla BCE.*

Key words: Unconventional monetary policy, Financial market, Realized Volatility, Multiplicative Error Model

---

Giampiero M. Gallo
Italian Court of Audits (Corte dei conti – disclaimer) and New York University in Florence, e-mail: giampiero.gallo@nyu.edu

Demetrio Lacava
University of Messina, e-mail: dlacava@unime.it

Edoardo Otranto
University of Messina, e-mail: eotranto@unime.it

1 Introduction

During the great recession, interest rates were brought close or beyond the zero lower bound, so that many central banks had to resort to unconventional monetary policy measures in order to stimulate the real economy. These policies consist of an expansion in the central bank's balance sheet - generally through asset purchase programs - which in turn affects the real economy by modifying inflation rate expectation during periods in which the so-called liquidity trap makes conventional policy no longer effective.
Coming after the Federal Reserve in the US and the Bank of England, the European Central Bank (ECB) implemented several unconventional monetary measures between 2009 and 2018. Even though the main concern is about their effects on the real economy, these policies have also unintended effects on financial markets that have been largely studied by recent literature. Less expectedly, to the best of our knowledge, not much attention was devoted to the impact of the quantitative easing on financial volatility as key research objective (for example Shogbuyi and Steeley, 2017), modelling volatility mainly through the GARCH family models.
In what follows, we analyze the impact of unconventional policies on stock market volatility within the Multiplicative Error Model framework (Engle and Gallo, 2006), by proxying for unconventional policies through the ratio between the securities held for unconventional policies purposes and ECB's total asset (D'amico, English, Lopez-Salido and Nelson, 2012). In carrying out our analysis we employ an annualized realized kernel volatility measure built on high frequency data, which should remove endogeneity arising when monetary policy decisions coincide with a stock price reduction. Finally, through a multi-step forecasting procedure it emerges how unconventional policy deploys its effects on volatility for a period between 25 and 60 business days, after which volatility converges to its long-run level.

2 The models

Multiplicative Error Models (MEM, Engle and Gallo, 2006) have proved effective in representing volatility dynamics as the product of a time-varying component, μ_t (following a GARCH–like process), and a positive random innovation, ε_t.[1] This has several advantages: it ensures positiveness without resorting to logs, and it exploits more accurate measurement of volatility. Within the MEM setup, Brownlees, Cipollini and Gallo (2012) as well as Otranto (2015) propose a new model specification, by decomposing the mean equation as the sum of two components, both evolving according to GARCH models. The approach provides a general framework within which the consideration of quantitative easing as an unobservable factor al-

[1] In what follows, we call AMEM the MEM with asymmetric effects, and AMEMX the AMEM including exogenous regressors (represented by x_t and Δ_t; see later) in the μ_t equation.

lows for an estimate of its impact on the evolution of volatility. In what follows we explore how unconventional policies can affect volatility both in an additive and a multiplicative way. In our specification, the first component of the mean equation evolves as a GARCH model (capturing the pure volatility dynamics) while the second one follows an autoregressive process with exogenous variables, to capture both the announcement and the implementation effects of unconventional measures.
One model we propose (following Brownlees, Cipollini and Gallo, 2012, we call it Composite AMEM–ACM) consists of four equations:[2]

$$\begin{aligned} &RV_t = \mu_t \varepsilon_t, \;\; \varepsilon_t | \Psi_{t-1} \sim Gamma(\vartheta, \tfrac{1}{\vartheta}) \\ &\mu_t = \varsigma_t + \xi_t \\ &\varsigma_t = \omega + \alpha RV_{t-1} + \beta \varsigma_{t-1} + \gamma D_{t-1} RV_{t-1} \\ &\xi_t = \delta E\left(x_t | \Psi_{t-1}\right) + \varphi \Delta_t + \psi \xi_{t-1} \end{aligned} \tag{1}$$

In this model, ς_t represents the proper volatility of the market, due to its intrinsic dynamics, which evolves as described by the third equation in (1); ξ_t represents the effect due to the unconventional policies and follows an AR(1) process with the addition of the impacts of the conditional expectation of an exogenous variables x_t made at time $t-1$ and of the mean shift impact in volatility on a day of a monetary policy announcement by the Central Bank (Δ_t is a dummy variable, taking value 1 in day characterized by the communication of policies news by ECB, 0 otherwise). The variable, x_t, represents the ratio between the amount employed in unconventional polices and total asset (UMP/TA), its expectation may naïvely be assumed to be x_{t-1}. Finally, as argued by Brownlees, Cipollini and Gallo (2012) (on the basis of Engle and Lee, 1999), the coefficient ψ in ξ_t process, is required to be $0 < \psi < \beta < 1$ so as to ensure the identification of the model by having the long run component more persistent than the short one. Such a model makes it possible to quantify and plot the effect of the unconventional ECB actions on the volatility RV_t, even in relative terms as $\frac{\xi_t}{\mu_t}$.
It is important to notice that, since the unconventional policy component is an unobservable signal, it remains to be seen whether its impact on the general level of volatility should be considered in an additive way, as in (1), rather than as a multiplicative component, in the spirit of Brownlees, Cipollini and Gallo (2011). In the latter case, we must have μ_t equal to ς_t multiplied by a component whose unconditional mean is equal to one as an identifying condition. In what follows, we discuss two different specifications of the Multiplicative ACM which ensure the compliance with this constraint. In the first specification we allow ξ_t to impact on RV_t through a logistic function. The model, called Logistic-ACM (L-ACM), differs from model (1) just in the second equation, which becomes $\mu_t = \varsigma_t \cdot 2(\exp(\xi_t)/1 + \exp(\xi_t))$.
Alternatively, we can specify $\mu_t = \varsigma_t \cdot \xi_t$ but we need to modify the above specification for ξ_t by adding a constant term in its equation. Within this specification, named Linear-ACM (Li-ACM), we have

[2] RV_t is the realized volatility at time t relative to a certain asset (index); D_t a dummy assuming value 1 when the return of the asset (index) at time t is negative, 0 otherwise (asymmetric effect).

$$\xi_t = \left(\frac{1 - \delta \bar{x} - \varphi \bar{\Delta}}{1 - \psi} \right) + \delta x_{t-1} + \varphi \Delta_t + \psi \xi_{t-1},$$

where $\bar{x}$ and $\bar{\Delta}$ represent mean values.

3 Estimation results

In this section, we discuss the estimation results according to the estimated models (tables are not reported for the sake of space, but they are available upon request). The first result is in line with other papers (e.g. Shogbuyi and Steeley, 2017): the sign of the coefficient φ of the dummy variable signals that there is a mean positive shift in volatility on days of monetary policy communication: such an increase in volatility is between 2.185 points for the most important market (Germany) and 3.694 points for Spain, which expectedly (together with Italy) is a sensible market to this kind of policy. The coefficient on our proxy is significant at 1% level and enters in the model with the expected negative sign. The higher impact is observed for the Southern countries: on average, the short run and the long run effects in Italy and Spain are higher than France and Germany by about 40% and 16%, respectively.

The impact of unconventional policies on stock market volatility is evident also by looking at the ratio between the corresponding component and the overall level of volatility: the estimated range of reduction of volatility caused by ECB's unconventional policies is between -3.19 and -3.86 (for Germany and Spain, respectively). This long run effect can be seen also in figure 1, which plots the evolution of the volatility components ς, the blue line, and ξ, across three red dotted-lines: the lowest represents the characteristic path of the component depending directly on unconventional policies; the highest, instead, represents volatility spikes due to the day of announcement effect; finally, the intermediate line, represents the level of volatility on days after a monetary policy announcement, when volatility comes back to its previous level. The effect is more evident starting from October 2014 - when the ECB implicitly communicated to the market it would purchase also private sector bond - because of the change in the slope of ξ equation (red line), which lasts for the entire period of the program. Results remain almost unchanged when we consider the multiplicative specifications. Once again our proxy enters in the model with the expected sign and with the highest level of significance. In both cases, coefficients are somewhat lower, also in view of the fact that in this case unconventional policies affect volatility in a multiplicative way. Note that in the additive version of ACM model generally ξ_t is negative - so that a higher proxy's coefficient (in absolute value) will reduce stock market volatility, whereas within the multiplicative versions ξ_t is always positive, but attain the reduction in volatility with coefficients associated to the proxies being less than 1.

Comparison between models is based on the information criteria (AIC and BIC) and loss functions for evaluating the forecasting power of the models (MSE and MAE). Focusing on information criteria, the ACM model is the best model in 3 out

of 4 cases. However, when we consider MSE and MAE, the L-ACM is the preferred one for Germany and Spain, while the additive ACM is the best model for France and Italy.
Finally, we rely on a multi-step out-of-sample forecasting procedure to assess how long unconventional policies reduce volatility. For this purpose, we consider the period from June 1, 2014 to August 24, 2017 as the estimation period, making forecasts for the following year. Results are summarized in Figure 2, where we can notice that volatility converges to a constant level. Of course the duration of the effects differs model by model and country by country: it lasts between 25 and 60 days, estimated via AMEMX, for the FTSEMib and the DAX30, respectively.

4 Conclusion

In this paper we examine how unconventional monetary policies by ECB affect realized volatility. The innovative feature of this study lies in the model we use, the Composite AMEM, which allows us to distinguish between a pure volatility mechanism and the part of volatility depending directly on quantitative easing policies. Results show how an increase in securities held by ECB for monetary policy purposes relative to ECB total asset reduces volatility in all markets for a period between 25 and 60 business days, with disrupted countries generally benefitting more. However, our proxies do not allow us to distinguish the specific effect of each policy, so that we cannot identify which of these extraordinary measures is more effective, nor the presence of spillovers among countries.

References

1. Brownlees, C. T., Cipollini, F., Gallo, G. M.: Intra-daily volume modeling and prediction for algorithmic trading. Journal of Financial Econometrics, Volume 9, Issue 3, Summer 2011, Pages 489–518 (2011)

2. Brownlees, C. T., Cipollini, F., Gallo, G. M.: Multiplicative error models. In: Bauwens, L., Hafner, C. M., Laurent, S., Eds., Handbook of Volatility Models and Their Applications. Hoboken, NJ: J. Wiley & Sons, pp. 281-310 (2012)

3. D'Amico, S., English, W. B., López-Salido, D. & Nelson, E.: The Federal Reserve's large-scale asset purchase programs: rationale and effects. Finance and Economics Discussion Series 2012-85, Board of Governors of the Federal Reserve System (US) (2012)

4. Engle, R. F., & Gallo, G. M.: A multiple indicatorsmodel for volatility using intradaily data. Journal of Econometrics, 131, 3-27 (2006)

5. Engle, R. F. & Lee, G. J.: A permanent and transitory component model of stock return volatility. In R. F. Engle & H. White, editors, Cointegration, Causality, and Forecasting:

A Festschrift in Honor of Clive W. J. Granger, pages 475 – 497. Oxford University Press, Oxford (1999)

6. Otranto, E.:Capturing the spillover effect with Multiplicative Error Models. Communications in Statistics - Theory and Methods, 44:15, 3173-3191 (2015)

7. Shogbuyi, A. & Steeley, J. M.: The effect of quantitative easing on the variance and covariance of the UK and US equity markets. International Review of Financial Analysis, Elsevier, vol. 52(C), 281-291 (2017)

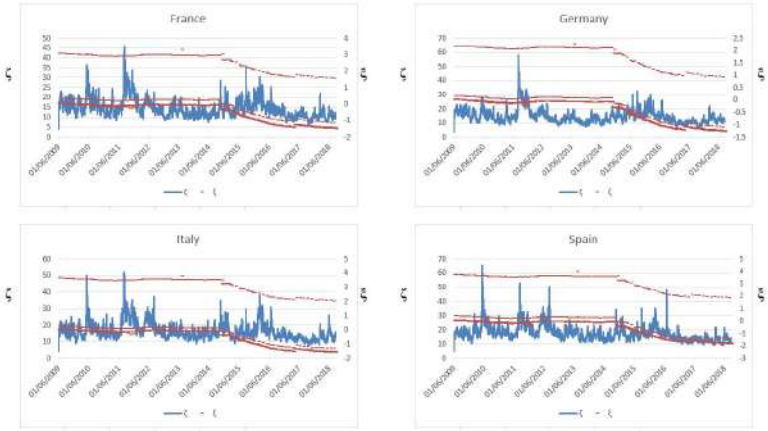


Fig. 1 CAC40, DAX30, FTSE MIB and IBEX35 zeta and csi functions obtained from ACM.

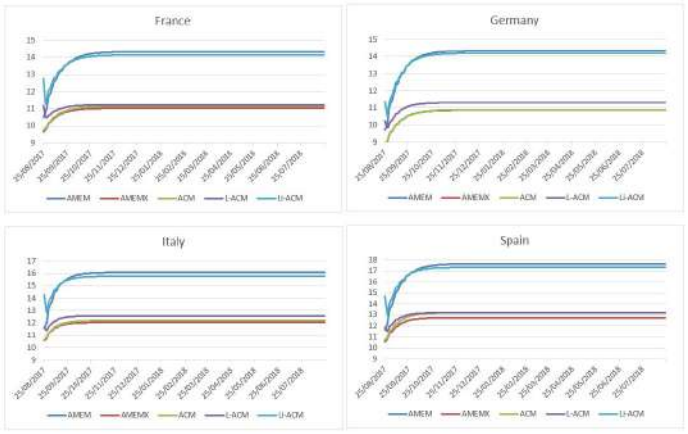


Fig. 2 Multi-step out-of-sample forecast. Estimation period: $June\,1,2009 - August\,24,2017$. Forecast Period: $August\,25,2017 - August\,24,2018$.

Multidimensional versus unidimensional poverty measurement

Misura multidimensionale versus unidimensionale della povertá

Michele Costa

Abstract This paper proposes a comparison between uni- and multidimesional approaches to the measurement of poverty. Traditional uni-, fuzzy uni- and fuzzy multidimensional indicators are compared by means of a rank correlation analysis. The robustness of the comparison is ensured by a simulation study. Our results stress that unidimensional indicators provide partial information on the poverty condition.
Abstract *Viene proposto un confronto tra gli approcci uni- e multidimensionale alla misura della povertá. Indicatori tradizionale uni, fuzzy uni e fuzzy multidimensionale sono confrontati grazie ad una analisi di correlazione sui ranghi. La robustezza del confronto é assicurata da uno studio di simulazione. I risultati sottolineano che l'approccio unidimensionale fornisce informazioni parziali sulla povertá.*

Key words: Multdimensional poverty, rank correlation analysis, fuzzy indicators

1 Introduction

If there is a general consensus on the multidimensional view of poverty, its empirical implications are still largely unexplored. An effective comparison between different approaches to poverty measurement is to be done by comparing the sets of poor units defined by each approach. If the poor sets are not particularly different, even distant methods would point out to the same poor units, and in this case the simplest approach is to be preferred. On the contrary, if we observed different poor sets, the distance in theory would be replicated and confirmed in the analysis of real data. In order to provide insights on the comparison between unidimensional and multidimensional poverty indicators, we develop a rank correlation analysis, focused on the correct definition of the set of poor units. The robustness of the results is ensured by a simulation study, developed on the framework of fuzzy set methods.

Michele Costa
Department of Economics, University of Bologna, e-mail: michele.costa@unibo.it

2 Unidimensional and multidimensional poverty indicators

Let be A the set of the n units $\{a_1, a_2, ..., a_n\}$ and B the subset of the poor units, our purpose is to compare different definitions of B achieved by means of unidimensional and multidimensional poverty indicators (see e.g. Deutsch and Silber (2004) and Duclos et al. (2006)). The set B represents the key point of the comparison between the unidimensional and the multidimensional approaches: if the two methods suggest the same set, the multidimensional extension is not strictly necessary, with the unidimensional indicator being fully informative, while, if the two approaches indicate two different sets, it is important to evaluate their different information.

To address this point, we carry out a correlation analysis on the ranks u_i of the list of the units, from to poorest to the richest, obtained from the unidimensional approach and on the ranks v_i of the list resulted from the multidimensional approach.

The analysis is carried out by referring to a set of correlation indexes (see e.g. Balakrishnan and Rao (1998a,b)) applied to the ranks (u_i, v_i): the Bravais-Pearson coefficient r, the Kendall's τ, the Spearman index S and the Gini rank correlation, or cograduation, index G. Rank correlation analysis allows to effectively compare the two lists, and also to investigate the similarity of specific subgroups of units, which is particularly useful in our case, as we aim to the comparison between the subgroups of the poor units B and also between subsets of B.

The key element in order to define B is the indicator function $\mu(X)$, which determines the degree of membership of the i-th unit to the set B with respect to a certain attribute X. The indicator function transforms the achievement of a unit to the deprivation of that unit. A low or null achievement implies a maximum deprivation, that is $\mu(X) = 1$, while an high achievement leads to an absence of deprivation, thas is $\mu(X) = 0$. With respect to the unidimensional poverty measurement, we refer first to the traditional head count ratio H, where $\mu(X)$ can assume only two values, 1 for the poor and 0 for the non poor units, X is the income, z a poverty line and n_i the sample weight of the i-th unit:

$$\mu(X) = \begin{cases} \mu_1 \text{ with } \mu_1 = 1 \text{ if } X < z \\ \mu_2 \text{ with } \mu_2 = 0 \text{ if } X \geq z. \end{cases} \tag{1}$$

$$H = \sum_{i=1}^{n} \mu(X)_i n_i / \sum_{i=1}^{n} n_i \tag{2}$$

Still on the framework of the unidimensional approach, we resort to fuzzy set methods, which allow to overcome the rigidity imposed by two groups. Many Authors analyzed poverty by means of fuzzy sets, among the others Betti et al. (2004), Dagum and Costa (2004), Lemmi and Betti (2006). In the variety of the meaningful contributions, a specific mention is due to the totally fuzzy approach by Cerioli and Zani (1990) and to the relative fuzzy approach by Cheli and Lemmi (1995). Within fuzzy sets, $\mu(X)$ is no longer only equal to 0 or 1, but ranges between 0 and 1, thus allowing intermediate or partial membership to B. In this way the indicator function takes on a more complex form than in (1):

$$\mu(X) = \begin{cases} \mu_1 \text{ with } \mu_1 = 1 \text{ if } X \in [X_{min}, X_1] \\ \mu_2 \text{ if } X \in [X_1, X_2] \\ \ldots \\ \mu_{k-1} \text{ if } X \in [X_{k-2}, X_{k-1}] \\ \mu_k \text{ with } \mu_k = 0 \text{ if } X \in [X_{k-1}, X_{max}]. \end{cases} \tag{3}$$

Fuzzy sets methods allow to develop a multidimensional framework for the poverty measurement. Given n units, m inequality factors and m indicator functions $\mu(X_j)$, we obtain m undimensional poverty indexes as

$$I(X_j) = \sum_{i=1}^{n} \mu(X_j)_i n_i / \sum_{i=1}^{n} n_i \tag{4}$$

and n multidimensional individual indexes as

$$I(a_i) = \sum_{j=1}^{m} \mu(X_j)_i w_j / \sum_{j=1}^{m} w_j.$$

The weights w_j measure the intensity of deprivation related to the j-th inequality factor (Esposito and Chiappero Martinetti, 2019): a factor not possessed by any unit has no effect on the social exclusion, while, if all the units but a few possess the factor, it represents a relevant source of inequality and social exclusion. A weighting system consistent with that is (Cerioli and Zani, 1990)

$$w_j = log(n / \sum_{i=1}^{n} \mu(X_j)_i n_i) \tag{5}$$

The synthesis of the m unidimensional indexes $I(X_j)$ and of the n multidimensional indexes $I(a_i)$ in a multidimensional poverty index can be achieved as

$$I = \frac{\sum_{j=1}^{m} I(X_j) w_j}{\sum_{j=1}^{m} w_j} = \frac{\sum_{i=1}^{n} I(a_i) n_i}{\sum_{i=1}^{n} n_i} = \frac{\sum_{j=1}^{m} w_j \sum_{i=1}^{n} \mu(X_j)_i n_i}{\sum_{j=1}^{m} w_j \sum_{i=1}^{n} n_i} \tag{6}$$

In order to ensure an high degree of robustness (see also Alkire and Santos (2014) and Ravallion (2011)) to the comparison between uni- and multidimensional approaches, we develop a simulation study where the scores of $\mu(X)$ are not a priori chosen, but endogenously determined by means of a Monte Carlo experiment. We randomly extract $\mu(X)$ from an uniform distribution, following a scheme such as

$$\mu(X_j) = \begin{cases} \mu_{j1} \text{ with } \mu_{j1} = 1 \\ \mu_{j2} \text{ with } \mu_{j2} \in [\mu_{j3} + h, (1-h)] \\ \ldots \\ \mu_{jk-1} \text{ with } \mu_{jk-1} \in [h, (1-h)/(k-2)] \\ \mu_{jk} \text{ with } \mu_{jk} = 0 \end{cases} \tag{7}$$

where k_j is the exogenous number of classes proposed for X_j and h is an exogenous minimum distance between two different values of $\mu(X)$.

The main purpose of our Monte Carlo experiment, achieved by comparing exogenous schemes as in (3) to simulated schemes as in (7), is to investigate the stability of the fuzzy indicators and to provide a more general basis for the rank correlation analysis aimed at evaluating the differences between different sets B of poor units.

3 Data and results

As a case study we resort to the income distribution of the Italian households for the 2016. The SHIW data, provided by the Bank of Italy, allow to derive the m poverty indicators based on both individual and household information; in the following we use, for the undimensional case, the household equivalent income X_1, and, for the multidimensional indicator, we add the educational level of the head of the household X_2, his/her age X_3 and the geographical area of residence X_4.

The list u_i related to the unidimensional framework is derived following three directions. First on the basis of the headcount ratio H, with z being the 60% of the median $\bar{x}_{me}$ of the equivalent income. Second, by referring to $I(X_1)$, obtained with $k = 4$, and the indicator function reported on the first two columns of Table 1. Third, we vary the scores of $\mu(X_1)$ by running a Monte Carlo experiment on the basis of the scheme outlined in (7). Observations presenting the same poverty index are sorted by increasing size of the equivalent income X_1. Overall we obtain one list u_i from H, one from $I(X_1)$ and the exogenous $\mu(X_1)$ of Table 1 and 10000 lists u_i from the 10000 $I(X_1)$ calculated from the simulated $\mu(X_1)$: all the lists u_i are identical, since the ranking is invariant with respect to the size of the score of $\mu(X_1)$.

The multdimensional list v_i related to the multidimensional poverty measurement is obtained first by using the $m = 4$ indicators and the indicator functions described in Table 1. Second, also for the multidimensional case we develop a Monte Carlo study on the basis of the scheme (7) with the aim of mitigating the subjectivity of the choice of $\mu(X_j)$. Observations presenting the same value of I are sorted by increasing size of the equivalent income X_1. Overall we obtain one list v_i from the multidimensional index I obtained by using the exogenous indicator functions reported on Table 1 and 10000 lists v_i from I calculated on the simulated $\mu(X_j)$.

Table 1 Indicator functions $\mu(X_j)$ proposed for the household equivalent income X_1, the educational level of the head of the household X_2, his/her age X_3, geographical area of residence X_4

$\mu(X_1)$	X_1	$\mu(X_2)$	X_2	$\mu(X_3)$	X_3	$\mu(X_4)$	X_4
1.0	$x_{1i} < 0.4\bar{x}_{1me}$	1.0	$x_{2i} =$ none	1.0	$x_{3i} > 65$	0.4	$x_{4i} =$ south
0.9	$0.4\bar{x}_{1me} \leq x_{1i} < 0.8$ $0.6\bar{x}_{1me}$		$x_{2i} =$ element.	0.3	$x_{3i} < 30$	0	$x_{4i} =$ center
0.5	$0.6\bar{x}_{1me} \leq x_{1i} < 0$ $0.8\bar{x}_{1me}$		$x_{2i} =$ other	0	$30 \leq x_{3i} \leq 65$	0	$x_{4i} =$ nord
0	$0.8\bar{x}_{1me} \leq x_{1i}$						

A first direct comparison between u_i and v_i shows that the highest ranks of the multidimensional indicator I generally correspond to the highest ranks of the unidimensional list, while the lowest ranks of I do not correspond to the lowest ranks of the unidimensional case.

Rank correlation indexes allow to carry out a more detailed and informative analysis. Table 1 reports, for r, S, G and τ, the rank correlations between the unidimensional u_i and the multidimensional v_i lists by referring to the six cases developed in the paper: the head count ratio by the multidimensional index (HI) and by the simulated multidimensional index HI^*, the unidimensional fuzzy by the multidimensional index $I(X_1)I$ and by the simulated multidimensional index $I(X_1)I^*$, the simulated unidimensional fuzzy by the multidimensional index $I^*(X_1)I$ and by the simulated multidimensional index $I^*(X_1)I^*$. Since the lists u_i are identical, the related correlations have the same values, that is $HI = I(X_1)I = I^*(X_1)I$ and $HI^* = I(X_1)I^* = I^*(X_1)I^*$.

The first column of Table 2 reports selected values of the cumulative distribution of the equivalent income X_1: the last row of Table 2 shows the indexes for the whole population ($F(X_1) = 100$), while the first row refers to the poorest 5% of total population ($F(X_1) = 5$). For the simulated values is reported the mean of the 10000 correlation indexes.

Moving from the bottom to the top of Table 2 it is clearly observable a strong decrease in rank correlation, thus suggesting that multidimensional and unidimensional approaches define two different subsets of poor units.

The results for the Bravais-Pearson index are similar to the indications provided by the Spearman index, while Kendall's and Gini's measures report a lower correlation between u_i and v_i. All indexes, however, clearly signal a decreasing rank correlation for decreasing values of $F(X_1)$, with a minimum value around $F(X_1) = 0.25$.

Table 2 Rank correlation between unidimensional and multidimensional poverty indicators, Bravais-Pearson coefficient, Spearman index, Gini index, Kendall *tau*, Italian households 2016

	Bravais-Pearson		Spearman		Gini		Kendall	
$F(X_1)$	HI	HI^*	HI	HI^*	HI	HI^*	HI	HI^*
5	0.16	0.18	0.34	0.36	0.10	0.12	0.22	0.23
10	0.23	0.27	0.41	0.44	0.15	0.18	0.20	0.24
25	0.13	0.26	0.29	0.39	0.08	0.17	0.13	0.20
50	0.38	0.46	0.41	0.48	0.25	0.32	0.29	0.33
100	0.61	0.62	0.58	0.59	0.46	0.47	0.44	0.46

4 Conclusions

The key point of the comparison between the unidimensional and the multidimensional approaches refers to the set B of poor units: if the two methods suggest the

same set B, the multidimensional extension can be considered as an elegant and theoretically useful improvement, but not strictly necessary, with the unidimensional indicator being fully informative.

When, on the other hand, the two approaches indicate two different sets of poor units, it is important to evaluate the different information contained in the two sets.

The reference to the set B of poor units is useful also in order to evaluate the effects of different specifications of the $\mu(X)$: to the extent that different $\mu(X)$ modify, or not, the set B, we derive an indication on the robustness of the multidimensional indicators.

To provide a satisfactory answer we develop a rank correlation analysis, which allows to demonstrate that the unidimensional and the multidimensional approaches define two different sets of poor households. Consequently any socio-economic policy to reduce poverty developed on the basis of income information is likely to no achieve its proposed goals, being addressed to socioeconomic units which, at least in part, are, in effect, non-poor.

Information provided by a multidimensional approach allows to correctly individuate the set of the poor and is therefore essential to formulate actions able to reduce poverty.

References

1. Alkire, S., Santos, M.E.: Measuring acute poverty in the developing world: robustness and scope of the multidimensional poverty index. World Development, **59**, 251–274 (2014)
2. Balakrishnan, N., Rao C.R.: Order statistics: theory and methods, Handbook of statistics n. 16. North Holland, Amsterdam (1998)
3. Balakrishnan, N., Rao C.R.: Order statistics: applications, Handbook of statistics n. 17. North Holland, Amsterdam (1998)
4. Betti, G., Cheli, B., Gambini, R.: A statistical model for the dynamics between two fuzzy states: theory and an application to poverty analysis. Metron, **62**, 391–411 (2004)
5. Cerioli, A., Zani, S.: A fuzzy approach to the measurement of poverty. In Dagum, C., Xenga, M.: Income and wealth distribution, inequality and poverty. Springer, Berlin, 272–284 (1990)
6. Cheli, B., Lemmi, A.: A totally fuzzy and relative approach to the multidimensional analysis of poverty. Economic Notes, **24**, 115–134 (1995)
7. Dagum, C., Costa, M.: Analysis and measurement of poverty. Univariate and multivariate approaches and their policy implications. In Dagum, C., Ferrari, G.: Household Behaviour, Equivalence Scales, Welfare and Poverty. Springer, Berlin, 221–271 (2004)
8. Deutsch, J., Silber, J.: Measuring multidimensional poverty: an empirical comparison of various approaches. Review of Income and Wealth, **51**, 145-174 (2005)
9. Duclos, J.Y., Sahn, D.E., Younger, S.D.: Robust multidimensional poverty comparisons. The Economic Journal, **116**, 943-968 (2006)
10. Esposito, L., Chiappero Martinetti, E.: Eliciting, applying and exploring multidimensional welfare weights: evidence from the field. Review of Income and Wealth, **65** (2019)
11. Lemmi, A., Betti, G.: Fuzzy set approach to multidimensional poverty measurement, Springer, Berlin (2006)
12. Ravallion, M.: On multidimensional indices of poverty. Journal of Economic Inequality, **9**, 235–248 (2011)

Multiple outcome analysis of European Agriculture in 2000-2016: a latent class multivariate trajectory approach

Analisi multivariata dei risultati dell'Agricultura Europea nel 2000-2016: un approccio con traiettorie multivariate a classi latenti

Alessandro Magrini

Abstract The consideration of the environmental and social dimensions, besides the economic one, is of great importance to highlight both development and sustainability objectives in the evaluation of the agricultural sector. However, multiple outcomes analysis is challenging due to not only methodological issues, but also quality and availability of time series data. In this paper, we propose a latent class multivariate trajectory model, which is applied to 31 European countries by considering a set of economic, environmental and social indicators in the period 2000-2016. The results support the effectiveness of our approach by clearly identifying four groups of countries with emblematic characteristics of the agricultural systems: mature and sustainable, mature but with weak social sustainability, developing with and without clear sustainable objectives.

Abstract *La considerazione delle dimensioni ambientali e sociali, oltre a quella economica, è di grande importanza per evidenziare obiettivi di sviluppo e sostenibilità nella valutazione del settore agricolo. Tuttavia, l'analisi multivariata dei risultati è difficile a causa non solo di problemi metodologici, ma anche di qualità e disponibilità di serie temporali. In questo articolo, si propone un modello con traiettorie multivariate a classi latenti, che viene applicato a 31 paesi europei considerando un insieme di indicatori economici, ambientali e sociali nel periodo 2000-2016. I risultati supportano l'efficacia del nostro approccio identificando chiaramente quattro gruppi di paesi con caratteristiche emblematiche del settore agricolo: maturo e sostenibile, maturo ma con debole sostenibilità sociale, in via di sviluppo con e senza chiari obiettivi sostenibili.*

Key words: European Agriculture, finite mixture modeling, group-based trajectories, multivariate growth curves, sustainability.

Alessandro Magrini
Dep. Statistics, Computer Science, Applications – University of Florence, Italy
e-mail: alessandro.magrini@unifi.it

1 Introduction

In the last two decades, economists have studied the performance of the agricultural sector mainly focusing on productivity or on the value of the output produced (see [2] for a review). Even if this approach is generally adopted in the literature, several authors have pointed out that productivity does not properly measure sustainability (e.g., [1]), motivating a line of research trying to account for the externalities and the social value of Agriculture in the computation of productivity (see, for example, [3]). Recently, several challenging objectives besides productivity compatible with the sustainable livelihood framework [5] have become central in the Horizon 2020 agenda, including the provision of safety and healthy food, the improvement of socio-economic conditions in rural areas, the conservation of natural resources, environmental quality and biodiversity.

In this paper, we address the multiple outcomes analysis of the agricultural sector. At this purpose, we propose a latent class multivariate trajectory model, which is applied to 31 European countries by considering a set of economic, environmental and social indicators in the period 2000-2016.

This paper is structured as follows. In Section 2, the data are described and the statistical model is detailed. In Section 3, the results are reported and discussed. Section 4 contains concluding remarks and purposes for future work.

2 Data and statistical model

We considered a set of indicators covering the economic, environmental and social dimensions of the agricultural sector compatibly with the availability of publicly released data. For the economic dimension, we considered Agricultural Total Factor Productivity (TFP) and Gross Value Added (GVA), describing, respectively, the efficiency in production and the ability to generate value of the sector. For the environmental dimension, we considered greenhouse gas emissions due to Agriculture and nitrogen nutrient balance in soil, describing the negative pressure of the sector on the environment. For the social dimension, we considered the rural at risk-of-poverty and unemployment rates, describing the ability of Agriculture to deal with inequality and abandonment in rural areas.

We downloaded data on the indicators for 31 European countries (the 27 European Union member countries plus Iceland, Switzerland, Norway and United Kingdom) in the period 2000-2016 from Eurostat database, excepting TFP, downloaded from USDA database, and greenhouse gas emissions, downloaded from Faostat. To make the indicators comparable, we transformed them into index numbers with fixed base (year 2000).

Our statistical model borrows from group-based trajectory modeling [4], a special case of latent class growth curves where the units in the same group have the same trajectory. To perform a multiple outcome analysis, we extended such technique by replacing the univariate Normal linear regression within each group with

a multivariate Normal one. Let $i = 1,\ldots,n$ denote the unit, $j = 1,\ldots,n_g$ the group and $t = 1,\ldots,n_t$ the period, thus $\mathbf{y}_{it}^{(j)}$ represents the multivariate observation of the p indicators on unit i at time t given group j. For a polynomial degree $d \geq 1$, our model assumes:

$$\mathbf{y}_{it}^{(j)} = (1, t, t^2, \ldots, t^d)' \beta^{(j)} + \varepsilon_{it}^{(j)} \tag{1}$$

where:

- $\varepsilon_{it}^{(j)} \sim \text{MVN}(0, \Sigma^{(j)})$ is the p-variate random error for $\mathbf{y}_{it}^{(j)}$, with $\Sigma^{(j)}$ the $p \times p$ covariance matrix of the random errors in group j;
- $\beta^{(j)}$ is the matrix $(d+1) \times p$ of the regression coefficients in group j.

The likelihood of the model is:

$$L(\beta^{(j)}, \Sigma^{(j)}, \pi) = \prod_{i=1}^{n} \left[\sum_{j=1}^{n_g} \pi_j \prod_{t=1}^{n_t} p(\mathbf{y}_{it}^{(j)} \mid \beta^{(j)}, \Sigma^{(j)}) \right] \tag{2}$$

where $p(\mathbf{y}_{it}^{(j)} \mid \beta^{(j)}, \Sigma^{(j)})$ is the multivariate Normal density of $\mathbf{y}_{it}^{(j)}$ according to the regression in equation 1, and $\pi = (\pi_1, \ldots, \pi_{n_g})$ is vector of the prior probabilities of the n_g groups.

Model parameters $\beta^{(j)}$, $\Sigma^{(j)}$ and π are estimated by maximizing the likelihood in equation 2 using an our own implementation of the Expectation-Maximization (EM) algorithm, and the countries are assigned to the groups according to the maximum posterior probability implied by the estimated parameter values.

Fig. 1 Trajectories for countries in Group 1: mature and sustainable agricultural sector.

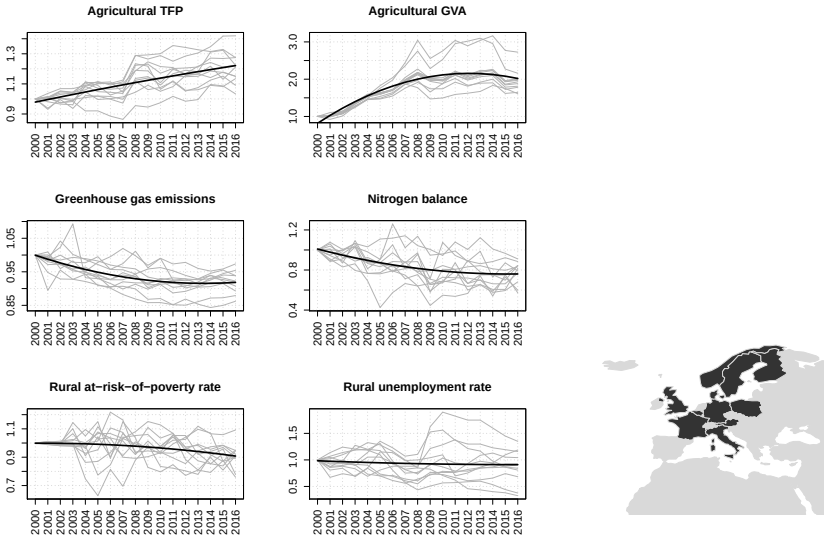


Fig. 2 Trajectories for countries in Group 2: mature agricultural sector but with weak social sustainability.

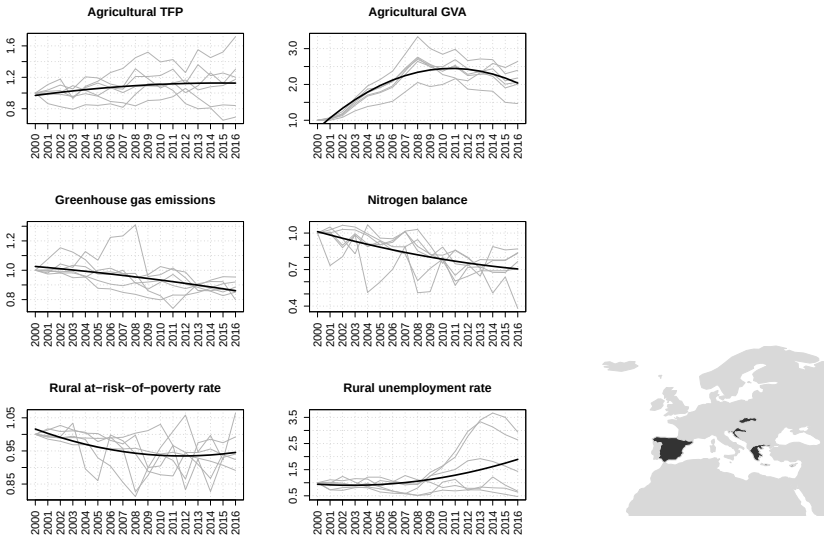
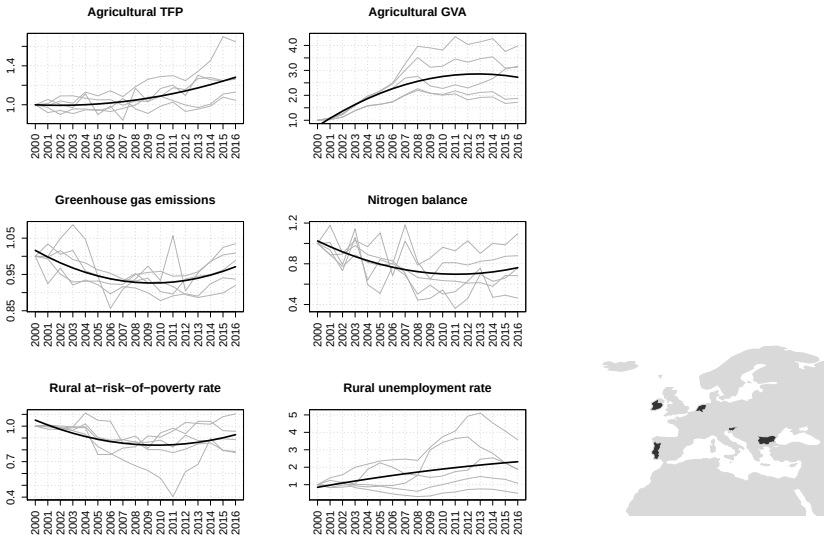


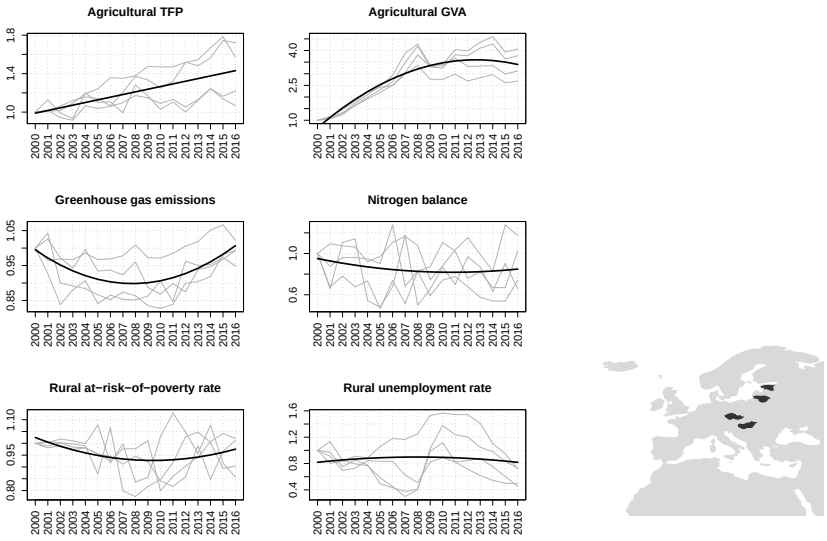
Fig. 3 Trajectories for countries in Group 3: developing agricultural sector with clear sustainability objectives.



3 Results and discussion

We fitted several latent class models, combining a number of polynomial degrees between 2 and 3 with a number of groups between 3 and 12. We selected the model

Fig. 4 Trajectories for countries in Group 4: developing agricultural sector without clear sustainability objectives.



with 2 degrees and 6 groups, as it showed the lowest value of the Bayesian Information Criterion (BIC). Results are described below and displayed in Figures 1-4.

Group 1 (Austria, Belgium, Denmark, Finland, France, Germany, Italy, Norway, Poland, Sweden, United Kingdom) is characterized by increasing TFP, increase then slowdown in GVA decreasing emissions and nitrogen balance, slightly decreasing rural poverty and unemployment (Figure 1). This group includes countries characterized by a mature and sustainable agricultural sector.

Group 2 (Croatia, Cyprus, Greece, Malta, Slovakia, Spain) is characterized by slightly increasing TFP, increasing GVA, decreasing emissions, nitrogen balance and rural poverty, increasing rural unemployment (Figure 2). This group includes countries characterized by a mature agricultural sector, but with weak social sustainability.

Group 3 (Bulgaria, Ireland, Netherlands, Portugal, Slovenia) is characterized by increasing TFP, increase then slowdown in GVA, increase then slowdown in emissions and nitrogen balance reduction, slightly decreasing of rural poverty and increasing in rural unemployment (Figure 3). This group includes countries characterized by a developing agricultural sector with clear sustainable objectives.

Group 4 (Czechia, Estonia, Hungary, Lithuania) is characterized by increasing TFP, increase then slowdown in GVA, U-shaped pattern (with null final balance) in emissions, stable nitrogen balance, rural poverty and unemployment. (Figure 4). This group includes countries characterized by a developing agricultural sector without clear sustainable objectives.

Group 5 (Iceland, Romania, Switzerland) and Group 6 (Latvia, Luxembourg) do not allow to deduce any clear tendency (results not shown).

We see that variability between trajectories is large for some groups, due to some outlier trends. Note that if variability is suspected to bias some group-based trajectories, it is possible to increase the number of groups to keep the R-squared values above a certain threshold, eventually at the cost of obtaining singleton groups.

4 Concluding remarks

In this paper, we proposed and applied latent a class multivariate trajectory model to perform a multiple outcome evaluation of the agricultural sector. Our model clearly identified four groups of countries with emblematic characteristics of the agricultural systems: mature and sustainable, mature but with weak social sustainability, developing with and without clear sustainable objectives.

Our approach is able to classify a set of countries with respect an arbitrary number of indicators, thus representing a valid alternative to the environmentally-adjusted productivity measures recently proposed in the literature [3].

The main limitation of our approach relies in quality and availability of data. Publicly released time series, especially for environmental and social indicators, are typically short or present structural breaks due to changes in classifications and nomenclatures, together with measurement errors and missing values.

References

[1] J. W. Bennett, T. G. Kelley and M. K. Maredia (2012). Integration of environmental impacts into ex-post assessments of international agricultural research: conceptual issues, applications, and the way forward. *Research Evaluation*, **21**: 216-228.

[2] K. Fuglie, M. Clancy, P. Heisey and J. McDonald (2017). Research, productivity and output growth in US Agriculture. *Journal of Agricultural and Applied Economics*, **49**(4): 514-554.

[3] V. N. Hoang and T. Coelli (2011). Measurement of agricultural total factor productivity growth incorporating environmental factors: a nutrients balance approach. *Journal of Environmental Economics and Management*, **62**(3): 462-474.

[4] D. S. Nagin (2005). Group-based modeling of development. Harvard University Press, Cambridge, US-MA.

[5] O. Serrat (2017). The sustainable livelihoods approach. In: Knowledge Solutions, Springer, pp 21-26.

Nowcasting GDP using mixed-frequency based composite confidence indicators

Previsioni a breve termine del PIL con indicatori compositi di confidence basati su frequenze miste

Maria Carannante, Raffaele Mattera, Michelangelo Misuraca, Germana Scepi and Maria Spano

Abstract Composite confidence indicators are widely used to nowcast GDP. In this paper, we aim to construct a new composite confidence indicator which weighting scheme reflects the impact of consumer and business confidence on economic conditions. While GDP is quarterly measured, confidence indicators are monthly recorded. Our approach allows us to deal with data sampled at different frequencies. In particular, we propose a weighting scheme estimation based on U-MIDAS regression techniques.

Abstract *Gli indicatori compositi di confidence sull'andamento dell'economia sono largamente utilizzati per la previsione a breve termine del PIL. In questo lavoro noi proponiamo l'utilizzo di un nuovo sistema di pesi per la costruzione di un indicatore composito di confidence. Il PIL è una variabile economica rilevata quadrimestralmente, mentre gli indicatori di confidence sono rilevati mensilmente. Per superare questo problema, il nostro approccio sviluppa uno schema di pesi dell'indicatore composito basato su tecniche di regressione U-MIDAS.*

Key words: economic forecasting, nowcasting, sentiment, mixed frequency, Unrestricted MIDAS

Maria Carannante
Department of Economics and Statistics - University of Naples "Federico II", e-mail: maria.carannante2@unina.it

Raffaele Mattera
Department of Economics and Statistics - University of Naples "Federico II", e-mail: raffaele.mattera@unina.it

Michelangelo Misuraca
Department of Business Administration and Law - University of Calabria, e-mail: michelangelo.misuraca@unical.it

Germana Scepi
Department of Economics and Statistics - University of Naples "Federico II", e-mail: germana.scepi@unina.it

Maria Spano
Department of Economics and Statistics - University of Naples "Federico II", e-mail: maria.spano@unina.it

1 Introduction

Traditional forecasting requires the explicit formulation of a model, including precise specification of the variables to be included. As pointed out by [4], confidence indicators (e.g. business and consumers) are important variables for predicting subsequent economic conditions. Indeed, if consumers and manufacturers feel confident about the current and the future economic conditions, they might increase their consumption and production respectively. However, we should mention that the issue of predicting the present (i.e. nowcasting) is a very complicated task since macroeconomic data are released with different frequencies. Information about the Gross Domestic Product (GDP) is quarterly available, some others (e.g. Industrial Production, Confidence Indicators and so on) monthly or even daily. For nowcasting a low-frequency variable, as the GDP, by using high frequency "leading indicators", not standard econometric methods have been developed (for a review, see [1]). One of the most recent approach is based on the use of composite confidence indicators as "leadings" ([3, 4]).

In this paper, we aim to construct a new composite confidence indicator for GDP nowcasting. In this framework, the European Commision has developed the European Economic Sentiment Indicator (ESI) [3]. The ESI is a survey-based composite indicator based on four sub-indicators related to business surveys (for the industrial, service, construction and retail sector) and one on consumers survey. The weights associated with each sub-indicator are based on the "relevance criterion" [3]: briefly, they reflect the relevance of the corresponding sector in the total economy as share of GDP. If the interest is nowcasting GDP, the weights for building the composite index could be obtained through a direct approach, by estimating the impact of each sub-indicator on the GDP. In this case, data are sampled at different frequencies (quarterly and monthly) so mixed frequency regression methods have to be used (for a review see [1]). In this paper, we propose to use a weighting scheme based on an Unrestricted-MIDAS regression [2] (U-MIDAS) and show its properties in our case study.

The paper is structured as follows. In the next Section, our proposal is presented in detail. In Section 3 we apply such methodology to Italian GDP nowcasting.

2 Methodology

When the aim is to build a composite indicator, we have to take into account different aspects. Firstly, it is necessary to clearly define the phenomenon under investigation by considering all its sub-components. For each component, it is important assessing its relative importance in explaining the phenomenon and defining its weight.

In this paper, our interest is focused on the public confidence (sentiment) on overall economic conditions in terms of GDP. Public sentiment indicators are obtained from confidence surveys conducted by National Statistical Institutes. In order

to build an overall composite confidence indicator, weighting accurately each of the sub-components is a very important issue. For comparing our results with the ESI ones, we consider four indicators of public confidence for different business sectors (manufacturing, construction, services and retail) and one confidence indicator for the consumers.

In this context, different weighting schemes have been proposed [3]. Most of the composite indicators are based on equally weighted sub-components. In this scheme all sub-components have the same importance in constructing the composite index. This scheme can be used also when not enough information is available about the contribution of each sub-component. Other approaches are based on the definition of weighting schemes based on expert opinions. The ESI scheme, instead, assigns weights according to the share of GDP produced by the sector over the total [3].

We propose to compute the weights by means of regression models. This family of methods are well established in the literature of forecast combinations [5] but not in the composite indicators one.

Suppose we want to nowcast GDP through a new monthly Composite Confidence Index based on an alternative weighting scheme (*AW*). We denote: GDP as the quarterly Gross Domestic Product, CCI the monthly Consumer Confidence Index, MS the monthly Manufacturing sector Confidence Index, CS the monthly Construction sector Confidence Index, SS the monthly Services sector Confidence Index and RS the monthly Retail sector Confidence Index. We propose to compute the weights of the Composite Confidence Index *AW* by performing the following U-MIDAS regression:

$$\begin{aligned}\Delta GDP_t = \beta_0 + \sum_{i=0}^{2} \beta_{i+1}^{(1)} \Delta CCI_{t-i/3}^{(3)} + \sum_{i=0}^{2} \beta_{i+1}^{(2)} \Delta MS_{t-i/3}^{(3)} + \sum_{i=0}^{2} \beta_{i+1}^{(3)} \Delta CS_{t-i/3}^{(3)} + \\ + \sum_{i=0}^{2} \beta_{i+1}^{(4)} \Delta SS_{t-i/3}^{(3)} + \sum_{i=0}^{2} \beta_{i+1}^{(5)} \Delta RS_{t-i/3}^{(3)} + \varepsilon_t\end{aligned} \tag{1}$$

under the constraints: $\sum_{j=1}^{5} \sum_{i=0}^{2} \hat{\beta}_{i+1}^{(j)} = 1$ and $\hat{\beta}_{i+1}^{(j)} \geqslant 0$

where Δ represents the first difference operator and $\beta_{i+1}^{(j)}$ ($i = 0, \ldots, 2$, $j = 1, \ldots, 5$) the parameters estimated *via* OLS. Therefore all the indicators involved in the analysis have to be stationary.

When the frequency mismatch is not so huge, Foroni *et al.* [2] showed that U-MIDAS is preferable to a classical MIDAS approach. In order to obtain a single weight $\hat{\beta}^{(j)}$ for each sub-component, I_j, we aggregate the inter-quarterly weights $\hat{\beta}_{i+1}^{(j)}$. For instance, in the case of CCI:

$$\hat{\beta}^{(1)} = \sum_{i=0}^{2} \hat{\beta}_{i+1}^{(1)}$$

At the end of the estimation process, the *AW* composite confidence indicator is computed as:

$$AW = \sum_{j=1}^{5} \hat{\beta}^{(j)} I_j \tag{2}$$

The use of a U-MIDAS approach gives the possibility to compute the relevance of each sub-component by considering both the inter-quarterly impact on GDP and their evolution over time. Therefore, we tested also a time-varying approach, in which the weights and the resulting composite indicator can change during the time.

3 Application to Italian GDP nowcasting

We downloaded the Italian GDP (Fig. 1) and the Confidence indicators data from the ISTAT website[1]. Since it is necessary that all the sub-components are available in the same interval, we considered data from 1st March 2005 to 1st June 2019.

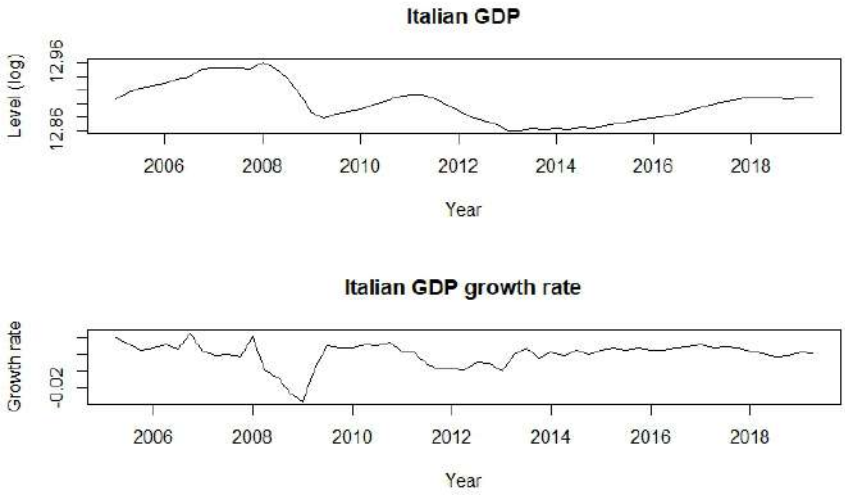


Fig. 1 Italian quarterly GDP: log-levels and growth rate (first difference)

[1] http://dati-congiuntura.istat.it/

The alternative weighting (AW) scheme returns very similar weights compared to ESI (Table 1).

Table 1 Composite Sentiment Indicator: weights

	AW	EW	ESI
Manifacturing sector Confidence	41.5%	20%	40%
Construction sector Confidence	25.9%	20%	5%
Services sector Confidence	1.8%	20%	30%
Retail sector Confidence	6.8%	20%	5%
Consumer Confidence	23.7%	20%	20%

Nevertheless, we noted that the weights associated with the construction and services sectors are overthrown. The AW scheme assigns an higher weight to the construction sector than the services one. This could mean that changes in confidence levels of construction firms have a much higher impact on GDP growth rate than changes in services confidence. This perspective is not taken into account by the ESI. Fig. 2 shows the composite indicators obtained with the three different weighting schemes.

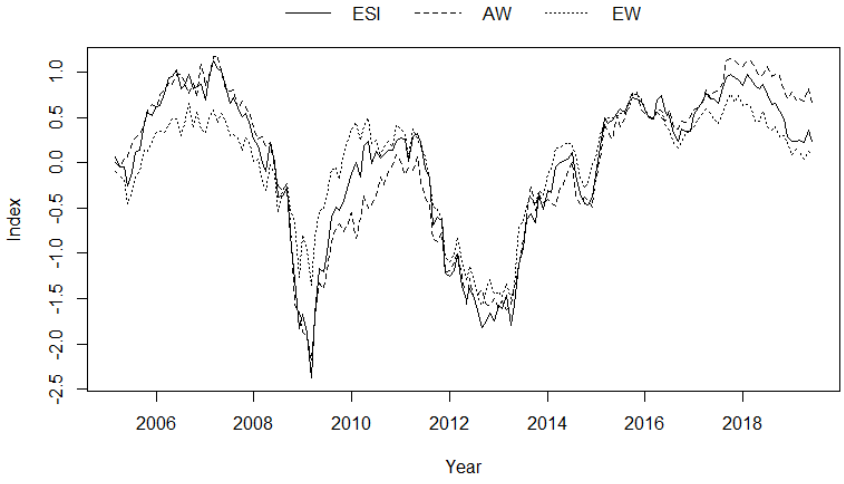


Fig. 2 Composite Confidence Indicators

We tested for stationarity but all the time series are not stationary. In order to assess the nowcasting power of the three composite confidence indicators, we performed the following U-MIDAS regression:

$$\Delta GDP_t = \alpha + \sum_{i=0}^{2} \beta_{i+1} \Delta CI^{(3)}_{t-i/3} + \sum_{i=0}^{2} \gamma_{i+1} \Delta X^{(3)}_{t-i/3} + \varepsilon_t \tag{3}$$

where α is a constant term, *CI* is the composite indicator, that is ESI, *EW*, *AW* or $AW^{\dagger}$ (*AW* with time-varying weights) as applicable, X_t is a matrix containing the other leading indicators, that is the Industrial Production Index and the Unemployment Rate, and β and γ are the respective weights.

Table 2 shows the nowcasting accuracy according to three different criteria: Mean Square Forecast Error (MSFE), Mean Absolute Forecast Error (MAFE) and Root-Mean Square Forecast Error (RMSFE).

Table 2 Prediction accuracy (MIDAS): results with rolling approach

	MSFE	MAFE	RMSFE	MSFE	MAFE	RMSFE
	(1)	(2)	(3)	(4)	(5)	(6)
AW	$1.26e^{-05}$	0.003086	0.003561	$6.23e^{-06}$	0.002007	0.002497
$AW^{\dagger}$	$1.26e^{-05}$	0.003029	0.003563	$6.15e^{-06}$	0.002077	0.002479
EW	$1.56e^{-05}$	0.003191	0.003954	$6.84e^{-06}$	0.002295	0.002616
ESI	$1.35e^{-05}$	0.003016	0.003681	$7.42e^{-06}$	0.002385	0.002724
X_t	NO	NO	NO	YES	YES	YES

According to the accuracy measures, the composite sentiment indicator with AW scheme has better power in nowcasting GDP than the other ones. However, we noted that ESI weighting scheme seems to be more accurate than the equally weighted one. If we consider additional leading indicators (column 4-6) the results are quite similar. The AW scheme still outperforms all the other approaches. In this case, the use of equal weights gives more accurate nowcasts for GDP than the ESI ones. In particular, we highlight that in both scenarios the $AW^{\dagger}$ approach is the best one.

Our results suggest that a data-driven approach for the weighting estimation returns better nowcastings for the Italian GDP than the other well-established alternatives.

References

[1] C. Foroni and M. Marcellino. A survey of econometric methods for mixed-frequency data. 2013.

[2] C. Foroni, M. Marcellino, and C. Schumacher. Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):57–82, 2015.

[3] S. Gelper and C. Croux. On the construction of the european economic sentiment indicator. *Oxford Bulletin of Economics and Statistics*, 72(1):47–62, 2010.

[4] D. Giannone, L. Reichlin, and S. Simonelli. Nowcasting euro area economic activity in real time: the role of confidence indicators. *National Institute Economic Review*, 210(1):90–97, 2009.

[5] A. Timmermann. Forecast combinations. *Handbook of economic forecasting*, 1:135–196, 2006.

On the tangible and intangible assets of Initial Coin Offerings

Panoramica degli asset tangibili e intangibili delle Initial Coin Offerings.

Paola Cerchiello and Anca Mirela Toma

Abstract The Initial Coin Offering (ICO) is one of the operations based on DLT or blockchain technology that allows fundraising activities for an entrepreneurial project, by issuing utility tokens instead of a security or an equity token. Based on a combination of different datasets with over 3000 observations and 30 features (such as social channels, amount raised, the duration, the rating, the community chat of the official telegram channel), the relative sentiment and the whitepaper characteristics) we provide an empirical approach how the wallet of tangible and intangible assets can be built in order to achieve the project's goals. By means of several statistical and text mining techniques we are able to exploit all the parameters in order to interpret the weights of features as ICO success or failure drivers.

Abstract *L'Initial Coin Offering (ICOs) è una delle possibili operazioni basate sulla tecnologia DLT o blockchain che consente a progetti imprenditoriali di raccogliere fondi, emettendo token di utilità anziché un token di sicurezza o di equità. Basato su una combinazione di diversi set di dati con oltre 3000 osservazioni e 13 variabili (come importo raccolto, durata, valutazione, chat della comunità del canale ufficiale di Telegram, relativo sentiment e caratteristiche del white paper) forniamo un approccio empirico su come costruire il portafoglio di beni materiali e immateriali per raggiungere gli obiettivi del progetto. Tramite diverse tecniche statistiche siamo in grado di sfruttare tutti i parametri al fine di interpretare i pesi delle variabili come driver di successo o fallimento delle ICO*

Key words: ICOs, sentiment analysis, textual analysis, whitepaper, ...

Paola Cerchiello
University of Pavia, e-mail: paola.cerchiello@unipv.it

Anca Mirela Toma
University of Pavia e-mail: ancamirela.toma01@universitadipavia.it

1 Introduction

The first Initial Coin Offering (ICO) was held in 2013. Since then, this new way of raising capital keeps investors and regulators around the world occupied as the number of held ICOs skyrocketed. An ICO (also called a token sale event) constitutes an alternative funding method for young and technically involved ventures. They offer a self created cryptocurrency to investors at the exchange of fiat money or established cryptocurrency. The sold currency, called a 'token' can meet different requirements. It can be constructed as a mean of payment, to enable access to services or else. It is however purely digital and the real world value depends on the success of the issuing venture. As the ICO environment grew big in a very small time frame, regulators around the world are still making a substantial effort to increase investor protection whilst fostering innovation in the Financial Technology (FinTech) environment. At the moment the regulation in terms of ICOs is characterized by the lack thereof. This leaves potential for scams and fraudulent activities[**?**]. Due to the continuing hype around the topic, gains for issuers of fraudulent ICOs can be astronomically high. In April 2018, a scam beyond the imaginary made heeadlines. The Vietnamese company Modern Tech JSC, had vanished into thin air with the generated ICO funds of USD 658 million. Over 32,000 investors were affected by the scam. Even though this was the biggest, scams in the environment are quite common. This work aims at addressing the specific characteristics of ICOs using relevant variables that play a key role in determining the success of the ICO.

As it stands there is no database with the information we are looking for, thus we have been building and constantly maintaining a dataset that is currently composed of 195 ICOs that occurred between October 2017 and November 2018. The database comprises companies from European countries namely France, Germany, Switzerland, Estonia, Latvia and non European countries such as Russia, United Kingdom, United States, Japan, Singapore, and Australia. The most common sectors in which ICOs operate are: high-tech services, financial services, smart contract, gambling platforms, marketplaces and exchanges.

2 Methodology

In this paper we leverage two kinds of information: structured and unstructured ones. Regarding the former, we take advantage of classical statistical classification models to distinguish the stauts of an ICO that made up of 2 classes, intended as follows:

- Success = 1: the ICO collects the predefined hard cap within the time horizon of the campaign;
- Failure/Scam = 0: the ICO does not collect the predefined hard cap within the time horizon of the campaign.

Logistic regression aims at classifying the dependent variable into two groups, characterized by a different status (1=scam vs 0=success or 1=success vs 0=failure) according to the following model:

$$ln(\frac{p_i}{1-p_i}) = \alpha + \sum_j \beta_j x_{ij}, \tag{1}$$

where p_i is the probability of the event of interest, for ICO i, $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})$ is a vector of ICOs-specific explanatory variables, the intercept parameter α, as well as the regression coefficients β_j , for $j = 1, ..., J$, are to be estimated from the available data. It follows that the probability of success (or scam) can be obtained as:

$$p_i = \frac{1}{1 + exp{-(\alpha + \sum_j \beta_j x_{ij})}}. \tag{2}$$

Considering the textual analysis of Telegram chats, we take advantage of quantitative analysis of human languages to discover common features of written text. In particular the analysis of relatively short text messages like those appearing on micro-blogging platform presents a number of challenges. Some of these are, the informal conversation (e.g. slang words, repeated letters, emoticons) and the level of implied knowledge necessary to understand the topics of discussion. Moreover, it is important to consider the high level of noise contained in the chats, witnessed by the fact that only a fraction of them with respect to the total number available is employed in our sentiment analysis.
We have applied a Bag of Word (BoW) approach, according to which a text is represented as an unordered collection of words, considering only their counts in each comment of the chat. The word and document vectorization has been carried out by collecting all the word frequencies in a Term Document Matrix (TDM). Afterwards, such matrix has been weighted by employing the popular TF-IDF (Term Frequency Inverse Document Frequency) algorithm. Classical text cleaning procedures have been put in place like stop-words, punctuation, unnecessary symbols and space removal, specific topic words addition. For descriptive purposes we have used word-clouds for each and every Telegram chat according to the general content and to specific subcategories like sentiments and expressed moods. The most critical part of the analysis relies on the sentiment classification. In general, two different approaches can be used:

- Score dictionary based: the sentiment score is based on the number of matches between predefined list of positive and negative words and terms contained in each text source (a tweet, a sentence, a whole paragraph);
- Score classifier based: a proper statistical classifier is trained on a large enough dataset of pre-labelled examples and then used to predict the sentiment class of a new example.

However, the second option is rarely feasible because in order to fit a good classifier, a huge amount of pre-classified examples is needed and this represents a particularly complicated task when dealing with short and extremely non conventional text like micro-blogging chats. Insofar, we decided to focus on a dictionary based approach, adapting appropriate lists of positive and negative words relevant to ICOs topics in English language.
The lexicons used are based on unigrams, i.e., single words, they contain many English words and the words are labeled with scores for positive/negative sentiment and also possibly emotions like joy, anger, sadness, and so forth. The NRC lexicon categorizes words in a binary fashion ("yes"/"no") into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The BING lexicon categorizes words into a binary manner into positive and negative categories. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. By applying the above described lexicons, we produce for each and every ICO a sentiment score as well as counts for positive and negative words. All these indexes are used as additional predictors within the logistic models.

3 Data

In this paper, we leverage on structured and unstructured information and empirically examine 195 ICOs starting from January 2017 till November 2018.
The first step in collecting data about each project is to gather information from the most used ICO related platforms as icobench, TokenData, coinschedule or similar. During such phase, we look for general characteristics such as the name, the token symbol, start and end dates of the crowdfunding, the country of origin, financial data such as the total number of issued token, the initial price of the token, the platform used, data on the team proposing the ICO, data on the advisory board, data on the availability of the website, availability of white paper, whitepaper characteristics and social channels.
Some of these data, such as short and long description, and milestones are textual descriptions. Others are categorical variables, such as the country, the platform, the category (which can assume many values), and variables related to the team members (name, role, group). The remaining variables are numeric, with different degrees of discretization.
As it concerns the unstructured data, insightful information can be derived from the white papers in terms of quality of the technical report and specific content. A white paper is a summary report that provides detailed information about the project, its originality and the benefits it can give to investors and users, about the technological features, team behind the project, project's background and future plans. The dimensions captured through the white paper features consists in checking the availability of the document for the ICO in question and additional

information about its scope (pages, sections, appendix). Furthermore, data on the disclosure of information regarding the issuers was collected (names and/or photographs of issuing and advisoring team members).
Social channels are more personal than every database, rating platform or websites, so they are a way to reach a wide range of users, to update them constantly about the evolution of the project and in the end to create a trusty environment that can finalize in a successful crowdfunding activity. In order to conduct the textual analysis, we enrich our database with the social channels data, such as the presence of a channel, the numbers of users as a proxy of the community engagement and as mentioned in the introduction the textual chat, retrieved backward till the creation of the chat, used to produce a sentiment based score of every ICO.

4 Results

The best model specified by AIC stepwise selection of features is reported in Table 1. The null deviance of the model is 210.680 on 194 degrees of freedom, the log-likelihood of the final model has a value of -28.821.

Table 1 Results of the logistic regression

	Dependent variable:
	class2
Telegram_du	-1.846^{**}
	(0.922)
Nr_Telegram	0.001^{**}
	(0.0002)
nr_team	0.282^{***}
	(0.103)
tw	3.884^{**}
	(1.743)
nr_adv	0.531^{***}
	(0.184)
Sent_NRC_sc	2.257^{***}
	(0.852)
name_team	1.670^{*}
	(0.876)
picture	-1.682^{*}
	(1.010)
Constant	-4.632^{**}
	(2.001)
Observations	195
Log Likelihood	-28.821
Akaike Inf. Crit.	75.641
Note:	*p<0.1; **p<0.05; ***p<0.01

It can be seen that the existence of a Telegram communication channel has a significant impact with a negative sign on the log odds of a successful ICO. The standard error for this term is 0.922. This means, that if all other variables are held constant, the existence of a Telegram channel has a negative effect of -1.846 on the log odds of a successful ICO. The number of people enrolled in the Telegram chat also has a significant effect on the dependent variable, however the effect is much smaller and has a positive sign. This means that with the amount of people participating in the communication channel, odds of a successful ICO rise. The number of members of the issuing team of an ICO has a highly significant effect with a positive sign on the log odds of a successful ICO. The existence of a twitter account dedicated to the ICO also has a highly significant impact with a positive sign. The dimension of this effect, all other variables held constant, is more than twice as high as the aforementioned effect of the existence of a Telegram channel. The number of advisors has a highly significant effect exhibiting a positive sign. The sentiment extracted from Telegram chats also has a high positive sign significant effect.

Acknowledgements This research has received funding from the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training program" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA). The European CommissionFurthermore we acnkowledge Federico Campasso and Anna Bartch for their contribution to the data collection.

References

1. Adhami S., Giudici G., Martinazzi S.: Why Do Businesses Go Crypto? An Empirical Analysis of Initial Coin Offerings. (2017) Available at SSRN: https://ssrn.com/abstract=3046209
2. BIS Annual Economic Report.: V. Cryptocurrencies: looking beyond the hype (2018)
3. Blaseg, D.: Dynamics of Voluntary Disclosure in the Unregulated Market for Initial Coin Offerings. (2018) Available at SSRN: https://ssrn.com/abstract=3207641 or http://dx.doi.org/10.2139/ssrn.3207641
4. Falk, M., Haler, J., and Reiss, R.: Laws of Small Numbers: Extremes and Rare Events. Springer (2010)
5. Fisch C.:Initial Coin Offerings (ICOs) to Finance New Ventures. Journal of Business Venturing, 34(1), 1–22. (2019)
6. Howell S. T., Niessner M., and Yermack D.: Initial coin offerings: Financing growth with cryptocurrency token sales. (No. w24774). National Bureau of Economic Research (2018)
7. Kotz, S. and Nadarajah, S.: Extreme Value Distributions. Theory and Applications. Imperial College Press, London (2000)
8. Zetzsche D.A. , Buckley R.P., Arner D.W., Föhr L.: The ICO Gold Rush:It's a Scam, It's a Bubble, It's a Super Challenge for Regulators. Univ.Luxembourg Law Working Paper No. 11 (2017)
9. Wright A. , De Filippi P.: Decentralized Blockchain Technology and the Rise of Lex Cryptographia. (2015) Available at SSRN: https://ssrn.com/abstract=2580664 or http://dx.doi.org/10.2139/ssrn.2580664

Seasonality variation of electricity demand: decompositions and tests

La variazione della stagionalità nella domanda di elettricità: scomposizioni e test

Luigi Grossi and Mauro Mussini

Abstract The change over time in seasonal concentration of electricity demand is broken down into two components. One component measures the contribution of the changes in the seasonal pattern of demand. The second component measures the contribution of the change in the magnitude of seasonality.
Abstract *La variazione nella concentrazione stagionale della domanda di energia elettrica è scomposta in due componenti. La prima componente misura il contributo dei cambiamenti nell'andamento stagionale. La seconda componente misura il contributo della variazione nell'intensità della stagionalità.*

Key words: decomposition**,** Gini index**,** seasonal concentration, electricity demand

1 Introduction

Liberalized electricity markets in the main economies of the world are organized by means of daily auctions where operators present their offers and bids. The most important session of the Italian electricity market (IPEX) is the day-ahead market (Mercato del Giorno Prima, MGP). In this market session, the equilibrium price and quantity for each hour (load period) of the next day is obtained by the intersection of the aggregated curves of demand and supply. Because of the intermittent nature of renewable sources, such as wind and solar, the regulator of the market (GME, Gestore

[1] Luigi Grossi, Dipartimento di Scienze Economiche, Università degli Studi di Verona; email: luigi.grossi@univr.it

Mauro Mussini, Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano Bicocca; email: mauro.mussini1@unimib.it

del Mercato Elettrico) provides the market with one-day-ahead predictions of the hourly electricity demand in order to ensure that the total power supply meets the global demand. It is then clear that the correct analysis and prediction of the dynamics of the demand is crucial in order to avoid shortages of supply and consequent black-outs [4]. Hourly time series of electricity demand are characterized by empirical regularities called "stylized facts" such as long-run cycles, presence of spikes and multiscale seasonality [6]. Seasonality can be thought as regular fluctuations with similar timing and magnitude from day to day, week to week and month to month.

Seasonality implies that electricity demand tends to be concentrated in certain periods of the day/year (peak periods), generating an unequal distribution of demand among the periods [5]. From this perspective, the degree of seasonality can be measured in terms of seasonal concentration by using a conventional concentration index, such as the Gini index or the Theil index.

Concentration indexes are descriptive statistical tools which were originally introduced in the literature on income inequality, but their field of application crossed that of the measurement of income inequality. Such indexes were used to quantify the level of seasonality in several studies examining the evolution of seasonality in tourist flows [1], [2]. The change over the days/years in seasonal concentration can be measured by observing the change in the concentration index. However, a basic element of seasonality cannot be monitored by examining solely the change in the concentration index. An essential feature of seasonality is that the intra-day/year fluctuations occur with the same timing from day to day and from year to year. However, the value of a concentration index does not vary if fluctuations have the same magnitude but a different timing. Since a conventional concentration index is invariant to changes in the seasonal pattern, a measure of seasonal stability is needed in addition to a measure of the change in seasonal magnitude when analysing the evolution of seasonality. In this paper, we show that these measures can be obtained by decomposing the change in seasonal concentration. In particular, seasonal concentration in a given day/year is measured by using the Gini index [3], a very popular measure of concentration. After measuring seasonal concentration in two different days/years, the change in seasonal concentration is split into two components. One component is a measure of seasonal stability, which tracks changes in the seasonal pattern. The second component measures the change in seasonal magnitude, assuming that the seasonal pattern is stable.

2 Decomposing the change in seasonal concentration

Let Y be a variable representing the forecasted demand of electricity in an area. Let $y_{1,t}, \cdots, y_{n,t}$ be the time series of electricity demand at time t, where t could be a generic day or a generic year; n is the number of periods in a day or in a year (e.g., $n = 24$, for hourly data, $n = 12$ for monthly observations) and $y_{i,t}$ is the electricity demand of period i in day/year t. To make the notation lighter, from now on, we will focus on the concentration of the hourly demand in a day, but the procedure might be

easily extended to the concentration of the monthly demand in a year. The Gini index measuring the concentration of electricity demand in t is

$$G_t = \frac{2cov[y_{i,t}, r(y_{i,t})]}{n\bar{y}_t}, \quad (1)$$

where $\bar{y}_t$ is the average electricity demand per period and $r(y_{i,t})$ is the rank of period i according to the increasing order of electricity demand.

Now suppose that the daily distribution of electricity demand is observed in day $t+k$. The change in the concentration of demand from day t to day $t+k$ is measured by the difference between the Gini index in $t+k$ and the Gini index in t:

$$\Delta G = G_{t+k} - G_t = \frac{2cov[y_{i,t+k}, r(y_{i,t+k})]}{n\bar{y}_{t+k}} - \frac{2co\ [y_{i,t}, r(y_{i,t})]}{n\bar{y}_t}. \quad (2)$$

However, ΔG does not necessarily capture all aspects of the change in seasonality. For instance, consider the numerical illustrations in figure 1 showing two hourly distributions of demand in days t and $t+k$. The twenty-four absolute figures in $t+k$ are equal to those in t but the ranking of hours by electricity demand has changed. The maximum value is observed at hour 14 at day *t*, while maximum demand at day *t*+*k* is at hour 12; hour 14 at day *t*+*k* has become the second highest demand value. Moreover, moving from day *t* to day *t*+*k*, hour 13 has changed its position from 23rd to 22nd. In such a situation, the Gini index in $t+k$ is equal to the Gini index in t. ΔG equals zero, suggesting that seasonality has remained unchanged since t, but the distribution of demand among hours has changed. To capture the changes in seasonality depicted in figure 1, a measure of seasonal stability and a "pure" measure of the change in seasonal magnitude are needed. These measures are obtained by decomposing ΔG.

Let $C_{t+k|t}$ stand for the concentration coefficient of electricity demand in $t+k$ obtained by sorting periods according to their ranking in t instead of that in $t+k$:

$$C_{t+k|t} = \frac{2cov[y_{i,t+k}, r(y_{i,t})]}{n\bar{y}_{t+k}}. \quad (3)$$

By adding and subtracting $C_{t+k|t}$ to the right-hand side of equation (2), the difference in the Gini index is broken down into two components:

$$\Delta G = \left\{\frac{2cov[y_{i,t+k}, r(y_{i,t+k})]}{n\bar{y}_{t+k}} - \frac{2cov[y_{i,t+k}, r(y_{i,t})]}{n\bar{y}_{t+k}}\right\} + \left\{\frac{2cov[y_{i,t+k}, r(y_{i,t})]}{n\bar{y}_{t+k}} - \frac{2cov[y_{i,t}, r(y_{i,t})]}{n\bar{y}_t}\right\}$$

$$\Delta G = R + M. \quad (4)$$

R in equation (4) is the re-ranking component measuring the change in concentration due to the position exchanges between periods in the ranking of periods from t to $t+k$. The re-ranking component is analogous to the re-ranking measure used for capturing the re-ranking of income receivers when decomposing the change in income inequality [5]. R equals zero if the ranking of periods in $t+k$ is the same as that in t. The component R reaches its highest value, that is $2G_{t+k}$, when the ranking

of periods in $t+k$ is completely reversed with respect to the ranking of periods in t. Since the re-ranking component captures any deviation in the seasonal pattern, R is seen as a measure of seasonal stability. The larger the re-ranking component, the more unstable the seasonal pattern.

The term M in equation (4) is the magnitude component measuring how much electricity demands in day $t+k$ are more (or less) concentrated on the periods with the highest ranks in day t. Since M is calculated by holding periods sorted according to their ranking in t, the seasonal pattern is supposed to be the same in t and $t+k$. The magnitude component is positive (negative) when the magnitude of seasonality has increased (decreased) over time. M is equal to zero if the electricity demand among periods has not changed or if the demand of all periods has changed in same proportion, leaving unchanged the relative disparities between values of demand.

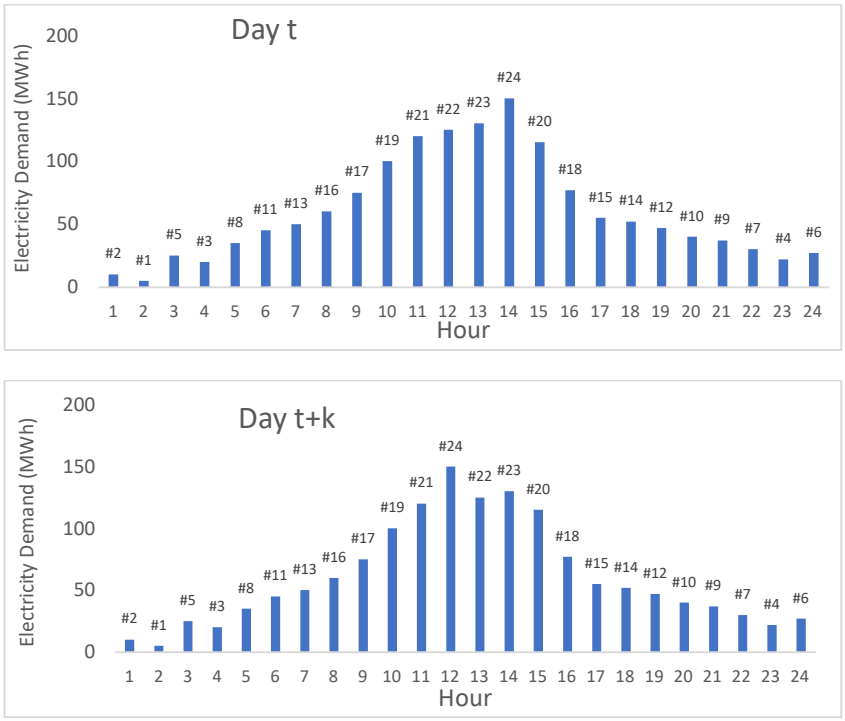


Figure 1: Hourly distributions of electricity demand in days t and $t+k$.

3 Application

The decomposition is used to analyse the evolution of seasonality in the average hourly electricity demand observed in the Italian electricity market called IPEX (see, https://www.mercatoelettrico.org). Although it is well known that the daily cycle of

hourly demand changes within days of the week and within days observed in different months [5], [6], the significance of this changes has never been tested. Moreover, it is very important to see if the change of the daily pattern can be significantly ascribed to one of the two components obtained in the previous section, or both. For instance, due to social behaviour and to the increase of light hours in the summer season the daily cycle of electricity consumption is smoother than in winter days and the peak of demand of electricity moves forward during the day. The first change should influence the magnitude component, while the second should affect the ranking component.

The change in the average daily seasonality over different months of the same year and of different years is examined using data observed in 2016 and 2018.

Table 1: decomposition of the change in seasonality in average daily cycles of different months.

	Indices		Variation	Components		Permutation test		
	G_t	G_{t+k}	ΔG	R	M	p(ΔG|H0)	p(R|H0)	p(M|H0)
January 2016 August 2016	0.10334	0.07505	-0.02829	0.0041	-0.0324	0.001	0.001	0.001
January 2018 August 2018	0.10556	0.07431	-0.03125	0.0051	-0.0363	0.001	0.001	0.001
August 2016 August 2018	0.07505	0.07431	-0.00074	0.0004	-0.0011	0.373	0.001	0.155

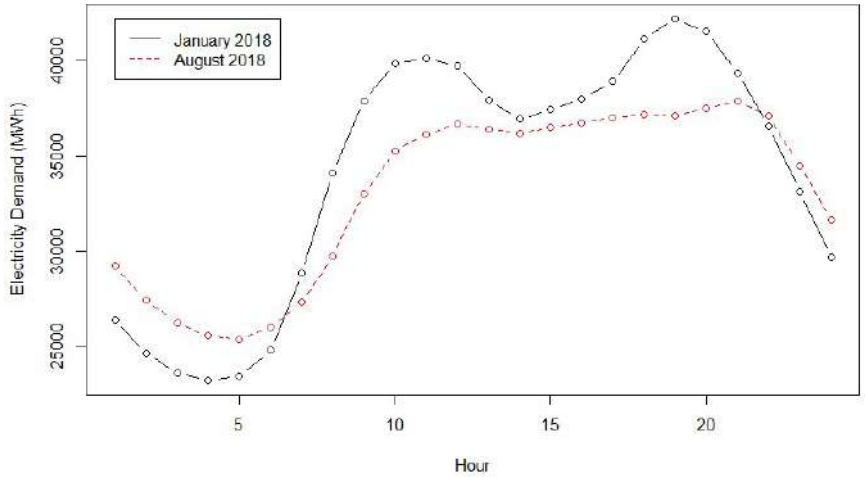


Figure 2: Average daily profile of electricity demand in January and August 2018. Data were downloaded from the website of the GME (Gestore Mercato Elettrico), www.mercatoelettrico.org.

Table 1 shows the output of the decomposition of the change in seasonal concentration comparing average daily cycles observed in different months. Last three columns report the *p*-values of the permutation test. First and second line of the table contains the results of the comparisons between a typical winter day profile (January) and a typical summer day profile (August) in 2016 and in 2018. The seasonal concentration

is lower in August because the demand of electricity in summer is more constant in the middle of the day. From the decomposition observed in the table, it is interesting to note that the reduction of seasonal concentration in summer days (ΔG) is due to a reduction of seasonal magnitude (component M is negative), while the contribution of the seasonal pattern is positive. The two average daily cycles in 2018 can be clearly observed in Figure 2. Although the seasonal concentration magnitude is reduced, the flatter pattern around midday goes in the opposite direction because of the change of many positions in the final ranking of hourly demand. Last row compares the seasonality of the average daily cycle in the same months (August) in 2016 and 2018. Although the global change of concentration is not significant, the seasonal pattern change, measured by R, is statistically significant.

4 Conclusion

In this paper a common approach used in the tourist literature to measure the seasonal concentration in tourist flows has been extended to the study of seasonality of the electricity demand. When measuring the change in seasonal concentration over time, observing the variation in the seasonal concentration index may lead to misleading conclusions since a small change in the index can be the outcome of the offsetting contributions of changes in the seasonal pattern and magnitude. The decomposition of the change in seasonal concentration sheds light on the different aspects of a change in the seasonal cycle of electricity demand, revealing the roles of changes in the seasonal pattern and magnitude. This decomposition provides regulators and other market operators with additional information on the change in the intra-day cycle of electricity consumptions. This is relevant in view of the massive penetration of renewable sources, especially photovoltaic plants, which follows the seasonal cycle of solar radiation.

References

1. Fernández-Morales, A.: Decomposing seasonal concentration. Annals of Tourism Research, 30, 942-956 (2003)
2. Fernández-Morales, A., Mayorga-Toledano, M.C.: Seasonal concentration of the hotel demand in Costa del Sol: A decomposition by nationalities. Tourism Management, 29, 940-949 (2008)
3. Gini, C.: Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. Bologna: Tipografia di P. Cuppini (1912)
4. Gianfreda A., Grossi L.: Forecasting Italian Electricity Zonal Prices with Exogenous Variables, Energy Economics, Vol. 34 (6), pp. 2228-2239 (2012).
5. Jenkins, S., Van Kerm, P.: Trends in income inequality, pro-poor income growth, and income mobility. Oxford Economic Papers, 58, 531-548 (2006)
6. Taylor, J.W.: Triple seasonal methods for short-term electricity demand forecasting, European Journal of Operational Research, Volume 204, Issue 1, Pages 139-152 (2010).
7. Weron, R.: Modeling and Forecasting Electricity Loads and Prices, John Wiley & Sons, Chichester (2006).

SMEs circular economy practices in the European Union: Implications for sustainability

Economia circolare nelle PMI Europee: implicazioni per la sostenibilità

Nunzio Tritto, Josè G. Dias and Francesca Bassi

Abstract This paper studies the willingness of small and medium-sized companies in the European Union to undertake circular economy practices. The dataset comes from a survey involving more than 10,000 SMEs in the EU. This hierarchical structure – companies within countries – was analyzed using a multilevel factor model that takes the heterogeneity between countries into account. Both at company and country levels, there are factors that explain the attitude towards CE. Factor scores at both levels suggest a division between Western and Eastern countries (with some exceptions) regarding willingness to undertake CE activities by SMEs, which identify regional consequences of the EU policies towards CE.

Abstract *Questo lavoro studia la intenzioni da parte delle piccole e medie imprese europee di adottare le pratiche di economia circolare. I dati sono stati raccolti presso più di 10.000 PMI, localizzate nei 28 paesi dell'Unione Europea. La struttura gerarchica delle informazioni è stata tenuta in considerazione stimando modelli fattoriali miltilivello. Si è rilevato che vi sono sia fattori a livello di impresa che a livello di paese che spiegano la propensione nei confronti dell'economia circolare. In particolare, si sono evidenziate marcate differenze tra i paesi dell'Europa Occidentale e quelli dell'Europa Orientale. Questi risultati identificano il verificarsi di conseguenze diverse a livello regionale delle politiche Europee che cercano di favorire l'adozione dell'economia circolare*

Key words: Circular economy, sustainability, European Union, small and medium-sized companies, multilevel models.

1 Introduction

The introduction of the concept of Circular Economy (CE) can be traced at the end of the 20th century when seminal papers were published and it attracted the attention

of many scholars (Lieder and Rashid, 2016). The concept of CE has evolved in the world of business in an attempt to find a compromise between economic growth and environmental protection. This concept is in contrast with the most used idea of linear economy, i.e. take-make-use-dispose. The EU business world has to face the environmental issue for ethical reasons but also because the European Union developed environmental policies for the product life cycle (European Commission, 2003; Dalhammar, 2015). Small and medium-sized enterprises are defined by the European Commission as companies with less than 250 employees and with an annual turnover that does not exceed 50 million euros, or a total annual balance that does not exceed 43 million euros (European Commission, 2003). For the European Union, SMEs are fundamental for CE because they are more active in sectors such as recycling, repair, and innovation; on the other hand, they have some difficulties in applying for funding and in sticking to CE principles if their activity is not directly involved (European Commission, 2015).

This paper aims to analyze the willingness of European SMEs to undertake specific activities related to Circular Economy (CE) and to identify the potential drivers of this behavior. Data are collected from a sample of SMEs operating in the 28 EU Member States. Country-level characteristics are also included and their impact on the overall willingness to undertake CE activities is evaluated. The collected data is hierarchical: SMEs are nested into countries; this required the specification and the estimation of multilevel models to evaluate the impact of factors, at company and country level, on the latent structure.

2 Data and methods

Data come from the Flash Eurobarometer 441 conducted in April 2016 and it contains 10,618 CATI interviews. The 28 countries included in the survey are (alphabetical order): Austria, Belgium, Bulgaria, Cyprus, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and United Kingdom. The data set contains five ordinal indicators on the implementation of CE practices in the past three years: 1. Re-plan of the way water is used to minimize use and maximize re-use; 2. Use of renewable energies; 3. Re-plan energy usage to minimize consumption; 4. Minimize waste by recycling or reusing waste or selling it to another company; and 5. Redesign products and services to minimize the use of materials or use recycled materials. The survey collects various company characteristics: the number of employees, total turnover in 2015, the age, the sector of economic activity, the type of goods and services sold, and the percentage of company's turnover in 2015 invested in Research and Development. Country-level covariates are available on the Eurostat website; however, there exists a problem of multicollinearity among indicators related to the same dimension (social, economic, environmental, institutional). For this reason, we selected only one variable per dimension: illiteracy rate, per capita GDP, generation of waste excluding major mineral wastes per GDP unit, corruption perception index.

Figure 1 summarizes the conceptual model of confirmatory factor analysis with a multilevel factor model (Hox et al., 2010). Two latent variables represent the willingness to undertake CE activities: one at company level (f^W), and one at country level (f^B). The H items Y_h correspond to the dependent variables of interest; the K variables X_k are the company-level covariates; and the set of M variables Z_m are the country-level covariates. This conceptual model represents a combination of a factorial model that links the latent constructs with the H items Y_h, and a linear model that regresses the two latent variables on the corresponding covariates at company and country level in a multilevel setting, respectively. Therefore, the final model assumes unidimensionality, i.e., a single individual latent variable that explains all observed items. Estimation is obtained with Maximum Likelihood (ML) method with Gaussian integration.

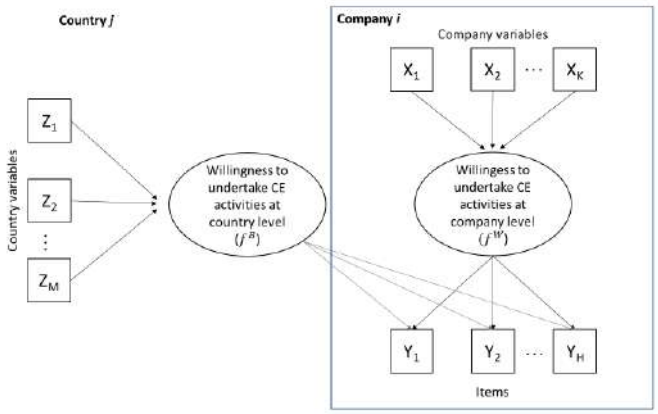


Figure 1: Conceptual model

3 Results

Confirmatory factor analysis (CFA) with one latent variable and five items has a good fit to the data: the RMSEA has a value lower than 0.05 (0.039); indexes CFI and TLI are greater than 0.95 (0.977 and 0.953, respectively). All factor loadings are greater than 0.7. The variances of the latent variables at the company and country level are both statistically significant, accounting for the presence of variability within and between countries. The following step of the analyses is the study of the impact of covariates at company and country level on the willingness to undertake CE practices, represented by the latent variable. We estimate two nested models (Table 1): the first one without country covariates (Model I) and the second one with these covariates (Model II). In general, it is possible to note that the differences between the two models in terms of the values of estimated coefficients, standard errors and p-values are negligible, as expected. The introduction of country-level variables in the model

does not affect other estimates. The estimated variance does not vary significantly as a result of the introduction of the upper level variables in the model. Values in Table 1 suggest that company size affects positively the willingness to undertake CE activities. The relationship between company's turnover in 2015 and undertaking of CE activities is linear and positive as well. Undertaking CE activities does not depend on the age of the company. The implementation of CE practices is more prevalent in companies that belong to more tangible sectors. Table 1 indicates a positive relationship between the percentage of turnover invested in R&D and the willingness to undertake CE practices.

Table 1: Company-level effects

	Model I			Model II		
	estimate	S.E.	p-value	estimate	S.E.	p-value
Sector of activity (ref: Man.)						
Manufacturing						
Retail	-0.239	0.091	0.008	-0.242	0.091	0.008
Services	-0.331	0.098	0.001	-0.336	0.098	0.001
Industry	-0.072	0.068	0.286	-0.075	0.068	0.266
Number of employees (ref: 1-9)						
10 to 49 employees	0.329	0.069	0.000	0.334	0.069	0.000
50 to 250 employees	0.620	0.097	0.000	0.631	0.097	0.000
Foundation (ref: < 1-1-2010)						
1-1- 2010 and 1-1 2015	0.012	0.058	0.838	0.012	0.058	0.832
After 1 January 2015	-0.097	0.206	0.638	-0.101	0.207	0.625
Total turnover (ref: > 25 000)						
> 25 000 to 50 000	-0.022	0.121	0.856	-0.025	0.120	0.832
> 50 000 to 100 000	0.010	0.106	0.928	0.003	0.105	0.976
> 100 000 to 250 000	0.094	0.121	0.435	0.086	0.118	0.470
> 250 000 to 500 000	0.257	0.123	0.036	0.245	0.120	0.041
> 500 000 to 2 million	0.308	0.122	0.012	0.293	0.120	0.015
> 2 to 10 million	0.418	0.122	0.001	0.400	0.120	0.001
> 10 million euros	0.719	0.165	0.000	0.700	0.163	0.000
Selling (multiple choice):						
Products to consumers	0.254	0.058	0.000	0.253	0.059	0.000
Products to companies o	0.184	0.072	0.010	0.183	0.072	0.011
Services to consumers	0.488	0.053	0.000	0.490	0.053	0.000
Services to companies	0.001	0.054	0.990	0.000	0.054	0.998
R & D (%) (ref: < 5%)						
From 5% to 9.9%	0.595	0.076	0.000	0.596	0.076	0.000
From 10% to 14.9%	0.708	0.057	0.000	0.709	0.057	0.000
From 15% to 19.9%	0.834	0.166	0.000	0.834	0.165	0.000
20% or more	0.666	0.129	0.000	0.668	0.129	0.000
Variance	1.962	0.215	0.000	1.967	0.216	0.000

Table 2 lists estimated coefficients for the two models without (Model I) and with covariates at country-level (Model II). Illiteracy rate has a significant and positive but low slope (0.024). Per capita GDP (log-transformed) is not statistically significant. The increase in the variable measuring generation of waste per GDP has a negative, but very low, impact on the latent variable. Finally, the Corruption Perception Index has a not statistically significant impact on undertaking CE activities. In the model without the upper level covariates, the between variance is equal to 0.435 and the ICC is 0.181; as expected, in the model with country-level covariates, the between

variability and the ICC are both lower (respectively 0.214 and 0.098). This means that the heterogeneity between countries is 9.8% of the total variance.

Table 2: Country-level effects

	Model I			Model II		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Illiteracy rate				0.024	0.007	0.001
GDP per capita (ln transformed)				0.230	0.318	0.470
Generation waste per GDP				-0.002	0.001	0.003
Corruption perception index				0.005	0.013	0.689
Variance	0.435	0.141	0.002	0.214	0.105	0.042
ICC	0.181			0.098		

4 Conclusions

This study focused on the willingness of small and medium-sized companies in the European Union to undertake CE practices. The first focus of the analysis was to synthesize the information of a set of variables related to CE practices creating a latent variable –willingness to undertake CE practices – through a Confirmatory Factor Analysis using ordinal data. Then we studied the effect of company-level and country-level covariates on this latent variable. The introduction of the macro-variables considerably reduces the intra-class correlation coefficient, which means that these variables give us important information about the differences between countries.

It is important to underline the limitations and possible developments of the analyses. The entire model has been thought in order to summarize the information about the five items of interest in one latent variable that measures the overall willingness and to study the impact of a small number of covariates at two levels, in order to have a parsimonious model. At company level, we chose to study objective covariates related to the characteristics of single companies, but it was possible also to introduce variables about the perception of the companies towards the CE concept. The choice of the country-level covariates was one possibility based on alternatives supported in the literature. As explained before, there are other indicators that may explain the latent variable, demanding a deeper research about it.

Another limitation involves the type of data analyzed. Usually surveys have the problem of social desirability bias, that is the trend of respondents to answer a question in order to give a better image of themselves, even if the survey is anonymous. This can negatively affect the validity of the survey and no instruments are available to avoid this snag. However, the introduction of the country-level covariates helps reduce this issue because the heterogeneity at the country-level is a good indicator about differences among respondents in cross-cultural research (Hoogendoorn et al., 2015).

5 References

1. Dalhammar, C. The application of 'life cycle thinking' in European environmental law: theory and practice. J Eur Environ Plann Law, 12(2), 97–127 (2015).
2. European Commission. Commission recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises 2003/361/EC. Official Journal of European Union (2003).
3. European Commission. Closing the Loop – An EU Action Plan for the CE. Available on https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52015DC0614 (2015).
4. Hoogendoorn, B., Guerra, D., & van der Zwan, P. What drives environmental practices of SMEs? *Small Business Economics*, 44(4), 759–781 (2015).
5. Hox, J. J., Maas, C. J., & Brinkhuis, M. J. S. The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64(2), 157-170 (2010).
6. Lieder, M., Rashid, A. Towards CE implementation: A comprehensive review in context of manufacturing industry. J Clean Prod, 115, 36–51 (2016).

Tax Incentives' Effect on the Provision of Occupational Welfare in Italian Enterprises

L'impatto delle agevolazioni fiscali sul welfare aziendale delle imprese italiane

Alessandra Righi

Abstract The goal of this paper is twofold; firstly, to describe the forms of occupational welfare in Italian enterprises and, secondly, to study the effect of recent tax incentives on the spread of these initiatives. Cluster analysis allows to verify to what extent the Small Manufacturing Enterprises (SMEs) in Italy have benefited from the new tax incentives and whether this has led to a greater diffusion of occupational welfare within these firms still having a low level of provision.

Abstract *Il lavoro intende descrivere le forme di welfare aziendale nelle imprese italiane e studiare l'effetto delle recenti misure di incentivazione fiscale sulla diffusione di queste iniziative. Un'analisi dei gruppi aiuta a capire se le PMI italiane hanno potuto beneficiare dei nuovi incentivi fiscali, consentendo una maggiore diffusione del welfare aziendale all'interno di queste imprese che offrono tali servizi ancora in misura limitata.*

Key words: cluster analysis, corporate welfare, Small Manufacturing Enterprises

1 Introduction

Occupational Welfare (OW) is the term used to describe the workplace schemes aimed at improving the well-being of the employees with the commitment of the parties involved in the labour relationship to improve the profitability or productivity of the company. This concept has been important through the history of industrial societies and its function has been to develop the human capital, through training and healthcare packages, pensions and other benefits, able to increase also the employee loyalty [3]. Nowadays, these provisions are used by companies for

* Alessandra Righi, Istat; email: righi@istat.it

arranging business interests, increasing their competitiveness, strengthening the collaboration with the workers and to compensate, at least in part, for the effects of the wage moderation regime practiced during the latest decades [8]. OW provision covers more than 20% of total employees in many European countries and, although in an unfavourable context (considering the large number of SMEs), this share increased in Italy since the 1990s [5].

Economic policies and collective bargaining proved to be important factors for the growth of the OW provision in Italy [8]. Recent legislation (e.g., Income Tax Code, the so-called TUIR) made more convenient for companies the supply of supplementary services compared to contractual salaries increases. Furthermore, the 2016 Italian Budget law abolished taxation on productivity bonuses when these are provided as welfare benefits, sustaining in this way the OW provision [6]. Besides, corporate bargaining has become increasingly important over the years and the OW initiatives become a source of legitimacy and power for the unions, which play a key role in the negotiation, the regulation and the administration of the OW schemes. [5].

A descriptive analysis of the forms of OW in Italian enterprises operating in the manufacturing industry and market services, through the most recent Italian official statistical sources, allows to study the characteristics of the provision and the effect of the tax incentives (in particular, of those introduced by the 2016 Italian Budget law) on the initiatives. Cluster analysis is aimed at identifying the characteristics of the enterprises who benefited from new tax incentives in order to find out if the new regulatory framework could favour the spread the OW provisions within small and medium enterprises, which still report low shares of these initiatives.

2 Data source and methods

Our data sources are the Harmonized Business Surveys[2], namely, monthly qualitative surveys conducted by the Italian Statistical Institute whose results are representatives of all the Italian enterprises. They are usually geared to measuring opinions of the operators regarding the trends of the most relevant economic variables, but an Ad hoc module surveyed the occupational welfare measures in February 2016 and 2017[3]. Similarly to Natali and Pavolini's approach [7], the survey considered as OW schemes all the benefits and services provided by social partners (employers and trade unions) to employees or his/her dependents. These

[2] Harmonized European Tendency Surveys use panels extracted from the Italian Statistical Register of Active Enterprises (ASIA) stratified (by firm size, sector and geographical area) according to the Robust Optimal Allocation with Uniform Stratum Threshold criterion of stratification and allocation for units with fewer than 1,000 employed person, whereas all the larger firms are considered. The Ad-hoc module has the same sample of the Business surveys and data are weighted using the ratio between the number of enterprises in the Business register (by each sampling stratum) and the number of enterprises in the sample.

[3] The samples consider 4,200 firms in manufacturing industry and 2,000 in market service.

measures refer to both collective occupational welfare schemes and company-based welfare schemes (regulated and managed by employers) not considering fringe benefits. Besides the firm structural variables (location, industry, size), the 2017 Ad hoc module surveyed the provision of the following OW forms by enterprises: training or retraining activities, initiatives for the reconciliation of lifetime of the personnel, support measures to the costs of services for early childhood or for special needs (e.g., health problems of the employees or his/her family members), revenue support measures, funding of supplementary health insurances and complementary pension schemes.

We classified the enterprises into different groups based on the set of our variables [1]. In the SAS software Cluster procedure we used, each observation begins in a cluster by itself and the two closest clusters are merged to form a new cluster that replaces the two old clusters. The measurement of the distance between observations in the agglomerative methods depends on the type of available data and different measures have been proposed to measure the 'distance' for a mixture of binary, categorical and quantitative data (as in our case) [2]. We used the median method to hierarchically clustering the observations in our data set. In this method, developed by Gower [4], the dissimilarity between cluster A and cluster B is represented by the distance between the determined median for the enterprises in cluster A and the median for the cases in cluster B. We clustered the enterprises using three structural variables (location of the firm in four territorial areas, firm size in five classes, 15 divisions of the Nace Rev. 2 classification of economic activities for manufacturing industry and four divisions for market services) and a variable indicating if the firm received tax incentives for the provision of OW measures and what kind of tax benefit the firm received.

3 Main results

Overall 58.5% of enterprises in the manufacturing industry and 64.4% in the market services are involved in at least one of the surveyed OW schemes. Training and retraining programmes[4] are the most commonly provided services (26.7% in market services and 22.4% in the manufacturing industry), besides the flexibility measures for the reconciliation of life of the personnel (22.3% in market service and 21.1% in the manufacturing industry). Funding for supplementary health insurances and complementary pension schemes are supplied by 8.8% and 7.2% of the manufacturing enterprises, and by 5.8% and 2.6% of market services firms, respectively. Revenue support measures for consumption (e.g., advances or loans) involve 9.6% of the manufacturing enterprises and 5.8% of the market services ones. The measures supporting the early childhood services costs (nursery schools, vouchers) are provided very rarely (2.2% in market services and 1.5% in manufacturing industry).

[4] This item excluded compulsory activities on safety matter.

We confirm the empirical literature findings related to the dualization in the OW provision by firm size and geographical area of location. As far as the size is concerned, large enterprises (250 employees and over) show a higher level of provision (73.9% in the manufacturing industry and 77.9% in the market services) than medium-sized enterprises (62.7% and 58.0%) or the small ones (55.5% and 50.5%). As regards as the area, the provision in the North (61.2% in the manufacturing industry and 67.0% in the market services) is higher than in the Centre (54.2% and 64.2%, respectively) or in the South (50.6% and 51.7%).

Coming to the study of the effect of tax incentives, OW initiatives undergone a significant development due to the tax incentives introduced by the 2016 Budget law. Enterprises who benefited from tax breaks on the basis of the new regulation on productivity bonuses in 2016 are 3.8% in the manufacturing industry and 1.5% in the market services. However, a further 3.2% of manufacturing firms and 1.8% of market services firms benefited from other (already existing) tax incentives for the adoption of OW schemes. The incentives provided by the 2016 Budget Law were mainly used by medium and large enterprises both in the market services (23.0% and 28.0%, respectively) and in the manufacturing industry (9.0% and 13.1%, respectively). This growth favoured initiatives like health insurance funding, flexi-time activities for the reconciliation of life and the complementary pensions funding within large firms; whereas small and medium-sized firms focused on the adoption of flexi-time measures, income sustains measures and pension benefits.

Cluster analysis allowed distinguishing three groups of enterprises, similar in both sectors, having provided occupational welfare measures in the considered period (from February 2016 to February 2017) but having had different access to tax incentives. Cluster 1 and 2 regroup enterprises having benefitted by tax incentives, but they slightly differ because the enterprises in cluster 1 (37.2% in the manufacturing and 64.9% in the market services) received only tax benefits envisaged by the 2016 Budget law, whereas those in cluster 2 (30.5% of the manufacturing firms and 24.0% of the market services) reported an extensive use of both tax benefits envisaged by the 2016 Budget law and other existing tax benefits (with a less extensive use in the case of enterprises in the market services). In Cluster 1 we found large enterprises, located in the North-east, operating in medium-high-technology manufacturing industry (in particular, in the Machinery and equipment not elsewhere considered, or Other manufacturing) besides large enterprises operating in the knowledge-intensive services sector, located in the Centre and South. As far as the human resources strategy, they have prevalently adopted the reduction of the employees; thus, they show decreasing trends in employment and a low level of employment turnover (less than 5% of newly employed out of total employed). The flexi-time initiatives and the child support measures are their prevalently provided OW activities, accompanied by the training activities and the support to the revenue in the case of market service firms (Table 1). In Cluster 2 are high-technology manufacturing medium enterprises (50-249 employed) located in the North-west (operating in Rubber and plastic or Fabricated metal, except machinery and equipment sectors) and knowledge-intensive service firms, operating in Information and communication and Support services. They declared a human resource strategy of pure replacement of the exits and show a

slightly increasing number of employed, with a low employment turnover (less than 5%) and a medium level of newly employed having a tertiary degree (10%-29%). Their prevalently provided OW initiatives are training programmes, child support, supplementary pension schemes and revenue support for manufacturing firms, and training programmes, child support and funding of health programs for market services enterprise.

Table 1: Characteristics of enterprises in the clusters by sector - Italy, 2017

	Manufacturing			**Market services**		
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Location	NE	NW	C/S	C/S	-	NW/NE
Size	250 and over	50-249	5-9	1000 and over	-	less than 50
Nace Rev 2 divisions	Machinery and equipment n.e.c., Other manufacturing	Rubber and plastic, Fabricated metal except machinery and equipment	Food-beverage, Textiles-wearing apparel, leather, wood	-	Information and communication, Support services to enterprises	Transportation and storage, Holiday and accommodation
Sectoral classification	Medium-high-technology	High-technology	Low-technology	-	Knowledge-intensive	Less knowledge-intensive
Employment trend	Decreased	Increased	Stable	Decreased	Increased	Stable
Newly empl./Total employed	Less than 5%	Less than 5%	15% and over	Less than 5%	5-15%	15% and over
Newly empl. with tertiary degree	>= 30%	10% - 29%	< 10%	< 10%	>= 30%	-
HR Strategy of the enterprise	reduction	replacement	replacement	reduction	replacement	replacement
OW initiatives provided	Flexi-time, Child	Training, Child, Pension, Revenue	Health	Training, Flexi-time, Child, Health, Revenue	Training, Child, Health	(Flexi-time)

Source: elaborations by the author on Istat, Ad Hoc module on Labour and business suveys, 2017

Enterprises in Cluster 3 provided OW initiatives without receiving any tax benefit (32.4% in the manufacturing industry and 11.1% in the market services). They are low-technology manufacturing small enterprises (operating in sectors like Food and beverage, Textiles and wearing apparel, Leather, Wood), located in the Centre and the South, and less knowledge-intensive services SMEs (operating in Transportation and storage ad Holiday and other short-stay accommodation sectors) located in the North. Having declared a human resource strategy of replacement of the exits, they show stable employment, a relevant employment turnover (15% and

over) with a low share of newly employed having a tertiary degree (< 10%). Their prevalently provided OW initiative is the funding of health programmes for manufacturing enterprises and the flexi-time initiatives for market services ones.

4 Conclusions

This analysis showed that a significant share of enterprises took advantage of the recent tax incentives. Large enterprises (those in Cluster 1), especially operating in the market services, have benefitted more of them. Unfortunately, they had already a high level of provision. However, it is comforting that the medium enterprises (Cluster 2) used all the tax incentive opportunities provided by the State, even if small enterprises in both sectors (being prevalently in Cluster 3) seem not having enjoyed from tax incentives. A second relevant finding is that we verified that enterprises benefitting from the tax benefits (Cluster 1) supplied a wider range of OW services than those not benefitting of any incentive (Cluster 3). Thus, the answer to our research question is that the undertaken political economy measures have, at least partially, succeeded in their goal of facilitating OW spread. It would, therefore, be appropriate to maintain the tax incentive measures at the medium term.

References

1. Aldenderfer, M.S., Blashfield, R.K.: Cluster Analysis. Sage Publications, California (1984)
2. Everitt, B.S., Landau, S. and Leese, M.: Cluster Analysis, Fourth edition, Arnold, London (2001)
3. Farnsworth, K. Occupational welfare. In: Greve, B. (ed) The Routledge Handbook of the Welfare State, pp.30-40. Routledge, London (2013)
4. Gower, J. C.: A comparison of some methods of cluster analysis. Biometrics, 23, 623–637 (1967)
5. Maino, F., Mallone, G.: Rise in Occupational Welfare Benefit Schemes. Eurofound, Dublin (2012) Available at https://www.eurofound.europa.eu/it/publications/article/2012/rise-in-occupational-welfare-benefit-schemes
6. Mallone, G.: Imprese e lavoratori: il welfare aziendale e quello contrattuale. In: Maino F., Ferrera M. (a cura di), Primo rapporto sul Secondo welfare in Italia. pp. 49-81 Centro di ricerca e documentazione Luigi Einaudi, Torino (2013)
7. Natali, D. Pavolini, E. with Vanhercke, B.: Occupational welfare in Europe: state of play, determinants and policy implications. In: Natali, D., Pavolini, E. with Vanhercke, B. Occupational Welfare in Europe: Risks, Opportunities and Social Partner Involvement, pp. 239-257. ETUI, Brussels, OSE, Brussels (2018)
8. Pavolini, E., Arlotti, M., Ascoli, U., Leonardi S., Raitano M.: The challenge of occupational welfare in Italy: between risks and opportunities. In: D. Natali, E. Pavolini, with Vanhercke, B. Occupational Welfare in Europe: Risks, Opportunities and Social Partner Involvement, pp. 173-192. ETUI, Brussels, OSE, Brussels (2018)

The determinants of eco-innovation: a country comparison using the community innovation survey.

Le determinanti della eco-innovazione: un confronto tra Paesi basato sui dati CIS.

Ida D'Attoma and Silvia Pacei

Abstract We study the driving forces behind the adoption of different types of eco-innovation in German manufacturing sectors compared to Romania over 2012-2014, through an empirical analysis of the community innovation survey (CIS). To this end, we consider a measure of eco-innovation performance that counts different types of eco-innovation that enterprises have undertaken. Moreover, we consider a wide and comprehensive set of potential drivers, including "technology push", "regulatory push-pull" and "firm specific factors" clusters of drivers. Due to the count nature of our dependent variable and to the large presence of zeros, we estimate a zero-inflated negative binomial model. Preliminary findings show interesting differences in the determinants of eco-innovation in the two countries.

Abstract *In questo lavoro si studiano le determinanti dell'eco-innovazione, mettendo a confronto il settore manifatturiero tedesco con quello rumeno negli anni 2012-2014. A questo scopo si usano i dati dell'indagine sull'innovazione della comunità europea (CIS), e si misura il fenomeno attraverso il numero delle diverse tipologie di eco-innovazione che l'impresa realizza. Si considerano diverse potenziali determinanti, tra cui fattori legati alla tecnologia, alla normativa e caratteristiche specifiche dell'impresa. Data la natura della variabile oggetto di studio e l'elevata presenza di zeri si considera un modello Zero Inflated Negative Binomial. I risultati mostrano differenze interessanti fra le determinanti dell'eco-innovazione nei due Paesi.*

[1] Ida D'Attoma, Department of Statistical Sciences – University of Bologna;
email: Ida.dattoma2@unibo.it
Silvia Pacei, Department of Statistical Sciences – University of Bologna;
email: silvia.pacei@unibo.it

Key words: eco-innovation, eco-innovation drivers, CIS data, zero-inflated negative binomial regression.

1 Introduction

Eco-innovation is commonly understood as innovations that are associated with environmental benefits. Differently from traditional innovation, eco-innovation 'emphasizes the firms' mitigation of negative impacts on the natural environment' [7]. It is acknowledged that a positive relationship between eco-innovation and firm performance and growth exists, thus motivating the identification and the analysis of trigger factors for eco-innovation. The early study of [8] found that the more enterprises eco-innovate, the greater the positive impacts on their performance. Firms implementing eco-innovation tend to be recognized in a positive sense [1] by consumers that are nowadays increasingly aware of environmental topics. Not only to obtain or enhance social legitimacy but also to react or pre-empt their rival's environmental moves, corporations are inclined to use eco-innovation strategies to exploit emerging opportunities [7].

The main purpose of this study was to analyse the main factors enhancing firms focusing on a different number of eco- innovations in Germany and Romania. We constructed a count measure of eco-innovativeness and then we analysed its determinants by estimating a zero-inflated negative binomial model [3,4].

This paper contributes to the previous literature in several ways. First, additional quantitative evidence is provided regarding the determinants for eco-innovation. Second, few studies have focused on a cross-country analysis. Instead, we considered the cases of Germany and Romania, thereby expanding the geographical scope of econometric research on eco-innovation in Europe, which has to date largely focused only on Germany. Third, differently from the large body of literature that used to consider a binary measure of eco-innovativeness, we considered a multidimensional eco-innovation construct that keeps track of all eco-innovations that enterprises simultaneously adopt.

2 Data and Variables

This study was based on firm-level data from the 2014 European Union Community Innovation Survey (CIS) of Germany and Romania in the period 2012-2014. The sample covers 3250 manufacturing firms with more than 10 employees across different industries for Germany and 4325 manufacturing firms for Romania.

The CIS defines an eco-innovation as "a new or significantly improved product (good or service), process, organizational method or marketing method that creates

environmental benefits compared to alternatives". The definition is not only confined to the technological sphere but it also encompasses organizational and market aspects.

The outcome of our model was the number of eco-innovations adopted by a firm and ranged from 0 (no eco-innovation strategy were in place) through 10 (all environmental strategies were implemented)[1]. For what concerns explanatory variables, we followed a relatively recent body of literature that has focused on four main clusters of drivers: technology-push, regulatory-push-pull, market-pull, and firm specific factors [9,6]. As *technology-push* factors we considered the *degree of cooperation* defined as the number of innovation cooperation partners type with which the enterprise undertakes an active cooperation, the *innovation intensity* described by the ratio of the total expenditure on innovation and *technological capabilities* proxied by qualified employees. As *regulatory push-pull* factors, we took into account whether the enterprise has received public support for its innovation activity. Then, because CIS data did not allow us to directly consider the role of environmental regulations, as a proxy of environmental regulation we adopted the logarithm of CO2 air emissions intensities expressed as ratios relating carbon dioxide emissions to value added by manufacturing industries (classified by NACE Rev. 2) in each country considered. As *firm-specific* factors we considered firm size, industry, whether the firm belongs to a group, the foreign market focus, the exporting behaviour, the employee growth rate and the turnover growth rate. No market-pull factors were available. Moreover, we envisaged the forms of innovation protection as an important trigger factor of eco-innovation. In fact, in light of the double externality problem, even when a firm successfully reaches the market with an eco-innovation, the profit appropriation could result still difficult especially in case the eco-innovation is accessible to potential imitators [5], thus requiring protection.

3 Methodology

Due to the count nature of our dependent variable Y and the large presence of zeros, we use the zero-inflated Negative Binomial regression model (ZINB) [4,5]. In our case in fact, the Negative Binomial model may not accurately assign probability to the outcome $Y = 0$. The ZINB model assumes that a separate process is simultaneously influencing the outcome. Therefore, the ZINB model is a "two-part" count data model, as it considers two data generation processes. For each observation i the first process, represented by d_i, gives zero counts with probability πi (when $d_i = 0$), and with the complementary probability $1 - \pi_i$ (when di=1) the

[1] The CIS 2014 includes ten types of eco-innovations: material use reduction; energy saving; air, water, noise or soil pollution reduction within enterprise; substitution of polluting and hazardous materials; substitution of fossil energy with renewable energy; waste, water and material recycling ; CO2 emissions reduction; air, water, noise or soil pollution reduction during the consumption by the end user; recycling of a product after use; extending product life through longer - lasting, more durable products

second process Y_i provides counts according to a negative binomial with mean λ_i. Also the second process may generate zero counts.

The model can be represented as follows:

$Y_i = 0$ with probability π_i
$Y_i \sim NB(\lambda_i)$ with probability $1 - \pi_i$

so that, the probabilities connected to the two possible outcomes are:

$Prob[Y_i = 0] = \pi_i + (1 - \pi_i)R_i(0)$
$Prob[Y_i = j > 0] = (1 - \pi_i)R_i(j)$

where

$R_i(y)$ = the negative binomial probability = $\Gamma(\theta + y_i)/[y_i!\,\Gamma(\theta)]u_i^{\theta}[1 - u_i]^{y_i}$
$\theta = 1/\alpha$, where α is the overdispersion parameter
$u_i = \theta/(\vartheta + \lambda_i)$
$\lambda_i = exp(\boldsymbol{\beta}'\boldsymbol{x_i})$

with $\boldsymbol{x}_i$ as the vector of covariates and $\boldsymbol{\beta}$ the vector of related coefficients.

For probability π_i we use a logistic probability model:

$\pi_i \sim Logistic(\nu_i)$

and we define $\nu_i = \tau log(\lambda_i) = \tau\boldsymbol{\beta}'\boldsymbol{x_i}$. This implies the assumption of the same determinants for the two processes, thus requiring the estimation of only one more parameter, τ.

The joint density can be written as follows:

$P(y_i, d_i|\boldsymbol{x_i}) = (1 - d_i)\pi_i + d_i(1 - \pi_i)R_i(y)$

while the conditional mean function is $(1 - \pi_i)\lambda_i$.

4 Results

Our response variable is a count variable and represents the number of eco-innovations adopted by a firm[1]. With reference to Germany, 11% of firms in the data reported only one type of eco-innovation, whereas about 70% reported two or more types, as many as 4.4% reported all 10 and about 20% reported zero eco-innovations. The presence of zeros is higher in the Romania case, where about 57%

[1] The Cronbach's Alpha coefficient of our dependent variable was 0.83 and 0.89 for Germany and Romania, respectively.

of manufacturing firms reported zero eco-innovations, about 5% only one type, about 36% reported two or more types and only about 2% reported all 10.

Table 1 reports results of the estimated zero inflated negative binomial model for Germany and Romania, respectively. The Vuong statistic is 7.3 for Germany and 4.4 for Romania. In both cases it favors the zero-inflated negative binomial model.

Table 1: The determinants of eco-innovation in Germany and Romania

Determinants	***Germany***		***Romania***	
	Partial Effect	***z***	***Partial Effect***	***z***
Degree of cooperation	1.5384***	2.81	1.3202**	1.99
Innovation intensity	1.4674**	2.27	2.0045	1.20
Qualified employees	.0020	0.18	0.0328	1.48
Qualified employees (squared)	-.0001	-0.05	-0.0004	-0.03
Public funding	-0.18514	-0.90	-0.3285	-1.07
Environmental regulation	0.1542***	2.68	0.0043	0.08
Export	-0.0405	-0.23	0.8078***	3.66
Group membership	-0.0561	-0.18	-0.0902	-0.46
Foreign multinational	0.0934	0.31	1.1302***	5.04
50-249 employees	0.5648***	3.40	-0.1499	-0.58
250-499 employees	1.1715***	3.73	0.2392	0.77
500 and more employees	1.3372***	4.15	0.4185	1.16
High-tech sector	-1.1507***	-4.25	-0.6061	-1.30
Medium-tech sector	-0.2384	-1.47	0.4868**	2.40
Employees growth rate	0.2711	1.30	-0.4399	-1.51
Turnover growth rate	.0010	0.03	0.0887	0.45
Level of protection	.3333***	5.13	-0.0267	-0.05

Notes: ***1% significance, ** 5% significance and * 10% significance.
Partial effects are the derivatives of the conditional mean function. Effects are averaged over units. To estimate our model, we use the software LIMDEP.

Source: Own elaboration of the CIS 2014 data.

The following variables were significant at 1% or 5% level for the Germany case: degree of cooperation, innovation intensity, environmental regulation, size, high-tech sector and level of protection. As far as the degree of cooperation is concerned, a partial effect of 1.5 indicates that a firm with a one unit change in the cooperation intensity is expected to have on average 1.54 more eco-innovations. This result might be due to the fact that eco-innovations are more complex and demanding than other types of innovation, thus requiring more resources picked up from external partners and sources of knowledge. In line with literature, innovation intensity was found to have a positive effect on eco-innovation. A firm with a one unit change in the internal and external R&D effort is expected to have 1.47 more eco-innovations. Environmental regulation was found to be an important stimulus to eco-innovation: if CO2 air emissions intensity increases of 1%, a firm is expected to have on average 0.15 more eco-innovations. The number of eco-innovations adopted by a firm increased as size classes increased (the smallest class dummy 1-49 employees was excluded from the model). Furthermore, for a one unit change in the level of protection, a firm is expected to have on average 0.33 more eco-innovations.

Conversely, firms in the high-tech sector are expected to have 1.15 less eco-innovation compared to the reference category (low-tech sector).

As far as Romania is concerned, similarly to Germany, a firm that experienced a one unit change in the cooperation intensity is expected to have more eco-innovations (+1.32). Differently from Germany, the following variables were significant at 1% or 5% level: export, foreign multinational and medium high-tech sector.

Disclaimer

The anonymous data of the Community Innovation Survey 2014 (CIS 2014) used in the analysis of this paper was provided by EUROSTAT. All results and conclusions are given by the authors and represent their opinion and not those of EUROSTAT, the European Commission or any of the national authorities whose data have been used. The responsibility for all conclusions drawn from the data lies entirely with the authors.

References

1. Buysse, K., Verbeke, A.: Proactive environmental strategies: A stakeholder management perspective. Strateg. Manag. J. , **24**(5), 453-470 (2003) doi: 10.1002/smj.299
2. Dìaz-Garcìa, C., Gonzàlez-Moreno, A., Sàez-Martìnez, F.J.: Eco-innovation: Insights from a literature review. Innov.: Manag., Policy and Pract., **17**(1), 6-23 (2015) doi: 10.1080/14479338.2015.1011060
3. Greene, W.: Models for count data with endogenous participation. Empir. Econ. 36: 133-173 (2009).
4. Hilbe, J:M.: Negative Binomial Regression. Cambridge University Press, Cambridge (2007).
5. Hojnik, J.: In Pursuit of Eco-innovation: drivers and consequences of eco-innovation at firm level. (2017). Doi: 10.26493/978-961-7023-53-4
6. Horbach, J., Rammer, C. and Rennings, K.: Determinants of eco-innovations by type of environmental impact- The role of regulatory push/pull, technology push and market pull. Ecolog. Econ. , **78**, 112-12 (2012) doi: 10.1016/j.ecolecon.2012.04.005.
7. Liao Y-C and Tsai K-H.: Innovation intensity, creativity enhancement, and eco-innovation strategy: The roles of customer demand and environmental regulation. Bus. Strat. and the Environ., 28, 316-326 (2019) doi: 10.1002/bse.2232
8. Russo, M.V. and Fouts, P.A: A resource-based perspective on corporate environmental performance and profitability. Academy of Manag. J., **40**, 534-559 (1997) doi: 10.2307/257052
9. Zulbeltzu-Jaka, Erauskin-Tolosa A., Heras-Saizarbitoria I.: Shedding light on the determinants of eco-innovation: A meta-analytic study. Bus. Strat. and the Environ., 27, 1093-1103 (2018) doi: 10.1002/bse.2054

World ranking of urban sustainability through composite indicators

Una graduatoria della sostenibilità urbana nel mondo basata su Indicatori Compositi

Elena Grimaccia, Alessia Naccarato and Silvia Terzi

Abstract Urban population worldwide is steadily increasing creating huge problems of sustainable development. The 2030 Agenda chose to devote a whole goal to urban sustainability: Goal 11 of the Sustainable Development Goals (SDG). Despite such attention, not many empirical studies are based on SDGs indicators, due to scarce international comparability of the available data.
This paper draws on data from the very SDG dataset, with the aim of computing an Urban Sustainable Development Index, using different normalisation, aggregation and weighting systems for the construction of composite indicators (CIs). Our attention will focus on the comparison among the rankings derived from each synthetic index, thus assessing the rank robustness of the CI.
Abstract. *La popolazione urbana nel mondo è in costante crescita, creando enormi problemi di sviluppo sostenibile. L'Agenda 2030 ha scelto di dedicare un intero obiettivo alla sostenibilità urbana: il Goal 11 degli Obiettivi di Sviluppo Sostenibile (SDG). Nonostante tale attenzione, non sono molti gli studi empirici basati su indicatori SDG, a causa della scarsa comparabilità internazionale dei dati disponibili. Il nostro lavoro attinge direttamente al set di dati SDG, con l'obiettivo di calcolare un Indice di sostenibilità Urbana, utilizzando diversi metodi di normalizzazione, aggregazione e ponderazione per la costruzione di indicatori compositi (IC). Lo studio si concentrerà sul confronto tra le classifiche derivate da ciascun indice sintetico, valutando così la robustezza delle graduatorie dell'IC.*

Key words: urban sustainability, composite indicators, sustainable development, principal component analysis.

[1] Elena Grimaccia; Istat, Italian National Institute of Statistics; email: elgrimac@istat.it
Alessia Naccarato; Roma Tre University; email: Alessia.naccarato@uniroma3.it
Silvia Terzi; Roma Tre University; email: Silvia.terzi@uniroma3.it.

1. Introduction

Sustainability is today at the centre of the scientific and public debate. Our current economic paradigm, based on transformation of material basis poses a serious threat in terms of resource availability and pollution (Matthews et al., 2000). Furthermore, since 2007, more than half the world's population lives in cities or urban settlements and UN estimates that by 2030, cities will be home to 60% of the global population, a share that will further increase to about 68.4% by 2050 (UN, 2018). This means that between 2010 and 2050, between 2.5 to 3 billion people will be added to the urban population worldwide; with the highest growth in East Asia, South Asia, and sub-Saharan Africa. Given the estimates, the question of sustainable urbanisation is critical for most of the countries. The Sustainable Development Goal 11 of the United Nations Agenda 2030 - known as the 'urban SDG: to make cities and human settlements inclusive, safe, resilient and sustainable' - addresses directly the urban level with 10 targets and 15 indicators developed by the UN (UN, 2017). Nowadays, cities contribute about 80% of global GDP. However, cities also account for about 70% of global energy consumption, 70% of global carbon emissions, as well as over 70% of resource use, making it even more compelling for policy makers to design sustainable policies for urban settlements.

Sustainability is a complex multidimensional phenomenon that needs to be synthetized into a simple and readable measure that policy makers can use. Composite indicators (CI) summarise complex and multidimensional phenomena, and are useful tools to ease interpretation and to allow benchmarking. However, care must be taken in their construction and interpretation to avoid misleading policy messages (Floridi et al., 2011).

CIs (also referred to as synthetic indices) aggregate standardised or normalised variables with the aim of capturing different and relevant aspects of a possibly latent multidimensional reality (Becker et al 2017). CIs are heavily linked to normative assumptions in variable selection and weighting. Here 'normative' is understood to be 'related to and dependent on a system of norms and values'. The statistical analysis of CIs is essential to provide stakeholders reliable instruments for policies (OECD 2008, Saisana et al., 2011, Luzzati and Gucciardi 2015).

The present paper aims at computing an Urban Sustainable Development Index, using different normalisation, aggregation and weighting systems for the construction of CI. Our attention will focus on the comparison among the rankings derived from each synthetic index, thus assessing the rank robustness of the CIs.

2. Methodological framework

The theoretical context of our CIs in based on the UN 2030 Agenda framework, benefitting from the huge work of the numerous High Level Committees, that have defined the goals and chosen the target indicators (Hak, 2015, UN, 2015).

Indicators may be measured in different measurement units and on different scales, so prior to any aggregation, normalisation of the data is required. Among the different normalisation methods (Talukoler et al., 2017, Seth and McGillivray, 2018), we used the rankings of each indicator, that is the simplest normalisation technique; it is not affected by outliers but it causes loss of information on the level of country performance in absolute terms. We also applied a standardisation (or z-scores), that converts indicators to a common scale with a mean of zero and unit standard deviation. Finally, we used a min-max method, by subtracting the minimum value and dividing by the range of the indicator value in order to have an identical range [0, 1].

In the following weighting and aggregation step, the standardised indicators are combined together. The weight assigned to an indicator in a CI reflects its relative importance. However, the real weight of the non-standardised variable also depends on the transformation. Consequently, even equal weighting allows the original variables to have smaller or greater effects on the composite indicator.

Most existing CIs are linear, i.e. weighted arithmetic averages (Paruolo et al., 2012), but linear aggregation rules have been criticised because weaknesses in some dimensions are compensated by strengths in other dimensions; this characteristic is called 'compensatory'. Non-compensatory and non-linear aggregate ranking rules have been supported by some literature on multicriteria decision making (Munda and Nardo 2009). Geometric aggregations are, therefore, better suited if the modeller wants some degree of non-compensability between individual indicators or dimensions. Furthermore, linear aggregations reward base-indicators proportionally to the weights, while geometric aggregations reward those countries with higher scores.

Mathematical weights derived from Principal Component Analysis (PCA) are commonly used to assign weights to single variables, often using the first factor as the 'best' composite indicator (Booysen, 2002, Krishnakumar and Nagar 2008). The index obtained accounts for the largest amount of total variance in the individual indicators, and the first factor will be also correlated with at least some of the individual indicators.

3. Composite Indicators of Urban Sustainability

The SDG 11 on Sustainable cities and communities, with the aim of making cities "inclusive, safe, resilient and sustainable" consists of a number of targets, including: adequate, safe and affordable housing; accessible and sustainable transport systems for all; inclusive and sustainable urbanisation; to reduce the number of people affected by disasters; to reduce the environmental impact of cities. We selected five variables, for the period 2014-2017, measuring five different aspects of urban sustainability: 1.proportion of urban population living in slums, informal settlements or inadequate housing ("Slums", negative polarity[1]); 2.number of deaths, missing persons and

[1] We speak of 'negative polarity' when the variable and the latent construct of urban sustainability, in the theoretical framework, have negative correlation.

directly affected persons attributed to disasters per 100,000 population ("Disasters", negative polarity); 3.proportion of urban solid waste regularly collected ("Waste") 4.annual mean levels of fine particulate matter ("Particulate", negative polarity); 5.proportion of local governments that adopt and implement local disaster risk reduction strategies, in relation with the Sendai Framework for Disaster Risk Reduction ("Sendai").

The analysis was developed with reference to UN countries. We applied various procedures to address missing data. We chose to include only indicators available for at least 90% of the 149 UN Member States. And we included in the data set Only countries having data for at least 70% of the indicators. To impute missing values, we performed a cluster analysis (based on the values of the 3 Human Development Index sub-indicators) and replaced missing values on target indicators with cluster average values.

The indicators included in the SDG11 are quite heterogeneous, which leads to weak correlations. Notice that the sign of the correlations are consistent with theorical expectations.

Table 1: Correlations among SDG11 indicators

SDG11 indicators	Slums	Disasters	Waste	Particulate	Sendai
Slums	1				
Disasters	0.2790	1			
Waste	-0.1938	-0.0161	1		
Particulate	0.4790	0.3963	-0.1547	1	
Sendai	-0.1223	0.0007	0.1133	-0.0679	1

Source: Authors' elaboration on UN data

4. Results

In order to build a CI of Urban Sustainability, we applied the standardisation and aggregation methodologies stated above, and we obtained five different CIs. The first one has been constructed by a linear aggregation of the rankings of the different variables; the second is the geometric aggregation of the rankings; the third uses a normalisation through the z-scores and then a linear aggregation; the fourth a min-max standardisation and then a linear aggregation, and the last applies PCA weights. For the variables where a larger value corresponded to a lower urban sustainability (negative polarity), the normalisation or standardisation had reverse sign. After normalization or standardisation were ranked in ascending order. Computing the median of the five rankings obtained according to the five different procedures, we ranked countries in descending order of Urban Sustainability. In figure 1, we plot, for each country, the median rank and its range.

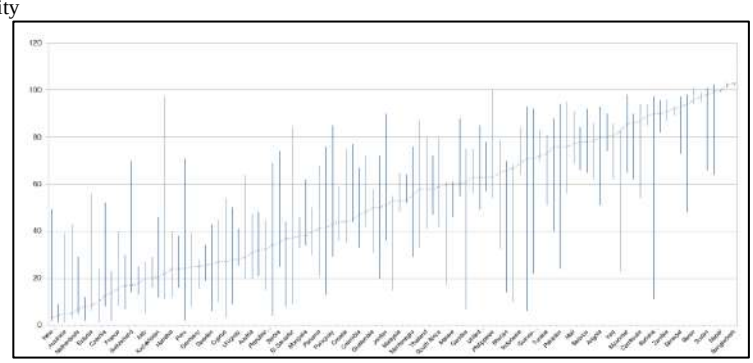


Figure 1: Composite indicators of Urban Sustainability: median, min and max of world Countries

Among the best performing countries, we found Japan, Oceanian countries and some North European countries, while China, South-Est Asian Countries, and Sub-Saharan African countries present the lowest levels of urban sustainability. However, the different rankings provided homogenous results only for some of the countries. In fact, most high and middle-range countries –in particular Lebanon, many Balkan countries, and some South American countries – present high variability across the different rankings. As expected, the higher correlations are between the two classical aggregation methods (standardisation and min-max) and among the linear and geometric aggregations of rankings. The CI based on PCA weights presents smaller rank correlations with the other methods and the greatest sum of the absolute distances from the median ranking. Also in terms of distances among the different rankings, we must conclude that the rankings are sensitive to transformation and aggregation procedures. This is probably due to low correlations among the original variables. However, a deeper insight would be useful.

Table 2: Rank correlations according to the different methodologies

	Linear aggregation of rankings	Geometric aggregation of rankings	Linear aggregation of min-max	Linear aggregation of z-scores	PCA weights
Linear aggreg. rankings	1				
Geometric aggreg. rankings	0.963	1			
Linear aggreg. min-max	0.819	0.783	1		
Linear aggregat. z-scores	0.879	0.832	0.967	1	
PCA weights	0.548	0.570	0.293	0.352	1
Deviation from the median ranking	*474*	*662*	*963*	*732*	*1987*
MAD (median absolute deviation)	*2*	*4*	*6.5*	*4*	*12*

Final remarks

The result obtained so far is to rank worldwide countries on the basis of the indicators selected by the UN for monitoring Goal 11 of the SDGs stated in the 2030 Agenda. Applying different normalisation, aggregation and weighting systems our aim was to assess the robustness of the suggested composite indicator. We conclude however that the rankings are sensitive to the chosen transformations and aggregations. The next step of our research will be to exploit model-based CIs or methods that account for association among original variables.

References

1. Becker W., Saisana M., Paruolo P., Vandecasteele I.: Weights and importance in composite indicators: Closing the gap. Ecological Indicators 80 (2017)
2. Booysen F., An Overview and Evaluation of Composite Indices of Development. Social Indicators Research 59, 115-151 (2002)
3. Floridi M., Pagni, S., Falorni S., Luzzati, T.: An exercise in composite indicators construction: Assessing the sustainability of Italian regions. Ecological Economics 70 1440–1447 (2011)
4. Hák, T., Janousková, S., Moldan, B.: Sustainable Development Goals: A need for relevant indicators. Ecological Indicators 60, 565–573 (2016)
5. Krishnakumar, J., Nagar A. L.: On Exact Statistical Properties of Multidimensional Indices Based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models. Social Indicators Research, 86:481–496 (2008)
6. Luzzati T., Gucciardi G. A non-simplistic approach to composite indicators and rankings: an illustration by comparing the sustainability of the EU Countries. Ecological Economics 113, 25–38 (2015)
7. Matthews E., Payne, R., Rohweder, M. and Murray S.: Pilot Analysis of Global Ecosystems: Forest Ecosystems, World Resources Institute, Washington, D.C. (2000)
8. Munda, G. and Nardo, M.: Non-compensatory/non-linear composite indicators for ranking countries: a defensible setting. Appl. Econ., 41, 1513–1523 (2009)
9. Organisation for Economic Co-operation and Development OECD: Handbook on Constructing Composite Indicators:Methodology and User Guide. Paris: Organisation for Economic Co-operation and Development (2008)
10. Paruolo, P., Saltelli, A., Saisana, M.: Ratings and rankings: Voodoo or Science?. Journal of the Royal Statistical Society Series A (Statistics in Society). 176. (2011).
11. Saisana, M., d'Hombres, B., Saltelli, A.: Rickety numbers: Volatility of university rankings and policy implications. Research policy 40(1):165{177} (2011)
12. Seth S., McGillivray M. (2018) Composite indices, alternative weights, and comparison robustness. Soc Choice Welf (2018) 51:657–679 https://doi.org/10.1007/s00355-018-1132-6
13. Talukder, B., Hipel K. W., vanLoon G. W.: Developing Composite Indicators for Agricultural Sustainability Assessment: Effect of Normalization and Aggregation Techniques
14. UN General Assembly: Resolution adopted by the General Assembly on 25 September 2015 n. 70/1. Transforming our world: the 2030 Agenda for Sustainable Development. New York (2015)
15. United Nations Statistical Commission: Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. UN Resolution A/ RES/71/313 (2017).
16. United Nations, Department of Economic and Social Affairs, Population Division: World Urbanization Prospects: The 2018 Revision, Online Edition. File 21: Annual Percentage of Population at Mid-Year Residing in Urban Areas by Region, Sub-region, Country and Area, 1950-2050 (2018)

Machine Learning and Data Science

A novel approach for Artificial Intelligence through Lorenz zonoids and Shapley Values

Un approccio basato sugli Zonoidi di Lorenz e gli Shapley Values nell'ambito dell'Intelligenza Artificiale

Paolo Giudici and Emanuela Raffinetti

Abstract In the era of big data and data science, Artificial Intelligence systems play a basic role. If on the one hand, Artificial Intelligence models deeply affect the decision process, on the other hand the lack of transparency of their black-box approaches may lead to distrust in Machine Learning methods. Explainability thus becomes a crucial issue, to understand how inputs in a model generate the outputs. In this paper, we provide a Shapley-Lorenz zonoid approach to assess the marginal contribution associated with each additional covariate in all the possible model configurations. In this way, the assessment of each effect contribution results more accurate and interpretable, leading to select the most parsimonious model with the minimum loss of information.

Abstract *Nell'era del data science, i sistemi di Intelligenza Artificiale svolgono un ruolo fondamentale. Se da un lato, i modelli di Intelligenza Artificiale influenzano profondamente il processo decisionale, dall'altro la mancanza di trasparenza dei loro approcci può pregiudicare la fiducia nei confronti dei metodi di Machine Learning. Di conseguenza, la spiegabilità rappresenta un aspetto cruciale soprattutto nello stabilire come gli input in un modello possano generare gli output. Mediante l'utilizzo degli Zonoidi di Lorenz e degli Shapley values, viene proposto un nuovo approccio per la valutazione del contributo marginale associato all'inclusione di ciascuna covariata con riferimento a tutte le possibili configurazioni di modello. In questo modo, la valutazione del contributo di ciascun effetto risulta maggiormente accurata e interpretabile, portando a selezionare facilmente il modello più parsimonioso che comporta la minima perdita di informazione.*

Key words: Shapley values, Lorenz zonoids, eXplainable Artificial Intelligence, model interpretability

---

Paolo Giudici
Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia, Italy e-mail: paolo.giudici@unipv.it

Emanuela Raffinetti
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano, Italy e-mail: emanuela.raffinetti@unimi.it

1 Introduction

Researchers in data science are paying close attention on the subject of the eXplainable Artificial Intelligence (XAI) and its possible fields of application. The notion of XAI is increasingly discussed and required by both public and private organizations in order to provide transparent and effective machine learning methods (see, e.g. [1], [2]). The idea is to introduce a suite of techniques that, on the one hand, allows to improve the interpretability of the models and, on the other hand, to preserve an adequate level of prediction accuracy.
The emerging aspect is that explainability represents a crucial issue for all practitioners to better understand, trust, and deal with powerful artificial intelligence systems. This led many scholars to promote new XAI methods in order to manage applications in the field of finance, among others, focused on credit risk management and default risk analysis (see e.g., [3] and [4]). For instance, in [4] an explainable AI model based on similarity networks (see [7] for more details) and Shapley values is proposed to measure credit risks associated to the use of credit scoring platforms. Shapley values were originally introduced by [8] as a solution concept in cooperative game theory. They correspond to the average of the marginal contributions of the players associated with all the possible orders of the players.
In the model explanation process, Shapley values can be exploited for measuring the average marginal contribution of each feature to the prediction of a machine-learning model (see, e.g. [9]). In line with these contributions, recently [5] introduced a novel model selection measure which is based on the employment of the Lorenz Zonoid tool (see, [6]) and on a mutual notion of variability more robust to the presence of outlying observations.
In this paper a new measure providing each explanatory variable marginal contribution in explaining the variability of the response variable in all the possible covariate model-configurations is proposed by combining the Lorenz Zonoid tool with the Shapley value based-approach. In this way, the assessment of each effect contribution results as more accurate and interpretable, leading to easily select the most parsimonious model with the minimum loss of information.

2 Background: Shapley values and Lorenz zonoids

An overview on the Shapley value and Lorenz zonoid tools is reported in order to illustrate the main features exploited for the construction of our proposal.

Shapley values
Shapley values were originally introduced by [8] as a pay-off concept from cooperative game theory. When referring to the machine learning models, the players of the cooperative game ,aimed at generating a pay-off, are replaced by the X_k variables within a model, aimed at generating predictions $\hat{f}(X)$. Following [9] and by using a notation coherent with that considered for the construction of our proposal, the

marginal contribution from variable X_k is expressed in the form of Shapley values as

$$\phi_S^{X_k}(\hat{f},X) = \sum_{X' \subseteq \mathscr{C}(X) \setminus X_k} \frac{|X'|!(n_k - |X'| - 1)!}{n_k!} [\hat{f}(X' \cup X_k) - \hat{f}(X')], \quad (1)$$

where $\mathscr{C}(X) \setminus X_k$ is the set of all the possible configurations of $m-1$ models excluding variable X_k, $|X'|$ denotes the number of variables included in the model, n_k is the number of the available variables, $\hat{f}(X' \cup X_k)$ and $\hat{f}(X')$ are the predictions associated with all the possible model configurations including variable X_k and excluding variable X_k, respectively. The quantity within the squared parentheses defines the contribution of variable X_k to the model prediction for any single observation.

Lorenz zonoids

The Lorenz zonoid was introduced by [6] as a generalization of the Lorenz curve in d dimensions. In the case of $d = 1$ the Lorenz zonoid corresponds with the Gini coefficient.

Suppose to consider a response variable Y and a set of explanatory variables $X_1, \ldots, X_j, \ldots, X_h$, with $j = 1, \ldots, h$. In order to evaluate the relationships between Y and the $X_1, \ldots, X_h$ explanatory variables, a linear regression model is applied, and the associated linear predicted values, denoted with $\hat{Y}_{X_1,\ldots,X_h}$, are determined. The Lorenz zonoids of Y and $\hat{Y}_{X_1,\ldots,X_h}$ can be defined as (see, e.g. [5])

$$LZ_{d=1}(Y) = \frac{2Cov(Y, r(Y))}{n\mu} \text{ and } LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_h}) = \frac{2Cov(\hat{Y}_{X_1,\ldots,X_h}, r(\hat{Y}_{X_1,\ldots,X_h}))}{n\mu}, \quad (2)$$

where n is the total number of observations, μ is the response variable Y mean value, $r(Y)$ and $r(\hat{Y}_{X_1,\ldots,X_h})$ are the rank scores corresponding to the Y and $\hat{Y}_{X_1,\ldots,X_h}$ variables. Given a sample data of size n, formulas in (2) can be defined as

$$LZ_{d=1}(y) = \frac{2Cov(y, r(y))}{n\bar{y}} \text{ and } LZ_{d=1}(\hat{y}_{y_1,\ldots,y_h}) = \frac{2Cov(\hat{y}_{x_1,\ldots,x_h}, r(\hat{y}_{x_1,\ldots,x_h}))}{n\bar{y}}, \quad (3)$$

where y and $\hat{y}_{x_1,\ldots,x_h}$ are the vectors of the observed and predicted values, $r(y)$ and $r(\hat{y}_{x_1,\ldots,x_h})$ are the ranks of the observed and predicted values, and $\bar{y}$ is the sample mean.

In [5], the Lorenz zonoids were exploited giving rise to new dependence measures suitable in assessing the contribution of each explanatory variable to the predictive power of a linear model. Specifically, a Marginal Gini Contribution (*MGC*) measure, allowing to measure the absolute explanatory power of any single covariate, and a Partial Gini Contribution measure (*PGC*), allowing to measure the additional contribution of a new covariate to an existing model, were developed as follows.

Let X_j be one of the h explanatory variables ($j = 1, \ldots, h$). The marginal contribution provided by the single covariate X_j is

$$MGC_{(Y|X_j)} = \frac{LZ_{d=1}(\hat{Y}_{X_j})}{LZ_{d=1}(Y)} = \frac{Cov(\hat{Y}_{X_j}, r(\hat{Y}_{X_j}))}{Cov(Y, r(Y))}. \quad (4)$$

Let $\hat{Y}_{X_1,\ldots,X_h}$ and $\hat{Y}_{X_1,\ldots,X_{h-1}}$ be the predicted values provided by a full linear regression model, including all the covariates, and a reduced linear regression model model, excluding covariate X_h. The additional contribution related to the inclusion of covariate X_h can be determined as

$$PGC_{Y,X_h|X_1,\ldots,X_{h-1}} = \frac{LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_h}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{h-1}})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{h-1}})}. \quad (5)$$

Sample and Covariance-based versions of formulas in (4) and (5) can be found in [5].

3 Proposal: a Shapley-Lorenz based approach

In line with the need of achieving both high predictive accuracy and interpretability, a new methodology which combines the Shapley value approach with the Lorenz zonoid tool is introduced. Specifically, the marginal contribution associated with each additional covariate in different model configurations can be formalized as

$$LZ^{X_k}_{d=1}(\hat{Y}) = \sum_{X' \subseteq \mathscr{C}(X) \setminus X_K} \frac{|X'|!(n_k - |X'| - 1)!}{n_k!} [LZ_{d=1}(\hat{Y}_{X' \cup X_k}) - LZ_{d=1}(\hat{Y}_{X'})], \quad (6)$$

where $LZ_{d=1}(\hat{Y}_{X' \cup X_k})$ and $LZ_{d=1}(\hat{Y}_{X'})$ describes the variability explained by the models including the $X' \cup X_k$ variables and the X' variables, respectively.
As shown in [5], $LZ_{d=1}(\hat{Y}_{X' \cup X_k})$ and $LZ_{d=1}(\hat{Y}_{X'})$ in equation (6) may be expressed as function of the covariance operator, leading to

$$LZ_{d=1}(\hat{Y}_{X' \cup X_k}) = \frac{2}{n\mu} Cov(\hat{Y}_{X' \cup X_k}, r(\hat{Y}_{X' \cup X_k})) \quad \text{and} \quad LZ_{d=1}(\hat{Y}_{X'}) = \frac{2}{n\mu} Cov(\hat{Y}_{X'}, r(\hat{Y}_{X'})). \quad (7)$$

Given a data sample of size n, denoting with $\bar{y}$ the sample mean, $\hat{y}_{X' \cup X_k(i)}$ and $\hat{y}_{X'(i)}$ the predicted values provided by the model including and excluding the X_k covariate and ordered in non-decreasing sense, the formulas in equation (7) become

$$LZ_{d=1}(\hat{y}_{X' \cup X_k}) = \frac{2}{n\bar{y}} Cov(\hat{y}_{X' \cup X_k}, r(\hat{y}_{X' \cup X_k})) = \frac{2}{n\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} i\hat{y}_{X' \cup X_k(i)} - \frac{n(n+1)}{2n} \right] \quad (8)$$

and

$$LZ_{d=1}(\hat{y}_{X'}) = \frac{2}{n\bar{y}} Cov(\hat{y}_{X'}, r(\hat{y}_{X'})) = \frac{2}{n\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} i\hat{y}_{X'} - \frac{n(n+1)}{2n} \right]. \tag{9}$$

Generally, the marginal contribution associated with all the possible model configurations including the explanatory variable X_k is specified as

$$\begin{aligned} LZ_{d=1}(\hat{Y}_{X'\cup X_k}) - LZ_{d=1}(\hat{Y}_{X'}) &= \frac{2}{n\mu} Cov(\hat{Y}_{X'\cup X_k}, r(\hat{Y}_{X'\cup X_k})) - \frac{2}{n\mu} Cov(\hat{Y}_{X'}, r(\hat{Y}_{X'})) \\ &= \frac{2}{n\mu} \left[Cov(\hat{Y}_{X'\cup X_k}, r(\hat{Y}_{X'\cup X_k})) - Cov(\hat{Y}_{X'}, r(\hat{Y}_{X'})) \right]. \end{aligned} \tag{10}$$

After some mathematical manipulations, the sample version of equation (10) can be re-written as

$$\begin{aligned} &LZ_{d=1}(\hat{y}_{X'\cup X_k}) - LZ_{d=1}(\hat{y}_{X'}) = \frac{2}{n\bar{y}} \left[Cov(\hat{y}_{X'\cup X_k}, r(\hat{y}_{X'\cup X_k})) - Cov(\hat{y}_{X'}, r(\hat{y}_{X'})) \right] \\ &= \frac{2}{n\bar{y}} \left[\frac{1}{n} \left(\sum_{i=1}^{n} i\hat{y}_{X'\cup X_k(i)} - \sum_{i=1}^{n} i\hat{y}_{X'(i)} \right) \right] = \frac{2}{n^2\bar{y}} \left[\sum_{i=1}^{n} i \left(\hat{y}_{X'\cup X_k(i)} - \hat{y}_{X'(i)} \right) \right]. \end{aligned} \tag{11}$$

Based on formula in (11), the marginal contribution due to the different model configurations including the covariate X_k is defined as

$$LZ_{d=1}^{X_k}(\hat{y}) = \sum_{X' \subseteq \mathscr{C}(X)\setminus X_K} \frac{|X'|!(n_k - |X'| - 1)!}{n_k!} \left\{ \frac{2}{n^2\bar{y}} \left[\sum_{i=1}^{n} i \left(\hat{y}_{X'\cup X_k(i)} - \hat{y}_{X'(i)} \right) \right] \right\}. \tag{12}$$

4 Application

As an applied example, we refer, without loss of generality, to the cryptocurrency data illustrated in [5]. The analysis focuses on the Coinbase Bitcoin as the target variable and Oil, Gold and SP500 as explanatory variables. Three different scenarios are built, according to the additional inclusion of each considered covariate in all the possible configuration models. First, the marginal contribution associated with the inclusion of variable SP500 and, subsequently, those related to the inclusion of variables Gold and Oil, are provided according to sample version of formula in (6).

For the sake of clarity, when considering the SP500 as the additional explanatory variable, the marginal contribution is derived as

$$LZ_{d=1}^{SP500}(\widehat{Coinbase}) = (1/3)(LZ(\hat{y}_{SP500,Gold,Oil}) - LZ(\hat{y}_{Gold,Oil}))$$
$$+ (1/6)(LZ(\hat{y}_{SP500,Gold}) - LZ(\hat{y}_{Gold})) + (1/6)(LZ(\hat{y}_{SP500,Oil}) - LZ(\hat{y}_{Oil}))$$
$$+ (1/3)(LZ(\hat{y}_{SP500})).$$

Analogously, we obtain the marginal contributions due to the Gold and Oil variables. Results are displayed in Table 1.

Additional covariate (X_k)	$LZ_{d=1}^{X_k}(\widehat{Coinbase})$
SP500	0.336
Gold	0.097
Oil	0.075

Table 1 Marginal contribution of each covariate in all the possible model configurations

From Table 1, variable SP500 provides the highest marginal contribution in explaining the daily Coinbase Bitcoin prices in different model configurations, while the minimum contribution is given by the variable Oil.

References

1. Arras, L., Horn, F., Montavon, G., Müller, K.-R., Samek, W.: "What is relevant in a text document?": An interpretable machine learning approach. PLoS ONE **12(8)**, 1–23 (2017)
2. Arrietaa, A.B., Dríaz-Rodríguezb. N, Del Sera,c,d, J., Bennetotb, A., Tabikg, S., Barbadoh, A., Garciag, S., Gil-Lopeza, S., Molinag, D., Benjaminsh, R., Chatilaf, R, Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (2019). Available at https://arxiv.org/pdf/1910.10045.pdf
3. Bracke, P., Datta, A., Jung, C., Shayak, S.: Machine learning explainability in finance:an application to default risk analysis, Staff Working Paper No. 816, Bank of England (2019). Available at https://ideas.repec.org/p/boe/boeewp/0816.html
4. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable AI in Credit Risk Management (2020)
5. Giudici, P., Raffinetti, E.: Lorenz Model Selection. J Classif (2020)
6. Koshevoy, G., Mosler, K.: The Lorenz Zonoids of a Multivariate Distribution. J Am Stat Accoc, **91(434)** , 873–882 (1996)
7. Mantegna, R.N.., Stanley H.E.: Introduction to econophysics: correlations and complexity in finance, Cambridge University Press (1999)
8. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**, 307–317 (1953)
9. Štrumbelj E., Kononenko, I.: An Efficient Explanation of Individual Classifications Using Game Theory Explaining prediction models and individual predictions with feature contributions. J Mach Learn Res. **11**, 1–18 (2010)

A warning signal for variable importance interpretation in tree-based algorithms

Un segnale di allarme per l'interpretazione dell'importanza delle variabili in algoritmi basati su alberi

Anna Gottard and Giulia Vannucci

Abstract Tree-based learning algorithms are largely utilized in several scientific domains. Each algorithm specifies a proper variable importance measure that captures the magnitude of the predictive importance of an explanatory variable. In applied studies, such measures are sometimes interpreted as importance relative to the data generating process, but this can be misleading. To prevent erroneous interpretation of the variable importance there is a need of a measure capable of detecting when the predictive variable importance differs from the explanatory one. We propose a measure that can act as a warning signal of this situation.

Abstract *Gli algoritmi basati su alberi sono ampiamente utilizzati in molte discipline scientifiche. Ogni algoritmo specifica una propria misura di importanza delle variabili, che cattura la grandezza dell'importanza preditiva di una variabile esplicativa. In alcuni studi, tali misure vengono talvolta interpretate come importanza relativa al processo generatore dei dati, anche se questo può essere fuorviante. Per prevenire un'interpretazione erronea dell'importanza delle variabili c'è bisogno di una misura in grado di rilevare quando l'importanza delle variabili predittiva è diversa da quella esplicativa. Proponiamo una misura che può agire da segnale di pericolo di questa situazione.*

Key words: Tree-based algorithms, Interpretable machine learning, Variable importance

Anna Gottard
DiSIA, University of Florence e-mail: anna.gottard@unifi.it

Giulia Vannucci
DiSIA, University of Florence, e-mail: giulia.vannucci@unifi.it

1 Introduction

Tree-based algorithms refer to a class of predictive models widely employed in several scientific domains. Besides being conceptually simple, capable of dealing with non-linearities and interactions and suitable also in high-dimensional settings, they are appreciated for their simple interpretation. Tree-based algorithm interpretation is usually based on a variable importance measure, that captures the magnitude of the contribution of each feature to the response prediction. The predictive importance is sometimes confused with an explanatory importance [4, 7]. Some examples of application of tree-based models to identify the relevant variables can be found, among many others, in [5] and [12]. The correct interpretation of these models raises a lot of interest in the scientific community. As a matter of facts, when tree-based algorithms are employed in real contexts, it is essential to know if they induce biases and prejudice, or lead to unfair and wrong decisions [8]. In [7], it has been shown that interpreting variable importance as representative of the direct effects in an explanatory sense can lead to incorrect scientific and practical conclusions. In this work, we explore more cases where this issue comes out, and we propose some measures that can be used as a warning signal of the misleading use of variable importance. These measures are based on the distance between the conditional and unconditional variable importance proposed, respectively, by [3] and [14].

2 An issue in variable importance interpretation

To show a situation where the variable importance can be misleadingly interpreted, we consider here two data generating processes described by the following system of equations and depicted by the two directed acyclic graphs in Figure 1.

$$\begin{cases} X_1 = \varepsilon_1 \\ X_j = aX_1 + \varepsilon_j \quad \text{for } j = 2,3,5 \\ X_4 = aX_2 + \varepsilon_4 \\ X_6 = aX_3 + \varepsilon_6 \\ Y \ = bX_4 + bX_5 + bX_6 + \varepsilon_Y. \end{cases} \tag{1}$$

$$\begin{cases} X_j = \varepsilon_j \quad \text{for } j = 1,2,3 \\ X_4 = aX_2 + \varepsilon_4 \\ X_5 = aX_1 + \varepsilon_5 \\ X_6 = aX_3 + \varepsilon_6 \\ Y \ = bX_4 + bX_5 + bX_6 + \varepsilon_Y. \end{cases} \tag{2}$$

These data generating processes concern a response variable Y, three variables, X_4, X_5 and X_6, having a direct effect on Y, and three variables, X_1, X_2 and X_3, having an indirect effect on Y. For both data generating processes $Y \perp\!\!\!\perp (X_1, X_2, X_3) \mid (X_4, X_5, X_6)$, but a different association is assumed among the background variables. In particular, the generating process (2) assumes X_1, X_2 and X_3 to be pairwise independent, while the generating process (1) assumes them correlated. We conducted a minimal simulation study on 100 data sets with size 500 generated from (1) and (2). All the datasets are sampled assuming independent errors ε_ℓ, $\ell \in \{1, \ldots, 6, Y\}$, with a standard normal distribution and $a = b = 3$.

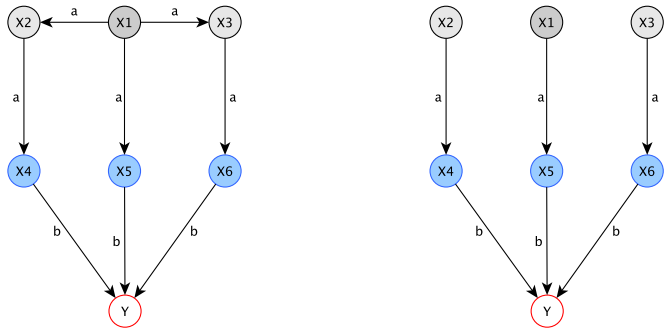


Fig. 1 Directed acyclic graphs relative to the data generating processes described in (1) on the left and (2) on the right

In Figure 2 we show the Monte Carlo distributions of the variable importance measures for the CART algorithm [2] (R package: `rpart`). This variable importance measure, proposed by Breiman [3] for the random forests, is based on the difference of the prediction accuracy before and after permuting the observed values of X_j. In the following, we will call this measure *unconditional variable importance* (VI^{u}).

$$VI^{\mathrm{u}}(X_j) = \frac{1}{H}\left\{\sum_{i=1}^{H}\left[Y_i - \widehat{Y}_i(X_{j,\mathtt{perm}})\right]^2 - \sum_{i=1}^{H}\left[Y_i - \widehat{Y}_i(X_j)\right]^2\right\} \tag{3}$$

where H is the size of the test set, $\widehat{Y}_i(X_{j,\mathtt{perm}})$ is the prediction on the test set when X_j is permuted, while $\widehat{Y}$ is the ordinary test set prediction. Following [3], in the plot the measures are rescaled to sum to 1. The simulation study indicates that for data generated by (2), X_4, X_5, X_6 are found to be the most important variables. Conversely, for data generated by (1), the background variable X_1 appears as the most important. Therefore, the mere presence of association among the background variables X_1, X_2 and X_3 induces an incorrect identification of the importance of variables to describe the data generating mechanism. This makes misleading to interpret the variable importance measures in an explanatory sense. As shown in [7], algorithms for tree construction based on a greedy splitting strategy exhibit a bias in the correct identification of relevant predictors, in the presence of irrelevant variables with high marginal association. In these situations, an explanatory variable which has an indirect effect on the response is preferred by the tree-based algorithms to forecast the response variable, inducing a discrepancy between the predictive and explanatory interpretation of feature importance. Our purpose is to give a measure that can signal this type of misleading interpretation.

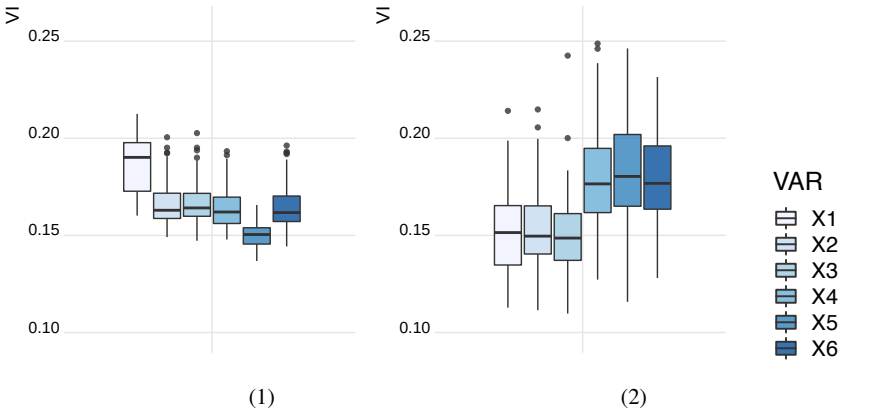


Fig. 2 Boxplot of the Monte Carlo distributions of the variable importance measures under the data generating processes described in (1) on the left and (2) on the right

3 Measures of warning

As the data generating process is usually unknown, it is not possible to know a priori if the observed set of variable importance measures can be safely interpreted in an explanatory sense. In this section, we propose some measures that can act as an alarm bell for possible misleading interpretation and potential unfairness in using tree-based algorithms. These measures of warning are based on the discrepancy between the variable importance measure in (3) and the variable importance measure proposed by [14], that permutes X_j only within a grid of values of the feature space.

$$VI^{\mathrm{c}}(X_j) = \frac{1}{H}\left\{\sum_{i=1}^{H}\left[Y_i - \widehat{Y}_i(X_{j,\mathtt{perm}|\boldsymbol{X}_{-j}})\right]^2 - \sum_{i=1}^{H}\left[Y_i - \widehat{Y}_i(X_j)\right]^2\right\} \tag{4}$$

where $\boldsymbol{X}_{-j}$ is an element of the permutation grid. As it can be seen in Figure 3 this measure of variable importance is more robust to the issue presented in Section 2.

A simple visual comparison of discrepancy between the two sets of variable importance measures can be used as a warning signal. As an alternative, one can compare variable importance ranks, checking for ranking inversions. However, only ranks inversions for the most important variables can be considered relevant, while inversions among variables with negligible importance can be due to sampling or permutation variability.

To obtain a proper measure of the discrepancy between these variable importance measures, notice that they have a different scale and cannot be directly compared. However, they carry interesting relative information if considered compositional, by scaling each vector to sum to 1. Let us denote $\boldsymbol{V}^{\mathrm{c}}$ and $\boldsymbol{V}^{\mathrm{u}}$ the vector of the compositional version of the conditional and unconditional variable importance in 4 and 3, respectively. The sample space of each vector is the standard simplex $\mathbb{S}^p = \{(v_1, \ldots v_p) \mid \sum_{i=1}^{p} v_i = 1, \text{ where } v_i \geq 0\}$. A Euclidean distance between these

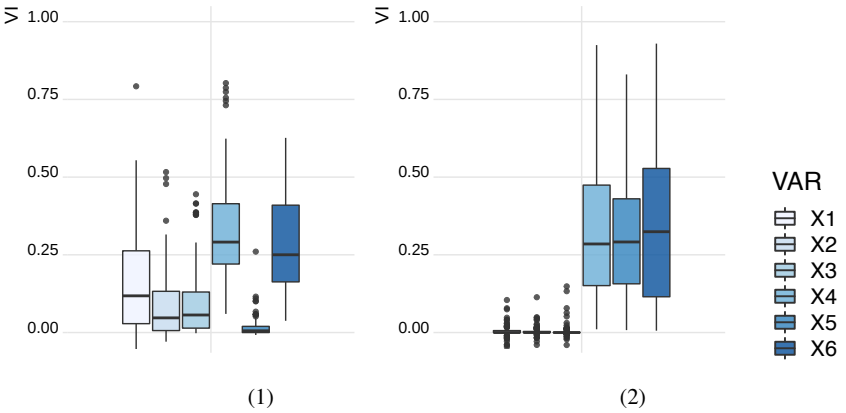


Fig. 3 Boxplot of the Monte Carlo distributions conditional variable importance measures under the data generating processes described in (1) on the left and (2) on the right

two vectors is ill-suited because of their compositional nature, and an Aitchison geometry [1] can be adopted instead. A possible distance between the two variable importance vectors is

$$d_a(\boldsymbol{V}^{\mathrm{u}}, \boldsymbol{V}^{\mathrm{c}}) = \sqrt{\frac{1}{2p}\sum_{i=1}^{p}\sum_{i=1}^{p}\left(\ln\frac{v_i^{\mathrm{u}}}{v_j^{\mathrm{u}}} - \ln\frac{v_i^{\mathrm{c}}}{v_j^{\mathrm{c}}}\right)}, \tag{5}$$

or a similar formula where the denominators are replaced by the geometric means, when using the so-called centred log-ratio transformation. These distances can be adopted whenever all the entries of the compositional vectors are strictly positive. At the presence of zero variable importance measures, [1] suggests to add small amounts to the zero entries. Even if this solution is not sub-compositionally coherent, it works quite well in our case, where the main issue concerns the most important variables. [6] and [9] transformations can be considered as another option. Alternative metrics could be used that treat compositional data as directional data. See for instance [13], [10] and [11] for this approach.

4 Concluding Remarks

Machine learning algorithms are powerful tools for prediction tasks. Interpretable machine learning is raising the attention of the scientific community, due to its spread out of the ordinary technological application field, guaranteeing fair and accurate predictions. Among the machine learning algorithms, the tree-based ones are considered the more easy interpretation, because of the tree diagrams and variable importance measures. However, these measures are more and more frequently interpreted in an explanatory sense. In this paper, we showed a situation in which this kind of interpre-

tation is misleading, in line with what suggested by [7] and proposed some measures that can act as an alarm bell of this issue. In the paper, we limited the discussion to the case of regression trees, but what presented is valid also for classification trees, random forests and boosted trees.

References

1. Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) 44.2 (1982): 139-160.
2. Breiman, L., Friedman, J., Stone, CJ, Olshen, RA: Classification and Regression Trees. CRC Press (1984).
3. Breiman, L.: Random forests. Machine Learning, 45(1) (2001): 5-32.
4. Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16.3 (2001): 199-231.
5. Fan, G. Z., Ong, S. E., & Koh, H. C.: Determinants of house price: A decision tree approach. Urban Studies 43.12 (2006): 2301-2315.
6. Fry, J. M., Fry, T. R., & McLaren, K. R.: Compositional data analysis and zeros in micro data. Applied Economics 32.8 (2000): 953-959.
7. Gottard, A., Vannucci, G., & Marchetti, G.M.: A note on the interpretation of tree-based regression models. Accepted in Biometrical Journal.
8. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) 51.5 (2018): 1-42.
9. Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V.: Zero replacement in compositional data sets. In Data Analysis, Classification, and Related methods (pp. 155-160). Springer, Berlin, Heidelberg (2000).
10. Scealy, J. L., & Welsh A. H.: Regression for compositional data by using distributions defined on the hypersphere. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.3 (2011): 351-375.
11. Scealy, J. L., & A. H. Welsh.: Fitting Kent models to compositional data with small concentration. Statistics and Computing 24.2 (2014): 165-179.
12. Schilling, C., Mortimer, D., Dalziel, K., Heeley, E., Chalmers, J., & Clarke, P.: Using Classification and Regression Trees (CART) to identify prescribing thresholds for cardiovascular disease. Pharmacoeconomics 34.2 (2016): 195-205.
13. Stephens, M. A.: Use of the von Mises distribution to analyse continuous proportions. Biometrika 69.1 (1982): 197-203.
14. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. : Conditional variable importance for random forests. BMC Bioinformatics 9.1 (2008): 307.

Assessment of the effectiveness of digital flyers: analysis of viewing behavior using eye tracking

Valutazione dell'efficienza dei volantini digitali: analisi del comportamento visivo tramite eye tracking

Gianpaolo Zammarchi[1], Claudio Conversano[2], Francesco Mola[3]

Abstract Digital flyers are commonly used by retailers to promote their products and special offers. However, only a small number of studies have explored how visual information presented by digital flyers is processed by consumers. We designed three tasks to evaluate differences in the way a group of university students recruited at the University of Cagliari looked at digital flyers of two different retailers advertising technology products. The viewing behavior was assessed using the eye tracking technology, which allows to register the sequence of observations made by a subject when looking at a visual stimulus. The analysis of metrics such as time to completion and number of fixations, as well as Markov Chain analysis conducted on fixations, showed differences in the performance of the two flyers. These differences will be further explored to detect relevant aspects to improve the effectiveness of digital flyers. Additional analyses are currently ongoing to identify clusters of subjects sharing similar viewing behavior as well as characteristics of the areas of interests (AOI) associated with a higher number of fixations during free observation.

Abstract *I volantini in versione digitale vengono sempre più utilizzati dai rivenditori per promuovere i propri prodotti ed offerte speciali. Tuttavia, solo un numero ridotto di studi ha esplorato il modo nel quale le informazioni visive presentate nei volantini digitali sono processate dai consumatori. Il nostro studio comprende tre task ideati per valutare le differenze nel modo in cui un gruppo di*

1 Gianpaolo Zammarchi, Dep. of Economics and Business Sciences, University of Cagliari (Italy); email: gp.zammarchi@unica.it

2 Claudio Conversano, Dep. of Economics and Business Sciences, University of Cagliari (Italy); email: conversa@unica.it

3 Francesco Mola, Dep. of Economics and Business Sciences, University of Cagliari (Italy); email: mola@unica.it

studenti dell'Università di Cagliari ha osservato i volantini digitali di due diversi rivenditori di prodotti tecnologici. Il comportamento visivo è stato valutato tramite l'eye tracking, una tecnologia che consente di registrare la sequenza di osservazioni durante l'osservazione di uno stimolo visivo. L'analisi dei tempi di completamento e del numero di fissazioni, così come un'analisi sulle fissazioni condotta tramite Markov Chain, ha permesso di evidenziare alcune differenze nelle performance dei due volantini. Queste differenze verranno ulteriormente esplorate per identificare accorgimenti che consentano di migliorare l'efficacia di un volantino digitale. Inoltre, ulteriori analisi sono in corso per identificare cluster di partecipanti che presentino un comportamento visivo simile oltre che caratteristiche delle aree di interesse (AOI) associate ad un maggior numero di fissazioni durante una fase di osservazione libera.

Key words: flyers, eye tracking, digital marketing, Markov Chain

1 Introduction

Nowadays almost every company exploits the potential of Internet using digital marketing techniques in order to reach a higher number of people and decrease costs. However, retailers keep using traditional approaches such as paper flyers (also known as door drop flyers) to advertise their products and special offers. Digital flyers represent a clever way to combine traditional and innovative approaches, through the creation of a flyer that shows a higher number of products and their characteristics compared to a classic internet ad, still having the chance to reach a higher number of people compared to paper flyers.

Eye tracking is a technology that allows to evaluate with high precision the sequence of observations made by a subject. This technology has successfully been used to evaluate how people interact with a web page [1,2,5,8]. However, only few studies have applied it to evaluate how visual information is processed by the consumers during the observation of flyers [3]. Data on visual observations might allow to identify differences in the observation of distinct visual stimuli (e.g. to measure differences in the effectiveness of two digital flyers) as well as clusters of subjects that share similar patterns of observations.

In this study we used eye tracking to analyze the viewing behavior of a sample of university students recruited at the University of Cagliari, Italy, aiming to identify differences in the performance of digital flyers from two different retailers. We followed a research path characterized by the four steps described in the following section.

2 Research path

2.1 Design of the tasks

This study aims to explore the differences between different types of flyers using the eye tracker. We selected the latest flyers from two major retailers promoting technology products. We created three tasks based on the products advertised, paying attention to choose pages that contained the same type of objects.

The tasks involved finding a specific product (unique for each task) that met the features communicated to the participant right before the test (Table 1). To complete the task, the participant had to search for the object satisfying the features on a specific page of one flyer and then on the other one (the two flyers were presented in random order). The search started immediately after communicating to the participant the criteria needed to choose the object and ended when the subject communicated the name of the right object. This procedure was repeated for the second flyer. In case of error in the identification of the object, the participant was told to keep searching and the number of errors was registered. During the test, the participants had to use only their eyes (no mouse, no keyboard) to locate the object. Before the task was started, the study procedures were explained to each subject.

Table 1: Description of the tasks

Task	***Description of the Task***
Task 1	Find the washer machine with the lowest price
Task 2	Find the smartphone with the best percentage of discount
Task 3	Find a TV screen measuring at least 40 inches, with the lowest original price

2.2 Sampling

Participants were recruited among university students, mainly from Economics and Law departments, randomly selected in group study rooms (in different days of the week and different times of the day). Each student performed the three tasks and information about age, gender, residency and university course were collected.

2.3 Assessment of eye movements

For each participant, eye movements during the task were gathered with a screen-based Tobii X2-60 Compact eye tracker, which captures gaze data at 60 Hz, applied to a 25-inch monitor. For each participant, the instrument was calibrated according to the specific height and distance from the screen. The velocity-threshold fixation identification (I-VT) algorithm implemented in Tobii Studio v. 3.3.1 was used to classify eye movements into different types (e.g. saccade, fixation, etc.). The algorithm classifies eye movements based on the velocity of the directional shifts of

the eye. Among different types of eye movements, we chose to focus on fixations, which are considered the metric of most interest in similar researches [7].

2.4 Statistical analysis

Normality of distribution of times and number of fixations during each task was evaluated using the Shapiro-Wilk test. Non-parametric tests were used in case of non-normal distribution. Correlation between times and number of fixations was checked using Pearson's or Spearman's correlation test. For each task, times and fixations between the flyers of the two retailers were compared using the paired samples t-test or Wilcoxon test. Analyses were adjusted for multiple testing according to Bonferroni ($p = 0.017$ i.e. 0.05 / 3 tasks). A qualitative analysis of the scanpaths for each participant was also conducted. To this aim, scanpaths for each participant were created using circles to plot the sequence of the different fixations. The dimension of each circle was proportional to the duration of the corresponding fixation. A preliminary analysis with Markov Chain was conducted on Task 3 to further analyze gaze transitions among AOIs. Seven AOIs were drawn around the main objects of the flyers and named with letters from A to G. These were considered to be the states of the Markov Chain. Using coordinates on the screen (X, Y) a letter was assigned to each fixation to obtain a sequence needed for the computation of transitions between states. Two stationary distributions were obtained and compared using the verify Homogeneity function in the Markov Chain R package [6], in order to assess the presence of differences regarding the viewing behavior within the two flyers. Analyses were conducted using R v. 3.6.2 [4].

3 Results

3.1 Comparison of the effectiveness of the two flyers

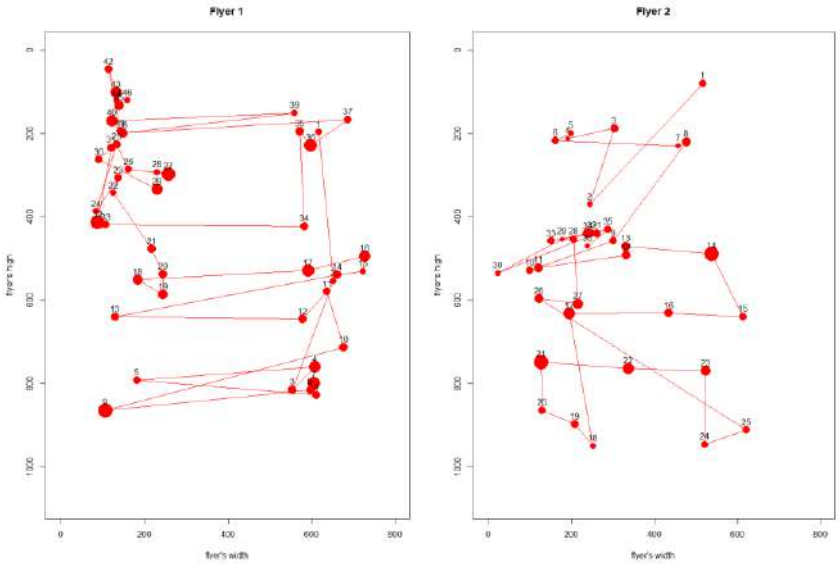
A total of 25 university students completed the three tasks. Three students were excluded due to technical problems during the task, leading to a final sample of 22 participants. In our sample, 41% of the participants were women and mean age (± standard deviation) was 23.86 ± 3.56. Distribution of times and number of fixations was normal for Task 2 and 3 but not for Task 1. Therefore, for the latter task non-parametric methods were used. For all tasks, a strong correlation between time and number of fixations was observed (Task 1: correlation coefficient: 0.84, $p < 0.001$; Task 2: correlation coefficient: 0.96, $p < 0.001$; Task 3: correlation coefficient: 0.71, $p < 0.001$). A better performance of Flyer 2 was observed during Task 1. Specifically, participants completed this task in a faster way and with a lower

number of fixations when looking at Flyer 2 compared to Flyer 1 ($p < 0.001$ for both). Additionally, participants completed Task 3 in a faster way when looking at Flyer 2 compared to Flyer 1, although this difference was not significant after adjusting for multiple testing ($p = 0.03$) . No difference was observed for Task 2. In the preliminary Markov Chain analysis, transition matrices between the different AOIs were computed using fixations made in the pages of the two flyers used for Task 3. The two stationary distributions were significantly different ($\chi^2 = 380.95$, $p < 0.001$), supporting the hypothesis that, on a global level, participants looked at the flyers in a different way while searching information.

3.2 Scanpaths analysis

Scanpaths for each participant were represented using filled circles to plot the sequence of the different fixations. As an example, scanpaths related to Task 1 for a participant are reported in Figure 1. The scanpaths show a lower number of deviations and changes of direction during the observation of Flyer 2 compared to Flyer 1, in line with the quantitative results showing a better performance of Flyer 2 as regard to time and number of fixations required to complete the task in the whole sample.

Figure 1: Representation of the scanpaths of a participant for Flyer 1 and 2 (Task 1)



4 Discussion

Our results showed significant differences in the performance of digital flyers from two retailers. These findings might be explained by differences in the layout of the two flyers. In fact, while both flyers showed a similar number of objects of the same type, the second flyer represented them using a more orderly design (the pictures of the promoted objects had similar dimensions and were evenly distributed across the page). Conversely, as regard to the task for which no differences were detected, the design of the pages selected for this task was similar among the two flyers. Additional analyses to identify clusters of participants with similar viewing behavior, as well as characteristics of areas of a page with higher density of fixations during free observation, are currently ongoing. These analyses will allow to further characterize which aspects of a digital flyer might be associated with higher effectiveness.

References

1. Bach, M.P.: Usage of social neuroscience in E-Commerce research – Current research and future opportunities, J. Theoretical Appl. Electron. Commer. Res. 13, I-IX (2018)
2. Hwang Y.M., Lee, K.C.: Using an eye tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. Int. J. Hum-Comput. Interact. 34,15-24 (2017)
3. Pentus, K., Ploom, K., Kuusik, A., Mehine, T.: How to optimize sales flyers – novel experiment design. Baltic J. of Manag. ISSN: 1746-5265 (2018)
4. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020)
5. Roth, S.P., Tuch, A.N., Mekler, E.D., Bargas-Avila J.A., Opwis, K.: Location matters, especially for non-salient features–An eye tracking study on the effects of web object placement on different types of websites. Int. J. Hum-Comput. Stud. 71, 228-235 (2013)
6. Spedicato, G.: Discrete Time Markov Chains with R, The R Journal, https://journal.r-project.org/archive/2017/RJ-2017-036/index.html (2017)
7. van der Lans, R., Wedel, M., Pieters, R.: Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm. Behavior Res. Methods. 43, 239-257 (2011)
8. Zammarchi, G., Mola, F.: Evaluation of the web usability of the University of Cagliari portal: an eye tracking study In: CLADAG 2019 Book of short papers. Centro Editoriale di Ateneo Università di Cassino e del Lazio Meridionale, Cassino, IT, ISBN: 978-88-8317-108-6 (2019)

At risk mental status analysis: a comparison of model selection methods for ordinal target variable

Analisi dello stato mentale a rischio: analisi di confronto dei metodi per la selezione di modelli per variabile target ordinale

Elena Ballante, Silvia Molteni, Martina Mensi, Silvia Figini

Abstract The presence of an ordinal variable poses several problems that are not deepen in the literature. In this work we aim to analyse a dataset on mental state at risk, a delicate and extremely debatable definition in the psychiatric field, in order to produce a model that is of double value. On the one hand it allows to understand which are the most influential characteristics on the mental state, on the other hand it allows to predict the correct diagnosis and possibly also to assess the risk of a possible transition to the psychotic state. With this aim we focus on model selection methods in the framework of ordinal target variable.
La presenza di una variabile ordinale pone diversi problemi che non trovano trattazioni approfondite in letteratura. In questo lavoro ci proponiamo di analizzare un dataset sullo stato mentale a rischio, definizione delicata ed estremamente dibattuta in ambito psichiatrico, al fine di produrre un modello che sia di duplice valenza. Da una parte che permetta di capire quali siano le caratteristiche maggiormente influenti sullo stato mentale, dall'altra che permetta di prevedere la corretta diagnosi ed eventualmente anche di valutare il rischio di una possibile transizione allo stato psicotico. Con questo obiettivo il lavoro si focalizza sulla fase di selezione del modello nell'ambito di variabile target ordinale.

Key words: ordinal classification, psychosis, model selection

---

Elena Ballante
Università di Pavia, Dipartimento di Matematica, e-mail: elena.ballante01@universitadipavia.it

Silvia Molteni
Università di Pavia, Dipartimento di Scienze del Sistema Nervoso e del Comportamento, e-mail: silvimolt@yahoo.it

Martina Mensi
IRCCS Fondazione Mondino, Pavia e-mail: martina.mensi@mondino.it

Silvia Figini
Università di Pavia, Dipartimento di Scienze Politiche e Sociali, e-mail: silvia.figini@unipv.it

1 Introduction

Despite advances in pharmacological and psychotherapeutic interventions over the last decades, psychotic disorders continue to be among the most severe disorders in medicine. In children and adolescents, schizophrenia is one of the ten main causes of disability-adjusted life years (DALYs) in 10 to 14-year-old boys and 15 to 19-year-old girls [15]). The identification of people at high-risk of developing psychosis is one of the most promising strategies to improve outcomes. Indeed, retrospective studies indicate that the onset of full psychosis is commonly preceded by a prodromal phase lasting up to several years [16, 24]. Recently, the importance of research in persons at high risk has been increasingly recognized to such an extent that Attenuated Psychosis Syndrome has been introduced in section III ("Emerging Measures and Models") of the Diagnostic and Statistical Manual of Mental Disorder, fifth Edition. Some concerns raised regarding the introduction of this new syndrome need to be addressed with special attention in children and adolescents, where research on the high risk state is still in its infancy [23]. In particular, criticism about pathologization of non-ill behaviours and experiences has been voiced. In fact, during adolescence, the assessment of psychiatric symptoms and disorders is challenging. Several authors have underlined the difficulty in discriminating between normal behaviours and psychiatric symptoms [26]. Normative adolescent experiences can make the clinical picture blurred and lead to false positive psychotic diagnoses [10].Overall, in children and adolescents research on the high risk state and attenuated psychotic symptoms is still in its infancy and the clinical validity of at risk criteria appears understudied. Furthermore, only few studies have evaluated the psychopathological and neuropsychological characteristics of adolescents with attenuated psychotic symptoms (APS).

In this context an accurate data analysis is fundamental for knowledge extraction and prediction: it is extremely important to deploy adequate models and to perform a suitable selection procedure.

If there are several models that are suitable for the analyses, it can be noted that the evaluation measures for model selection are few and inadequate to the problem under analysis.

Performance indicators can be used to assess the performance of a model in terms of accuracy, discriminatory power and stability of the results. The choice of indicators to made model selection is a fundamental point and many approaches have been proposed over the years (see e.g. [6, 18]).

Multi-class classification models are generally evaluated averaging binary classification indicators (see [22]) and in the literature there is not a clear distinction among them with respect to multi-class nominal and ordinal targets (e.g. [13]).

While in the model definition stage for ordinal target variable there are different approaches in the literature (see [1, 3]), for the model selection there is a lack of adequate tools ([9]).

In our opinion, performance indicators should consider the nature of the target variable, especially when the dependent variable is ordinal. In medical application is also fundamental to take into account the uncertainty that the model assign to the

prediction in order to obtain the maximum of interpretability. This leads us to apply and compare different measures to select the best model to predict the mental status, contexts characterized by a multi-class ordinal target variable.

The rest of the paper is organized as follow: Section 2 describes data at hand and the analysis performed, Section 3 describes preliminary results obtained.

2 Data description and analysis

The dataset is composed by 240 observation (corresponding to 240 patients under examination). The target variable is the membership to one of three categories: subject not at risk, at risk, psychotic. This variable is clearly ordinal and in this context is extremely important to consider the order of levels.
The covariates considered are some personal information like age at onset, sex, ethnicity, some medical history of the patients and familiarity to mental illnesses, some information about symptoms, duration of the psychotherapy and of the drugs assumption, the IQ index (Intelligence Quotient) and the SOFAS (Social and Occupational Functioning Assessment Scale).
On this dataset, five different models are implemented and compared: Ordinal logistic regression [20], Classification tree [7], Support vector machine [12], Random forest [8], k-Nearest Neighbour [11].
In the model selection step we compare the information given by an index for ordinal target proposed in [4] with standard indexes used in literature that are AUC (Area Under the ROC curve), accuracy (expressed in terms of correct classification) and MSE (Mean Square Error) (see [13] and [19] among others), another index for ordinal target variable proposed in [9] and the total misclassification cost used in [21]. The index proposed in [4] is defined basing on a classification function, i.e. a function which represents the actual classification made by the model under evaluation, compared with an exact classification function that is the goal of each model. This index takes into account the ordinal structure of the target variable and the probability assigned from the model at each observation. This first aspect has obvious advantages in this context, whereas the second aspect is extremely useful in medical application, when we need to consider the uncertainty of the prediction as well as the prediction itself. In the rest of the paper we refer to this index as OPI (Ordinal Probability Index). The AUC for multi-class classification is defined in [17] as a generalization of the AUC (based on the probabilistic definition of AUC); it suffers of different weaknesses also in the binary classification problem ([14]) and it is cost-independent, assumption that can be viewed as a weakness when the target is ordinal.
Accuracy (percentage of correct predictions over total instances) is the most used evaluation metric for binary and multi-class classification problems ([25]), assuming that the costs of the different misclassifications are equal.
Mean square error (MSE) measures the difference between prediction values and

observed values in regression problems using an Euclidean distance. MSE can be used in ordinal predictive models, converting the classes of the ordinal target variable y in integers and computing the difference between them and it does not takes into account the ordering in a predictive model characterized by ordinal classes in the response variable. Furthermore, it is well known that in imbalanced data characterized by under-fitting or over-fitting the mean square error could provide trivial results (see [22]).
Total misclassification cost is simply defined as the sum of absolute values of the differences between the real class and the predicted class, transformed integer values as in MSE. It was used in [21] to prune a new classification tree algorithm in ordinal framework.
Ordinal Index proposed in [9] is based on confusion matrix and on the concept of non-discordant pair of points, i.e. when the relative order of the predicted classes of two observations is the same of the relative order of the real classes. The advantage of this method with respect to AUC, accuracy and MSE is that consider the ordinal structure of the target variable, but it does not take into account the probabilities assigned to the prediction like the first index proposed.

3 Preliminary results

Data at hands are composed by 240 observations and fifteen covariates both qualitative and quantitative as possible explanatory variables and predictors. The target variable has three ordered level (not at risk, at risk, psychotic). On this dataset six different predictive models are implemented:

- Ordinal logistic regression (Ord Log),
- Classification tree (Tree),
- Support vector machine (SVM),
- Random forest (RFor),
- k- Nearest Neighbour (kNN),
- Naive Bayes (NBayes).

In order to select the best model a 5-fold cross validation is implemented. The models are compared in terms of out of sample performance on the basis of OPI, AUC, accuracy, MSE, the ordinal index (Ord Ind) and misclassification cost (Misc Cost). Table 1 reports the mean values of the metrics under comparison derived from the cross validation exercise. For sake of clarity, Table 2 shows the resulting ranks for the models, using the results obtained for the four metrics under comparison.

The performances achieved on the model under comparison underline that there is a group of models with worst performances (Ord Log, Tree, kNN) and a group of model with better performances (SVM, RFor, NBayes). On the basis of OPI, Accuracy, MSE, Ord Ind and Misc Cost the Random Forest model is the best one. Looking at AUC the best model is Naive Bayes but the performance exactly the

Model	OPI	AUC	Accuracy	MSE	Ord Ind	Misc Cost
Ord Log	0.176	0.847	0.713	0.316	0.372	12.6
Tree	0.215	0.850	0.685	0.399	0.410	14.6
SVM	0.150	0.987	0.760	0.310	0.335	11.2
RFor	0.138	0.902	0.774	0.282	0.315	10.4
kNN	0.351	0.735	0.624	0.404	0.451	16.4
NBayes	0.165	0.912	0.756	0.301	0.341	11.2

Table 1 Model selection.

Model	OPI	AUC	Accuracy	MSE	Ord Ind	Misc Cost
Ord Log	4	5	4	4	4	4
Tree	5	4	5	5	5	5
SVM	2	2	2	3	2	2.5
RFor	1	3	1	1	1	1
kNN	6	6	6	6	6	6
NBayes	3	1	3	2	3	2.5

Table 2 Results in terms of ranking.

same of Random Forest in terms of De Long test ($p > 0.2$).
On the basis of the results obtained in Table 1 a further analysis considers the results obtained using Random Forest. A necessary aspect to consider is the analysis of features involved in the classification process. The next steps will be divide the data in 60% and 40% to deploy and train the single model selected on the dataset, on this model the variable importance can be analyse in order to understand which aspects have more weight in the classification. Random Forest model produces as output an index of variable importance; on the basis of the data at hand the variable with higher degree of importance are also interesting in the clinical domain, as for example SOFAS score, CGIS score, depressive disorders. Also SES score is one of the variables with greater importance, underlining the influence of the socio-economic level on mental status.
A further analysis will consider longitudinal behaviour of the patient thus providing the opportunity to model also time dependent covariates. Moreover in a different dataset we have information about the transition to psychosis of some patient included in this study and further analysis will focus on this groups of patients that make the transition to psychosis in order to evaluate the performances of the model selected only on this category.

References

1. Agresti, A.: Analysis of ordinal categorical data. John Wiley & Sons (2010)
2. American Psychiatric Association, American Psychiatric Association, DSM-5 Task Force: Diagnostic and statistical manual of mental disorders: Arlington, American Psychiatric Association (2013)

3. Ahmad, A., Brown, G.: Random ordinality ensembles: ensembles methods for multi-valued categorical data. Information Sciences **296**, 75–94 (2015)
4. Ballante, E., Uberti, P., Figini, S.: A new approach in model selection for ordinal target variables. arXiv:2003.02761 (2020)
5. Bartels-Velthuis, A.A., Wigman, J.T.W., Jenner, J.A., Bruggeman, R., van Os, J.: Course of auditory vocal hallucinations in childhood: 11-year follow-up study. Acta Psychiatr. Scand. (2016)
6. Bradley, A.P.: The use of the area under the ROC curve in evaluation of machine learning algorithms. Pattern Recognition, **30**, 1145–1159 (1997)
7. Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Taylor & Francis (1984)
8. Breiman L.: Random Forests. Machine Learning **45**, 5–32 (2001)
9. Cardoso, J., Sousa, R.: Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence, **25**, 1173–1195 (2011)
10. Carol, E.E., Mittal, V.A.: Normative adolescent experiences may confound assessment of positive symptoms in youth at ultra-high risk for psychosis. Schizophr. Res. **166**, 358–359 (2015)
11. Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, **13**, 21–27 (1967)
12. Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., Vapnik, V.: Support vector regression machines. In Advances in neural information processing systems, 155–161 (1997)
13. Gaudette, L., Japkowicz, N.: Evaluation Methods for Ordinal Classification. In: Gao Y., Japkowicz N. (eds) Advances in Artificial Intelligence, 207–210 (2009)
14. Gigliarano, C., Figini, S., Muliere, P.: Making classifier performance comparisons when ROC curves intersect. Computational Statistics and Data Analysis, **77**, 300–312 (2014)
15. Gore, F.M., Bloem, P.J.N., Patton, G.C., Ferguson, J., Joseph, V., Coffey, C., Sawyer, S.M., Mathers, C.D.: Global burden of disease in young people aged 10–24 years: a systematic analysis. Lancet Lond Engl, **377**, 2093–2102 (2001)
16. Hafner, H., Löffler, W., Maurer, K., Hambrecht, M., an der Heiden, W.: Depression, negative symptoms, social stagnation and social decline in the early course of schizophrenia. Acta Psychiatr. Scand. **100**, 105–118 (1999)
17. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning **45**, 171–186 (2001)
18. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning **77**, 103–123 (2009)
19. Huang, J., Ling, C.X.: Constructing New and Better Evaluation Measures for Machine Learning. Proc. 20th International Conference on Artificial Intelligence (IJCAI2007), 859–864 (2007)
20. McCullagh, P.: Regression Models for Ordinal Data. Journal of the Royal Statistical Society: Series B (Methodological) **42**, 109–127 (1980)
21. Piccarreta, R.: Classification trees for ordinal variables. Computational Statistics. **23**, 407–427 (2008)
22. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process **5**, 171–186 (2015)
23. Schimmelmann, B.G., Walger, P., Schultze-Lutter, F.: The significance of at-risk symptoms for psychosis in children and adolescents. Can. J. Psychiatry **58**, 32–40 (2013)
24. Schultze-Lutter, F., Ruhrmann, S., Berning, J., Maier, W., Klosterkötter, J.: Basic symptoms and ultrahigh risk criteria: symptom development in the initial prodromal state. Schizophr. Bull. **36**, 182–191 (2010)
25. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence. Lecture Notes in Computer Science **4304** (2006)
26. Welsh P., Tiffin P.A.: Attitudes of patients and clinicians in relation to the at-risk state for psychosis. Early Interv Psychiatry **7**, 361–367 (2013)

Categorical Encoding for Machine Learning

Quantificazione delle variabili qualitative per il Machine Learning

Agostino Di Ciaccio

Abstract In recent years, interest has grown in addressing the problem of encoding categorical variables, especially in deep learning applied to big-data. However, the current proposals are not entirely satisfactory. The aim of this work is to show the logic and advantages of a new encoding method that takes its cue from the recent word embedding proposals and which we have called Categorical Embedding. Both a supervised and an unsupervised approach will be considered.

Abstract *Negli ultimi anni è cresciuto l'interesse nell'affrontare il problema della quantificazione delle variabili qualitative soprattutto nel deep learning applicato a grandi insiemi di dati. Le soluzioni proposte non sono però del tutto soddisfacenti. Obiettivo di questo lavoro è mostrare la logica e i vantaggi di un nuovo metodo di codifica che, prendendo spunto dalle recenti proposte di word embedding, abbiamo chiamato Categorical Embedding. Sarà considerato sia un approccio supervisionato che non-supervisionato.*

Key words: categorical encoding, deep learning, word embedding

1 Introduction

Usually, Big-Data include tens or hundreds of variables, which have mixed measurement levels with many categorical variables, sometimes with high cardinality. The treatment of many categorical variables, especially when combined with quantitative variables, is a complex topic that has no easy solutions. The problem has been considered in many areas of classical statistics (see for example Azzalini 2001)

This theme is particularly relevant in machine learning applied to large datasets. In fact, the only method that can naturally handle many variables with a mixed

[1] Agostino Di Ciaccio, University of Rome "La Sapienza"; agostino.diciaccio@uniroma1.it

measurement level is the decision tree (although software often does not take this potential into account).

The purpose of this work is to show the logic and the advantages of a new encoding method that takes its cue from the recent proposals of *word embedding* in Natural Language Processing (NLP) (Bengio et al. 2003) and which we call *categorical embedding*. Some applications to real datasets will show the interest of our proposal.

2 Encoding Categorical variables

Applying neural networks to categorical data requires some form of encoding. Perhaps the most used method is *one-hot* encoding, i.e., for each category, adding a new binary feature indicating it. However, in the case of high cardinality variables, such technique leads to a large number of new features. Moreover, the new variables are perfectly independent, and this is unrealistic. The categories can have relationships and similarities that could be extracted from the context. An example is the variable "*day of the week*" which is a cyclic ordinal feature. If we represent the days of the week through 7 one-hot vectors (or at least 6, but in Machine Learning it is not necessary to drop one dummy) we obtain a spatial representation of the variable in 7 dimensions in which every pair of categories is at Euclidean distance $\sqrt{2}$ from each other. This representation is not meaningful: it could be more coherent to represent the days of the week on a circumference in a smaller two-dimensional space.

On the other hand, even a 'circular' representation does not take into account that the days of the week can be distinguished between 'working' and 'non-working' and this distinction is often more relevant for the analysis. In table 1 we have reported the distances between the days of the week obtained by considering the encodings given by the Glove's word-vectors (Pennington et al. 2014) obtained by analysing millions of documents by an NLP model. If our analysis concerns, e.g., the sales forecast of a supermarket or the level of particulates in the air of a great city, this is certainly a coherent coding.

Table 1: Distances between the days of the week in the 50-dimensional Glove word-vectors

	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	0.002	0.004	0.002	0.007	**0.086**	**0.074**
Tuesday		0.002	0.002	0.009	**0.087**	**0.077**
Wednesday			0.003	0.008	**0.079**	**0.069**
Thursday				0.007	**0.085**	**0.074**
Friday					**0.072**	**0.062**
Saturday						0.009

It is therefore not obvious that the 'natural' order of the categories should be kept in our coding. If we want to maintain the natural order, we could use the matrix of "*hot-vectors of order*" as suggested in the Optimal Scaling approach (Gifi 1981, Di Ciaccio 1988).

In general, also categorical variables with high cardinality can have a satisfactory representation in a small space, correctly representing their relationship. Consider, for

example, the 102 Italian provinces: the one-hot encoding would provide a representation in a big dimensional space while we know that a representation in a two-dimensional space, using for example the geographical position of the principal city, could be sufficient for our analysis. In general, there is no encoding which is optimal independently of the objective of the analysis and the model applied. The biggest distinction between the encoding methods is based on the approach that can be supervised or unsupervised.

An old but interesting proposal is Optimal Scaling (Gifi 1981) which can generate quantifications both in a supervised and unsupervised approach. In an unsupervised approach, two well-known equivalent methods can be derived: Homogeneity Analysis (HA, Gifi 1981) and Multiple Correspondence Analysis (MCA, Benzecri 1973). By these methods, the categories are 'optimally' encoded maximizing the eigenvalues of the correlation matrix. In the French approach the problem is solved analytically, while in the Gifi approach the problem is solved numerically. This numerical variant offers a large amount of flexibility.

Let m the number of categorical variables and j the index of the generic variable ($j=1,2,\ldots,m$), k_j the number of categories of the j-*th* variable, $\mathbf{G}_j = \left[\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{k_j}\right]$ the indicator matrix of dimension $n \times k_j$. Let p be the number of dimensions, which needs to be fixed a priori, and r the generic dimension ($r=1,2,\ldots,p$). Each variable can be associated with a matrix $\mathbf{C}_j$ of dimension $k_j \times p$ containing the *category quantifications*. The quantification for the j-*th* variable on the r-*th* dimension is given by

$$\mathbf{v}_{jr} = \mathbf{G}_j \mathbf{c}_{jr} = \sum_{h=1}^{k_j} c_{jhr} \mathbf{g}_{jh} \qquad (1)$$

Hence, the vector of the quantified data is a linear combination of the indicator variables and the set of possible quantifications defines a subspace R^{k_j}. In fact, the quantification $\mathbf{v}_{jr}$ is a linear combination of an orthogonal base of R^{k_j}. If the variable is ordered, the constraints on the quantifications define a polyhedric convex cone (Gifi 1981) and a similar approach based on b-splines can manage also quantitative data (Di Ciaccio 1990). Define the matrix $\mathbf{X}$ containing the so-called *object scores* with dimension $n \times p$, to obtain the scores and the quantifications, we can minimize the following equation by an *alternating least squares* algorithm:

$$\sigma(\mathbf{X}, \mathbf{C}_1, \ldots, \mathbf{C}_m) = \sum_{j=1}^{m} tr(\mathbf{X} - \mathbf{G}_j\mathbf{C}_j)'(\mathbf{X} - \mathbf{G}_j\mathbf{C}_j) = \sum_{j=1}^{m} \left\|\mathbf{X} - \mathbf{G}_j\mathbf{C}_j\right\|^2 \qquad (2)$$

subject to the normalization constraints $\mathbf{X}'\mathbf{X}=n\mathbf{I}$ and $\mathbf{J}\mathbf{X}=\mathbf{X}$, where $\mathbf{J}$ is the projector on the subspace orthogonal to the unit vector.

In a classical, supervised, regression model we want to predict a quantitative response variable Y using m predictor variables. If the predictor variables are categorical, we can use the expression (1), with $p=1$, to quantify the variables obtaining the loss function:

$$\left\|\mathbf{y} - \sum_{j=1}^{m} \boldsymbol{\beta}_j \mathbf{G}_j \mathbf{c}_j\right\|^2 = min \qquad (3)$$

Assuming $\varphi(\mathbf{y}) = \mathbf{y}$, with the appropriate constraints, this loss function corresponds to the Morals model (Gifi 1981) and the solution can be identified through an alternating least square procedure, which is sometimes called backfitting.

Morals calculates a single quantification of each categorical variable. A way to extend (3) to consider p quantifications is:

$$\left\|\mathbf{y} - \sum_{r=1}^{p}\sum_{j=1}^{m}\boldsymbol{\beta}_{rj}\mathbf{G}_j\mathbf{c}_{rj}\right\|^2 = min \qquad (4)$$

Other encoding methods for ML have been proposed in the literature (see for example Potdar et al. 2017) mainly for the supervised case. The *Scikit-learn* software library allows to apply 15 different methods.

3 Categorical encoding

In this paper we propose a method to encode categorical variables that uses dense matrices of reduced size through an '*embedding*' of the categories in a low dimension space. A well-known form of embedding is *word-embedding* (Bengio et al. 2003). Embedding in NLP is a vector representation of the words in such a way that the words that frequently appear in similar contexts are close to each other.

With the most used software libraries for Neural Networks (e.g. Tensorflow, www.tensorflow.org) it is now available the '*Embedding layer*' for NLP. This layer transforms a sequence of words into their vectorial representation introducing an array of quantification parameters. This operation can be defined as in (1) in which the one-hot vectors identify the different words in a vocabulary and the vectors of parameters $\mathbf{c}_r$ are the corresponding vectorial representations in a p-dimensional space. The *Embedding layer* is the first layer in a predictive model and the parameters are determined by backpropagation, trying to maximize the fit of the model to the target.

This layer performs the same kind of transformation as the Optimal Scaling. The difference lies in the objective function and in the method of estimating the parameters given by the *gradient descent*. Let $t = \sum_{j=1}^{m} k_j$, using the approach of (4), it is possible to search for a vectorial representation of the t categories inside a p-dimensional space to optimize the prediction of a target Y. The embedding of categories was considered also by two previous works (Guo et al. 2016, Stefanini 2020) in which the authors proposed the '*entities embedding'*. This approach consists of a distinct embedding phase for each categorical variable where the encoding is defined as in (1), each feature embedding is optimized independently and could have a different dimension.

From a computational point of view, in our approach there is no need to create the inefficient indicator super matrix $G = (G_1| \ldots |G_m)$ with dimensions $n \times t$ and sparsity equal to $1-m/t$. It is sufficient to create the t-dimensional '*vocabulary*' of the categories and index it. Then all the categories in the data will be substituted by the corresponding numerical index. At this point we can introduce the array $\mathbf{C} = (C_1| \ldots |C_m)$ of parameters with dimension $t \times p$ containing the quantification of the categories in the *vocabulary*. The value of p can be small (≤ 10) also with high cardinality variables. It is an hyperparameter in the fitting of the model.

In an unsupervised approach, the categorical variables can be quantified by other NLP algorithms. The best-known method is *word2vec* (Mikolov et al. 2013) but also more recent proposals are available. Without a target variable, it is possible to estimate the array $\mathbf{C}$ by defining a Neural Network model that, for each unit, predict the category of a variable by knowing the categories of the other variables. The obtained

quantifications can be then used in a supervised model. This approach, which has allowed great progress in NLP, can manage big-data and categorical variables with high cardinality. One advantage of the unsupervised approach is that it does not necessarily require the analysis of many data, since the encodings can be obtained indirectly also on other data sets. Furthermore, the encodings do not use the target and therefore there is not risk of overfitting, but, in predictive models, these encodings are not optimized for the specific target that is analysed.

4 A comparison

Some of the most used encoding techniques have been compared with our proposal. Each of these methods has important drawbacks. Target encoder has a risk of '*target leakage*', in fact, it uses some information from target to predict the target itself, increasing the chance of overfitting on the training data. One-Hot Encoder, in the case of high cardinality leads to a large number of orthogonal features. Furthermore, this method alters the relationship between categorical and quantitative variables and can create memory problems. The Helmert contrast encoder can lead to overfitting and requires that the levels of the categorical variable are ordered and target is quantitative. In computational terms, we can remark that target-based methods are the fastest, while one-hot and categorical encoding take longer to calculate, depending on the cardinality of the categorical variables.

We considered the application to two different well-known datasets. The *Human Resources dataset* consists of 54,808 examples, 14 categorical and quantitative features and one binary target. The *Allstate Claims Severity* dataset is composed of 131 features, including 116 categorical variables and one continuous target observed on 188,318 examples.

In tab. 1 we have compared some encoders in a supervised approach. Considering a predictive aim, in each analysis we added to the encoder the same simple neural network with 4 internal layers. To regularize the network, dropout layers have been introduced.

Table 1: Comparison of encoding methods in two supervised analysis

Allstate Claim Severity data	MSE on Train (70%)	MSE on Test (30%)
One-Hot encoder	0.6584	0.6581
Helmert encoder	0.6583	0.6580
Target encoder	0.6582	0.6581
Categorical Encoder	**0.2688**	**0.3199**

Human Resources data	AUC on Train (70%)	AUC on Test (30%)
One-Hot encoder	0.9181	0.8991
Helmert encoder	0.9345	0.8929
Target encoder	0.8053	0.8070
Categorical Encoder	**0.9085**	**0.9033**

The usefulness of each encoder can therefore be assessed based on the predictive capacity of the model obtained. We can see that the categorical encoding (with p=10) is more effective in both cases, but mostly on the dataset with many categorical variables and high cardinality.

5 Conclusions

In Neural Networks, the One-Hot is the most common encoding technique for categorical data and, often, it leads to good results. On the other hand, this encoding has many drawbacks that can degrade the predictive performance of the Neural Networks model. Also, the alternative methods proposed in the literature have not proved particularly effective. The approach we have shown in this paper seems to bring clear advantages when there are many categorical variables with high cardinality. Using this approach, we can also build unsupervised models to analyse the relationship between categorical variables as in MCA. Some early applications in this regard show encouraging results.

References

1. Azzalini, A.: Inferenza statistica - Una presentazione basata sul concetto di verosimiglianza. Springer (2001)
2. Bengio Y., Ducharme R., Vincent P., Janvin C.: A neural probabilistic language model, The Journal of Machine Learning Research 3, 1137-1155 (2003).
3. Benzécri J.P.: L'Analyse Des Données: Tome II: L'Analyse Des Correspondances. Dunod, Paris (1973).
4. Di Ciaccio A.: Some considerations on the quantification of categorical data, Research-Report, Department of Data Theory, University of Leiden, (1988).
5. Di Ciaccio A.: Analisi simultanea dei caratteri qualitativi e quantitativi, Metron, vol.XLVIII, n. 1-4, (1990).
6. Gifi A.: Nonlinear Multivariate Analysis. Technical report, University of Leiden, Department of Data Theory, Leiden (1981).
7. Guo C., Berkhahn F.: Entity Embeddings of Categorical Variables, available: https://arxiv.org/abs/1604.06737 (2016)
8. Mikolov T., Chen K, Corrado G., Dean J.: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 (2013).
9. Pargent F.: A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling. Master Thesis in Statistics at LMU Munich (2019)
10. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014), doi: 10.3115/v1/D14-1162
11. Potdar, K., Pardawala, T.S., Pai, C.D.: A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications 175, 4 (2017).
12. Stefanini, E.: Deep Embedding per variabili categoriche. Master's thesis, dept. of Statistics, university of Rome, La Sapienza (2020).

Dynamic Quantile Regression Forest
Regression Forest Quantilica Dinamica

Mila Andreani[1] and Lea Petrella[2]

Abstract The potential of machine learning algorithms in the assessment of market risks has not been completely investigated in the literature, such as in the forecasting Value-at-Risk (VaR). In this paper we introduce the Dynamic Quantile Regression Forest, a model combining Quantile Regression Forests with a dynamic VaR. The model is dynamic as the quantile prediction of the previous random forest becomes part of the training set used to train the next random forest. Thus, it is possible to estimate the response variable conditional distribution by accounting for the evolution of the quantile over time among other covariates.

Abstract *Le potenzialità degli algoritmi di machine learning per la valutazione dei rischi di mercato sono ancora poco conosciute, in particolar modo per quel che concerne il calcolo del Value-at-Risk (VaR). Lo scopo di questo lavoro, dunque, è quello di introdurre la regression forest quantilica dinamica, un modello che unisce le regression forest con il calcolo dinamico del VaR, ossia tenendo conto dell'evoluzione del quantile nel tempo: in questo senso il modello è definito dinamico in quanto permette di stimare la distribuzione condizionata della variabile tenendo conto, fra le altre covariate, anche dell'evoluzione del quantile nel tempo.*

Key words: Value-at-Risk, Random Forest, Quantile Regression.

[1] Mila Andreani
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9, 000161, Rome, Italy, e-mail: mila.andreani@uniroma1.it

[2] Lea Petrella
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9, 000161, Rome, Italy, e-mail: lea.petrella@uniroma1.it

1 Introduction

Numerous contributions in literature point out the high predictive accuracy of machine learning algorithms in the economic and financial fields [1,8]. However, it is necessary to account for the complexity of such tools when it comes to compare them to traditional econometric models in terms of computational effort. As a matter of fact, many machine learning models require a long time to be tuned before being implemented, whereas traditional models might deliver similar results in a more timely manner.

Yet, the Random Forest algorithm introduced by [4], represents an efficient compromise between predictive accuracy and computational effort. Random Forest is an ensemble machine learning tool composed of decorrelated decision trees. Each tree is trained with a subset of the training set sampled with replacement from the original dataset with a method called *bagging*. The term "random" is related not only to the random sampling of the training subset, but also on the randomly chosen variable used to make the splits in each decision tree. The final prediction is computed as the aggregation of each tree prediction. The main advantage of random forests is related to the high training speed and the low number of parameters to be tuned in order to deliver reliable predictions (*number of trees, node size, number of features to choose from for each split*).

Random forest can be used both for classification and regression tasks, and the latter model has been further developed by [6], who introduced the quantile regression forest. This model allows to predict quantiles by estimating the whole conditional distribution of the variable, and, in this sense, it is possible to use this model to forecast the Value-at-Risk, representing the quantile of the conditional distribution. However, other econometric models have been developed in order to compute the Value-at-Risk in a dynamic setting.

In particular, [7] proposed a model to compute the Conditional Autoregressive Value-at-Risk (CAViaR) via nonlinear quantile regression. This model is based on directly modelling the quantile as nonlinear function of its past values and other factors instead of estimating the entire conditional distribution of the variable. By doing so, it is possible to account for changes in the variable's distribution over time and to compute a dynamic Value-at-Risk. Despite the advantages of random forest, the state-of-the-art in the machine learning field doesn't account for the application of quantile random forests in computing VaR. Moreover, the only technique by which is possible to train machine learning tools in a dynamic setting is via "incremental learning". This method allows to train a model on several batches of a given dataset, but it doesn't allow to account for an autoregressive structure of the model, which is the idea behind the CAViaR model. For this reason, in this paper a new kind of random forest is proposed in order to account for a dynamic setting for forecasting Value-at-Risk. Moreover, the aim of this paper is also to design a model to compete with traditional econometric techniques in terms of predictive accuracy and computational effort.

2 The model

The main feature of machine learning is that the training phase is performed on a static training set. This means that a given dataset is divided into two parts, a training set and a test set. Subsequently, the algorithm is trained on the training set and tested on the test set. Thus, it is not possible to account for changes in the training set and, for instance, to consider an autoregressive structure in the model. Consequently, in order to forecast a dynamic quantile, it is necessary to change the way in which machine learning algorithms are trained and to develop a model that allows to update the training set each time a new prediction is made.

The dynamic quantile regression forest has been developed for this purpose: this model relies on updating the training set whenever the regression forest produces a prediction and subsequently training a new random forest on the updated training set. Thus, it is possible to account for an autoregressive structure as proposed by [7]. However, the random forest models the entire distribution instead of directly modelling the quantile. Specifically, given a set of observations Y_i with $i = 1, \dots, T$, and a set of regressors $X_{k1} \dots, X_{kT}$ with $k = 1, \dots, K$, a subset $Y_1 \dots, Y_t$ (in which t represents the window width) is used in order to initialize the random forest and to produce t quantile predictions via the CAViaR model. To do that consider one of the possible CAViaR models proposed by [5] i.e. the Symmetric Absolute Value:

$$q_t(\beta_1) = \beta_1 + \beta_2 q_{t-1}(\beta) + \beta_3 |y_{t-1}| \qquad (1)$$

The first random forest is trained on the subset $Y_1 \dots, Y_t$ using the $q_1 \dots, q_t$ estimated from the CAViaR in (1) with covariates given by $X_{k1}, \dots, X_{kt}$ with $k = 1, \dots, K$. The one-step-ahead q_{t+1} is computed implementing the quantile regression forest and added to the dataset. Subsequently, an additional quantile regression forest is fitted using $Y_1 \dots, Y_{t+1}$, q_{t+1} and $X_{k1} \dots, X_{k(t+1)}$ with $k = 1, \dots, K$. The procedure is repeated until the entire dataset $Y_1, \dots, Y_T$ is used to train the last random forest, which is fitted on the whole dataset comprehensive of the previous quantile predictions. Consequently, the one-step-ahead quantile prediction will be a nonlinear function of its past values and the regressors.

3 Results

The model has been tested on T=1350 daily returns of the S&P 500 starting from 02-09-2010 and the VaR at 5% is predicted. The covariates are: Hang Seng index, CAC 40 index, FTSE 100 index, Dow Jones Index, NASDAQ 100 index, VIX, CBOE Equity VIX on Apple, CBOE Equity VIX on Amazon. The window width is equal to t=350 and the number of trees is set to 100. The first 350 observations are used to initialize the random forest using the CAViaR model in (1) and the remaining 1000

are used to estimate the quantile regression forest's predictions. The model proposed results to be:

-*accurate*: to measure the model accuracy in estimating the VaR, we use backtesting procedures like the conditional and unconditional coverage tests [2,3]. We show that our dynamic quantile regression forest produces both correct and independent exceedances, as shown in Table 1 (the expected exceedances are 50):

Table 1: Backtesting results

	Values
Actual exceedances	42
Unconditional coverage p-value	0.2304129
Conditional coverage p-value	0.1514733

-*competitive in terms of accuracy and computational effort:* the running time for our model is equal to 1 minute and 23 seconds and the prediction for the 1001-th quantile is equal to -0.023. Moreover, we compare the dynamic quantile regression forest with the four CAViaR models proposed in [7] (i.e., Symmetric absolute value, Asymmetric slope, Indirect GARCH (1,1), Adaptive) and the static quantile regression forest as proposed in [6]. The table below shows the quantile loss values for the models being compared and the ratio between our model's quantile loss, the CAViaR models' quantile loss and the static quantile regression forest quantile loss. Results show that our model produces a sensibly smaller quantile loss with respect to the Indirect GARCH (1,1) model and the static quantile regression forest, whereas it delivers similar results to the other three CAViaR models. Results are summarised in Table 2:

Table 2: Accuracy results

	% Loss	***Loss***
Dynamic quantile regression forest	-	0.0010227
Symmetric absolute value	108%	0.0009463
Asymmetric slope	107%	0.0009545
Indirect GARCH (1,1)	29%	0.0035005
Adaptive	108%	0.0009467
Static quantile regression forest	43%	0.0024011

4 Comparison with the static quantile regression forest

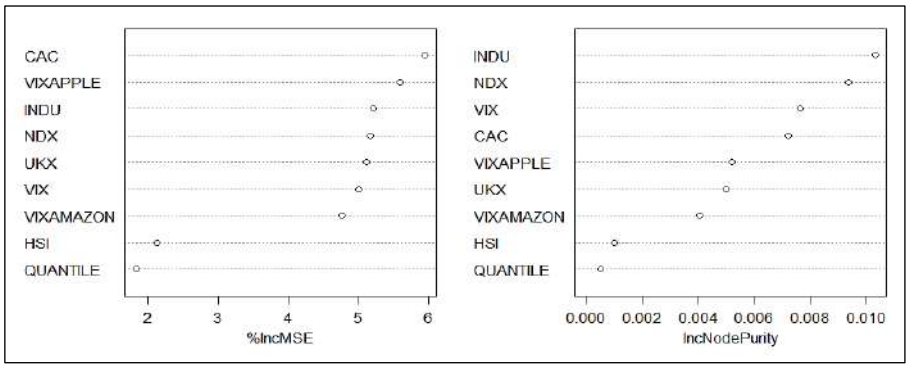
Moreover, according to the backtesting results the static quantile regression forest as proposed in [6] produces inaccurate estimates for the 5% VaR as shown in Table 3:

Table 3: Backtesting results for the static quantile regression forest

	Values
Actual exceedances	322
Unconditional coverage p-value	0
Conditional coverage p-value	0

As a matter of fact, Figure 1 highlights the relevance of introducing quantile predictions as regressors in the dynamic quantile regression forest. The figure shows that the quantile covariate positively affects the dynamic quantile regression forest's estimates in terms of Mean Squared Error and Node Impurity. As described in [5], the t-th node impurity measures the sum of squared errors between the observation value y_i and its mean in the t-th node $\mu(t)$ (i.e., the variance in the t-th node).

Figure 1: Dynamic Quantile Regression forest variable importance



References

1. A. E.Khandani, A. J.Kim, A. W.Lo , Consumer Credit-Risk Models Via Machine-Learning Algorithms , Journal of Banking and Finance 34 (2010): 2767-2787, 2010.
2. Christoffersen, P., Evaluating Interval Forecasts, International Economic Review, 39(4), 841-862, 1998.
3. Kupiec, P.H., 1995. Techniques for Verifying the Accuracy of Risk Management Models, The Journal of Derivatives 3(2), 73-84, 1995.
4. Leo Breinman, Random Forests, Machine Learning, 45, 5–32, 2001
5. M. Segal, Y. Xiao, Multivariate Random Forest, Data Mining and Knowledge Discovery, 2011.
6. Nicolai Meinshausen, Quantile Regression Forests, Journal of Machine Learning Research 7, 983–999, 2006.
7. R.F. Engle, S Manganelli, CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles, 2004.
8. S.I. Lee, S.J. Yoo, Multimodal deep learning for finance, The Journal of Supercomputing, 2019.

Estimating the UK Sentiment Using Twitter
Stima del sentiment nel Regno Unito attraverso Twitter

Stephan Schlosser, Daniele Toninelli and Michela Cameletti

Abstract In the era of social networks and big data Twitter represents a tremendous and cheap source of data able to provide valuable information about any possible topic. Such a source requires techniques to transform text into sensible numerical indexes. In this paper we consider the daily sentiment score measured by two lexicons (AFINN and Bing) on tweets collected for UK from January 15th to February 15th, 2018. As the analysed daily tweets are geolocated at the NUTS sub-area level, we focus on the comparison of the two score distributions across regions and serially. Results show that the two lexicons perform very similarly. However, our analysis shows that the sentiment estimated using the tweets doesn't correlate with the individual well-being estimated using data from an European survey (ESS).

Abstract *Nell'epoca dei social network e dei big data Twitter rappresenta una fonte di dati straordinaria ed a basso costo, capace di fornire preziose informazioni riguardo ad ogni possibile argomento. Questa fonte di dati richiede, però, metodi in grado di tradurre il testo in indicatori numerici. In questo lavoro si prendono in considerazione i punteggi stimati attraverso due dizionari (AFINN e Bing) su tweet geolocalizzati a livello di NUTS raccolti per il Regno Unito tra il 15 Gennaio ed il 15 Febbraio 2018. Confrontando le distribuzioni regionali dei punteggi e la loro evoluzione storica, si notano risultati molto simili per i due dizionari. Tuttavia non sembra esserci correlazione tra i livelli rilevati mediante sentiment analysis ed il livello di benessere emerso usando i dati dell'indagine ESS.*

Key words: Twitter, sentiment analysis, lexicon, spatial distribution, ESS

[1] Stephan Schlosser, Methodenzentrum Sozialwissenschaften, Georg-August-Universität Göttingen; email: stephan.schlosser@sowi.uni-goettingen.de.

Daniele Toninelli, Dipartimento di Scienze aziendali, economiche e metodi quantitativi, Università degli Studi di Bergamo; email: daniele.toninelli@unibg.it.

Michela Cameletti, Dipartimento di Scienze aziendali, economiche e metodi quantitativi, Università degli Studi di Bergamo; email: michela.cameletti@unibg.it.

1 Introduction and literature review

Social media are considered, at least potentially, new important data sources for research on various topics and for decision making in several fields (Kapoor *et al.*, 2018; Stieglitz *et al.*, 2018; Batrinca & Treleaven, 2015), despite currently affected by several issues of different types (Peng *et al.*, 2018; Bello-Orgaz *et al.*, 2015).

Currently, Twitter is one of the main big data sources: it is one of the most spread social network, worldwide, with about 330 million monthly active users (2019, first quarter; https://www.statista.com), 145 million daily users (http://shorturl.at/jnPS0) and 1.3 billion of accounts (http://shorturl.at/hszRW). Twitter users produce 500 million of tweets, daily (https://business.twitter.com/). These short posts (up to 140 characters long) are already used in several field, such as: political studies (Jungherr, 2015; Bose *et al.*, 2019), public health (Annice et al., 2013; Bates *et al*., 2014) and sentiment analysis (Mäntylä *et al*., 2018; Saura *et al*., 2019).

In a recent research, we experimented methods to retrieve tweets fully geolocated (Schlosser *et al*., 2019; Schlosser *et al*., *forth.* 2020). These methods allow us to retrieve all tweets sent in a wide geographical area (United Kingdom, excluding Northern Ireland, i.e. UK, from here on) and to assign them to NUTS sub-areas (the NUTS 1 level of UK; https://ec.europa.eu/eurostat/web/nuts/background). In particular, in this paper we analyse tweets collected in UK from January 15th, 2018 to February 15th of the same year, using a method defined as Method 2 (or M2; for details see Schlosser *et al*., 2019), that resulted as the best approach for tweet collection and geolocation among the three tested alternatives.

The main purpose of this paper is twofold. On the one hand we want to apply two different lexicons (AFINN and Bing) in order to estimate the level of sentiment observed on the collected tweets; in particular, we aim at understanding if the results obtained with the two lexicons are comparable, both considering the main statistics and the distribution of the output scores and the longitudinal perspective (i.e. analysing the relative changes in the sentiment time series). On the other hand we want to check whether the two lexicons are able to produce results comparable to the ones obtained using a more traditional data source, such as the European Social Survey (ESS; link: http://www.europeansocialsurvey.org/).

2 Data & Method

For this research we use tweets collected by means of a method based on the "theory of circle" (see M2 in Schlosser *et al*., 2019). By setting circles in order to cover in the most efficient way the area of a country (in our case UK), we are able to geolocate at the NUTS level all the collected tweets, allowing us to develop a spatial analysis of the sentiment observed in the different UK NUTS. In this paper we use 40,330,747 tweets collected in 2018 from January 15th to February 15th.

After a preliminary standard phase of cleaning (deleting stop words, hashtags, links, and so forth), we apply to all tweets two different lexicons: AFINN (Nielsen, 2011) and Bing (Bing, 2015). The current version of the first method (AFINN-en-165) is based on a dictionary including over 3,300 words; it ranks the words included in a tweet and belonging to the dictionary using a score between -5 and +5 and 0 neutral terms. The second lexicon, based on a dictionary including 6,788 words, assigns a score equal to -1 (for negative worlds), 0 (for neutral terms) or +1 (for positive worlds). Thus, we compute the sentiment of each tweet by summing the scores assigned by each of the two lexicons. This score estimates the level of sentiment of each tweet.

In addition to this, we also work on ESS data, analysing, in particular, the results of a question answered by a representative sample of 2,201 UK respondents that participated in the 9th wave of the survey (reference year: 2018). This question is no. B27: "*All things considered, how satisfied are you with your life as a whole nowadays?*". The evaluation scale ranges from 0 (="Extremely dissatisfied") to 10 (="Extremely satisfied"). Our scope is to understand whether the AFINN & Bing sentiment scores are able to reproduce the levels observed from the ESS data.

3 Main results

As **first step**, we analysed the sentiment score distribution by NUTS and by lexicon: this leaded us to identify four different regions, shown in Figure 1 together with the corresponding average Bing scores (whereas the average scores by NUTS are shown in Table 1). The AFINN lexicon produces very similar results (summary statistics reported in Table 1 are different, but the relative positions of the NUTS are very similar, as shown later), thus we do not show them here.

Figure 1: Regions based on the Bing score and corresponding NUTS (with unweighted avg. of Bing sc.).

Purple region (Bing avg: **0.215**) includes:
- UKM (Scotland)

Orange reg. (Bing avg: **0.243**) includes:
- UKG (West Midlands);
- UKF (East Midlands);
- UKE (Yorkshire and the Humber);
- UKD (North West);
- UKC (North East).

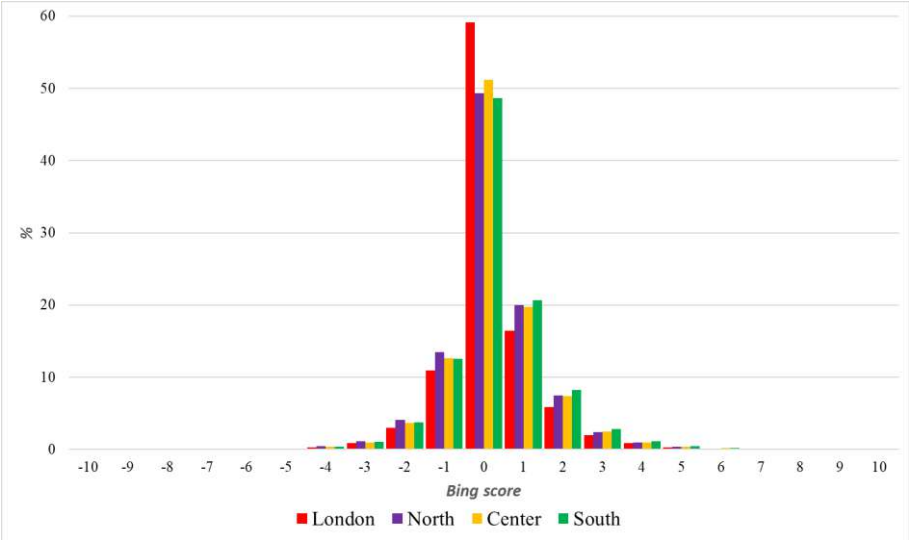
Green region (Bing avg: **0.284**) includes:
- UKL (Wales);
- UKK (South West);
- UKJ (South East);
- UKH (East of England).

Red region (Bing avg: **0.185**) includes:
- UKI (London).

The distribution of the Bing and AFINN scores enables us to identify four well-defined regions: the lowest average levels (Bing: 0.185; AFINN: 0.598) are observed for London, whereas the highest levels (Bing: 0.284; AFINN: 0.861) are observed in the Southern region (Wales, South West, South East and East of England).

The percentage distribution of the Bing score by region is shown in Graph 1. There are no big differences in the Bing score between the four regions. However, we notice a higher concentration of neutral tweets (score = 0) for the London area, whereas negative scores are more common in the North and positive scores are more likely to be observed in the South (as already highlighted in Figure 1). Similar patterns can be observed studying the AFINN scores.

Graph 1: Distribution of the Bing score by region (percentages).



As **second step**, we analyse if, using the two lexicons, we observe a similar behaviour of the sentiment over time. In particular, we were mainly interested in understanding if the two methods were able to similarly measure relative movements of the sentiment over time. Thus, we studied the relative changes observed between consecutive time points. Analysing this variable, we observed very high and significant correlations (all $p<0.001$) between AFINN and Bing scores daily relative changes for all the four zones; the Pearson correlation coefficient is equal to: 0.9530 for London, 0.9557 for North, 0.9816 for Center and 0.9830 for South. This means that the relative changes detected by using the two lexicons are very similar: the two lexicons seem to confirm each other's results.

Finally, as **third step** we compare, globally and by NUTS, the scores obtained applying the sentiment analysis and the scores from the ESS question no. B27. The summary statistics by NUTS are shown in Table 1, as long as the corresponding NUTS ranking.

Table 1: Sentiment scores (AFINN & Bing lexicons) and ESS average score by NUTS

	AFINN			*Bing*			*ESS*		
NUTS	***Mean***	***Std.dv.***	***Rank***	***Mean***	***Std.dv.***	***Rank***	***Mean***	***Std.dv.***	***Rank***
UKL	0.859	3.035	2	0.284	1.259	2	7.585	1.943	1
UKK	0.907	3.023	1	0.300	1.272	1	7.442	2.010	2
UKJ	0.822	2.970	5	0.271	1.256	4	7.330	2.031	7
UKH	0.856	3.031	3	0.281	1.272	3	7.427	1.864	4
UKG	0.764	2.966	7	0.253	1.220	6	6.804	2.210	11
UKF	0.702	2.893	10	0.226	1.193	9	7.430	1.963	3
UKE	0.829	3.067	4	0.263	1.264	5	7.374	1.938	6
UKD	0.711	2.926	8	0.227	1.197	8	7.222	2.208	9
UKC	0.771	3.128	6	0.246	1.256	7	7.259	2.370	8
UKM	0.710	3.028	9	0.215	1.243	10	7.380	2.276	5
UKI	0.598	2.676	11	0.185	1.114	11	7.072	2.291	10
ALL	**0.769**	**2.978**		**0.247**	**1.231**		**7.313**	**2.095**	

AFINN and Bing lexicons provide with very similar results, showing a Pearson correlation coefficient between scores equal to 0.9895 (p<0.001). The Spearman correlation coefficient computed on the ranks confirms these results (0.9727; p<0.001). Nevertheless, using ESS data we do not obtain a significant correlation both taking into account the average scores and the ranks. The Pearson correlation coefficient is 0.493 (p=0.123) vs the AFINN score and 0.433 (p=0.183) vs the Bing score. The Spearman correlation coefficients computed on ranks highlight nonsignificant linkages as well: 0.555 (p=0.077) vs the AFINN rank and 0.536 (p=0.089) vs the Bing rank. It seems that the scores obtained by analysing tweets are not able to reproduce the levels obtained using the ESS data. But this was partially expected, given that the ESS data are referred to the whole year, whereas the tweets are linked to just a specific month.

4 Discussion, limitations and further research

Our main results highlight similar performances of the AFINN and of the Bing lexicons, considering the score distribution, the relative changes of the sentiment time series and the NUTS ranking as well. This seems to confirm that our method of collection is robust with respect to the lexicon, leading to significantly non-different results. Nevertheless, if we compare the sentiment level with the scores observed from the ESS data, it seems that we are not able to reproduce the same relative levels in the geographical distribution. This is, of course, a consequence of our analysis: first, by definition, sentiment does not correspond to the subjective well-being that the ESS question we consider aims at measuring. Thus, this preliminary analysis should be refined, for example focusing on just one or few of the dominions that characterize the subjective well-being. Moreover, the comparison relies on two

different reference periods. Studying tweets, we are able to produce daily time series, but we just consider one month, whereas the ESS data are referred to the overall year. The relative levels of sentiment observed in a specific month could be completely different from the levels referred to the full year.

In order to try to overtake such limitations, we suggest, as further research ideas, to include a pre-screening or a filtering of tweets by scope (e.g. advertising vs private posts) and/or by specific topics (work, health, safety, and so forth). In order to fix the issue of the different time periods, we also propose to develop a more general analysis relying on annual average levels of the sentiment that should guarantee a higher level of comparability.

References

1. Annice, E.K., Hansen, H.M., Murphy, J., Richards, A.K., Duke, J., Allen, J.A.: Methodological Considerations in Analyzing Twitter Data. JNCI Monographs 2013(47), 140—146 (2013). https://doi.org/10.1093/jncimonographs/lgt026
2. Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A., Escobar, G.: Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. 33(7), 1123--1131 (2014). https://doi.org/10.1377/hlthaff.2014.0041.
3. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. AI & Soc 30, 89--116 (2015). https://doi.org/10.1007/s00146-014-0549-4
4. Bello-Orgaz, G., Jung, J., Camacho, D.: Social Big Data: Recent achievements and new challenges. Information Fusion 28, 45--59 (2015). https://doi.org/10.1016/j.inffus.2015.08.005
5. Bing, L.: Sentiment analysis and opinion mining. Cambridge University Press, New York (2015)
6. Bose, R., Dey, R.K., Roy, S., Sarddar, D.: Analyzing Political Sentiment Using Twitter Data. In: Satapathy, S., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies 107. Springer, Singapore (2019)
7. Jungherr, A.: Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research. Springer (2015) https://doi.org/10.1007/978-3-319-20319-5.
8. Kapoor, K.K., Tamilmani, K., Rana, N.P., Patil, P., Dwivedi, Y.K., Nerur, S.: Advances in Social Media Research: Past, Present and Future. Inf Syst Front 20, 531--558 (2018). https://doi.org/10.1007/s10796-017-9810-y
9. Mäntylä, M.V., Graziotin, D., Kuutila, M.: The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. Computer Science Review 27, 16—32 (2018). https://doi.org/10.1016/j.cosrev.2017.10.002
10. Nielsen, F. Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, Keraklion, Crete, Greece, 93-98 (2011).
11. Peng, S., Yu, S., Mueller, P. Social networking big data: Opportunities, solutions, and challenges. Future Generat. Comp. Syst. 86, 1456--1458 (2018). https://doi.org/10.1016/j.future.2018.05.040
12. Saura, J.R., Palos-Sanchez, P., Grilo, A.: Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining. Sustainability 11(917), 1--14 (2019). https://doi.org/10.3390/su11030917
13. Schlosser, S., Toninelli, D., Cameletti, M.: Comparing Methods to Collect and Geolocate Tweets (*forthcoming*, 2020)
14. Schlosser, S., Toninelli, D., Fabris, S.: Looking for Efficient Methods to Collect and Geolocalise Tweets. In Arbia, G., Peluso, S., Pini, A., Rivellini, G. (eds.) Smart Statistics for Smart Applications, pp. 1057-1062. Pearson (2019). http://shorturl.at/ijxKW
15. Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C.: Social media analytics - Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management 39, 156—168 (2018). https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Forecasting local rice prices from crowdsourced data in Nigeria

Previsione sui prezzi locali del riso dai dati di crowdsourcing in Nigeria

Ilaria Lucrezia Amerise and Gloria Solano Hermosilla

Abstract The ability to anticipate food prices is important in developing economies for decision-making by governments, agri-food chain actors and other institutions. Farmers and agri-businesses decide based on price expectations, while well-being of households is affected by food price changes, since food makes up a large portion of household budget in developing regions. This study assesses the use of crowdsourcing in price forecasting. The Reg-ARMA approach is used to build simultaneous prediction intervals for daily retail prices of local rice in Nigeria submitted by citizens using a mobile app. Results show that the proposed approach generates effective and efficient point and interval forecasts, which can be of help to policy makers, managers, marketers, planners and researchers.

Abstract Nelle economie in via di sviluppo, la capacitá di previsione dei prezzi dei prodotti alimentari é importante nel processo decisionale dei governi, degli attori della catena agroalimentare e di altre istituzioni. Questo studio valuta l'uso del crowdsourcing nelle previsioni dei prezzi. L'approccio Reg-ARMA viene utilizzato per costruire intervalli di previsione dei prezzi al dettaglio giornalieri del riso locale in Nigeria, inviati dai partecipanti attraverso un'app mobile. I risultati mostrano che questo approccio genera previsioni puntuali e intervallari in modo efficace ed efficiente, che possono essere di aiuto per i politici, i manager, i responsabili marketing e i ricercatori.

Key words: Simultaneous prediction intervals, Regression with time series errors, food price forecasting, mobile app crowdsourcing, citizen volunteers

---

Ilaria Lucrezia Amerise
Department of Economics, Statistics and Finance, University of Calabria, via P. Bucci, 87036, Rende (CS). e-mail: ilaria.amerise@unical.it

Gloria Solano Hermosilla
European Commission Joint Research Centre (JRC), Directorate D - Sustainable Resources, Economics of Agriculture, Edificio Expo Inca Garcilaso, 3, E-41092 Seville, SPAIN e-mail: Gloria.SOLANO-HERMOSILLA@ec.europa.eu

1 Introduction

The ability to anticipate food price changes is important in developing regions for governments and other organizations to plan and adjust food security and market interventions, but also for food supply chains actors. Farmers and agri-businesses decide based on price expectations, while well-being of households in developing countries can be greatly affected by food prices changes, since food makes up a significant portion of household expenditure. But many developing countries lack the ability to track food prices with sufficient frequency and granularity to anticipate changes [5]. Today, alternative data sources and new technologies around the world offer big potential to complement official statistics [2] and have given rise in Africa to several ICT and/or citizen science projects to track geo-located food prices in real-time. Prior research has explored the use of web scraping (e.g. twitter) and google trend indices to predict food prices [5] yet more limited has been the attention to crowdsourcing. The objective of this paper is to evaluate whether the use of daily crowdsourced market prices can help to produce accurate forecasts. We use 16-months daily local rice prices contributed by citizen volunteers on the Food Price Crowdsourcing Africa (FPCA) platform[7] in Kano and Katsina states (Nigeria) to estimate Reg-ARMA processes. We forecast rural and urban daily observations 4/6 months ahead (120-180 days). Then, we compute the monthly average, which can be compared with the National Bureau of Statistics (NBS) data. Nigeria is the main rice producer in Africa and a major consumer [8]. Its government is trying to encourage rice production to promote self-sufficiency and food security and since 2015 banns imports. Katsina state, however, is still under food security stress (IPC v3.0). Small-scale farmers that account for 80% of rice production could benefit from improved price predictions when negotiating and selling their rice. In the rest, Section 2 describes the FPCA; Section 3 introduces the time-series forecasting methods. Section 4 includes the empirical application and Section 5 presents the discussion and conclusions.

2 Data

We use data collected in 2018-19 by the European Commission Joint Research Centre FPCA platform [7]. The crowd source volunteers using a mobile app in Kano and Katsina states (urban and rural) in northern Nigeria submitted daily market prices of several commodities for a reward. Volunteers owned a smartphone with GPS and followed web instructions. Radio spots and flyers served to publicize the FPCA. The system was built based on Open Data Kit and deployed on a compatible server. Automatic routines were in place to process, quality check and validate data, which fed into a web-dashboard updated twice a day. The 737 volunteers submitted 40,000 valid local rice retail prices that translated into more than 400 urban/rural daily observations. The consistent data flow reveals nuances of price data, prior to and after harvest, which hints to reliability of the crowdsourcing system. Main price declines

occurred during harvests, yet around Oct-19, prices were unusually high compared to previous year.

3 Methods

3.1 Assessing point forecast accuracy

The main problem with assessing the reliability of forecasts is that the magnitude of forecast errors cannot be evaluated until the actual values have been observed. To simulate such a situation, we split local rice time series into two parts: the "training" period, which ignores a number of the most recent time points and the "validation" period, which comprises only the ignored time points. Both the time series have undergone a pre-stage consisting: 1) imputation of some missing values by using na.interpolate included in the R package forecast by Hyndman [4]; 2) Rescaling a few outliers by using the method discussed in [1]. The prices observed in the period from 10/11/2019 to 07/12/2019 ($H=28$ days) are set-aside to serve as a benchmark for forecasting purposes. Local rice prices from 21/09/2018 to 09/11/2019 act as training period. There are a number of indices that assess predictive accuracy. For our current purpose, we prefer an index that varies in a fixed interval and makes good use of the observed residuals. This is the relative absolute error of forecast (RAEF)

$$RAEF = 100\left[1 - H^{-1}\sum_{h=1}^{H}\frac{|A_{n+h} - \widehat{A}_{n+h}|}{|A_{n+h}| + |\widehat{A}_{n+h}| + \varepsilon}\right] \quad h = 1,2,\cdots,H \quad t = 1,2,\cdots,n \tag{1}$$

where ε is a small positive number (e.g. $\varepsilon = 0.00001$) which constitutes a safeguard against division by zero. Coefficient (1) is independent on the scale of the data and, due to the triangle inequality, ranges from zero to 100. The maximum is achieved in the case of perfect forecasts: $L_{n+h} = \widehat{L}_{n+h}$ for each h. The lower the $RAEF$ is, the less accurate the model is. The minimum stands for situations of inadequate forecasting such as $\widehat{H}_{n+h} = 0$ or $\widehat{L}_{n+h} = -L_{n+h}$ for all h.

3.2 Simultaneous prediction intervals

In predicting local rice prices for marketing, planning, development and policy-making it is not only important to generate point forecasts for several steps ahead, but providing an assessment of the uncertainty associated with forecasts can be equally important. The problem is thus to integrate point forecasts with prediction intervals (PIs) which apply simultaneously to all possible future values of the predictors.

A reasonable strategy can be as follows: Given the availability of H future values, we can construct two bands such that, under the condition of independent Gaussian distributed random residuals, the probability that consecutive future prices $A_{n+h}, h = 1,2,\cdots,H$ lie simultaneously within their respective range is at least is γ

$$P\left[\bigcap_{h=1}^{H}\left(A_{1,h,\gamma} \leq A_{n+h} \leq A_{2,h,\gamma}\right)\right] \geq \gamma \tag{2}$$

where

$$\begin{cases} A_{1,h,\gamma} = \widehat{A}_{n,h} - \theta_{H,\nu,\gamma}\widehat{\sigma}_h \\ A_{2,h,\gamma} = \widehat{A}_{n+h} + \theta_{H,\nu,\gamma}\widehat{\sigma}_h. \end{cases} \tag{3}$$

The multiplier $\theta_{H,\nu,\gamma}$ is the γ-th quantile of the of the maximum absolute value $|t|$ of the H-variate Student t probability density function with ν degrees of freedom. See [3]. In short, $\theta_{H,\nu,\gamma}$ is the solution of

$$\int_{-\theta}^{\theta}\int_{-\theta}^{\theta}\cdots\int_{-\theta}^{\theta} f(t_1,t_2,\cdots,t_h;\nu)\,dt_1 dt_2\cdots dt_H = \gamma \tag{4}$$

The critical values can be found solving iteratively $u_{H,\nu,\gamma}$ using the command *pmvt* of the *R* package *mvtnorm*, which provides the multivariate t probability. See [6].

The most important characteristic of PIs is their actual coverage probability (PIAC). We measure PIAC by the proportion of true prices of the validation period enclosed in the bounds

$$PIAC_\gamma = 100H^{-1}\sum_{h=1}^{H} c_{h,\gamma} \quad \text{where } c_{h,\gamma} = \begin{cases} 1 & \text{if } A_{n+h} \in \left[A_{1,h,\gamma}, \quad A_{2,h,\gamma}\right] \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

If $PIAC_\gamma \geq \gamma$ then rural and urban prices tend to be covered by the constructed bounds, but this may also imply that the estimates of the variances in the forecast errors are positively biased. A $PIAC_\gamma < \gamma$ indicates under-dispersed forecast errors with overly narrow prediction intervals and unsatisfactory coverage behavior.

All other things being equal, narrow PIs are desirable as they reduce the uncertainty associated with forecasts. However, high accuracy can be easily obtained by widening PIs. A complementary measure that quantifies the sharpness of PIs might be useful in this context. Here, we use the score function.

$$R_{h,\gamma} = \left(\frac{\gamma}{2}\right)\frac{\left(C^2_{h,\gamma} - C^1_{h,\gamma}\right)}{A_{n+h}}, \quad h = 1,2,\cdots,H. \tag{6}$$

This expression reflects a penalty proportional to the narrowness of the intervals that encompass the true values at the nominal rate. The penalty increases as γ decreases, to compensate for the tendency of prediction bands to be broader as the confidence level increases. Of course, the lower $R_{h,\gamma}$ is, the more accurate PI will be. The average value of the score width across time points

$$ASW_\gamma = \frac{1}{H}\sum_{h=1}^{H} R_{h,\gamma} \tag{7}$$

can provide general indications of PIs performance.

3.3 Empirical analysis

Table 1 shows the results obtained with the Reg-ARMA method applied to the data of the training period for the prices in the validation period. The Reg-ARMA approach offers several improvements over OLS. The auto-correlation is almost inexistent because the p-value of the LB statistics is now larger that 99%. Also, the quality of the fitting is increased as proven by a lower AICc and a higher $\bar{R}^2$ than those observed for OLS.

Table 1 OLS and Reg-ARMA estimation and forecasting.

	β_0	t	$\gamma_{1,1}$	$\gamma_{2,1}$	$\gamma_{1,2}$	$\gamma_{2,3}$	$\gamma_{1,5}$	RAEF	PIAC	ASW
A										
Reg-A	4491.4	4785.9	-989.0	-1745.4	878.8	619.8	257.0			
$\bar{R}^2$:	0.979	AICc:	2431.9	LB:	1.3			84.5%	88.1%	11.6
B										
Reg-A	3797.9	8488.9	-1202.1	-2135.2	284.7	-287.1				
$\bar{R}^2$:	0.981	AICc:	2458.8	LB:	6.7			77.7%	65.0%	14.1

In both cases the amelioration is quite substantial. In addition, as can be readily seen from Figure 1, the coverage rate is now significantly higher compared to OLS. The cost of these enhancements is a larger width of the simultaneous forecast intervals, which, as it is well known determine broader brackets than marginal (and wrong) classical OLS intervals. Furthermore, the stability of the RAEF index across the two estimation methods, is a demonstration that the predictive accuracy does not deteriorate when Reg-ARMA method is used. Nonetheless, we must remark that, both OLS and Reg-ARMA methods, yield prediction intervals whose actual coverage rate resulted to be less than the nominal level (90%) (the latter are better than the former). We explain this result as due to a change in the evolutionary trajectory of prices, which in the last year, has not been following the trends of previous years. This is an obvious consequence of the need of bracketing future values within the same scheme used for past observations. Thus a constraint is imposed on the forecasting tool: strong local fluctuations or outliers cannot appear in the set of future values even if we know that they are there; thus, some failures are inevitable.

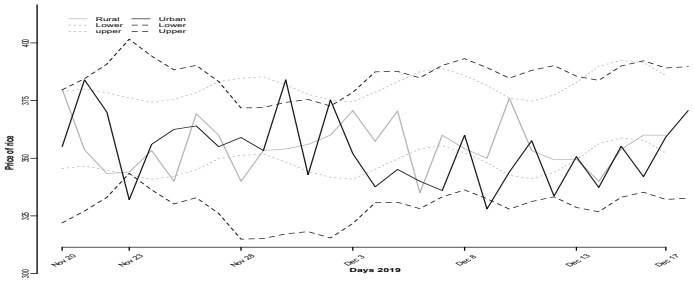


Fig. 1 Prediction intervals (PIs) of daily urban and rural prices of local rice in Kano and Katsina states (Nigeria)

4 Discussion and conclusion

Our method is an example of the use of crowdsourcing for accurate price predictions that could help empowering smallholder farmers and low-income consumers in developing countries, but also governments and other organisations. In this paper, we have assumed that the regression residuals arise from an auto-regressive moving average stochastic model. This way, we have not only eliminated serial correlation from the regression residuals, but also constructed valid simultaneous prediction intervals to contain future local rice prices. Further research could exploit the combination of crowdsourced prices with other determinants of staple food prices. Further, we believe that including this functionality in price monitoring platforms could help to enhance their possibilities to support better decision-making.

References

1. [1] Amerise, I.L., Tarsitano, A.: An L1 smoother for outlier cleaning of time series. *Journal of Statistical and Econometric Methods*, **9**, 47 – 66, (2020).
2. [2] DGINS: Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics), (2018).
3. [3] Hahn, G. J. : Simultaneous prediction intervals for a regression model. *Technometrics* **14**, 203 – 214, (1972).
4. [4] Hyndman R., Athanasopoulos G., Bergmeir C., Caceres G., Chhay L., O'Hara-Wild M., Petropoulos F., Razbash S., Wang E., Yasmeen F. : forecast: Forecasting functions for time series and linear models. *R package version 8.12*, (2020).
5. [5] Kim, J., Cha, M., Lee, G. : Nowcasting commodity prices using social media. *PeerJ Comput. Sci* **14**, 126, (2017).
6. [6] Genz, A., *et al.*: *M*ultivariate Normal and t Distributions. R package version1.0-11, https://CRAN.R-project.org/package=mvtnorm, (2019).
7. [7] Solano-Hermosilla, G., Adewopo, J., Micale, F., Gorrn Gonzlez, C., Arbia, G. and Nardelli, V. : Food Price Crowdsourcing Agrica on DataM. *European Commission, Joint Research Centre (JRC)* , https://datam.jrc.ec.europa.eu/datam/public/pages/index.xhtml, (2019).
8. [8] USDA Foreing Agricultural Service, *GAIN Report Nigeria Grain and Feed Annual 2019*, (2019).

Generalized Mixed Effects Random Forest: does Machine Learning help in predicting university student dropout?

Random forest generalizzati a effetti misti: il machine learning aiuta a prevedere l'abbandono degli studenti universitari?

Massimo Pellagatti, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Abstract We develop a new statistical method, called Generalized Mixed Effects Forest (GMEF), which embeds a Random Forest (RF) regression algorithm in a Generalized Mixed Model. Our model is new to the literature and can deal with any type (continuous or discrete) of both response variable and covariates and it does not assume any parametric function on the fixed-effects part of the model. GMEF is able to model hierarchical data, which is very important in the case of a nested structure of the observation, which might affect not only the final response itself, but also the effect that other covariates have on it. We apply this method to data from Politecnico di Milano to predict university student dropout, aiming to open a new perspective to university tutoring systems.

Abstract *In questo lavoro, sviluppiamo un nuovo modello statistico, chiamato Random Forest a effetti misti generalizzato (GMEF), che innesta una random forest di classificazione in un modello di regressione generalizzata a effetti misti. Il modello GMEF è nuovo in letteratura, può trattare variabili risposta e covariate sia continue che discrete e non assume nessuna funzione parametrica sulla parte degli effetti fissi del modello. Inoltre, GMEF si adatta a dati annidati, tenendo in considerazione la loro struttura gerarchica, che può influenzare non solo la risposta, ma anche l'effetto delle covariate del modello. Nel caso studio, applichiamo il modello GMEF ai dati del Politecnico di Milano, per prevedere l'abbandono degli sudenti e aprire nuove prospettive per il sistema di tutoraggio dell'università.*

Massimo Pellagatti
MOX - Politecnico di Milano, via Bonardi9, 20133 Milano
e-mail: massimo.pellagatti@mail.polimi.it

Chiara Masci
MOX - Politecnico di Milano, via Bonardi 9, 20133 Milano e-mail: chiara.masci@polimi.it

Francesca Ieva
MOX - Politecnico di Milano, via Bonardi 9, 20133 Milano e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
MOX - Politecnico di Milano, via Bonardi 9, 20133 Milano e-mail: anna.paganoni@polimi.it

Key words: Random forest, Mixed-effects models, University student dropout.

1 Introduction and motivation

The Student Profile for Enhancing Tutoring Engineering (SPEET) [9] project is an international ERASMUS$^+$ project that aims at determining and categorizing the different profiles for Engineering students across Europe and at opening new perspective to university tutoring systems. Politecnico di Milano is one of the six partners of the project and our aim is to extract useful insights from educational data collected by Politecnico di Milano (PoliMI) and use them to predict student dropout. This data are collected in big databases that, for each student, provide many information regarding their collateral information, their previous studies and their complete career at PoliMI. Our aim is to predict student dropout as soon as possible, by means of this information. Students at PoliMI are also naturally nested within different engineering degree programs in which the student dropout phenomenon might be different and might have different drivers. While investigating the learning process, it is necessary to disentangle the effects given by each level of data hierarchy.

In the light of this aspects, multilevel models take into account the hierarchical nature of data and are able to quantify the portion of variability in the response variable that is attributable to each level of grouping [3]. In this perspective, Generalized Linear Mixed Models (GLMM) fit a multilevel model on a binary response variable, imposing a linear effect of covariates on a transformation of the response variable [1]. The goal of our work is to exploit the advantage of generalized mixed-effects models to handle nested data, but developing a method that does not impose a parametric functional form in the relationship between the response and the covariates, since the educational context is a complex process where many factors interact among each others. Tree-based methods such as the CART model and its ensemble methods, i.e. random forest, learn the relationship between the response and the predictors by identifying dominant patterns in the training data [6, 7]. In addition, these methods allow a clear graphical representation of the results that is easy to communicate. We propose a novel method able to preserve the flexibility of classification RF model and to extend it to a clustered data structure, where multiple observations can be viewed as being sampled within groups: the generalized mixed-effects random forest (GMEF). In the literature this is not the first time in which tree-based methods are adopted to deal with longitudinal and clustered data. In [4, 8] regression tree methods for longitudinal or clustered data are proposed, but they both deal with a Gaussian response variable and they are not suitable to a classification problem. In [5] the approach presented in [4] is extended to non-gaussian data and a generalized mixed effects regression tree (GMERT) is proposed, but embedding a simple classification tree instead of a more robust method as a RF and a similar work is proposed in [2]. Indeed, this is the first time in the literature in which we extend a classification tree-based ensemble method to the case of a nested data structure.

2 Methodology

Let considering a generic GLMM. This model is an extension of a generalized linear model that includes both fixed and random effects in the linear predictor [1]. Therefore, GLMMs handle a wide range of response distributions and a wide range of scenarios where observations are grouped rather than being completely independent. For a GLMM with a two-level hierarchy, each observation j, for $j = 1,\dots,n_i$, is nested within a group i, for $i = 1,\dots,I$. Let $y_i = (y_{1i},\dots,y_{n_i i})$ be the n_i-dimensional binary[1] response vector for observations in the $i-$th group. The GLMM model is the following:

$$\begin{aligned}
&\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{b}_i] \qquad\qquad j = 1,\dots,n_i, \quad i = 1,\dots,I \\
&g(\mu_{ij}) = \eta_{ij} \\
&\eta_{ij} = \sum_{p=1}^{P+1} \beta_p x_{ijp} + \sum_{q=1}^{Q+1} b_{iq} z_{ijq} \\
&\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \Psi)
\end{aligned} \tag{1}$$

where $\mu_{ij} = p_{ij}$, $g(\mu_{ij})$ is the logit link function, i.e. $g(\mu_{ij}) = logit(\mu_{ij}) = logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$; β is the $(P+1)-$dimensional vector of coefficients and $\mathbf{x}_{ij}$ is the $(P+1)-$dimensional vector of fixed-effects covariates (including 1 for the intercept); $\mathbf{b}_i$ is the $(Q+1)-$dimensional vector of coefficients relative to the $i-$th group and $\mathbf{z}_{ij}$ is the $(Q+1)-$dimensional vector of random-effects covariates (including 1 for the intercept).

In our GMEF model, the fixed-effects part is not linear as in model (1) but it is replaced by the function $f(X_i)$ that is estimated through a RF algorithm. Thus, the model formulation is the following:

$$\begin{aligned}
&\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{b}_i] \qquad\qquad j = 1,\dots,n_i, \quad i = 1,\dots,I \\
&g(\mu_{ij}) = \eta_{ij} \\
&\eta_{ij} = f(\mathbf{x}_{ij}) + \sum_{q=1}^{Q+1} b_{iq} z_{ijq} \\
&\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \Psi).
\end{aligned} \tag{2}$$

The algorithm to implement GMEF is inspired by the GMET algorithm [2], but the main difference is that we use a random forest instead of a simple classification tree.

[1] We consider here the case of a binary response, but the model can handle any response variable in the exponential family.

3 Case study: GMEF for student dropout prediction

The PoliMi dataset within the SPEET project consists of 24,536 concluded careers[2] in Bachelor of Science (BSc) that began between A.Y. 2010/2011 and 2015/2016. Students are nested within I = 19 degree programs. A descriptive analysis shows that a high percentage - about 30% - of students leaves the PoliMI before obtaining the degree. Therefore, our goal is to find out which student-level indicators could discriminate between dropout and graduated students. We consider that student j is nested within degree program i. The model takes the form:

$$\begin{aligned} &\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{b}_i] \qquad\qquad j = 1,\dots,n_i, \quad i = 1,\dots,19 \\ &g(\mu_{ij}) = \eta_{ij} \\ &\eta_{ij} = f(\mathbf{x}_{ij}) + b_i \\ &\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2_\Psi). \end{aligned} \tag{3}$$

where the response variable Y is the career status, a two-level factor we code as a binary variable: status = 1 for careers definitely completed with graduation and status = 0 for careers definitely concluded with a dropout; the random-effects structure simplifies to a random intercept, b_i, that represents the value-added, either positive or negative, of the $i-$th degree program. The set of fixed-effects covariates is shown in Table 1. We select only the career information at the first semester of the first year because our aim is to obtain a good prediction as soon as possible[3].

3.1 Results

We consider 70% of the sample as training set and 30% as test set. We apply GMERF model, setting 0.02 as convergence tolerance for the random-effects and 500 as number of trees for the RF (selected by cross-validation). The algorithm converges in 8 iterations. Figure 1 reports the importance rankings of the fixed-effects covariates extracted from the RF, the partial dependence plots of the three most important ones and the 19 estimated random intercepts with their confidence intervals. We observe that the information about student career at the university results to be the most importance predictor. In particular, looking at the partial dependence plot, the higher is the `WeightedAvgEval1.1`[4], the lower is the dropout probabil-

[2] We exclude active careers from the sample, i.e. we consider only students who graduated or dropped their studies (at any time during the follow-up time).

[3] A comparison analysis shows that the gain in prediction accuracy that we obtain if using the whole first year instead of the first semester information is too low to justify the waiting of the prediction of at-risk students until the end of the first year.

[4] We let the mean of the exams score start from 0 because we take into consideration students who did not pass any exam, whose mean is therefore 0.

Variable	Description	Type of variable
`Sex`	gender	factor (2 levels: M, F)
`Nationality`	nationality	factor (Italian, foreigner)
`PreviousStudies`	high school studies	factor (*Liceo Scientifico*, *Istituto Tecnico*, Other)
`WeightedAvgEval1.1`	weighted average of the evaluations during the first semester of the first year	real number
`AvgAttempts1.1`	average number of attempts to be evaluated on subjects during the first semester of the first year (passed and failed exams)	real number
`TotalCredits1.1`	number of ECTS credits obtained by the student during the first semester of the first year	natural number

Table 1 List and explanation of the student-level covariates included in the GMEF model.

ity; the higher is the `AvgAttempts1.1`, the lower is the dropout probability; the higher is the `TotalCredits1.1`, the lower is the dropout probability. Regarding the random intercepts, degree programs with estimated $\hat{b}_i$ whose confidence interval is totally positive (or negative) have on average students more (or less) likely to dropout, all else equal. The GMEF model applied to PoliMI data has, on the test set, Accuracy index equal to 0.9101 and a Sensitivity index equal to 0.8133.

4 Conclusions

We propose a multilevel tree-based ensemble model for a non-gaussian response (GMEF algorithm) and we apply it to PoliMI data to predict student dropout at the university. The GMEF model is new to the literature and can be applied to different classification problems relative to data with a hierarchical structure. In the case study, it provides easily interpretable results and has very good predictive power.

References

1. A. Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.
2. L. Fontana, C. Masci, F. Ieva, and A. M. Paganoni. *Performing Learning Analytics via Generalized Mixed-Effects Trees*. MOX-report n. 43/2018, 2018.
3. H. Goldstein. *Multilevel statistical models*, volume 922. John Wiley & Sons, 2011.
4. A. Hajjem, F. Bellavance, and D. Larocque. Mixed effects regression trees for clustered data. *Statistics & probability letters*, 81(4):451–459, 2011.

5. A. Hajjem, D. Larocque, and F. Bellavance. Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126:114–118, 2017.
6. T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
7. T. Hastie, R. Tibshirani, and J. Friedman. Random forests. In *The elements of statistical learning*, pages 587–604. Springer, 2009.
8. R. J. Sela and J. S. Simonoff. Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86(2):169–207, 2012.
9. R. Vilanova, J. Vicario, M. Prada, M. Barbu, M. Dominguez, M. J. Pereira, M. Popdora, U. Spagnolini, P. Alves, and A. Paganoni. Speet: software tools for academic data analysis. In *EDULEARN 2018-International Conference on Education and New Learning Technologies*, 2018.

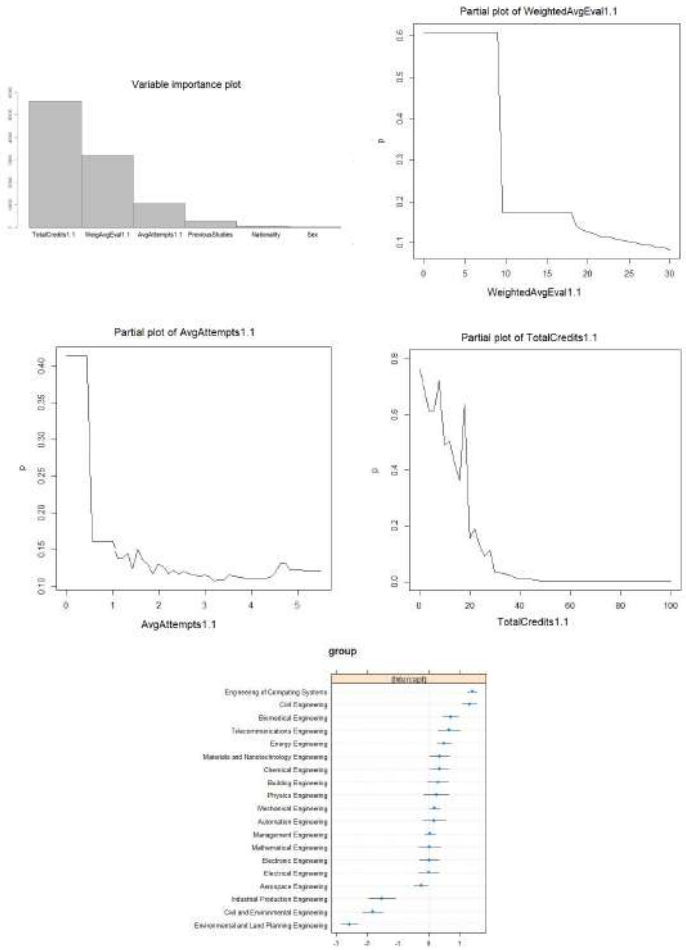


Fig. 1 Variable importance plot, partial dependence plots of the three most important variables selected by GMEF and the 19 estimated random intercepts with their confidence intervals.

HateViz: a textual dashboard Twitter data-driven

HateViz: una Dashboard Twitter data-driven

Emma Zavarrone[1], Maria Gabriella Grassia[2], Marina Marino[3], Rocco Mazza[4], Nicola Canestrari[5]

Abstract The paper introduces *HateViz* dashboard, an interactive platform Twitter based about hate speech against women. Starting from texts collected from Twitter, using R packages, the aim of dashboard is to provide to the community of decision-makers an easy-to-use tool to monitor this important phenomenon. The dashboard mixes three methods: textual mining, latent topic models and textual network analysis and proposes a new approach for explaining the network based on topics and terms. Joint usage of topic modelling and textual network approach results in a better description of semantic content of each topic.

Abstract *Il paper vuole introdurre la dashboard HateViz, una piattaforma interattiva che si appoggia a Twitter, riguardante l'hate speech contro le donne. Partendo dalla raccolta da Twitter di testi scaricati tramite pacchetti di R, l'obiettivo della dashboard è di fornire alla community dei decision-makers uno strumento semplice da utilizzare per monitorare un fenomeno così importante. HateViz unisce tre metodologie: textual mining, latent topic model e analisi delle reti testuali, e propone un nuovo tipo di approccio per spiegare il network formato da topic latenti e lemmi. L'uso congiunto degli approcci di topic modelling e delle misure di rete fornisce una migliore interpretazione del contenuto di ogni topic.*

Key words: *HateViz*, Dashboard, Text mining, Textual network analysis

[1]Emma Zavarrone, Università IULM, emma.zavarrone@iulm.it

[2]Maria Gabriella Grassia, Università Federico II, mariagabriella.grassia@unina.it

[3]Marina Marino, Università Federico II, marina.marino@unina.it

[4]Rocco Mazza, Università Federico II, rocco.mazza@unina.it

[5]Nicola Canestrari, Università IULM, nicolacanestrari.nc@gmail.com

1 Introduction

The paper focuses on the data visualization of the hate speech against women inside the Italian Twitter community. *HateViz* is the proposed shiny dashboard, to represent an alternative way for reading the social-political changes (Fig. 1). Hate speech based communication is increasing with the massive production of user generated content on social network. The literature defines hate speech as content that disparages a person or a group on the basis of some characteristic such as ethnicity, gender, sexual orientation (Davidson et al., 2017; Malmasi and Zampieri, 2018). Under this perspective, to provide the community of decision-makers with an easy-to-use tool to monitor these changes becomes more crucial. Starting from tweets, collected through API using R, the *HateViz* allows to explore the mining of contents extracted and study the lexical structure that link the principal discussion topics. The contribution at the state of the art is a new point of view studies that focus on communities' discourse and a development of a representation tool for our results. This paper is structured as follows: section 2 show the methods and the data visualization tools used, section 3 more specifically discusses the results of the research, and in section 4 there are the future works.

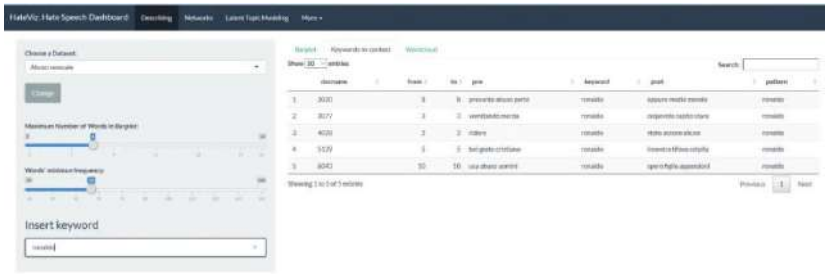


Figure 1: The dashboard user interface

2 Methodological approach

The *HateViz* mixes three methods: textual mining, latent topic models and social network analysis and proposes a new approach for explaining the network based on topics and terms. Figure 2 shows the dashboard's flowchart: (1) content extraction and corpus pre-processing; (2) descriptive study of texts: most frequent words and co-occurrence network analysis; (3) applying a model to extract and identify the latent topics within the contents collected; (4) using network analysis's both a better interpretation of each topic and the semantic relationships between the extracted topics and documents' terms. The corpus was built with a random sample of tweets in Italian language and the pre-treatment operations (bag of words approach) have been applied. A final Document-Term Matrix has been generated without sparse words and empty documents. DTM allows to describe the corpus of tweets through common visualizations such as barplot of most frequent words and wordcloud. In addition, the

DTM can be read like an affiliation matrix in which to analyse the relationships between words and the texts. We converted this in co-occurrence matrix to transform the collection of text into a visual maps of words, a similar approach has been applied by Segev (2020). In this step *HateViz* uses the networks and its centrality measures (Faust, 1997): degree and closeness. The search of semantic structure has been realized from a DTM through the Latent Dirichlet Allocation model (Blei et al., 2003; Griffiths and Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001). LDA is the method used both to extract latent topics and to construct the *terms-topics* matrix. The model is a generative and Bayesian inferential model and it allows to infer the latent structure of topics through by recreating the documents in the corpus considering the relative weight of the topic in the document and the word in the topic, in an iterative way. At the base of the LDA we find the following assumptions: a) documents are represented as mixtures of topics, where a topic is a probability distribution over words; b) the topics are partially hidden, latent precisely, within the structure of the document (Steyvers and Griffiths, 2007). At first glance, the main methodological challenges faced lie in the construction of a two-mode matrix of reduced order able to represent the terms topics network. The original contribution is the network construction, we operated a selection of terms based on a probability threshold derived from the model (LDA) results.

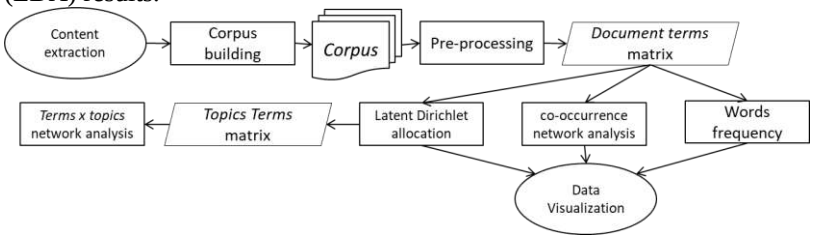


Figure 2: The *Hateviz* dashboard's flowchart

3 Results

The main outcome of the dashboard (Fig. 2) is to show results from the analysis in an easy and interpretable way. *HateViz* has an intuitive layout with a clear division between the control panel on the left and the plotting space on the right. The interactive control panel allows the selection of the datasets from different keywords and the customed plot output from different analysis: 1. Barplot of frequencies; 2. Keywords in context; 3. Co- occurrence network analysis; 4. Wordcloud of terms; 5. Latent Topics; 6. Terms Topics Network (A beta version of shiny web app is already on line at link: https://rccmazza.shinyapps.io/Donne4/). The dashboard has been developed on 403.612 tweets extracted from July 2018 to May 2019. The complete database is composed by seven sub-datasets, each one selected from specific keywords. *HateViz* allows an analysis for each textual dataset built with extraction keywords. The wordcloud in Figure 4 shows three semantic dimensions: the first one is banally referable to the cases of news commented by users on the social; the second

one to an institutional and regulatory dimension, in fact there are words referred to the need for stronger penalties; in the last one we find the ways in which violence against women can be realized. From the representation of the network and the centrality measure, it is possible to make some observations. Looking at the position in the graph, terms in a central position like “donne” (women), “stupro” (rape), “accuse” (charges) or “vittime” (victims) indicates a strong link and relations with most of the other lemmas, meaning that these words are the main core of the network.

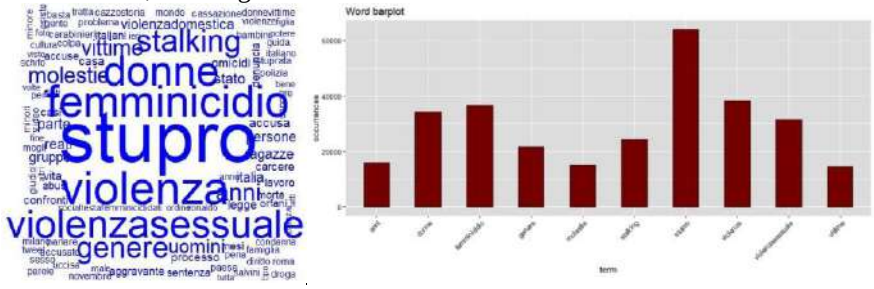


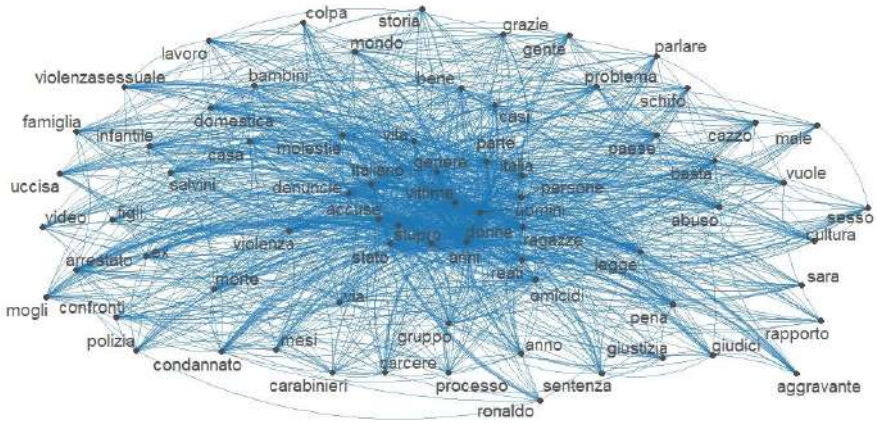
Figure 4 a. Wordcloud **b.** Barplot

Terms	***Abs. degree***	***Nor. Degree***	***Terms***	***Abs. degree***	***Nor. Degree***
abuso	70	1.000	colpa	65	0.929
aggravante	65	0.929	sentenza	64	0.914
legge	68	0.971	bene	63	0.900
gruppo	67	0.957	genere	62	0.886
stato	66	0.943	pena	61	0.871

Table 1: Degree centrality values calculated on total co-occurrence matrix

Even marginal terms, arranged along the borders of the graph, for example “aggravante” (aggravating), have lot of connections within the network and with central terms.

Figure 5 *Co-occorence network (high sparsity cut)*



The topics extracted with the LDA model are 5. The dimensions emerged from model take up some latent themes. Using jointly topic modeling and SNA allows to better define the content of each topic. Specifically, it emerges five main themes:

1. First topic: a reference to the implementation of judgement, some examples of terms in this topic are “polizia” (police), “arrestato” (arrested), “denuncie” (denunciation/charge);
2. Second topic: referring to the public attention on social media, involving terms like “tweet”, “foto” (pictures), “gente” (people), “visto” (seen);
3. Third topic: a reference to the cultural dimension of the phenomenon, with terms like “genere” (gender), “uomini” (men), “donne” (women), “amore” (love);
4. Forth topic: a reference to the institutional dimension, involving terms like “ministro” (minister), “salvini”, “italia” (italy), “italiani” (italian)
5. Fifth topic: referred to legal and trial aspects, with terms like “reato” (crime), “omicidi” (murder), “pena” (sentence).

Through the words-topic network (Zuo, Zhao and Xu, 2015) it is possible to observe how the terms are associated with the referred topic. The network is composed by latent topics identified through the LDA technique giving the opportunity to examine how the terms from corpus are associated with them. A topic that takes a central position in the network represents the main semantic area identified in the dataset. A term, which represents a node connecting different topics indicates that is not only present in both thematic groups to which it is connected, but also represents a connection between semantic areas associated with each topic.

Terms	***Closeness centrality***
Vittime	0.699
Molestie	0.694
Ragazze	0.689
Stato	0.694

Table 2: Closeness centrality values calculated on two mode matrix

The words-topics graph shows how topic1 (implementation of judgement) and topic3 (cultural dimension) are positioned at the centre of the network, which means that they represent the most relevant semantics in the dataset. Its position in the network makes topic1 a bridge, a node connecting topic3 and topic5 (legal and trial aspects) with topic2 (public attention) and topic4 (institutional dimension), since topic1 is the node that connects the other topics, and without that node the whole network would collapse. It is possible to notice terms that belong to more than one topic, connecting them: topic1 and topic4 are linked by the term "ragazze" (girls); topic2 and topic1 are linked by the terms "stato" (country); topic1 and topic3 by the word "molestie" (harassment); topic3 and topic5 are linked by the term "vittime" (victims). In Tab. 1 there are the closeness centrality values for this hub terms.

4 Future works

This paper represents a work in progress, In the future we will extend the functionality of the application by implementation of machine learning tools that will intercept statements containing hate speech and can made a semantic map of it. The work presents this possible further development:

- Deep learning algorithm to report hate speech contents
- Control measure about the correct content identification

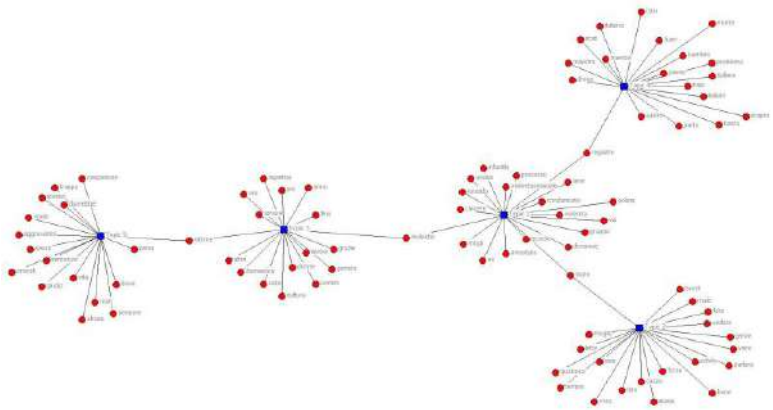


Figure 6 *Words and topics networks*

References

1. Blei, D. M., Ng, A. Y., Jordan, M. I., Latent dirichlet allocation, Journal of machine Learning research, 3(Jan), 993-1022, (2003).
2. Broy, M.: Software engineering --- from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.) Software Pioneers, pp. 10-13. Springer, Heidelberg (2002)
3. Charu R., Amit V., Bharath Y., Rama K., Kiran S., API-FICATION, Hcl Technologies, https://www.hcltech.com/sites/default/files/apis_for_dsi.pdf, 2014.
4. Davidson, T., Warmsley, D., Macy, M., & Weber, I. Automated hate speech detection and the problem of offensive language, in Eleventh international aaai conference on web and social media. (2017).
5. Dod, J.: Effective substances. In: The Dictionary of Substances and Their Effects. Royal Society of Chemistry (1999) Available via DIALOG.
6. Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for Computer Algebra. Kluwer, Boston (1992)
7. Griffiths, T. L., Steyvers, M.: A probabilistic approach to semantic representation. In: Proceedings of the annual meeting of the cognitive science society, Vol. 24, No. 24 (2002).
8. Griffiths, T. L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235, (2004).
9. Griffiths, T. L., Steyvers, M.: Prediction and semantic association. In: Advances in neural information processing systems, pp. 11-18, (2003).
10. Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. \textbf{169}, 321--354 (1995)
11. http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
12. Malmasi, S., & Zampieri, M. Challenges in discriminating profanity from hate speech. Journal of Experimental & Theoretical Artificial Intelligence,30(2), pp. 187-202. (2018).
13. Puschmann, C., & Ausserhofer, J. (2017). Social Data APIs: Origin, Types, Issues.
14. Shi, L., Zhong, H., Xie, T., & Li, M. An empirical study on evolution of API documentation. In International Conference on Fundamental Approaches To Software Engineering (pp. 416-431). Springer, Berlin, Heidelberg. (2011, March).
15. Segev, E. Textual network analysis: Detecting prevailing themes and biases in international news and social media. Sociology Compass, 14(4), e12779. (2020).
16. Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. (2000) doi: 10.1007/s001090000086

How to perform cyber risk assessment via cumulative logit models

Valutazione del cyber risk con modelli logit cumulativi

Silvia Facchinetti, Silvia Angela Osmetti and Claudia Tarantola

Abstract Cyber risk assessment is a field with a growing amount of research. Information on monetary losses due to a cyber attack is difficult to gather due to the sensitive nature of the data. An evaluation of the severity of the attack on an ordinal scale is easier to obtain. We apply a cumulative logit model to identify the variables that mostly affect the severity of an attack. A systemic risk indicator based on social network analysis is proposed and included in the model.

Abstract *La valutazione del cyber risk è un tema di ricerca molto attuale. Informazioni sulle perdite monetarie dovute a un attacco informatico sono difficili da reperire a causa della natura sensibile dei dati. È più facile ottenere una valutazione della gravità dell'attacco su scala ordinale. Applichiamo un modello logit cumulativo per identificare le variabili che influenzano maggiormente la gravità di un attacco. Un indicatore di rischio sistemico basato sull'analisi dei social network è proposto ed incluso nel modello.*

Key words: criticality index, cumulative logit model, cyber risk, social network

Silvia Facchinetti
Department of Statistical sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 - Milano, e-mail: silvia.facchinetti@unicatt.it

Silvia Angela Osmetti
Department of Statistical sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 - Milano, e-mail: silvia.osmetti@unicatt.it

Claudia Tarantola
Fintech laboratory, Department of Economics and Management, University of Pavia, Strada Nuova 65 - Pavia e-mail: claudia.tarantola@unipv.it

1 Modelling cyber risk data

Institutions should be encouraged to collect cyber incident data in order to employ statistical techniques to prevent such type of incidents and estimate their impact/gravity. Although quantitative loss data are rarely available, sometimes it is possible to obtain a qualitative appraisal of the attack severity based on an ordinal scale.

We consider a set of $2,679$ observations regarding serious cyber attacks occurred worldwide in the period 2017-2018. The data have been collected by the researchers of the *Hackmanac Project* and described in the Clusit Report on ICT Security[1]. An attack is classified as "serious" if it had a significant impact on the victims in terms of economic losses, damages to reputation and/or dissemination of sensitive data. Cyber risk severity (S) is described by an ordinal variable assuming values 1 (medium severity), 2 (high severity) and 3 (critical severity).

We apply a cumulative logit model to express the severity of a cyber attack as a function of a vector of covariates $\mathbf{x}$

$$logit[P(S \leq s)] = \alpha_s - \mathbf{x}^T \boldsymbol{\beta} \qquad s = 1,2 \tag{1}$$

where α_s is a specific intercept and $\boldsymbol{\beta}$ is the vector of the regression parameters. The larger the value of $\mathbf{x}^T \boldsymbol{\beta}$, the higher the probability to obtain an elevated level of cyber risk. Model (1) implies a constant relationship between the cumulative probabilities and the covariates. For a specific set of of covariates, the logit is altered only by the intercepts which are different for each category. This is known in the literature as the Proportional Odds (PO) assumption.

As covariates we consider the following categorical variables classified on a nominal scale: Type of Attack, Attack Technique, Continent. We use social network analysis and its centrality measures to construct a quantitative indicator (Closeness) of the diffusion of cyber attacks among the different victims. The nodes of the network correspond to the victims of cyber attacks, while the edges indicate the strength of their connections, see Figure 1.

The network in Figure 1 is constructed as follows. For each victim we calculate the weekly time series of the Criticality Index by [3], then we estimate the network based on the partial correlation matrix among the previous series. Variable Closeness corresponds to the network centrality measure proposed by the authors of [7].

2 Empirical results and discussion

Table 1 shows the estimates of the model coefficients obtained adapting model (1) to our data[2]. As expected, Closeness is significant and affects positively the severity

[1] Clusit is an Italian association for cyber security based in Milan.

[2] The baseline categories for the categorical variables Type of Attack, Attack Technique and Continent are Cybercrime, SQL Injection and Multiple continents, respectively.

Table 1 Cumulative logit model fitted to the cyber risk data.

Covariates		value	std.error	z.value	p.value
Type of Attack	Espionage/Sabotage	1.879	0.140	13.431	0.000
	Hacktivism	0.252	0.162	1.558	0.119
	Information Warfare	1.083	0.199	5.441	0.000
Attack Technique	0-day	-0.568	0.854	-0.665	0.506
	Multiple Threats	-1.470	0.761	-1.932	0.053
	Trivial Threats	-1.933	0.740	-2.613	0.009
	Unknown	-1.935	0.743	-2.606	0.009
Continent	Africa	0.367	0.377	0.974	0.330
	America	0.756	0.097	7.749	0.000
	Asia	0.883	0.143	6.161	0.000
	Australia/Oceania	1.197	0.292	4.093	0.000
	Europe	0.743	0.125	5.928	0.000
Closeness		0.985	0.452	2.179	0.029

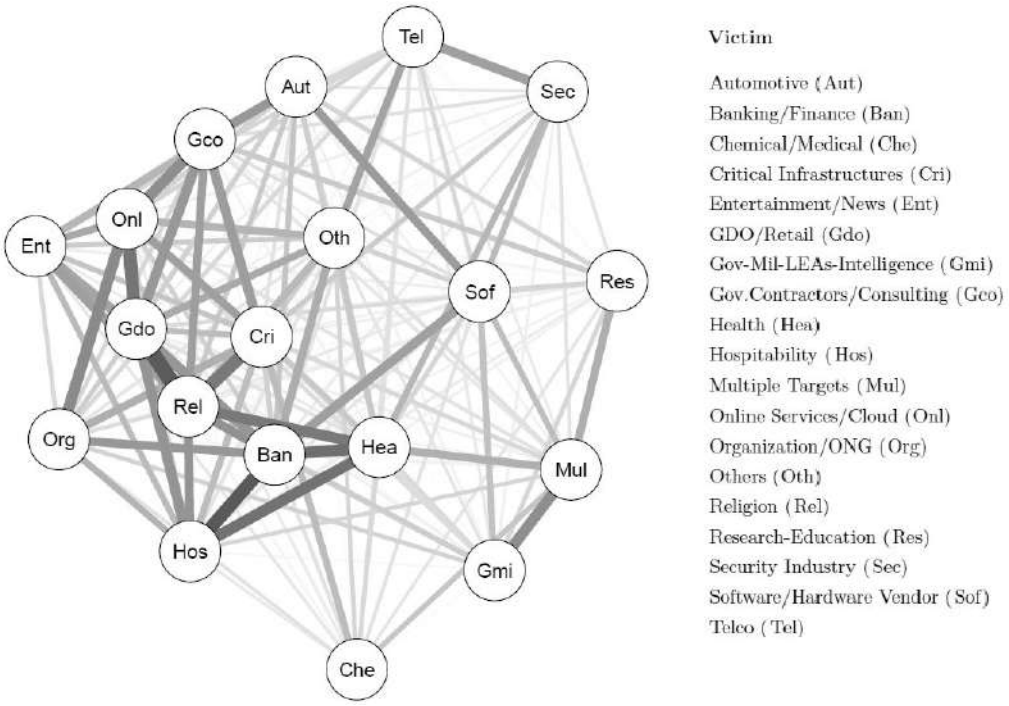


Fig. 1 Network model among the victims of cyber attacks.

level. In terms of real-world implications, if a victim strongly connected suffer a critical attack, it has an influence on the attack severity level of the other connected victims, with respect to the ones not strongly connected. Hence, strongly connected victims should have a collaborative approach to prevent critical cyber attacks. Otherwise, the gravity of the attacks they suffered and the related consequences could increase.

The Lipsitz global test [5] shows a good fit for the model, and the likelihood ratio test indicates a significant improvement over the baseline intercept-only model. For a detailed analysis of this data see [4].

We conclude this section with a discussion on the validity of the PO assumption. Testing the PO assumption is challenging both in a frequentist and Bayesian frameworks, especially when many categorical variables are involved. Usually the score test by [8] is used to check the PO assumption. However, as discuss in [2], it tends to reject the null hypothesis even when the PO assumption is reasonable. Hence, we computed the score test for each category of the covariates, separately (see [4]). These tests suggest that the PO assumption could be rejected for the categories Unknown and Espionage/Sabotage. A reasonable strategy for investigating whether these categories are really problematic could be to fit three separate logistic regressions, calculate the corresponding estimated Odds Ratios (ORs) and compare them with the OR obtained from the cumulative logit model.

The analysis performed suggest that the ordinal PO model is a fair summary of the patterns in the data in relation to the severity levels. However, it is possible to relax the PO assumption by specifying a more complicated cumulative logit model without proportional odds [1], or a partial proportional odds model [8]. For a recent contribution in the Bayesian setting see [6].

References

1. Agresti A. (2010) *Analysis of Ordinal Categorical Data*, 2*nd* ed., J.Wiley & Sons, Hoboken.
2. Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression, Biometrics, 46: 1171-1178.
3. Facchinetti, S., Giudici, P., and Osmetti, S.A. (2019). Cyber risk measurement with ordinal data, Statistical Methods and Application, https://doi.org/10.1007/s10260-019-00470-0.
4. Facchinetti, S., Osmetti, S.A., and Tarantola, C. (2020). A statistical approach for assessing cyber risk via ordered response models, submitted.
5. Lipsitz, S.R, Fitzmaurice, G.M., and Molenberghs, G. (1996). Goodness-of-Fit Tests for Ordinal Response Regression Models, J. of the Royal Statistical Society (Series C), 45: 175-190.
6. McKinley, T.J., Morters, M., and Wood, J.L.N. (2015). Bayesian Model Choice in Cumulative Link Ordinal Regression Models, Bayesian Analysis, 10: 130.
7. Opsahla,T. , Agneessensb, F. and Skvoretzc, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths, Social networks, 32, 245-251.
8. Peterson, B., and Harrell F.E.Jr (1990). Partial Proportional Odds Models for Ordinal Response Variables, Journal of the Royal Statistical Society (Series C), 39: 205-217.

Machine learning prediction for accounting system

Classificazione predittiva per la contabilizzazione delle fatture elettroniche

Chiara Bardelli, Silvia Figini

Abstract Electronic invoicing has become mandatory for Italian companies since January 2019. Each invoice has to be structured in a predefined xml template which facilitates the accountants work. Thanks to this measure the information included in the invoice can be easily extracted and analysed giving the possibility to automate part of the accounting process: all sent or received invoices of a company have to be classified into specific accounting codes. This task can be easily translated into a machine learning multiclass classification problem where the accounting code is the target variable of our problem. In this work we propose two different approaches and compare them in terms of prediction accuracy. We remark that better performances can be achieved considering the hierarchical structure of the accounting codes, our target variable.

Abstract *Da gennaio 2019 la fatturazione elettronica è diventata obbligatoria per qualsiasi azienda italiana. Questa misura ha permesso di snellire il processo di contabilizzazione delle fatture, uniformando le informazioni contenute in esse in un template xml. La parte più corposa del processo di contabilizzazione rimane comunque la classificazione delle fatture utilizzando il piano dei conti econimici (un elenco di conti che descrivono la natura economica delle transazioni effettuate). Questa fase del processo può essere tradotta in un problema di classificazione di machine learning nel quale vogliamo predire il conto economico associato ad ogni linea di fattura. Proponiamo quindi due possibili approcci per trattare questo tipo di problema confontandoli in termini di accuratezza nella predizione. La struttura gerarchica a livelli del piano dei conti suggerisce l'applicazione di un algoritmo di classificazione gerarchico come lavoro futuro.*

Key words: multiclass classification, text mining

Chiara Bardelli
Università di Pavia, Dipartimento di Matematica, e-mail: chiara.bardelli01@universitadipavia.it

Silvia Figini
Università di Pavia, Dipartimento di Scienze Politiche e Sociali, e-mail: silvia.figini@unipv.it

1 Introduction and motivation

The accounting profession is faced with the numerous challenges of the digitalization era. Modern technologies lead to the automation of part of the workload of accountants. As a consequence, repetitive routine tasks can be replaced by algorithms that give the possibility to accountants to focus instead on all those activities that require human critical thinking and creativity [4]. One of the tasks which can be easily automated is the process of recording financial transactions contained in supplier or customer invoices. When an invoice is received or issued by a company, the accountant has to decide which codes should be charged. This task can be easily translated into a predictive classification problem: input variables can be extrapolated from the invoices and accounting codes can be considered as the target variable.

Machine learning techniques have been extensively used for extracting features and information from scanned invoices by combining the classical OCR (Optical Character Recognition) methods to deep neural networks [5], [8]. Yet, since January 2019 all Italian companies have been asked by law to issue electronic invoices in a structured and fixed xml template. This aspect solves the main issue of extracting information from scanned documents allowing to focus on the automation of the classification of transactions included in the invoice ([1],[2]).

In this paper we apply a machine learning model to the content of the electronic invoice to predict the accounting codes. Each line of the invoice is represented by a numeric vector that combines textual description and other codes related to the line. We test performances of our classifier on sample data provided by a company which develops software for professional accounting. The main challenges of this problem are: (i) the reconstruction of the training set, and (ii) the complex structure of the accounting codes which are organized in a hierarchical taxonomy based on 5 levels for a total of 90 different labels leading to an imbalanced classification problem.

The rest of the paper is organized as follows: Section 2 defines the problem and the dataset; Section 3 introduces the methods applied to reconstruct the training set and to solve our predictive problem; in Section 4 the application and the most relevant results are displayed; finally, Section 5 contains some concluding remarks.

2 Description of the dataset

Given the information about a line of an invoice and the characteristics of the companies which are involved, the aim is to predict the accounting code related to the line. Multiple lines of an invoice can be associated to the same accounting code. For sake of simplicity, we assume that the lines inside an invoice are independent ignoring the grouping term which could influence the predictive output. In our classification task, we construct the prediction rule given the training sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$ with:

- y_i, $i = 1,...,N$, categorical observations which represent the accounting codes associated to the i-th line of invoice

- $\mathbf{x}_i$, the vector of predictors related to the content of the invoice

The dataset considered in this study is the combination of two data sources: electronic invoices in xml format and accounting outputs result of the recording of the invoices. The match of these two different sources is possible only at document level. On the other hand, it is not possible to recreate directly the relation between a single line of an invoice and the associated accounting code. This problem of data reconstruction can be addressed exploiting the information about the amounts which are included both in the xml invoices (detailed amounts) and in the accounting outputs (aggregated amounts). Thanks to this information it is possible to translate it into a knapsack problem with equality constraints as explained in Section 3.1.

In our classification problem features consider information related to the content of the invoice and characteristics of the companies. In particular, we consider:

- textual description of the line of the invoice
- codes associated to the line (like VAT code)
- information about activities performed by companies:
  - ATECO code (classification of economic activity) provided by ISTAT
  - ISA categories based on the level of fiscal reliability

The dataset contains information about 9422 electronic invoices which provide more than 60000 lines to classify.

3 Methods

We describe methods applied for the reconstruction data problem, for the preprocessing of textual data and for the choice of the predictive classification algorithm.

3.1 Knapsack problem

The reconstruction data problem is represented as a multi knapsack problem with equality constraints [7] considering a single invoice at a time. Let $i = 1,...,N$ be the lines of the invoice and $j = 1,...,M$ be the index that identifies the accounting codes. The problem can be formulated as follows:

$$\max \sum_{j=1}^{M} \sum_{i=1}^{N} p_i z_{ij} \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} c_i z_{ij} = b_j, \quad \forall j \tag{2}$$

$$\sum_{j=1}^{M} z_{ij} = 1, \quad \forall i \tag{3}$$

where c_i is the detailed amount of the i-th line and b_j is the aggregated total of the j-th accounting code. The vector of weights p_i is considered equal to 1 for each i. The value of $z_{i,j}$ is 1 if line i-th is associated to the j-th accounting code, 0 otherwise.

The problem has been resolved through an heuristic which stops when the first feasible solution is found. First, the heuristic orders the amounts of lines in a vector in decreasing order and the aggregated amounts in an increasing order vector. Then, the algorithm starts to match first values in the first positions of the two vectors.

The algorithm has successfully matched the 88% of the lines of our initial dataset, which has been used for the training phase and the evaluation of the performances.

3.2 Data preprocessing

Textual data included in lines of invoices have been previously cleaned through standard preprocessing steps [6]. Finally, textual information has been tokenized to create array of words. In order to transform textual data information into a suitable feature space we compare two different procedures:

- Bag of Words (BoW) approach [10], a simple way to encode the array of words into a binary vector. The main drawback is that the length of the feature vector grows linearly with the number of distinct words, leading to infeasible sizes.
- Word2Vec algorithm [9], a language modeling technique which maps similar sentences into similar numeric vectors of fixed size.

This two different procedures have been combined to two different classification algorithms described in subsection 3.3.

3.3 Classification algorithm

The different nature of sent and received invoices made us split them in 2 different sub-datasets and analyse them separately to obtain 2 classification algorithms.

For each sub-dataset two different classification algorithms have been compared in this study: Support Vector Machine classifier (SVM) [10] and Random Forests (RF) [3]. To adopt the SVM classifier for multiclass classification, the one-vs-rest technique has been used. This procedure has a very high computational cost since the number of classes is very large (70 distinct classes for the received invoices).

4 Application to real dataset

To evaluate the performance of the two approaches the dataset has been split into training set (80 %) and test set (20 %). Tables 1 and 2 shows performances of the two

different combinations applied respectively to sent invoices and received invoices. As we expected, the algorithm trained on received invoices shows an accuracy rate lower than sent invoices, due to the fact that first are more diversified in terms of content and accounting codes associated. As far as the classification algorithm is concerned, Random Forests combined to Word2Vec performs better both for sent and received invoices. This can be explained by the limited vocabulary used in SVM combined with BoW: only 200 words have been considered because of the high computational costs.

Table 1 Performances evaluated on test set (sent invoices)

	Precision	Recall	F1-score
RF + Word2Vec	91 %	92 %	91%
SVM + BoW	87 %	84 %	86 %

Table 2 Performances evaluated on test set (received invoices)

	Precision	Recall	F1-score
RF + Word2Vec	82 %	85 %	82%
SVM + BoW	78 %	74 %	75 %

To investigate the hierarchical structure of the target variable, Random Forests algorithm has been trained for each level of the hierarchy in order to show the improvement of the performances and to motivate a deeper study on hierarchical classification algorithms. Figure 1 shows the accuracy rate for the different levels of the target variable. Both for sent invoices and received invoices the accuracy rates improve at lower levels (accounting codes associated to lower levels are more generic and easier to predict). This aspect can be exploited to guide the classification output through the structure of the accounting codes.

5 Conclusion

Part of the bookkeeping process deals with the classification of the transactions reported in invoices into specific accounting codes. This is often a time-consuming process, which can be translated in a machine learning classification algorithm. In this work we presented a possible procedure to handle with accounting data from the reconstruction of the dataset to the training of a machine learning classifier. We tested two different approaches and we noticed that Random Forests combined to Word2Vec can be considered a good candidate to solve our problem.

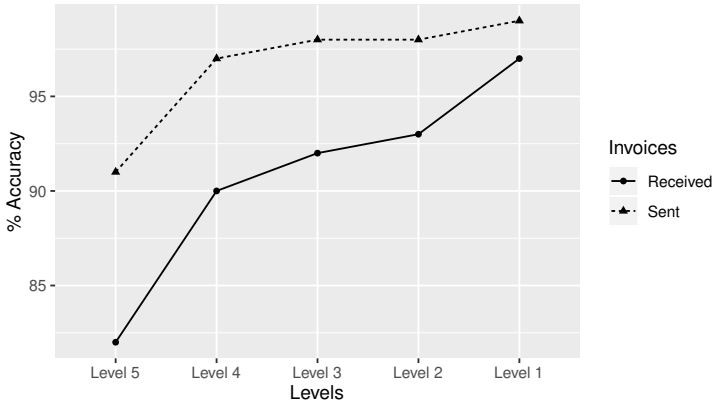


Fig. 1 Accuracy computed at different levels of the target variable with RF and Word2Vec.

One of the main difficulty is to handle the high number of different labels used in this dataset. A possible approach to solve this problem is to consider a hierarchical classification framework: the target variable is structured in a predefined hierarchy which takes the form of a rooted tree. This is also motivated by the results obtained for the classification at different levels of our target variable. We consider the application of hierarchical classification algorithm [11] as a possible future work.

References

1. Bengtsson, H., Jansson, J.: Using classification algorithms for smart suggestions in accounting systems. Master Thesis, Chalmers University of Technology, Gothenburg (2015)
2. Bergdorf, J.: Machine learning and rule induction in invoice processing: Comparing machine learning methods in their ability to assign account codes in the bookkeeping process. Master Thesis, Kth Royal Institute of Technology, Stockholm (2018)
3. Breiman, L.: Random forests. Machine Learning. **45.1**, 5–32 (2001)
4. Gulin, D., Hladika, M., Valenta, I.: Digitalization and the Challenges for the Accounting Profession. IRENET-Society for Advancing Innovation and Research in Economy, (2019)
5. Holt, X., Chisholm, A.: Extracting structured data from invoices. In: Proceedings of the Australasian Language Technology Association Workshop 2018. pp. 53–59 (2018)
6. Khan, A.,Baharudin, B., Lee, L., Khan, K.: A review of machine learning algorithms for text-documents classification. Journal of advances in information technology. **1.1**, 4–20 (2010)
7. Kozanidis, G., Melachrinoudis, E., Solomon, M.: The linear multiple choice knapsack problem with equity constraints. International Journal of Operational Research. **1**, 52–73 (2005)
8. Palm, R. B., Winther, O., Laws, F.: Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) **1**, 406–413 (2017)
9. Rong, X.: Word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 (2014)
10. Joachims, T., Learning to classify text using support vector machines. Springer Science & Business Media (2002)
11. Silla, C. N., Freitas, A. A.: A survey of hierarchical classification across different application domains.Data Mining and Knowledge Discovery. **22** 31–72 (2011)

Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory

Un modello di valutazione della conoscenza per l'insegnamento della statistica: l'esperienza del progetto Erasmus+ ALEAS

Cristina Davino[a], Rosa Fabbricatore[a], Carla Galluccio[b], Daniela Pacella[a], Domenico Vistocco[a], Francesco Palumbo[a]

[a] University of Naples Federico II
[b] University of Florence

Abstract In recent years, there is an increasing need for technology-based platforms to assist traditional learning methodologies. However, it is challenging to set up a common assessment framework to evaluate user knowledge. To address this issue, we propose an approach to teaching undergraduate statistics that makes use of the psychometric Item Response Theory based on latent class categorization to evaluate the user ability based on the European learning outcomes - the Dublin descriptors. Additionally, we enclose the user assessment workflow in a formalized structure using the principles of Knowledge Space Theory to track the current user knowledge state adaptively. The methodological framework serves as a base for the app developed within the ALEAS ERASMUS+ Project.

Abstract *Di recente, è stata evidenziata una maggiore necessità di piattaforme tecnologiche come ausilio alle metodologie di apprendimento tradizionali. Tuttavia, è difficile istituire un quadro di valutazione comune per le conoscenze degli utenti. Pertanto, proponiamo un approccio all'insegnamento della statistica universitaria che si avvalga della teoria psicometrica Item Response Theory basata sulla categorizzazione in classi latenti, allo scopo di valutare la capacità dell'utente sulla base di obiettivi di apprendimento europei, i descrittori di Dublino. Inoltre, racchiudiamo il flusso di valutazione degli utenti in una struttura formalizzata utilizzando i principi della Knowledge Space Theory per tracciare in modo adattivo l'attuale stato della conoscenza dell'utente.*

Key words: Item Response Theory; Knowledge Space Theory; Statistics; Technology-enhanced learning; Intelligent tutoring systems

1 Introduction

The problem of assessing the student on a multidimensional level is one of the most challenging tasks in education (Deonovic et al., 2018). The student behaviour and responses should be evaluated on the one hand in comparison with the student population, on the other hand in comparison with the individual performance changes over time, so to measure the learning outcomes.

Recently, there is rising attention towards teaching statistics using technology. Although there are several technologies developed for teaching and assessing statistics, there are not many applications specifically targeting undergraduate students in non-scientific courses. Recently, Lopez Lamezon and colleagues demonstrated the advantages of using a virtual environment for teaching statistics in Medicine degree courses although not grounding their learning assessment upon specific educational theories (López Lamezón et al., 2018).

A complete assessment framework should be able to:

- detect the overall ability level of the student sample and to adapt to it;
- personalize the student experience selecting the most appropriate set of topics, questions and items to present according to each student's knowledge progress.

Concerning the first requirement, we exploit the Item Response Theory (IRT) (Rasch, 1960a). IRT allows modelling the probability that each student answers correctly to a particular item given her or his ability. It also allows estimating the difficulty of each item given the sample of answers. We categorize the items of each statistics subtopic according to the Dublin descriptors of the learning outcomes. Concerning the second requirement, we resort to Knowledge Space Theory (KST) (Doignon and Falmagne, 1985), widely used in the field of expert systems to the end of organizing the full knowledge required to master a specific subject into a direct acyclic graph structure. Nodes of the graph represents a competency. Each student is classified into a state and the system decides the following direction on her/his learning path. The proposal describes the core of ALEAS (Adaptive LEArning in Statistics), a learning platform developed as the intellectual output of the homonymous Erasmus+ project. ALEAS aims to be a complement to traditional teaching methodologies. It focuses on the assessment phase, primarily for university students enrolled in non-scientific degree curricula.

Over the knowledge structure for basic statistics knowledge, a directed acyclic graph consisting of several statistics subtopics is defined. Students' performance assessment is based on the multidimensional latent class IRT model with dichotomous items (Bartolucci, 2007).

The Dublin descriptors are the framework for qualifications of higher education courses in the European area; they qualify the expected outcome of any learning process. Our proposal focuses on the assessment of hard skills, and in particular statistics, and hence we consider the following three descriptors (Gudeva et al., 2012):

- *Knowledge and understanding (K)*: the ability to demonstrate knowledge and understanding with a theoretical, practical and critical perspective on the topic;
- *Applying knowledge and understanding (A)*: the ability to apply the knowledge identifying, analysing and solving problems sustaining an argument;

- *Making judgments (J)*: the ability to gather, evaluate and present information exercising appropriate judgment.

2 Methodology

A description of both the general KST framework and the IRT methodology will be provided in the following subparagraphs.
KST is a methodological framework that analyzes the performance over time of individual students (Doignon and Falmagne, 1985, 2016). It is particularly useful for the current knowledge assessment of the students, and also to adaptively select the most tailored items to present to the student given his or her ability. KST formalizes a structure of the domain into nodes of a directed acyclic graph aiming to assess each learner according to which subsets of the domain they master. The use of KST induces a twofold issue:

- to construct the structure of the knowledge domain investigating the relationship among the different subsets of the domain;
- to select the most appropriate questions that allow assessing the current knowledge subsets the student has mastered.

Several algorithms implement the progression rule from one to the next knowledge state. The probabilistic model considers each knowledge state having a likelihood function whose parameters are updated according to the student answers; while the expert model that we proposed for the ALEAS platform aims to construct a knowledge graph with the help of experts. We divided statistical knowledge into nodes, and we set constraints to the student progress path. In particular, to progress from one node to the other, a student has to obtain mastery - considered as the highest level of ability - in all the required nodes. The level of ability is evaluated through the multidimensional IRT model described below (Fabbricatore et al., 2019).
IRT is a model-based theory that aiming at estimating the probability that each student will answer correctly to each item of a set. IRT models founded on the idea that the probability that an individual provides a certain response to a certain item can be described as a function of the person's latent trait, plus one or more parameters that characterize a specific item. A latent trait is typically described by a continuous normal probability distribution (Bartolucci et al., 2015).

According to the traditional logistic IRT models with dichotomous items, the probability that the s-th subject (with $s = 1, \ldots, S$) will respond correctly to the item i (with $i = 1, \ldots, I$) in the most general form (4-parameter logistic model; 4PL) can be formalized as following:

$$P(X_{si} = 1 | \theta_s, \Gamma_i) = c_i + \frac{1 - d_i - c_i}{1 + e^{a_i(\theta_s - b_i)}} \tag{1}$$

where X_{si} is the response of the s-th subject on the i-th item with realization $x_{si} \in [0,1]$, $\theta_s \in \mathbb{R}$ is the ability of the s-th subject, and $\Gamma_i = (a_i, b_i, c_i, d_i)$ is the whole set of parameters of the i-th item. In particular, $a_i \in \mathbb{R}$ is the item discrimination parameter,

$b_i \in \mathbb{R}$ represents the item difficulty, and $c_i, d_i \in \mathbb{R}$ are the guessing and the ceiling error parameters respectively (Noventa et al., 2019). So, logistic IRT models are based on the idea that the correct response probability follows a logistic curve called the Item Characteristic Curve, and it is always directly proportional to the ability of the student (Rasch, 1960a; Birnbaum, 1969). Each item has its specific curve, which is defined by a set of parameters, between one and four. The four parameters refer to: *(i)* discriminating power (slope), *(ii)* item difficulty (location), *(iii)* guessing (lower asymptote) and *(iv)* ceiling (upper asymptote). Reduced models can be obtained constraining one or more parameters. This is the case of 2 parameter logistic IRT ($c_i, d_i = 0$ in equation 1) and 1 parameter logistic IRT ($a_i = 1$ and $c_i, d_i = 0$ in equation 1) models. In particular, only the error parameters (c_i, d_i) are considered for the model with 2 parameters. Such parameters take into account the case incidence on the probability of the correct answer. Models with 1 parameter also assume that all items have the same discriminative power (Rasch model; Rasch (1960b)).

In ALEAS, to perform the categorization of the items we adopted the model with 2 parameters, disregarding the guessing and the ceiling parameters. The choice was made taking into account that each item has four possible answers, lowering the impact of guessed answers. The model is applied to student responses, where we only take into account if the given response is correct or incorrect.

3 The proposed model for item and user categorization

The extension of traditional IRT models proposed by Bartolucci (2007) with the introduction of multidimensional latent class IRT models provides a suitable statistical tool to pursue our goals. Introducing the concepts of multidimensionality and discreteness of the latent traits, these models allow releasing both the IRT constraints of unidimensionality and continuous nature of the latent trait. In this view, we can simultaneously consider more latent traits each of which is represented by a discrete distribution with $\xi_1, \ldots, \xi_k$ support points defining k latent classes with weights $\pi_1, \ldots, \pi_k$. Latent class weights π_c (with $c = 1, \ldots, k$) represent the probability that a subject belongs to class c, and can be expressed as:

$$\pi_c = p(\Theta_s = \xi_c) \tag{2}$$

with $\sum_{c=1}^{k} \pi_c = 1$ and $\pi_c \geq 0$, where Θ_s ($s = 1, \ldots, S$) represents the discrete random variable of the latent trait of the sth subject. Multidimensionality, in this model, is introduced through a simple formulation in which each item is related only to one latent trait. In particular, items are divided in different subsets I_d (with $d = 1, \ldots, D$) based on D different dimensions, which, in our model, are represented by the three Dublin descriptors considered: K (knowledge and understanding), A (Applying knowledge and understanding), and J (making judgments).

Once the model has been defined, the next step is to estimate the parameters. We exploit a Maximum Marginal Likelihood approach making use of the Expectation-Maximization (EM) algorithm (Bartolucci et al., 2014).

In particular, the model tries to estimate (Expectation step) the probability of each individual belonging to one of the latent classes given his or her vector of responses; during the Maximization step it maximizes the log-likelihood that the individual belongs to that latent class. These two steps are repeated until convergence.

In order for the parameters of the model to be identified, a full matrix of individuals and responses must be supplied without missing values. Additionally, the model requires the constraint that, for each latent trait, one discriminating index is equal to 1 and one difficulty parameter is equal to 0 (Bartolucci et al., 2014).

Each user will thus be categorized into a latent class according to their response configuration. The number of latent classes k for the data can be estimated using information criteria such as the Bayesian Information Criterion or the Akaike information criterion (Gnaldi, 2017), or they can be defined according to the theoretical framework. For the present work, we established the number of classes to be $k = 3$, consistently with the number of dimensions defined by the Dublin descriptors, and hence with the theorized levels of ability. The model is applied using the R package `MultiLCIRT` (Bartolucci et al., 2014). The IRT model is applied separately for every node of the KST. The flow of information within each of the knowledge states of the KST is shown in Figure 1.

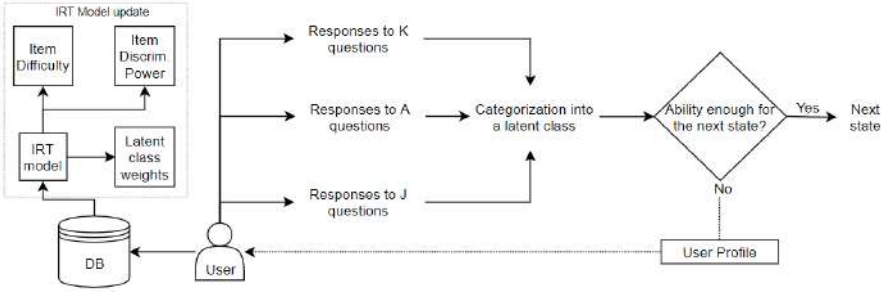


Fig. 1 Flow of information within a KST node and progress using the ability evaluated by the IRT model.

4 Conclusion

We proposed a methodological framework for the development of an intelligent tutoring system for assessing knowledge of statistics. The model targets undergraduate students and assesses them on a multidimensional level. To do so, we have presented a possible integration of a classical statistical and psychometric methodology - the

Item Response Theory, the Dublin descriptors of learning outcomes and a computational knowledge formalization - the Knowledge Space Theory for the structure of the user model.

References

Bartolucci, F. (2007). A class of multidimensional irt models for testing unidimensionality and clustering items. *Psychometrika 72*(2), 141.

Bartolucci, F., S. Bacci, and M. Gnaldi (2014). `MultiLCIRT`: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis 71*, 971–985.

Bartolucci, F., S. Bacci, and M. Gnaldi (2015). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Chapman and Hall/CRC.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology 6*(2), 258–276.

Deonovic, B., M. Yudelson, M. Bolsinova, M. Attali, and G. Maris (2018). Learning meets assessment. *Behaviormetrika 45*(2), 457–474.

Doignon, J.-P. and J.-C. Falmagne (1985). Spaces for the assessment of knowledge. *International journal of man-machine studies 23*(2), 175–196.

Doignon, J.-P. and J.-C. Falmagne (2016). Knowledge spaces and learning spaces. *New Handbook of Mathematical Psychology 2*, 274–321.

Fabbricatore, R., C. Galluccio, C. Davino, D. Pacella, D. Vistocco, and F. Palumbo (2019, September). The effects of attitude towards statistics and math knowledge on statistical anxiety: A path model approach. In M. Carpita and L. Fabbris (Eds.), *Statistics for Health and Well-being*, pp. 97–100.

Gnaldi, M. (2017). A multidimensional irt approach for dimensionality assessment of standardised students tests in mathematics. *Quality & Quantity 51*(3), 1167–1182.

Gudeva, L. K., V. Dimova, N. Daskalovska, and F. Trajkova (2012). Designing descriptors of learning outcomes for higher education qualification. *Procedia-Social and Behavioral Sciences 46*, 1306–1311.

López Lamezón, S., R. Rodríguez López, L. M. Amador Aguilar, and L. M. Azcuy Lorenz (2018). Social significance of a virtual environment for the teaching and learning of descriptive statistics in medicine degree course. *Humanidades Médicas 18*(1), 50–63.

Noventa, S., A. Spoto, J. Heller, and A. Kelava (2019). On a generalization of local independence in item response theory based on knowledge space theory. *Psychometrika 84*(2), 395–421.

Rasch, G. (1960a). Probabilistic models for some intelligence and attainment tests. studies in mathematical psychology. *Copenhagen: Danmarks Paedagogiske Institut*.

Rasch, G. (1960b). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

The weight of words: textual data versus sentiment analysis in stock returns prediction

Il peso delle parole: dati testuali versus sentiment analysis per predire i rendimenti azionari

Riccardo Ferretti and Andrea Sciandra

Abstract The focus of this paper is to understand whether the words contained in a text corpus improves the explained variance of stock returns better than the use of the polarity of the same texts, obtained through a sentiment analysis using a generic ontological dictionary. The empirical analysis is based on the content of a weekly column in the most important Italian financial newspaper, which published past information and analysts' recommendations on listed companies. The use of textual data clearly increases the explained variance of stock returns but, through comparisons between data mining techniques, we observed minor differences in terms of MSE, by adding a selection of specific terms as features. In this context, the text mining approach proved to be very useful to improve the explanatory power of forecasting models, while it emerged the limited explanatory power of an automatic sentiment analysis based on a generic lexicon.

Abstract *Il focus di questo contributo è capire se le parole contenute in un corpus di testi migliorano la varianza spiegata dei rendimenti azionari rispetto all'uso della polarità degli stessi testi, ottenuta mediante una sentiment analysis utilizzando un dizionario ontologico generico. L'analisi empirica si basa sul contenuto di una rubrica settimanale del più importante quotidiano finanziario italiano, che ha pubblicato informazioni già note e raccomandazioni di analisti per una selezione di aziende. L'utilizzo di dati testuali chiaramente migliora la varianza spiegata dei rendimenti azionari ma, mediante il confronto tra diverse tecniche di data mining, abbiamo osservato modeste differenze in termini di MSE, quando abbiamo aggiunto alle variabili esplicative una selezione di parole. In questo contesto, l'approccio del text mining ha mostrato la sua utilità per migliorare la potenza esplicativa in modelli di previsione, mentre è emersa la scarsa capacità esplicativa di una sentiment analysis automatica basata su un dizionario ontologico generico.*

Key words: Text Mining, Sentiment Analysis, Stock Returns, Data Mining

[1] Riccardo Ferretti, Department of Communication and Economics, University of Modena and Reggio Emilia; email: riccardo.ferretti@unimore.it

Andrea Sciandra, Department of Communication and Economics, University of Modena and Reggio Emilia; email: andrea.sciandra@unimore.it

1 Introduction

This paper is part of a wider project whose main goal is to analyse the market reaction to the dissemination of analysts' recommendations published in print media, in order to explain the market reaction to 'buy' advice. Specifically, the hypothesis under consideration is that past analysts' recommendations induces abnormal movements in stock prices and returns.

Previous research found a positive market reaction to the publication of the past information when analysts grade the stock as a good opportunity. In particular, Cervellati et al. (2014) showed that the asymmetric market reaction supports the Barber and Odean (2008) Attention-Grabbing Hypothesis (AGH), assuming that naïve investors' behaviour affects the market. Therefore, AGH predicts positive and significant abnormal returns for positively recommended stocks and no reaction for negative ratings. In other words, the market reaction is motivated by an attention-grabbing phenomenon, because only the publication of positive recommendations induces a significant (positive) price movement.

Moreover, previous literature showed that investor mood varies systematically across calendar months and weekdays, with possible shifts in investor attention (Hirshleifer et al., 2020), affecting financial decision-making and asset prices. Hirshleifer et al. (2020) have also documented the 'day of the week' effect, which would explain how aggregate stock markets tend to do better at the end of the week than at the beginning of the week. In this regard, they found consistent results with the mood-based theory, by documenting several mood recurrence and reversal effects across calendar months and weekdays.

Our empirical analysis is based on the content of two similar weekly columns in the most important Italian financial newspaper, which published past information and analysts' recommendations on listed companies. One, named 'The Stock of the Week' appears on Saturday and the other, named 'Letter to the investor" appears on Sunday. It's important to stress that the these columns have the same author, the same content (past balance sheet and P&L data; single analyst recommendations, consensus forecasts, company's profile), and their characteristics have remained unchanged during our observation period.

The focus of this paper is to understand whether the words contained in this particular text corpus improves the explained variance of stock returns better than the use of the polarity of the same texts, obtained through a sentiment analysis using a generic ontological dictionary. Although with very different methods, this approach has long existed, as Li et al. (2014) have documented the use of textual analysis for studies on the influence of news on stock markets.

2 Data collection and processing

We collected all the 'Stock of the Week' and 'Letter to the investor' columns published from January 2005 to March 2010 that were devoted each week to a

domestic company listed on the Italian Stock Exchange. The final dataset consists of 214 records.

In order to explain the variance of the average returns (AR) on the stock exchange opening day following the publication of the column, we added some explanatory variables, including: the number of quoted analysts, the natural logarithm of the order size, the natural logarithm of market capitalization, a dummy variable indicating the presence of any confounding effect, a dummy variable indicating the day of the week of the column (Saturday or Sunday), the turnover ratio, the price-to-book and the past performance (applying the absolute value or not).

2.1 Pre-processing

The pre-processing phase of the column texts involves the following steps:

- text cleaning, in order to normalize text encoding, remove punctuation, handling capitalized words, etc.
- Stop words removing in Italian language (articles, prepositions, pronouns, etc.).

In this phase we used `TextWiller` package (Solari et al., 2019), one of the few R libraries for the Italian language.

Figure 1 shows a word cloud of the most frequent words after pre-processing.

Figure 1: Most frequent words (word cloud)

2.2 Textual analysis

A textual analysis allowed us to select some terms to be used as features in the predictive models and to calculate the tf-idf weighting (Salton & Buckley, 1988). This index should show how important a word is to distinguish each weekly column in our corpus. In particular, some simple text mining procedures allowed us to create the document-term matrix we used for the regressions. In this phase we chose not to

analyse the main multiword expressions (n-grams). We worked within the "bag of words" framework, as only tf-idf weighting was applied to the words and we don't assume any Natural Language Processing rule.

Following, we used an ontological dictionary to obtain a measure of polarity for each text. Among the few resources available for the Italian language, we chose the NRC[1] lexicon (Liu, 2012), through which we extracted a polarity score for each column. The new variable containing the sentiment of each text had 41 different levels and, with reference to the sign (polarity), a positive sentiment had been attributed to 90% of the texts (a text is considered positive if the number of "positive" words is higher than the number of "negative" words).

Since we had a big sparse document-term matrix (7840 terms), we decided to allow maximal sparsity at 75%, or 25% in relation to document frequency. The resulting matrix contains only 113 terms, since we removed all the terms which have at least a 75% of empty elements (terms occurring 0 times in a document).

3 Data analysis

Concerning the aims presented above, the data analysis phase involved applying two stepwise regressions - with and without the use of words as features - to compare the two approaches in terms of explained variance. Subsequently, some data mining methods (Hastie et al., 2009) will be tested to compare the predictive power of the basic model with only the polarity feature and the one with words as features in addition (113 new features). In both cases we expect, through the use of words, an improvement in terms of R-squared (also adjusted) and Mean Squared Error (MSE).

We chose the following fitting techniques that can lead to better predictive accuracy and interpretability: Stepwise regression, Principal Components Regression (PCR), Partial Least Squares (PLS), Elastic-net, and Lasso. We estimated by cross-validation the number of components in the subset selection procedures (PCR and PLS) and the regularisation parameter in the shrinkage method (Elastic-net).

The dataset was randomly divided into two parts: a training set, used to fit the models, made up of a sample of 75% of the observations, and a validation set, used to estimate the prediction error for model selection, made up of the remaining 25% of the observations.

[1] The NRC lexicon is a list of words and their associations with eight emotions and two sentiments: negative and positive. This lexicon includes sentiment values for 13,901 words and has translations for just over 40 languages (see https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm).

4 Results

A summary of the performances of the two stepwise regressions is reported in Table 1. The R-squared indices showed a remarkable increase (about 2.5 times) of explained variance by introducing textual variables among the features (21 out of 113 terms were selected in the second model).

Table 1: Explained variance of the two models

Stepwise regression model	***R-squared***	***Adjusted R-squared***	***F-statistic***	***p-value***
M1: basic explanatory variables	0.1901	0.1667	8.099	7.178e-08
M2: adding text variables	0.4793	0.4005	6.082	7.489e-15

The variable that defines the polarity of the texts of the column was not significant in the basic model, while it was significant at 5% ($p = 0.044$) with a negative sign (estimated coefficient = -0.048) in the model that included the terms.

Instead, through comparisons between data mining models, we observed minor differences in terms of MSE, as shown in Table 2.

Table 2: Performances of predictive models

Technique	***MSE (basic features)***	***MSE (adding textual data)***
Stepwise	5.003575	9.434886
PCR	5.293135	5.293065
PLS	5.285215	5.284997
Elastic-net	4.788059	4.344863
Lasso	4.477144	4.702495

Lasso showed the lowest MSE for the basic model, while Elastic-net had the best performance for the model including textual variables, although the difference in terms of MSE is limited in favour of the second (4.48 vs 4.34).

5 Conclusion

The text mining approach has proven to be very useful to improve R-squared with respect to our dependent variable, the average returns. In addition, the poor ability to be a good predictor of a sentiment feature, automatically obtained from a generic lexicon, has emerged. In these cases, it would be preferable to use a supervised learning method with human tagging (Ceron et al., 2017) or, at least, a thematic lexicon.

If we observed a clear improvement in terms of R-squared, data mining techniques did not show substantial improvements in terms of MSE, by introducing some terms as features.

The use of a generic lexicon is surely a limitation for this work, as well as the presence of a quite small sample (214 records), which may affect the stability of the estimates for the data mining methods, as we were sampling just over 50 cases (training set) to compare the accuracy of various regression techniques.

References

1. Barber, B.M. and Odean, T.: All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors, The Rev. of Financial Stud., 21,785–818 (2008) doi: 10.1093/rfs/hhm079
2. Ceron, A., Curini, L., Iacus, S.M.: Politics and Big Data: Nowcasting and Forecasting Elections with Social Media. Routledge, New York (2017) doi: 10.1080/23248823.2019.1619298
3. Cervellati, E.M., Ferretti, R., Pattitoni, P.: Market reaction to second-hand news: Inside the attention-grabbing hypothesis. Appl. Econ, 46(10), 1108-1121 (2014) doi: 10.1080/00036846.2013.866206
4. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer, New York (2009) doi: 10.1111/j.1751-5823.2009.00095_18.x
5. Hirshleifer, D., Jiang, D., Meng, Y.: Mood beta and seasonalities in stock returns. J. of Financial Econ. (2020) doi; 10.1016/j.jfineco.2020.02.003
6. Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., Chen, Y.: The effect of news and public mood on stock movements. Information Sci, 278, 826-840 (2014) doi: 10.1016/j.ins.2014.03.096
7. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. on Hum. Lang. Technol., 5(1), 1-167 (2012) doi: 10.2200/S00416ED1V01Y201204HLT016
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Process. & Manag. 24(5), 513-523 (1988) doi: 10.1016/0306-4573(88)90021-0
9. Solari, D., Sciandra, A., Finos, L.: TextWiller: Collection of functions for text mining, specially devoted to the Italian language. J. of Open Source Softw., 4(41), 1256, (2009) doi: 10.21105/joss.01256

Unsupervised Energy Trees: clustering with complex and mixed-type variables

Energy Trees non supervisionati: clustering con variabili complesse e di tipo misto

Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti

Abstract In the spirit of the recently developed and successful Object Oriented Data Analysis, we introduce Energy Trees as a model to perform classification and regression using complex and mixed-type covariates. Energy Trees may be seen as a generalization of Conditional Trees, where the testing procedures that characterize both variable selection and stopping criterion are here performed by means of Energy Statistics. The use of Energy Statistics allows to compare variables that need not to be defined on the same space, thus permitting to simultaneously model complex and mixed-type covariates. In this contribution we show how, adapting the main scheme offered in the literature to perform unsupervised learning using tree-like methods, Energy Trees can also be used to perform clustering on structured and mixed-type data, giving rise to the proposed Unsupervised Energy Trees.

Abstract *Nello spirito della recente e popolare Object Oriented Data Analysis, presentiamo gli Energy Trees, un modello per fare classificazione e regressione con variabili complesse e di tipo misto. Gli Energy Trees sono una generalizzazione dei Conditional Trees, in cui i test che caratterizzano sia la selezione delle variabili che i criteri di stop sono implementati usando le Energy Statistics. L'utilizzo di queste ultime permette il confronto tra variabili che non sono definite sullo stesso spazio, consentendo di modellare contemporaneamente variabili complesse e di tipo misto. In questo contributo spieghiamo come, adattando lo schema principale presente in letteratura per fare apprendimento non supervisionato tramite alberi, gli Energy Trees possono essere utilizzati anche per fini di clustering su dati strutturati e di tipo misto, dando origine alla proposta di Energy Trees non supervisionati.*

Key words: Object Oriented Data Analysis, Statistical Learning, Decision Trees, Clustering

---

Riccardo Giubilei
Sapienza University of Rome, Rome, e-mail: riccardo.giubilei@uniroma1.it

Tullia Padellini
Imperial College London, London, e-mail: t.padellini@imperial.ac.uk

Pierpaolo Brutti
Sapienza University of Rome, Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction

In many data analysis contexts, the quantities of interest are complex objects that live in non-Euclidean spaces. Examples include curves, graphs, shapes, images and strings. In these cases, one might be tempted to represent data using feature vectors that live in the Euclidean space in order to apply standard statistical techniques. However, it is not always easy to find a natural representation of this kind: sometimes the choice is made arbitrarily, other times it is not possible at all. For these reasons, the framework called *Object Oriented Data Analysis* (OODA) has been introduced with the idea of directly analysing the original data objects, however complex they can be. In the spirit of this framework, we define *Energy Trees* as a model capable of performing supervised classification and regression using complex and mixed-type variables as covariates. In this work, we show how Energy Trees can also be used for unsupervised classification (i.e. clustering) purposes, giving rise to the *Unsupervised Energy Trees* model.

2 Motivation

The need of analysing arbitrarily complex variables is increasing to the point that new and unifying frameworks of data analysis become necessary [24, 10]. Among these, Object Oriented Data Analysis (OODA), in the sense defined by [24], provides a thought-provoking approach to deal with populations of complex objects. The statistical scientific community has shown renewed and growing interest in OODA since the overview proposed in [12].

Object Oriented Data Analysis is defined as the *statistical analysis of populations of complex objects* [12]. In this framework, the most basic elements to be analyzed are called "data objects". The core idea of OODA is that *by analyzing data objects directly we avoid loss of information that occurs when data objects are transformed into numerical summary statistics* [11]. An additional advantage of this approach is that the arbitrary choices necessary to translate complex objects into feature vectors are avoided.

OODA has been first applied to tree-structured objects [24], and then has become particularly successful in the field of Functional Data Analysis [16, 17, 13]. Several types of data objects have been explicitly analyzed within this framework, e.g. trees [2, 3, 19], graphs [11, 8], persistence diagrams [4, 14], shapes [7], manifolds (see [7] for references) and even sounds [23].

Despite the wide variety of applications, so far contributions within the OODA framework have been focusing only on single-type data objects. Among other consequences of this, many of the techniques used to tackle data objects are domain-specific and do not favour cross-pollination between different fields. Attempting to fill this gap, we introduce Energy Trees, a model for classification and regression using complex variables that perfectly fits in the OODA framework, with the additional advantage of accommodating variables of mixed type. We believe that this

possibility is an important innovation within the OODA framework, widening and enhancing its applicability.

Since classification and regression are not the only tasks one may want to perform on data objects, in this work we focus on how Energy Trees can be used in order to achieve unsupervised classification, i.e. clustering. We call the resulting model *Unsupervised Energy Trees*.

3 Energy Trees

The building blocks upon which Energy Trees are founded are Conditional Trees and Energy Statistics. For this reason, before passing onto describing the Energy Trees model, each of them will be shortly introduced.

Conditional Trees [9] are a special class of recursive binary partitioning models where a constant model is fitted in each region of the resulting partition. Conditional Trees address two well-known shortcomings of traditional trees, namely overfitting and a selection bias towards covariates with many possible splits, embedding a class of conditional inference statistical tests known as *permutation tests* [20] to perform the splits. The conditional distribution of statistics measuring the association between the response and the covariates is used to solve the selection bias problem. Moreover, multiple test procedures are performed to establish whether there is significant association between any of the covariates and the response, and, if not, the recursion stops. In other words, the overfitting problem is addressed by construction, without necessitating post-pruning techniques. Therefore, Conditional Trees do not suffer from the problems of other tree-like models, nonetheless it has been shown [9] that they can be implemented without comparative loss of prediction accuracy.

Energy Statistics are defined as functions of distances between statistical observations in metric spaces [21]. In this framework, Energy dependence coefficients are characterized by the important feature of measuring all types of dependence; e.g., the distance correlation coefficient equals zero if and only if the variables are independent [21]. In our context, it is worth to underline the implications of defining Energy Statistics as functions of *distances* between statistical observations. Indeed, based on this idea, when the variables (i.e. data objects) are complex, it is possible to use their real-valued nonnegative distances for inference. In other words, the power of this kind of statistics is that they focus on a single variable at a time, defining the distances between observations based on that variable, regardless of its nature. Once this is done for every variable, the resulting distances are compared, making the variables themselves comparable even if they are not of the same type.

Energy Trees may be seen as Conditional Trees where data objects of various types are made comparable using Energy Statistics. In particular, the statistical tests of association that characterize Conditional Trees are here performed using Energy tests of independence, that are based on Energy dependence coefficients. Subsequently, the algorithm of Energy Trees may be seen as an adaption of that of Conditional Trees, which is in turn the following:

1. test the global null hypothesis of independence between the response and all the covariates using a permutation test: if the null hypothesis is not rejected, stop;
2. select the covariate that shows the strongest association with the response;
3. search the best split point for that covariate using a permutation test;
4. partition the data;
5. repeat steps 1, 2, 3 and 4 for both of the new partitions.

where the global test in step 1 is performed separately testing the independence of the response with each covariate, and then the global null hypothesis is not rejected only if all the single null hypothesis are not rejected. To take into account this multiple testing, p-values are adjusted using Bonferroni correction.

In Energy Trees, the tests in steps 1 and 3 are implemented using Energy tests of independence, i.e. after evaluating the observations in terms of Energy Statistics, and in particular using distance correlation. However, the use of distance correlation is not straightforward, since it requires the definition of an appropriate distance for each type of complex variable. In addition to this, the search for the split point is not trivial for complex data objects. In the model, we include the possibility to choose between two alternative procedures to perform the split after selecting the most associated covariate. The first one is the translation of data objects into feature vectors, when possible; then, the feature vector that is more associated with the response is chosen for splitting, and the split point is searched in a traditional way with respect to that feature vector. The second one is clustering, where at each split two medoids are selected, and then data are partitioned minimizing the distances from the medoids.

Energy Trees are implemented in the R package called `etree`. Current version covers functional data, graphs and persistence diagrams, as well as numerical or nominal variables, as covariates.

4 Unsupervised Energy Trees

Energy Trees is a tree-structured model with a sound statistical foundation where data objects are complex and possibly mixed. Although it was initially tailor-made only for performing supervised classification and regression, it can be used also for unsupervised classification purposes. To achieve this goal, we follow and adapt the scheme of Unsupervised Random Forests (URF) [18].

URF are based on a simple but clever intuition: if the data are characterized by any structure, it should be recognizable from a randomly generated version of itself [18]. As brilliantly explained by [1], following this idea

> a synthetic dataset is randomly generated from the original dataset, and together they form a two-class classification problem that is then modeled using classical (supervised) RF. If the subsequent analysis results in a meaningful classification model then the interpretation of a RF model, using a proximity matrix, is what gives it the ability of being used as an unsupervised learning method. This proximity matrix assesses the number of instances in which two cases (i.e. [...] observations) are adopted into the same child node of a RF tree.

These instances are then averaged across the number of trees, producing a proximity matrix where only original observations are considered. These proximity scores are then used to perform a powerful unsupervised classification analysis.

Unsupervised Energy Trees (UET) are a model that exploits the basic principles used by URF in order to perform unsupervised classification using Energy Trees. However, since Energy Trees are structurally different from CART learners, the scheme must be adapted in an appropriate way.

The main difficulty is that in Energy Trees the covariate used for splitting is selected using an independence test between the response and each covariate. On the other hand, in the URF scheme the response variable is artificially attached to the dataset, resulting in no association at all with any covariate. Thus, another solution must be found.

The core idea of Unsupervised Energy Trees is to exploit the spirit of the URF scheme, where classification is made possible since in the synthetic dataset any relationship between original variables is removed [1]. Therefore, in UET, variable selection is carried out performing a distance correlation test of independence between every two covariates, using only the observations from the original dataset; then, the couple of covariates with the strongest association is selected for splitting.

Consequently, the split point search is performed in a similar manner to traditional Energy Trees, i.e. using clustering. However, clustering is performed on the complete dataset formed by the original and the synthetic datasets, and for each of the two variables selected. Then, the two resulting partitions are compared in terms of node impurity with respect to the response variable, and the best partition in this sense is selected as the actual one.

The stop criterion is the same as in Conditional Trees and Energy Trees. Once it is met, observations in terminal nodes are classified using the majority voting rule. Finally, a proximity matrix is calculated using the classification predictions provided by trees in an ensemble method such as bagging or Random Energy Forest (i.e. a Random Forest of Energy Trees).

5 Conclusion

In this contribution we showed how Unsupervised Energy Trees can be obtained starting from Energy Trees. Energy Trees are a model which perfectly fits in and possibly enhances the Object Oriented Data Analysis framework. Up to now, Energy Trees were mainly used to perform supervised classification and regression with complex and mixed-type covariates. However, it is also possible to apply them in clustering problems where the data objects may be both traditional, i.e. numeric or nominal, and complex, such as functions and graphs. This is obtained at the cost of suitably adapting the scheme of Unsupervised Random Forests to Energy Trees. The resulting model, Unsupervised Energy Trees, may be applied in a number of cases where other classical clustering techniques cannot be used, as it may be the case when data objects are not only traditional.

References

1. Afanador, N.L., Smolinska, A., Tran, T.N., Blanchet, L.: Unsupervised random forest: a tutorial with case studies. J. Chemometrics 30, 232–241 (2016)
2. Aydın, B., Pataki, G., Wang, H., Ladha, A., Bullitt, E., Marron, J.S.: New Approaches to Principal Component Analysis for Trees. Stat Biosci 4, 132–156 (2012)
3. Barden, D., Le, H., Owen, M.: Central limit theorems for Fréchet means in the space of phylogenetic trees. Electron. J. Probab. 18(25) (2013)
4. Bendich, P., Marron, J.S., Miller, E., Pieloch, A., Skwerer, S.: Persistent Homology Analysis of Brain Artery Trees. The annals of applied statistics 10(1), 198–218 (2016)
5. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)
6. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, California (1984)
7. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis, with Applications in R (Second Edition). John Wiley and Sons, Chichester (2016)
8. Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S., Kolaczyk, E.D.: Hypothesis testing for network data in functional neuroimaging. Ann. Appl. Stat. 11 (2), 725–750 (2017)
9. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics 15(3), 651–674 (2006)
10. Jain, B.J., Obermayer, K.: Structure Spaces. J. Mach. Learn. Res. 10, 2667–2714 (2009)
11. La Rosa, P. S., Brooks, T. L., Deych, E., Shands, B., Prior, F., Larson-Prior, L. J., Shannon, W. D.: Gibbs distribution for statistical analysis of graphical data with a sample application to fcMRI brain images. Statist. Med. 35, 566–580 (2016)
12. Marron, J.S., Alonso, A.M.: Overview of object oriented data analysis. Biom. J. 56, 732–753 (2014)
13. Menafoglio, A., Secchi, P.: Statistical analysis of complex and spatially dependent data: A review of Object Oriented Spatial Statistics. EJOR 258(2), 401–410 (2017)
14. Patrangenaru, V., Bubenik, P., Paige, R.L., Osborne, D.: Challenges in Topological Object Data Analysis. Sankhya A 81, 244–271 (2019)
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo, California (1993)
16. Sangalli, L.M., Secchi, P., Vantini, S., Veneziani, A.: A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. Journal of the American Statistical Association 104(485), 37–48 (2009)
17. Shen, D., Shen, H., Bhamidi, S., Maldonado, Y.M., Kim, Y., Marron, J.S.: Functional Data Analysis of Tree Data Objects. J Comput Graph Stat., 23(2), 418–438 (2014)
18. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. J. Comput. Graphical Stat. 15(1), 118–138 (2006)
19. Skwerer, S., Bullitt, E., Huckemann, S., Miller, E., Oguz, I., Owen, M., Patrangenaru, V., Provan, S., Marron, J.S.: Tree-Oriented Analysis of Brain Artery Structure. J Math Imaging Vis 50, 126–143 (2014)
20. Strasser, H., Weber, C.: On the Asymptotic Theory of Permutation Statistics. Mathematical Methods of Statistics 8, 220–250 (1999)
21. Székely, G.J., Rizzo, M.L.: Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference 143(8), 1249–1272 (2013)
22. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing independence by correlation of distances. Ann. Statist. 35(6), 2769–2794 (2007)
23. Tavakoli, S., Pigoli, D., Aston, J.A.D., Coleman, J.S.: A Spatial Modeling Approach for Linguistic Object Data: Analyzing Dialect Sound Variations Across Great Britain. Journal of the American Statistical Association 114(527), 1081–1096 (2019)
24. Wang, H., Marron, J.S.: Object oriented data analysis: Sets of trees. Ann. Statist. 35(5), 1849–1873 (2007)

Using anchoring vignettes to adjust self-reported life satisfaction: a nonparametric approach leading to a Semantic Differential scale

Le anchoring vignettes nell'analisi della soddisfazione per la propria vita: un approccio non parametrico che conduce ad una scala sul Differenziale Semantico

Sara Garbin, Serena Berretta, Maria Iannario and Omar Paccagnella

Abstract Anchoring vignettes were introduced in the literature as a way to adjust self-evaluations from the Differential Item Functioning (DIF) problem, that leads incomparability of the individual answers based on an ordinal scale. In this article we introduce a new nonparametric approach to rescale and correct for DIF the original ratings using data collected by anchoring vignettes. Self-assessments based on the original Likert scale can be rescaled in a new variable based on a Semantic Differential scale, showing a larger number of ordinal values. An empirical study on life satisfaction in an old population (data are carried out from the Survey of Health, Ageing and Retirement in Europe) illustrates the importance of accounting for differences in the use of ordinal scales, by showing how our inferences about interpersonal comparisons may change as a function of the assumptions we consider.
Abstract *Le anchoring vignettes sono uno strumento introdotto in letteratura per correggere le valutazioni soggettive dal problema del Differential Item Functioning (DIF), che rende incomparabili le risposte degli individui derivanti da una scala ordinale. Questo articolo introduce un approccio non parametrico per riscalare e correggere dal DIF le valutazioni originarie sfruttando i dati raccolti per mezzo delle anchoring vignettes. In questo modo, le auto-valutazioni basate su una scala Likert possono essere riscalate in una nuova variabile costruita come una scala del Dif-*

Sara Garbin
Demetra opinioni.net s.r.l., Via Andrea Costa 34/C - Venice, e-mail: sara.garbin.1@studenti.unipd.it

Serena Berretta
Department of Mathematics, University of Genoa, Via Dodecaneso 35 - Genoa, e-mail: berretta@dima.unige.it

Maria Iannario
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò 22 - Naples, e-mail: maria.iannario@unina.it

Omar Paccagnella
Department of Statistical Sciences, University of Padua, Via Cesare Battisti 241 - Padua, e-mail: omar.paccagnella@unipd.it

ferenziale Semantico, che presenta un numero maggiore di valori ordinali rispetto alla variabile originaria. Un'applicazione empirica sulla soddisfazione verso la propria vita in una popolazione anziana, con dati raccolti attraverso l'indagine SHARE, illustra l'importanza di tener conto delle differenze nell'utilizzo di scale ordinali e mostra come i confronti interpersonali possano cambiare a seconda delle assunzioni adottate.

Key words: Anchoring vignettes, Ordinal data, Life satisfaction, Response scales.

1 Introduction

Life satisfaction is now widely used to evaluate people well-being [4]. Life satisfaction and other subjective psychological domains are usually measured by rating scales. However, whenever people are asked to rate themselves, the presence of individual heterogeneity leads respondents to interpret, understand or use the response categories for the same questions differently. This phenomenon is called *Differential Item Functioning* (DIF) by the psychometric literature [6].

[7] introduced the approach of the anchoring vignettes for adjusting self-evaluations by DIF and enhancing comparability across countries or groups. Anchoring vignettes are additional questions in the domain under investigation, where a hypothetical individual, manifesting just the trait of the concept of interest, is described. The same question is addressed to all respondents, but usually more anchoring vignettes are proposed, changing the level of severity, condition or context of the depicted scenario.

Anchoring vignette data may be analysed by means of a parametric and a nonparametric approach [8]. In particular, the idea of the nonparametric solution is to recode self-ratings relative to the answers collected in the corresponding set of anchoring vignettes. According to this nonparametric approach, our work aims at providing a novel mixture model to fit data coming from multi-ratings, recoding a variable based on a Likert scale to a new, DIF-free, variable based on a Semantic Differential (SD) scale [10]. This work is organised as follow: Section 2 introduces the new nonparametric solution dealing with anchoring vignettes. Section 3 briefly explains data used in the analysis proposed in Section 4.

2 The new adjusted measure

In the original nonparametric approach, self-evaluation is adjusted according to its relative position with respect to the set of the anchoring vignettes: self-rating is judged as better, equal or worse than the anchoring vignette ratings, regardless of the value of such evaluations. However, this solution implies the validity of the so-called *global ordering* assumption: the majority of the individuals belonging to each

group in the sample (i.e. people living in the same country) defines the same ranking of the anchoring vignettes as the one thought by the vignette designers according to the content of the vignettes (that is, the severity of the problem described in each scenario).

Our proposal for a DIF-free measure relies on the comparison of the self-evaluation with only one anchoring vignette. Assuming this vignette as the individual benchmark for the own response scale, the difference of the ratings between self-evaluation and the related anchoring vignette is computed. Self-assessments based on the original Likert scale can be rescaled in a new variable based on a Semantic Differential scale [10]. Indeed, the rescaled variable rates on a bipolar scale, where the position marked as "0" can be labelled as *neutral* (individual evaluates herself in the same way of the scenario depicted in the anchoring vignette). In addition to the standard nonparametric approach, our solution allows to calculate how much self-assessment is close or not to the benchmark described by the scenario. This is a typical feature of the SD scale, because it is able to measure both directionality of the reaction (i.e. positive vs negative) and intensity (i.e. slight vs extreme).

Let m the number of answer categories in the original self-reported and vignette questions and let the original Likert scale vary from 1 to m. According to the difference, our adjusted variable will assume a total of $2m-1$ values, from $-m+1$ to $(m-1)$. In a conventional ordinal way, the rescaled G-variable may be defined as:

$$G = m + d \tag{1}$$

where d is the value of the new variable according to the SD scale (that is, the magnitude of the difference between self- and vignette ratings).

This approach has several advantages with respect to the original nonparametric solution. First, it can be applied whenever anchoring vignettes are collected in a survey, regardless their number. This could be very important in longitudinal studies, where it is likely to collect just one anchoring vignette to retrieve information on the individual response scales varying over time. Second, there is no longer need of the *global ordering* assumption. Third, our solution transforms a Likert in a SD scale increasing the number of categories to be used for modelling the responses (as an example, if the original scale has 5 answer categories, the G-scale is composed by 9 values). This allows to obtain more information from the collected data.

3 Data

Data come from the second wave of SHARE (Survey of Health, Ageing and Retirement in Europe), collected in 2006/2007 [2]. For methodological details, see [3]. SHARE is a panel survey that collects detailed cross-national information on health, socio-economic status and family networks of people aged 50 and over from a large set of European countries. The sample is composed by 7,654 respondents, living in 11 countries (Belgium, Czech Republic, Denmark, France, Germany, Greece,

Italy, the Netherlands, Poland, Spain and Sweden). It is mainly characterised by female respondents (55.4%), aged 64.3 years on average, low educated and retired; however, a large cross-country heterogeneity is present. However, other than standard demographic information a wide set of explanatory variables is available, from socio-economic status (job status, income, wealth and so on) to health (chronic diseases, limitations with activities, cognitive abilities and so on).

The proposed self-reported question on life satisfaction is: "*How satisfied are you with your life in general?*"; an example of anchoring vignette (on the same domain) is: "*Carry is 72 years old and a widow. Her total after tax income is about 1,100 per month. She owns the house she lives in and has a large circle of friends. She plays bridge twice a week and goes on vacation regularly with some friends. Lately, she has been suffering from arthritis, which makes working in the house and garden painful. How satisfied with her life do you think Carry is?*". For both questions, respondents answered according to a 5-point Likert scale: 'Very dissatisfied'; 'Dissatisfied'; 'Neither satisfied, nor dissatisfied'; 'Satisfied'; 'Very satisfied'.

4 Results

In the second wave of SHARE, two anchoring vignettes on life satisfaction were collected. For this work, the one reported in Section 3 (Carry) was exploited, which results to be the best in terms of fitting when regressed on the original self-rating (AIC=16047.13 with respect to AIC=16093.52 for the other anchoring vignette). The G-variable was then computed. Frequency distributions of the original and the rescaled items are displayed in Figure 1.

When the observed variable Y is an m-category ordinal response variable, one of the candidate models to analyse this rating is the cumulative link model [9]. Alternatively, researchers may use the CUP model: introduced by [11] to account for uncertainty in the response, it is a mixture (Combination) of a discrete Uniform random variable U and a Preference random variable Y which mimics the standard cumulative models.

In this analysis, a CUP model is fitted to the data. Here we assume that $Y_i^* = j$ when $\alpha_{j-1} < Y_i^* \leq \alpha_j$, with latent regression for the preference part.

The estimate of the uncertainty parameter on the G-variable is close to one ($\hat{\pi} = 0.99$). Taking into account that the subject propensity to adhere to a well-structured response behaviour, rather than to a random choice, is modelled by mixing the preference components via the uncertainty parameter π, results reveal no uncertainty in the response process. Our rescaling approach has removed DIF from the collected data and the rescaled rating is an entirely reasoned assessment. As a consequence, we continue with our analysis of the standard cumulative model with proportional assumption (POM) [1]. Results for the final estimated model on the G-variable are reported in Table 1.

The number of statistically significant country dummies is large: controlling for individual characteristics, cross-country variability in life satisfaction reporting is

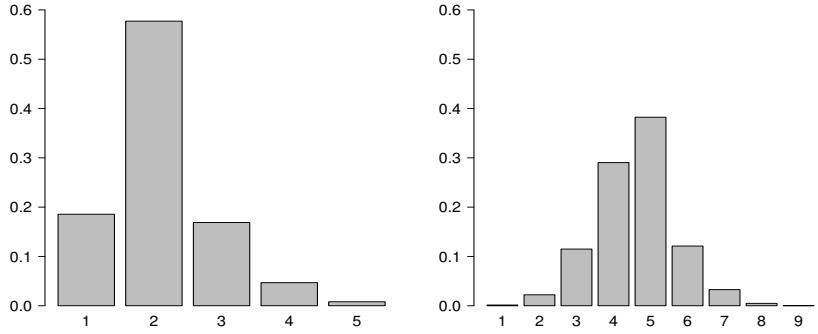


Fig. 1 Left: Frequency distribution of the original ordinal variable collected on a 5-point scale (1=Very satisfied; 5=Very dissatisfied). Right: Frequency distribution of the rescaled ordinal variable on a 9-point scale (1=High satisfaction; 9=High dissatisfaction).

Table 1 POM estimation of the rescaled life satisfaction variable ($k = 9$).

	Covariate	Satisfaction parameters	AIC
Thresholds	1|2	$\hat{\alpha}_1 = -7.959\,(0.359)$	21120.99
	2|3	$\hat{\alpha}_2 = -5.129\,(0.208)$	
	3|4	$\hat{\alpha}_3 = -3.232\,(0.196)$	
	4|5	$\hat{\alpha}_4 = -1.621\,(0.193)$	
	5|6	$\hat{\alpha}_5 = \ \ 0.326\,(0.193)$	
	6|7	$\hat{\alpha}_6 = \ \ 1.943\,(0.199)$	
	7|8	$\hat{\alpha}_7 = \ \ 4.015\,(0.251)$	
	8|9	$\hat{\alpha}_8 = \ \ 6.940\,(0.732)$	
Main model	DE	$\hat{\gamma}_1 = -0.317\,(0.065)$	
	SE	$\hat{\gamma}_2 = -0.745\,(0.093)$	
	NL	$\hat{\gamma}_3 = -0.823\,(0.089)$	
	FR	$\hat{\gamma}_4 = -0.544\,(0.104)$	
	DK	$\hat{\gamma}_5 = -0.228\,(0.070)$	
	CZ	$\hat{\gamma}_6 = \ \ 0.358\,(0.074)$	
	Age	$\hat{\gamma}_7 = -0.014\,(0.003)$	
	Single	$\hat{\gamma}_8 = \ \ 0.377\ \ (0.055)$	
	Retired	$\hat{\gamma}_9 = -0.295\,(0.064)$	
	Employed	$\hat{\gamma}_{10} = -0.348\,(0.067)$	
	Income	$\hat{\gamma}_{11} = -0.176\,(0.062)$	
	Wealth	$\hat{\gamma}_{12} = -0.029\,(0.007)$	

Note: (s.e. in parenthesis)

still important. As expected, life satisfaction is positively and significantly correlated with the socio-economic status. On the other hand, it is surprising the lack of any health covariates, given the well-known negative relationship between health status and satisfaction with one's own life [5]. However, this finding can provide an interesting interpretation of such relationship: the *G*-variable is a DIF-free measure of individual life satisfaction. Health variables may strongly affect the individual

response scales. Correcting for DIF, we are able to report all self-evaluation on a common scale, health status may have a role weaker than the socio-economic one.

Acknowledgements

This study was partially supported by the project 'Care, Retirement & Wellbeing of Older People Across Different Welfare Regimes' (CREW). This paper uses data from SHARE Wave 2 (DOIs: 10.6103/SHARE.w2.700). SHARE data collection has been funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N.211909, SHARE-LEAP: GA N.227822, SHARE M4: GA N.261982) and Horizon 2020 (SHARE-DEV3: GA N.676536, SERISS: GA N.654221) and by DG Employment, Social Affairs & Inclusion. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

References

1. Agresti, A.: *Analysis of Ordinal Categorical Data*, 2^{nd} ed. Hoboken: Wiley (2010)
2. Börsch-Supan, A.: *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2*, Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w2.700 (2019)
3. Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S. and Zuber, S.: Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE), *International Journal of Epidemiology*, **42**, 992–1001 (2013)
4. Bruni, L. and Porta, P.: *Economics and happiness. Framing the analysis*, Oxford: Oxford University Press (2005)
5. Easterlin, R.A.: Happiness of women and men in later life: nature, determinants, and prospects. In Sirgy M.J., Rahtz D.R. and Samli A.C. (eds), *Advances in Quality-of-Life Theory and Research*, Dordrecht: Kluwer Academic Publishers, 13–26 (2003)
6. Holland, P. and Wainer, H.: *Differential Item Functioning*, Hillsdale, NJ: Lawrence Erlbaum (1993)
7. King, G., Murray, C., Salomon, J. and Tandon, A.: Enhancing the validity and cross-cultural comparability of measurement in survey research, *American Political Science Review*, **98**, 191–207 (2004)
8. King, G. and Wand, J.: Comparing incomparable survey responses: New tools for anchoring vignettes. *Political Analysis*, **15**, 46–66 (2007)
9. McCullagh, P.: Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980)
10. Osgood, C.E., Suci, G.J. and Tannenbaum, P.H.: *The measurement of meaning*, Urbana: University of Illinois Press (1957)
11. Tutz, G., Schneider, M., Iannario, M. and Piccolo, D.: Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification*, **11**, 281–305 (2017)

Variable selection for robust model-based learning from contaminated data

Selezione di variabili nella stima robusta di modelli per dati contaminati

Andrea Cappozzo, Francesca Greselin, and Thomas Brendan Murphy

Abstract Several contributions to the recent literature have shown that supervised learning is greatly enhanced when only the most relevant features are selected for building the discrimination rule. Unfortunately, outliers and wrongly labelled units may undermine the determination of relevant predictors, and almost no dedicated methodologies have been developed to face this issue. In the present paper, we introduce a new robust variable selection approach, that embeds a classifier within a greedy-forward procedure. An experiment on synthetic data is provided, to underline the benefits of the proposed method in comparison with non-robust solutions.

Abstract *Recenti risultati in letteratura hanno dimostrato che l'apprendimento supervisionato migliora notevolmente quando si scelgono le variabili più rilevanti per la costruzione della regola discriminante. La presenza di valori anomali e di unità erroneamente classificate nel learning set può severamente minare la determinazione dei predittori rilevanti e sfortunatamente quasi nessuna metodologia affronta questo problema. Il presente contributo propone un nuovo approccio robusto, che incorpora un classificatore all'interno di un metodo incrementale di selezione delle variabili. Risultati simulativi mostrano i vantaggi del nuovo metodo, in comparazione con soluzioni non robuste.*

Key words: Variable Selection, Model-Based Classification, Label Noise, Outliers Detection, Wrapper approach, Impartial Trimming, Robust Estimation

Andrea Cappozzo, Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: a.cappozzo@campus.unimib.it, francesca.greselin@unimib.it

Thomas Brendan Murphy
School of Mathematics & Statistics and Insight Research Centre, University College Dublin e-mail: brendan.murphy@ucd.ie

1 Introduction

Nowadays, hundreds or thousands of variables on each sample are available in fields like chemometrics, computer vision, engineering and genetics, and many other scientific domains. Feature selection techniques have been introduced in data analysis, mainly aiming at building simpler models, easier to interpret by researchers/users, with shorter training times. Models based on the right selection of variables allow to avoid the *curse of dimensionality*, reduce overfitting, and prevent identifiability problems that may arise in high dimensional spaces. This has been known for a long time, as demonstrated by the specific literature reviews on the topic in the fields of machine learning, data mining, bioinformatics, genomic, and statistics. Surprisingly, the impact that outliers and wrongly labelled units cause on the determination of relevant predictors has received far less attention. Indeed, contaminated data can heavily damage a classifier performance [6], and most variable selection methods rely on the implicit assumption of dealing with an uncontaminated training set.

The present paper aims at filling this gap. We propose a new robust variable selection method for model-based classification, by embedding a robust classifier, recently introduced in the literature, in a greedy-forward stepwise procedure for model selection. Section 2 recalls the problem of variable selection in model-based discriminant analysis, and the Robust Eigenvalue Decomposition Discriminant Analysis (REDDA), and then introduce the robust variable selection technique. Section 3 presents the comparison of several feature selection procedures within a simulation study in an artificially contaminated scenario. A discussion of our results concludes the paper, outlying some remarks and future research directions.

2 Robust variable selection in model-based classification

Model-based discriminant analysis is a probabilistic framework for supervised classification, in which a classifier is built from a complete set of N learning observations (i.e., the training set):

$$(\mathbf{x},\mathbf{l}) = \left\{(\mathbf{x}_1,\mathbf{l}_1),\ldots,(\mathbf{x}_N,\mathbf{l}_N);\mathbf{x}_n \in \mathbb{R}^P,\, \mathbf{l}_n = \{l_{n1},\ldots,l_{nG}\}' \in \{0,1\}^G;\, n=1,\ldots,N\right\} \tag{1}$$

where $\mathbf{x}_n$ is a P-dimensional continuous predictor and $\mathbf{l}_n$ is its associated class label, such that $l_{ng}=1$ if observation n belongs to group g and 0 otherwise with, clearly, $\sum_{g=1}^{G} l_{ng} = 1\,\forall n = 1,\ldots,N$. We assume that the prior probability of class g is $\tau_g > 0$ and $\sum_{g=1}^{G}\tau_g = 1$. The gth class-conditional density is modeled with a P-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_g \in \mathbb{R}^P$ and covariance matrix $\boldsymbol{\Sigma}_g \in PD(P)$: $\mathbf{x}_n|\mathbf{l}_n = g \sim N_P(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$. Therefore, the joint density of $(\mathbf{x}_n,\mathbf{l}_n)$ is given by:

$$p(\mathbf{x}_n,\mathbf{l}_n;\boldsymbol{\theta}) = p(\mathbf{l}_n;\boldsymbol{\tau})p(\mathbf{x}_n|\mathbf{l}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) = \prod_{g=1}^{G}\left[\tau_g\phi(\mathbf{x}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right]^{l_{ng}} \tag{2}$$

where $\phi(\cdot;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ denotes the multivariate normal density and $\boldsymbol{\theta}$ represents the collection of parameters to be estimated, $\boldsymbol{\theta}=\{\tau_1,\ldots,\tau_G,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_G,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_G\}$. Eigenvalue Decomposition Discriminant Analysis (EDDA) is a family of classifiers developed from the probabilistic structure in (2), wherein different assumptions about the covariance matrices are considered. Particularly, EDDA is based on the following eigenvalue decomposition:

$$\boldsymbol{\Sigma}_g=\lambda_g\boldsymbol{D}_g\boldsymbol{A}_g\boldsymbol{D}_g^{'} \tag{3}$$

where $\boldsymbol{D}_g$ is an orthogonal matrix of eigenvectors, $\boldsymbol{A}_g$ is a diagonal matrix such that $|\boldsymbol{A}_g|=1$ and $\lambda_g=|\boldsymbol{\Sigma}_g|^{1/p}$. Allowing each parameter in (3) to be equal or different across groups a family of 14 patterned models arises. To protect parameter estimates against label noise and outliers, [1] introduced a robust version of EDDA, called REDDA, by means of the maximization of a *trimmed mixture log-likelihood* [4], in which an impartial trimming level γ is enforced in the estimation procedure.

The next step is therefore to include a robust variable selection procedure within REDDA. We proceed in a stepwise manner, by considering the inclusion of extra variables into the model, and also the removal of existing variables from the model, one at a time, conditioning on their discriminating power. We start from the empty set and then, in each step of the algorithm, we partition the learning observations $\mathbf{x}_n$, $n=1,\ldots,N$, into three parts $\mathbf{x}_n=(\mathbf{x}_n^c,x_n^p,\mathbf{x}_n^o)$, where:

- $\mathbf{x}_n^c$ indicates the set of variables currently included in the model
- x_n^p the variable proposed for inclusion
- $\mathbf{x}_n^o$ the remaining variables

To decide whether to include the proposed variable x_n^p, we compare the following two competing models:

- *Grouping* ($\mathscr{M}_{GR}$): $p(\mathbf{x}_n|\mathbf{l}_n)=p(\mathbf{x}_n^c,x_n^p,\mathbf{x}_n^o|\mathbf{l}_n)=p(\mathbf{x}_n^c,x_n^p|\mathbf{l}_n)p(\mathbf{x}_n^o|x_n^p,\mathbf{x}_n^c)$
- *No Grouping* ($\mathscr{M}_{NG}$): $p(\mathbf{x}_n|\mathbf{l}_n)=p(\mathbf{x}_n^c,x_n^p,\mathbf{x}_n^o|\mathbf{l}_n)=p(\mathbf{x}_n^c|\mathbf{l}_n)p(x_n^p|\mathbf{x}_n^r\subseteq\mathbf{x}_n^c)p(\mathbf{x}_n^o|x_n^p,\mathbf{x}_n^c)$

where $\mathbf{x}_n^r$ denotes a subset of the currently included variables $\mathbf{x}_n^c$. The Grouping model specifies that x_n^p provides extra grouping information beyond that provided by $\mathbf{x}_n^c$; whereas the No Grouping model specifies that x_n^p is conditionally independent of the group membership given $\mathbf{x}_n^r$. We consider $\mathbf{x}_n^r$ in the conditional distribution because x_n^p might be related to only a subset of the grouping variables $\mathbf{x}_n^c$ [3]. The differences between the two models are graphically illustrated in Figure 1. The model structure of $p(\mathbf{x}_n^o|x_n^p,\mathbf{x}_n^c)$ is assumed to be the same for both grouping and no grouping specification, and we let $p(\mathbf{x}_n^c,x_n^p|\mathbf{l}_n)$ and $p(\mathbf{x}_n^c|\mathbf{l}_n)$ be a normal density with parsimonious covariance structure. Additionally, we assume $p(x_n^p|\mathbf{x}_n^r\subseteq\mathbf{x}_n^c)$ to be a normal linear regression model, as a result from conditional multivariate normal means. The selection of which model to prefer is carried out employing a robust approximation to the Bayes Factor $\mathscr{B}_{GR,NG}$, given by the ratio between the integrated likelihood of the two competing models. Along the lines of [5], twice the logarithm of $\mathscr{B}_{GR,NG}$ can be approximated with

$$2\log(\mathscr{B}_{GR,NG})\approx BIC(Grouping)-BIC(No\ Grouping) \tag{4}$$

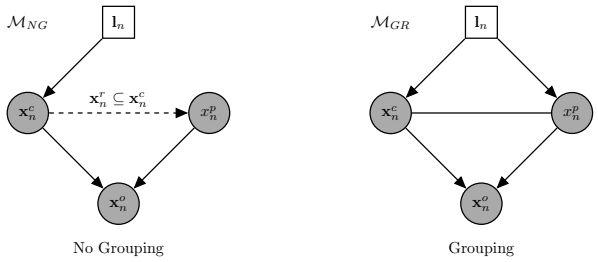


Fig. 1: Graphical representation of the Grouping and the No Grouping models

and a variable x_n^p with a positive difference in $BIC(Grouping) - BIC(No\ Grouping)$ is a candidate for being added to the model. A robust version of the BIC is employed here, for avoiding the detrimental effect that class and attribute noise might produce in the variable selection procedure. The Trimmed BIC (TBIC), firstly introduced in [4], is employed as a robust proxy for the quantities in (4). Let us define:

$$TBIC(Grouping) = \underbrace{2\sum_{n=1}^{N}\zeta(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G} l_{ng}\log\left(\hat{\tau}_g^{cp}\phi(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\mu}}_g^{cp}, \hat{\boldsymbol{\Sigma}}_g^{cp})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n)} + \\ - v^{cp} log(N^*) \tag{5}$$

$$TBIC(No\ Grouping) = \underbrace{2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G} l_{ng}\log\left(\hat{\tau}_g^{c}\phi(\mathbf{x}_n^c; \hat{\boldsymbol{\mu}}_g^{c}, \hat{\boldsymbol{\Sigma}}_g^{c})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, \mathbf{l}_n)} - v^c log(N^*) + \\ \underbrace{+2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\log\left[\phi\left(x_n^p; \hat{\alpha} + \hat{\boldsymbol{\beta}}' \mathbf{x}_n^r, \hat{\sigma}^2\right)\right]}_{2\times\text{trimmed log maximized likelihood of } p(x_n^p | \mathbf{x}_n^r \subseteq \mathbf{x}_n^c)} - v^p log(N^*). \tag{6}$$

The penalty terms v^{cp} and v^c indicate the number of parameters for a REDDA model respectively estimated on the set of variables $\mathbf{x}_n^c, x_n^p$ and $\mathbf{x}_n^c$; while v^p accounts for the number of parameters in the linear regression of x_n^p on $\mathbf{x}_n^r$. The 0-1 indicator functions $\zeta(\cdot)$ and $\iota(\cdot)$ identify the subset of observations that have null weight in the trimmed likelihood under the Grouping and No Grouping models, with $N^* = \sum_{n=1}^{N}\zeta(\mathbf{x}_n) = \sum_{n=1}^{N}\iota(\mathbf{x}_n)$. Accordingly, at each iteration of the procedure that leads to the final robust estimates, we discard the $\lfloor N\gamma \rfloor\%$ of the sample with the lowest contribution to the conditional likelihood, under the no grouping model. Once the Concentration step is enforced, the set of parameters $\{\alpha, \boldsymbol{\beta}, \sigma^2\}$ for the regression part is robustly estimated via ML on the untrimmed observations,

in which a stepwise method is employed for automatically choosing the subset of regressors $\mathbf{x}_n^r$.

After each addition stage, we make use of the same procedure described above to check whether an already chosen variable in $\mathbf{x}_n^c$ should be removed: in this case x_n^p takes the role of the variable to be dropped, and a positive difference in terms of TBIC implies the exclusion of x_n^p to the set of currently included variables. The procedure iterates between variable addition and removal stage until two consecutive steps have been rejected, then it stops. Notice that, whenever $\gamma = 0$, BIC and TBIC coincide and the entire approach reduces to the methodology described in [3].

3 Simulation study

The aim of this simulated example is to numerically assess the effectiveness of the new methodology, whilst investigating the effect that a (small) percentage of contamination has on standard variable selection procedures. We adopt the data generating process (DGP) in [3], and add some attribute and class noise to the original experiment. A total of $B = 100$ Monte Carlo (MC) experiments are conducted as follows. From the DGP outlined in [3], $N = 500$ units are generated and their group membership retained for constructing the training set; while $M = 5000$ unlabelled observations compose the test set. Subsequently, label noise is simulated by wrongly assigning 20 units coming from the fourth group to the third class. In addition, 5 uniformly distributed outliers, having squared Mahalanobis distances from $\boldsymbol{\mu}_g$ greater than $\chi^2_{3,0.975}$ $\forall g \in \{1,2,3,4\}$, are appended to the training set, with randomly assigned labels. These contaminations produce, in each MC replication, a total of 25 adulterated units, that account for slightly less than 5% of the entire learning set. We validate the performance of our novel method in correctly retrieving the relevant variables, the comparison being carried out considering the following methods:

- TBIC: new robust stepwise greedy-forward approach via TBIC
- SRUW: stepwise greedy-forward approach via BIC [3]
- SelvarMix: variable selection in model-based discriminant analysis with a regularization approach [2].

Once the important variables have been identified, the associated classifier (i.e., REDDA for the robust variable selection criteria, with trimming level $\gamma = 0.05$; and EDDA for the non-robust ones) is trained on the reduced set of predictors and the classification accuracy is computed on the test set. Lastly, for providing benchmark values on the relevance of feature selection, both EDDA and REDDA classifiers are also fitted on the original set with $P = 16$ variables. Table 1 and Figure 2 show that the misclassification error for TBIC is always lower than for non-robust procedures. As expected, the best prediction accuracy is obtained via the forward selection algorithm with TBIC selecting 3 variables. Interestingly, the EDDA classifier, coupled with (non-robust) variable selection via either SelvarMix or SRUW, shows on average a higher misclassification error than REDDA learned on the entire set of features. That is, the harmful effect of adulterated observations is increased

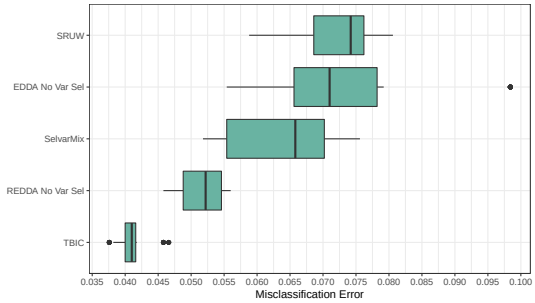


Fig. 2: Boxplots of the misclassification error, varying variable selection and model-based classification methods.

Table 1: Average misclassification errors, followed by their standard deviations.

Method	TBIC	REDDA NoVarSel	EDDA NoVarSel	SRUW	SelvarMix
Misc Error	0.0409	0.051	0.073	0.072	0.0639
	(0.0026)	(0.0026)	(0.0026)	(0.0037)	(0.0028)

by the presence of noisy variables, also shown by the poor performance of EDDA with no feature selection. Further research will be devoted to the development of a methodology that automatically assesses the contamination rate present in a sample, as the a-priori specification of the trimming level still remains an open issue in this field, particularly delicate for high-dimensional data.

References

[1] A. Cappozzo, F. Greselin, and T. B. Murphy. A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification*, aug 2019.

[2] G. Celeux, C. Maugis-Rabusseau, and M. Sedki. Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Advances in Data Analysis and Classification*, 13(1):259–278, 2019.

[3] C. Maugis, G. Celeux, and M. L. Martin-Magniette. Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, 102(10):1374–1387, 2011.

[4] N. M. Neykov, P. Filzmoser, R. I. Dimova, and P. N. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308, sep 2007.

[5] A. E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[6] X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, nov 2004.

Variable Selection in Text Regressions: Back to Lasso?

Selezione delle variabili nelle regressioni testuali: ritorno al Lasso?

Marzia Freo and Alessandra Luati

Abstract This study aims to explore the utility of text data as predictors in parametric regression models. Using the description of items on e-commerce platforms, we compare standard lasso, variable screening-based and randomization-based models in their ability to identify words really contributing to explain prices. Lasso models optimized for prediction attain high Predictive R^2, at the cost of using too many variables. The stability selection models provide an excellent trade-off between Predictive R^2 and the number of selected features. Sure Independence Screening and Lasso reach comparable results, when the number of features to select is set equal to the one of Stability Selection. Interestingly, Lasso behaves as well as computational intensive methods, when the number of selected variables is limited.

Abstract *Questo studio esplora l'utilità dei dati testuali nel ruolo di predittori in modelli di regressione parametrica. Usando la descrizione degli articoli nelle piattaforme di commercio online, vengono confrontati il lasso, i modelli basati sul variable screening e sulla randomizzazione, in termini di abilità di identificare le parole influenti per spiegare i prezzi. Il lasso ottimizzato per la previsione mostra un alto R^2 predittivo ma utilizza troppe variabili. Lo Stability Selection ha un eccellente trade-off tra R^2 predittivo e numero di parole selezionate. Sure Independent Screening e Lasso producono risultati confrontabili se il numero delle parole è fissato a priori pari allo Stability Selection. È interessante notare che il Lasso è competitivo con metodi computazionali intensivi quando il numero di parole selezionate è limitato.*

Key words: Text Mining, Lasso, Variable screening. Randomization methods, Online Prices.

[1] Marzia Freo, University of Bologna; email: marzia.freo@unibo.it
Alessandra Luati, University of Bologna; email: alessandra.luati@unibo.it

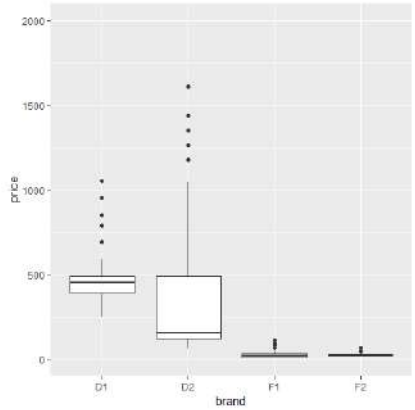
Problem and data

In the last decades, new technologies and the development of online social platforms have made available a large amount of text data. Consequently, several studies have been carried out with the aim of exploiting the informative content of text data, in favour of business and economic insights. Currently, texts are used as data in a variety of applications concerning authorship, sentiment, nowcasting, policy uncertainty, media slant, market definition, and other topics, and can be analysed by means of many different approaches (Gentzkow et al., 2019).

This study aims to explore the usefulness of text data as explanatory variables in parametric models. It investigates the analytics that allow one to exploit the content of unstructured text and its explanatory power. Are text data the new data for regression models? In what extent do they substitute the more often used, quantitative, data in economics? Which methodologies allow one to extract as much information as possible for text regressions?

We consider an application focused on the task of pricing fashion items available on producers' e-commerce platforms. The aim is to model regular prices, by avoiding prices on sale affected by markdown strategies which vary among brands. We use data scraped from UK e-commerce websites of four fashion producer brands. The brands are chosen to be heterogeneous with respect to the average price of items on sales (Figure 1): two of them are large fast-fashion chains selling at very low prices (F1 and F2), the remaining two (D1 and D2) found their business on design of enhanced fashion items. Fast fashion and enhanced design brands are discussed in Cachon and Swinney (2011). Each record collects information on price, category, brand and a field of description. The analysis is carried out for the knitwear category. Further categories have been analysed with coherent results, not shown here for sake of brevity. It is known that text is inherently high dimensional and sparse. Moreover, in this circumstance, it is reasonable to assume that only a small number of predictors contribute to the dependent variables. Before affording the issue of sparse modelling, we process some preliminary steps to map the raw text into a numerical matrix, the Document Term Matrix (DTM), whose cells indicate the count of the *j*-th word or token in the *i*-th document. Some text preparation operations are required in order to prepare data and to reduce meaningless dimensionality. First, we cancel out non-words elements (like numbers, punctuations, and proper names); then very common words are collected in a stop words list and deleted; finally, words are replaced by their roots through stemming. We respectively consider single words and single words plus bigrams, formed by all (unordered) pairs of words. In the end, the DTM contains 382 documents and 1245 features, of whom 229 are single words and 1016 are bigrams. Assuming that each single word or bigram represents a relevant feature for explaining the price of an item, our purpose is to identify a linear function relating prices to features. Note that, as prices are highly asymmetric it might be convenient to transform them (Box and Cox, 1964; Atkinson et al. 2019). However, we proceed with untransformed data as in our context "*approximate inference about the most meaningful parameters is clearly preferable to formally "exact" inference about parameters whose definition is in some way artificial*" (Box and Cox, 1964, pp. 214).

Figure 1. Distribution of prices by brand



Models and design of the study

Variable selection in regression analysis is an age-old problem in statistics, which currently encountered a renewed interest due to the increasing availability of high dimensional data. In this new setting, and in particular for text data, it is hard to disentangle the few meaningful variables that play a major role for interpretation purposes, from the redundant and noisy remaining variables. Among the large variety of variable selection techniques, the most applied method is certainly the Lasso (Tibshirani, 1996), both for its predictive performance and for its computational feasibility. Lasso does not come without limitations. For instance, it tends to select noisy variables if the tuning parameter is chosen to optimize prediction (Meinshausen and Bühlmann, 2006); it may exhibit poor variable selection results, if highly correlated predictors are present (Bach, 2008). Two main classes of methods have been proposed to overcome these limitations: variable screening and randomization. Variable screening aims to rank predictors by relevance; subsequently, the dimensionality will be reduced by selecting the highly ranked predictors and by running the standard variable selection Lasso (or its variants) over selected variables. A prominent method which belongs to this class is the Sure Independence Screening (from here in afterwards SIS, Fan and Lv, 2008). The rationale of randomization is to execute the Lasso, or another variable selection algorithm, over multiple training sets, generated by bootstrap or resampling, and to average over multiple results, so that the instability of running a selection algorithm only once can be overcome. Among the most significant randomization algorithms, we mention Bolasso (Bach, 2008), Stability selection (Meinshausen and Buhlmann, 2010), and Random Lasso (Wang et al., 2011).

The aim of this study is to identify which of these methods are more useful in finding the relevant variables to explain prices of items sold online. To this purpose, we compare some specifications of previously presented methods. To evaluate to

which extent true variables have been detected is a hard task, as the true model is unknown but in simulations. Hence, we generate 100 bootstrap replications from the original dataset. Each replicated dataset is divided into a training set and a hold-out sample. We define the selected variables as truly influential variables on the basis of 1) an a priori judgement, such as accepting features describing materials or decorations; 2) their explanatory power over the hold-out sample, measured in terms of Predictive R^2.

Main results and conclusions

For each bootstrap dataset, we first select variables over training data, using a pool of models, then we estimate a linear regression model by using previously selected variables as predictors, over the hold-out sample and finally we measure the Predictive R^2. The analyses have been carried out, for each generated dataset, by departing from the set of single words and bigrams, and the subset of only single words, respectively. Two variants of Lasso are presented: *lasso-min*, which is attained by minimizing the cross-validation error, and *lasso-1se*, which provides the most parsimonious model such that error is within one standard error of the error of *lasso-min*. To evaluate the performance of randomization, we choose the Stability Selection method, by imposing different thresholds of the selection probability. The results are very similar by changing the selection probability and here we present those attained with the thresholds 0.7, named *stabs*. As a screening method, we run the SIS algorithm and consider the performance produced by different number of selected predictors, from 1 to 15, labelled from *sis-k1* to *sisk15*. Finally, we evaluate the performance of the simple Lasso obtained by imposing different number of selected predictors, from 1 to 15, labelled from *lasso-k1* to *lasso-k15*. The main results, in terms of number of selected variables and Predictive R^2, are displayed in Fig.2 and 3.

We first observe that, in this type of application, the explanatory power of bigrams does not play a relevant role. Indeed, models based on bigrams perform only slightly better than corresponding models estimated over single words. Thus, we proceed by discussing results only for the analyses run over single words.

Lasso-min performs as expected: it always uses the highest number of predictors (in median around 55) to reach the highest Predictive R^2 (in median higher than 0.75). Lasso-1se is more parsimonious than lasso-min. It selects in median 15 predictors to explain 60% of price variability. Stability Selection produces in median a Predictive R^2 just lower than 0.5, lower than the ones of lasso, but using in median only 6 predictors. Focusing on the SIS method with the same median number of predictors, *P=6*, we note that it produces very similar results to the Stability Selection. The same pattern may be observed when the simple Lasso is used by fixing the number of selected variables to *P=6*.

Table 1 compares the six most frequently selected features over the one hundred bootstrap replications by the three methods, as well as the number of times in which they are selected. The column type indicates whether the feature may be classified as a “true” feature according to the a priori judgment. Both in term of persistence and

relevant variables Stability Selection outperforms the other methods, even if not heavily. In conclusion, in term of out-of-sample explanatory power, our findings show that both Stability Selection and SIS are able to improve the Lasso optimized through standard criteria finalized to prediction purposes. Moreover, it is worth noting that by limiting the number of selected variables, Lasso may attain comparable results.

Figure 2. Number of selected feature (left) and Predictive R^2, by method over bootstrap data – Single words (229 predictors)

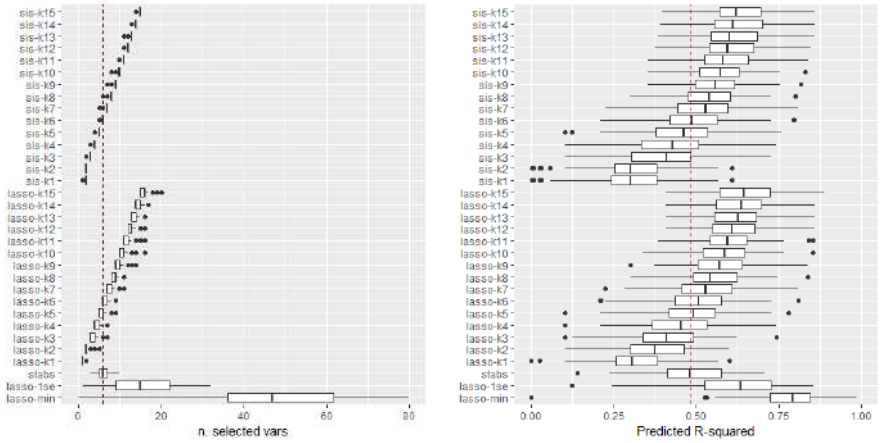
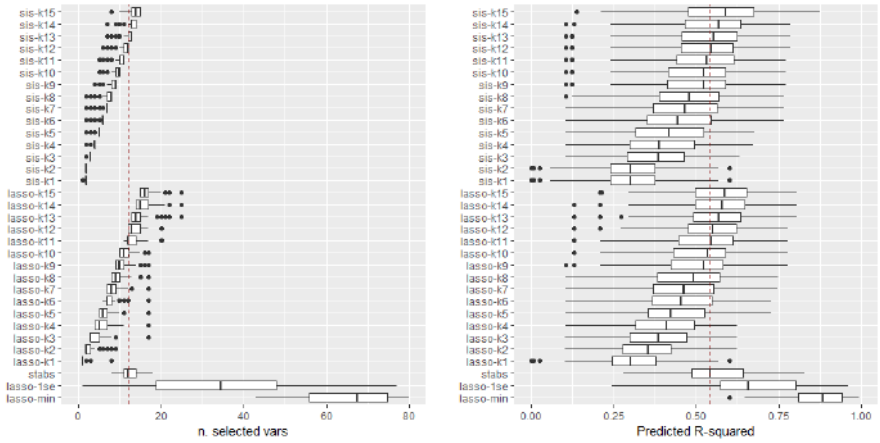


Figure 3. Number of selected feature (left) and Predictive R^2, by method over bootstrap data – Single words and bigrams (1245 predictors)



Concerning the ability to select relevant explanatory predictors, Stability Selection slightly outperforms other methods, which, notwithstanding, exhibit good performance. In our opinion, the most interesting finding is that Lasso behaves well when the number of selected variables is limited. So, our recommendation for future

works is to pose more emphasis on how to get better insights on stopping rules or Information Criteria.

Table 1: The six most frequently features selected over 100 datasets by method. Single words.

stabs			***sis-k6***			***lasso-k6***		
feature	***present***	***type***	***feature***	***present***	***type***	***feature***	***present***	***type***
cashmere	100	M	cashmere	100	M	cashmere	100	M
wool	72	M	turtleneck	71	D	turtleneck	61	D
turtleneck	62	D	silk	47	M	wool	55	M
blend	53	M	wool	44	M	beach	51	O
sweater	49	D	belted	39	O	belted	43	O
beach	40	O	beach	38	O	handknit	40	D

Legend of Type. M:= materials, D=: decorations; O:= other.

References

1. Atkinson A.C., Riani M., Corbellini A., (2019) The analysis of transformations for profit-and-loss data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* Article in Press.
2. Bach F.R. (2008) Bolasso: Model consistent lasso estimation through the bootsrap, *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, ACM Press, pp. 33–40.
3. Box G.E.P., Cox D.R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), pp.211-252.
4. Cachon G.P., R. Swinney, (2011) The Value of Fast Fashion: Quick Response, Enhanced Design, and Strategic, Consumer Behavior. *Management Science* 57(4):778-795. https://doi.org/10.1287/mnsc.1100.1303
5. Fan, J., and Lv, J. (2008), Sure Independence Screening for Ultra-High Dimensional Feature Space, *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911.
6. Gentzkow, M., B. Kelly, and M. Taddy. (2019) Text as Data, *Journal of Economic Literature*, 57 (3): 535-74. DOI: 10.1257/jel.20181020
7. Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34, 1436–1462.
8. Meinshausen, N. and Bühlmann, P. (2010), Stability selection, *Journal of the Royal Statistical Society, Ser. B*, 72, 417–473.
9. Tibshirani, R. J. (1996), Regression Shrinkage and Selection via the LASSO, *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288
10. Wang H. (2009) Forward Regression for Ultra-High Dimensional Variable Screening, *Journal of the American Statistical Association*, 104:488, 1512-1524, DOI: 10.1198/jasa.2008.tm08516

Web Usage Mining and Website Effectiveness
Il Web Usage Mining e l'Efficacia dei Siti Web

Maria Francesca Cracolici and Furio Urso

Abstract Using usage mining techniques, the study explores the effectiveness of a tourism business website focussing on the differences between desk and mobile user browsing behaviour. Results indicate important direction for a rethink of the website design.
Abstract *Nel seguente articolo sono state utilizzate tecniche di Usage Mining al fine di valutare l'efficacia del website di un'azienda di ricettività turistica distinguendo tra comportamento di navigazione di utenti PC e utenti mobile. I risultati ottenuti hanno evidenziato problematiche relative alla progettazione del sito in questione.*

Key words: online browsing behaviour, tourism 2.0, web usage mining

1. Introduction

Nowadays, the effectiveness of a website is a key factor in the competitiveness of a company. A successful website should be easy to use, graphically appealing and match the mission of the company. In order to evaluate effectiveness we have to understand how users interact with the site by applying Web Usage Mining (WUM) techniques which analyse web generated Big Data.
In recent years, because of advances in mobile technologies, empirical studies have been carried out on the browsing behaviour of web users and how it may be influenced by the physical characteristics of the access device, such as the ability to conduct transactions at any time and place (Sumita and Yoshii, 2010); or the fact that the limited manageability of the touchscreen in some websites makes it difficult to use long-term channels for m-commerce (Ipsos, 2015). Furthermore, since the use of reduced screens requires a large number of scrolling movements, the type and amount of information that the user is able to manage is limited (Ghose et al., 2012).

Maria Francesca Cracolici, University of Palermo; email: mariafrancesca.cracolici@unipa.it
Furio Urso, University of Palermo; email: furio.urso@community.unipa.it

Moreover, PCs differ when it comes to consulting different sources of information at the same time (Chae and Kim, 2004).
By using Data from the website PalermoTravel[1], a Sicilian company operating in the tourism sector which offers rental apartments and auxiliary tourism services, our study explores whether the browsing behaviour of PC users differs from that of mobile device users. Tourism sector is an interesting case study because consumers in Tourism 2.0 (Sparks et al., 2013) obtain information anytime and anywhere to better plan their trips; hence travel services must be customizable and web site constantly improved.
The rest of the paper is structured as follows. Section 2 includes data description and WUM methods used to explore the surfing behaviour of users. Section 3 presents empirical results and some concluding remarks.

2 Data and Methods

As mentioned above, data on access to the website of PalermoTravel has been used to explore the browsing behaviour of its users. The website has 5 thematic areas: Attractions, Accommodation, Services, Experiences and Events. In addition, the site provides general information on the company and its staff, on partners and access to bloggers and user reviews.
The dataset consists of web server log data in *IIS-W3Cex Extended* format which was collected for the months of September to December 2017. It consists of 2,487,802 lines divided into 17 log fields.
Web Usage Mining enable us to identify and analyse user behaviour patterns in the flow of clickstreams (a sequential list of access requests). These patterns are generally represented by sets of pages, objects or resources that have been viewed and used by user groups with common needs and interests. In the Web Usage Mining process data is collected, cleaned and divided into sets of user transactions that represent the activity of the individual users during visits to the site (Liu e Kaselj, 2007). In Table 2.1., we can see an example of a section of a Path Matrix, where each row is a user session:

Table 2.1: *Path Matrix example*

	V1	V2	V3	V4
1.0.153.1.0	Homepage			
1.129.97.1.0	Homepage	Accommodation	Events	
1.132.110.1.0	Homepage	Accommodation	Accommodation	Events
1.136.106.1.0	Homepage	Attractions		

Data is then processed in order to obtain the hidden patterns that reveal the behaviour of the users, and indices are calculated that are representative of users, sessions and site components; patterns and statistics are further processed and

[1] A pseudonym has been used, instead of using the real name of the company

aggregated to be used by Data Mining algorithms such as association rules and the Markov chain (Cooley, 2000).

Association rules analysis (Agrawal et al., 1993) in the context of WUM allows identification of groups of objects purchased or pages visited (more frequently in an observed period) providing a greater understanding of users' tastes.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items and $T = \{t_1, t_2, \ldots, t_n\}$ a set of transactions where t_i is a set of items such that $t_i \subseteq I$. An association rule is an implication of the form:

$$X \rightarrow Y, where\ X \subset I, Y \subset I\ and\ X \cap Y = \emptyset$$

Using support and confidence measures, the effectiveness of the rule can be assessed. The support of a rule is the percentage of transactions in *T* that contains $X \cup Y$; while the confidence of a rule is a percentage value that measures how many of the transactions in *T* which contain *X* also contain *Y*, and consequently it determines the predictability of the rule. Additionally, the lift measure can be used to check whether the association rules are effective in predicting user behaviour. In our case, the lift is the ratio between the probability that users will display a page in a target category, conditional on having visited pages in other categories and the probability that a page of another target category will be displayed.

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{P(Y)}$$

Lift values above (below) the unit indicate that the user's display of other page categories increases (decreases) the probability of moving on to the target category. An association rule mining algorithm, such as *apriori* algorithm (Agrawal and Srikant, 1994), determines the set of rules on the basis of minimum threshold values for confidence and support.

The Markov chain, however, enables us to explore the user navigation path within a website. This model is particularly suitable for modelling and predicting Data based on contiguous sequences of events; viz. each page-view can be seen as a precise state and the transition probability between two states can represent the probability that a user will move from one page to another during navigation. The Markov model allows one to predict the user probability of viewing a certain page given a certain route, to identify the most frequent navigation routes and to forecast subsequent visualizations conditional on the route.

A time homogeneous Markov chain of order *k* is a stochastic process $X^{(n)}$ that, at each time *n* takes state s_n from a finite set *S*, with probability that is independent of time *n* and that depends only on the states attained in the previous *k* times. This process can be described by transition matrices $P^{(k)}$, where the generic element $P_{i,j}^{(k)}$ is the probability of transitioning from state *i* at time $n-k$ to state *j* at time *n*.

Considering Usage Mining applications, the probability for a transition to either of the possible next states (pages) depends on user navigation behaviour, that can be

seen by considering the user's last *k* states (Moe 2003). Transition probability matrixes are estimated using Ching, Huang, Ng and Siu model (2013).

$$X^{(n+k+1)} = \sum_{i=1}^{k} \lambda_i Q_i X^{(n+k+1-i)} \;\; ; \;\; \sum_{i=1}^{k} \lambda_i = 1, \;\; \lambda_i \geq 0 \;\; \forall i$$

$$Q_i = \left[Pr_{j,h}^{(i)} \right]_{m \times m}$$

Where Q_i is an $m \times m$ lag-specific transition probability matrix and λ_i is the weight for lag i.

3 Results

Data was cleaned and pre-processed to obtain a suitable dataset consisting of 95,201 lines by IP address. 43,182 User Sessions were identified in order to follow the path of users within the site in the form of a sequence of pages viewed.

Table 3.1: *PC and Mobile access*

	Average page number	**Average page duration**	**Average session duration**
PC	6.488	60.129	182.606
Mobile	4.412	67.405	151.486

Table 3.1 shows that users accessing PCs view more pages on average and spend more time per session than users accessing via mobile. In contrast, mobile users spend more time on each page.
The association rules analysis, as regards PC access, reveals that many lift values are above 1 (Table 3.2). As for access via mobile device, we have only one significant lift value of 1.404, indicating an increase in the probability of viewing "Accommodation" after visiting the "Experience" area. Moreover, for access via mobile, we observe that few rules satisfy the threshold established for the support and confidence (Table 3.3.). Specifically, we applied the *apriori* algorithm by selecting the support threshold level at 5% and the confidence threshold level at 30%. We also considered the visualization of "Accommodation" as the rule *consequent*, distinguishing between PC and mobile users, using the R "arules" package.
It is noticeable that in case of PC access, viewing another area (except "Homepage" and "Attraction") increases the likelihood that the "Accommodation" area will be visited.
Users who access via mobile tend to view fewer areas, and joint visualizations decrease the probability of displaying a page of the "Accommodation" category. For both PC and Mobile access, the viewing the "Experiences" area seems to increase

the probability of exploring “Accommodation”, and indeed in the case of mobile access it is the only area that leads to viewing "Accommodation".

Table 3.2: *Association rules with Accommodation consequent for PC access data*

Rule	support	confidence	lift
{Events} => {Accommodation}	0.05	0.59	1.43
{Services} => {Accommodation}	0.06	0.556	1.337
{Info} => {Accommodation}	0.057	0.444	1.065
{Experiences} => {Accommodation}	0.097	0.596	1.433
{Homepage} => {Accommodation}	0.136	0.372	0.893
{Attractions} => {Accommodation}	0.187	0.392	0.94
{Experiences,Homepage} => {Accommodation}	0.06	0.851	2.044
{Attractions,Experiences} => {Accommodation}	0.065	0.758	1.82
{Attractions, Homepage} => {Accommodation}	0.092	0.67	1.61

Table 3.3: *Association rules with Accommodation consequent for Mobile access data*

Rule	support	confidence	lift
{Experiences} => {Accommodation}	0.051	0.531	1.404
{Homepage} => {Accommodation}	0.218	0.383	1.012
{Attractions, Homepage} => {Accommodation}	0.113	0.326	0.86

By applying the Markov chains, we tracked the paths of the users in probabilistic terms. The estimation of transition matrices performed on PC and mobile users did not show significant differences, so we clustered them all together.

Figure 3.1: *Transition probability matrixes: order 1 and 6*

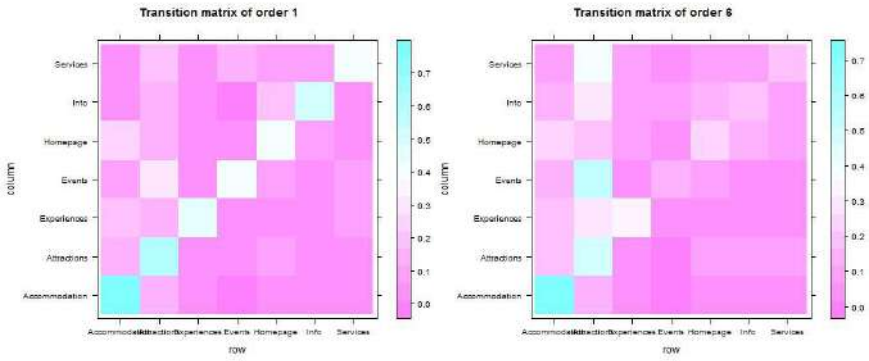


Figure 3.1 shows clearly that users tend to remain in the same subject area. Users who view a page in the "Accommodation" category have a 75% chance of moving to another page in the same category, a 12.9% probability of viewing the "Attractions" category and roughly a 6% probability for other areas.
Similar figures have been observed for users who view the "Attractions" category, while in the other categories the probability of moving to another page in the same area varies from between 38% and 54%.

The "Homepage" category is a little different. Those visiting the Homepage have a 23% chance of moving to "Accommodation " and a 16% probability of moving to "Attractions". The latter are the most visited areas coming from each of the categories. By increasing the order of transition matrices, for the "Accommodation" and "Attractions" categories, the situation remains almost the same.
Summing up, the association rules analysis showed that users seem to be more interested in obtaining general information about a holiday in Palermo rather than looking for accommodation. The obtained transition matrices have highlighted the monothematic nature of user paths; users tend to view pages sequentially in the same category.
Some critical considerations emerge. The homepage provides the user with a list of image links related to the "Attractions" area with the clear intent of stimulating curiosity. However, access to an attraction page leads to further attractions, as each page belonging to a thematic area contains image links limited to the same area. Since the core business of the company is its accommodation service, we suggest that the interest shown by users for the "Attractions" area pages be exploited to include image links of accommodation near the tourist attraction being described on each page.

References

1. Agrawal, R., Imieliski, T., Swami, A., (1993). "*Mining association rules between sets of items in large databases*", In Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD-1993).
2. Agrawal, R., Srikant., R., (1994). "*Fast algorithms for mining association rules",* In Proceedings of International Conference on Very Large Data Bases (VLDB1994).
3. Chae, M., Kim, J. (2004). "*Do size and structure matters for mobile users? An empirical stady of the effects of screen size, information structure and task complexity on user activities with standard web phones*", Behav. Inf. Technol. 23, 165-181.
4. Ching, WK, Huang, X, Ng, M, Siu, TK (2013). "Markov Chains: Models, Algorithms and Applications". 2nd edition. Springer-Verlag
5. Cooley, R., Mobasher, B., Srivastava, J. (2000). "*Web Mining: Information and Pattern Discovery on the World Wide Web"*, SIGKDD Exploration, Vol.1 Department of Computer Science and Engineering, University of Minnesota, Minneapolis.
6. Ghose, A., Goldfarb, A., Han, S.P. (2012). "*How is the mobile Internet different? Search costs and local activities"*, Inf. Syst. Res. 24, 614-631.
7. Ipsos (2015). *Ipsos Study for Paypal reveal Drivers and Barriers of Mobile Shopping Around the World.*
8. Liu, H. & Keselj, V. (2007). "*Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests",* Data and Knowledge Engineering, 61(2), 304-330.
9. Moe W (2003). "*Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream*", Journal of Consumer Psychology, 13(1–2), 29– 39.
10. Sparks, B., H. Perkins & R. Buckley (2013). "*Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behaviour*", Tourism Management, 39: 1-9.
11. Sumita, U., Yoshii, J. (2010). "*Enhancement of e-commerce via mobile accesses to the Internet"*, Electron. Commerce Res. Appl. 9, 217- 227.

Models and methods - Categorical, Ordinal, Rank Data

Aberration for the analysis of two-way contingency tables

L'aberrazione per l'analisi di tabelle di contingenza a due vie

Roberto Fontana and Fabio Rapallo

Abstract The aberrations are quantities usually computed in the context of Factorial Experiments. In this work, we introduce the use of the aberrations in the framework of contingency table analysis, and we propose a test of independence for 2×2 tables based on the aberrations. With a simple simulation study, we compare its performance with the well known chi-square test and Fisher's exact test.

Abstract *Le aberrazioni sono quantità usualmente definite per i piani fattoriali. In questo lavoro introduciamo l'uso delle aberrazioni nell'ambito dell'analisi delle tabelle di contingenza e studiamo un test di indipendenza per tabelle* 2×2 *basato su queste. Mediante uno studio di simulazione ne valutiamo le performance confrontandole con quelle del test chi-quadro e del test esatto di Fisher.*

Key words: Factorial designs, Independence test, Replicates, Aberration

1 Introduction

When two binary random variables X and Y are observed on a sample of size n, the collected data are summarized in a 2×2 contingency table. If we denote with -1 and $+1$ the levels of both X and Y, the joint counts are $n_{-1,-1}$, $n_{-1,1}$, $n_{1,-1}$, and $n_{1,1}$. Moreover, we denote with $n_{-1,+}$ and $n_{1,+}$ the marginal counts of the variable X, and similarly with $n_{+,-1}$ and $n_{+,-1}$ the marginal counts of the variable Y. In the literature there are several techniques to test the independence of the two variable, or to measure the extent of the connection. For a review on 2×2 tables, see [1].

Roberto Fontana
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, e-mail: roberto.fontana@polito.it

Fabio Rapallo
Dipartimento di Economia, Università di Genova, via Vivaldi 5, 16126 Genova, e-mail: fabio.rapallo@unige.it

In this work we consider the chi-square test of independence and the Fisher's exact test. The chi-square test of independence is based on the test statistic

$$C = \sum_{(i,j)\in\{-1,1\}^2} \frac{(n_{i,j}-\hat{n}_{i,j})^2}{\hat{n}_{i,j}} = \sum_{(i,j)\in\{-1,1\}^2} \frac{(n_{i,j}-\frac{n_{i,+}n_{+,j}}{n})^2}{\frac{n_{i,+}n_{+,j}}{n}} \tag{1}$$

which is compared with the chi-square distribution with 1 degree of freedom. When the sample size is small, the Fisher's exact test should be preferred. In such a test, the exact hypergeometric probability of all tables with the same marginal counts as those of the observed table are computed. Given the value $n_{-1,-1}$ of the cell in position $(-1,-1)$, the value of $n_{-1,-1}$ together with the margins completely determine all the counts, and the corresponding hypergeometric probability is

$$\begin{aligned} &p(n_{-1,-1};n,n_{-1,+},n_{+,-1}) = \\ &= \frac{n_{-1,+}!n_{1,+}!n_{+,-1}!n_{+,1}!}{n!n_{-1,-1}!(n_{-1,+}-n_{-1,-1})!(n_{+,-1}-n_{-1,-1})!(n_{+,1}-n_{-1,+}+n_{-1,-1})!}. \end{aligned} \tag{2}$$

Then, the p-value of the test is obtained by summing up all the probabilities of the tables more extreme than the observed one, i.e., with hypergeometric probability less than or equal to the probability of the observed table. Note that the Fisher exact test can be used in both the one-sided and in the two-sided setting. The idea behind the Fisher's exact has been extended to multi-way tables in the framework of Algebraic Statistics, see e.g. [6]. Another common measure of association is the log-odds-ratio $\ell = \log\left(\frac{n_{1,1}n_{-1,-1}}{n_{-1,1}n_{1,-1}}\right)$ which is related to independence because under independence one has $\ell = 0$.

In this work we use aberrations to analyse two-way binary contingency tables. Aberrations are usually computed for Fractional factorial designs, which are commonly used in Design of Experiments. More specifically, un the case of binary designs, an important object associated to a design is its Word-Length Pattern (WLP). The WLP is used to discriminate among different designs through the Minimum Aberration (MA) criterion, which is based on the sequential minimization of the WLP. The MA criterion was introduced in [4] for binary designs and then extended with the name of Generalized Minimum Aberration to non-regular multilevel designs in [7]. The WLP is computed using the aberrations of all the main effects and interactions.

In general a $I \times J$ contingency table can be interpreted as a factorial design with two factors with I and J levels respectively. With a slight abuse of notation, we denote the two factors again with X and Y. In this context the values $n_{i,j}$ in the table, represent the number of replicates of the point (i,j) in the design, $i = 1,\ldots,I$ and $j = 1,\ldots,J$.

In our study, given a 2×2 contingency table we use the aberrations computed for the corresponding factorial design to build a statistic that measures the departure from independence, i.e., the distance of the observed table to the table that would have been obtained under the hypothesis of independence.

2 The case of 2×2 contingency tables

Let us consider a 2×2 contingency table

$$\begin{pmatrix} n_{-1,-1} & n_{-1,1} \\ n_{1,-1} & n_{1,1} \end{pmatrix}$$

where the entries $n_{i,j}$ are non-negative integers, $i, j \in \{-1, 1\}$. There is a natural 2^2 factorial design corresponding to this table

X	Y	n
1	1	n_{11}
1	2	n_{12}
2	1	n_{21}
2	2	n_{22}

In the context of the design above, $n_{i,j}$ represents the number of replicates corresponding to the points (i, j), $i, j \in \{-1, 1\}$. A commonly used measure associated to factorial design are *aberrations*. Using Prop. 1 in [2], we can compute the aberrations of the design in the previous table. We obtain:

$$a_X = a_{10} = \frac{(n_{1,+} - n_{-1,+})^2}{n^2}, \quad a_Y = a_{01} = \frac{(n_{+,1} - n_{+,-1})^2}{n^2},$$
$$a_{XY} = a_{11} = \frac{(n_{-1,-1} + n_{1,1} - n_{-1,1} - n_{1,-1})^2}{n^2}.$$

In [2] it was proved that under independence the difference δ between the second-order aberration a_{11} and the product of the first-order aberrations $a_{10} \cdot a_{01}$, $\delta = a_{11} - a_{10} \cdot a_{01}$ is zero.

We use δ as a measure of the distance of the observed table to table would have been obtained under independence.

More specifically we consider the case in which all the margins $n_{-1,+}$, $n_{1,+}$, $n_{+,-1}$, and $n_{+,1}$, are supposed to be fixed. For example this situation occurs when a fixed number of participants for two categories ($n_{-1,+}$, and $n_{1,+}$) are observed until a predefined number $n_{+,-1}$, of events are observed.

In this case, under the null hypothesis of independence, $n_{-1,-1}$ follows an hypergeometric distribution, $n_{-1,-1} \sim \text{Hypergeometric}(n, n_{-1,+}, n_{+,1})$. Then the probability mass function underlying the Fisher's exact test is given in Eq. (2) mentioned in the previous section.

The two-sided p-value for the Fisher's exact test, p_F is computed as

$$p_F = P(|n_{-1,-1} - n_I| \geq |n^{\text{obs}}_{-1,-1} - n_I|)$$

where $n^{\text{obs}}_{-1,-1}$ is the observed value of $n_{-1,-1}$, n_I is the expected value of $n_{-1,-1}$ under independence, $n_I = \frac{n_{-1,+} n_{+,-1}}{n}$ and $n_{-1,-1} \sim \text{Hypergeometric}(n, n_{-1,+}, n_{+,-1})$.

In a similar way, using the statistic δ we define the corresponding p-value p_δ as

$$p_\delta = P(|\delta - \mu_0| \geq |\delta^{\text{obs}} - \mu_0|)$$

where δ^{obs} is the observed value of δ and μ_0 is the mean of δ under the null hypothesis of independence. We notice that δ is a function of $n_{-1,-1}$ and then its density f_δ can be written as

$$f_\delta(d;n,n_{-1,+},n_{+,-1}) = \sum_{n_{-1,-1}:\delta(n_{-1,-1})=d} p(n_{-1,-1};n,n_{-1,+},n_{+,-1})$$

where $p(n_{-1,-1};n,n_{-1,+},n_{+,-1})$ is again the hypergeometric probability defined in Eq. (2). We obtain that the mean of δ under the null hypothesis of independence is

$$\mu_0 = \sum_{a=\max(0,n_{-1,+}+n_{+,-1}-n)}^{\min(n_{-1,+},n_{+,-1})} \delta(a)p(a;n,n_{-1,+},n_{+,-1}).$$

We compare the behaviour of these two tests and the chi-square test through a simulation study. We consider 2×2 tables with sample size $n = 20$. In the first scenario we consider the independence model. We choose the marginal probabilities as $p_{-1,+} = 0.5$ and $p_{+,-1} = 0.4$. It follows $p_{1,+} = 0.5$ and $p_{+,1} = 0.6$. We compute the cell probabilities $p_{i,j}$ as $p_{i,+}p_{+,j}$ with $i,j \in \{-1,+1\}$. In the second scenario, using the same marginal probabilities we consider a model close to the lower Fréchet-Hoeffding bound of the Fréchet class of the bivariate Bernoulli with means equal to 0.5 and 0.6:

$$p_{-1,-1} = \varepsilon, p_{-1,1} = 0.5 - \varepsilon, p_{1,-1} = 0.4 - \varepsilon, p_{1,1} = 0.1 + \varepsilon$$

where ε is chosen equal to 0.001. In the third scenario we define an intermediate situation defining the cell probabilities as a the average of the cell probabilities defined in scenario 1 and 2. We obtain

$$p_{-1,-1} = 0.1 + \varepsilon/2, p_{-1,1} = 0.4 - \varepsilon/2, p_{1,-1} = 0.3 - \varepsilon/2, p_{1,1} = 0.2 + \varepsilon$$

For all scenarios we run $1,000$ simulations. In each simulation we randomly generate a contingency table as a sample from a multinomial distribution with parameters $n = 20$ and $p_{-1,-1}, p_{-1,1}, p_{1,-1}, p_{1,1}$ defined as described above. We use the `RandMultinomial` function of SAS/IML [5]. Then for each simulation we compute the p-values corresponding to the two-sided Fisher's exact test (p_F), to the chi-square test (p_χ), and to the δ statistic (p_δ).

For each scenario and for each test, varying the significance level α between 0.010 and 0.200 we determine the corresponding *observed significance level* α^{obs} as the ratio between the number of simulations where the p-value is less than or equal to α and the total number of simulations ($1,000$ for each scenario):

$$\alpha^{\text{obs}} = \frac{\#(p\text{-value} \geq \alpha)}{1,000}$$

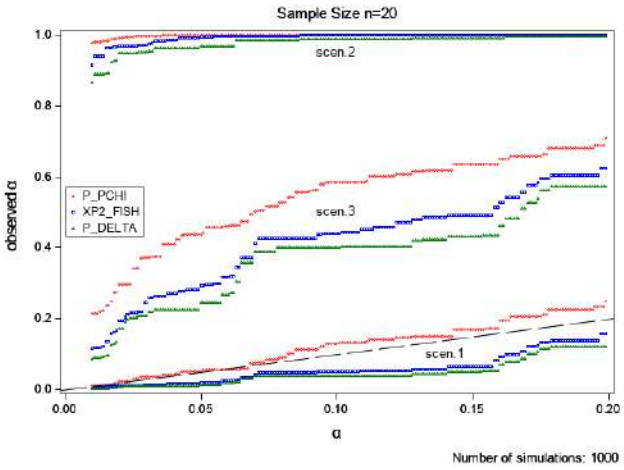


Fig. 1 Comparison between the significance level and the observed significance level. Sample size $n = 20$.

The results are summarized in Figure 1 . The scenario 1 (independence) shows that the Fisher's exact test and the delta-based test behaves similarly and they both look conservative. The chi-square test performs quite well if the significance level is chosen to be less than 7% approximately. The scenario 2 (absence of independence) exhibits a very good power for all the tests even if the delta-based test and the Fisher's exact test performs less well than the chi-square test. The scenario 3 shows an intermediate behaviour as expected.

The simulation study has been repeated reducing the total sample size to $n = 12$. The results are summarized in Figure 2. The overall performance decreases as expected but the observations made for the case $n = 20$ remain the same.

3 Extensions and future work

The use of aberrations in the context of contingency tables is new and looks promising. Nevertheless the simulation study shows that further research should be done to improve the performance of the test based on the δ statistic. Exploiting the results in [3], the results for the 2×2 case can be extended to the 3×3 case. In fact, the aberrations can be easily derived and straightforward computations show that, under independence, all aberrations of order two are equal to the product of the

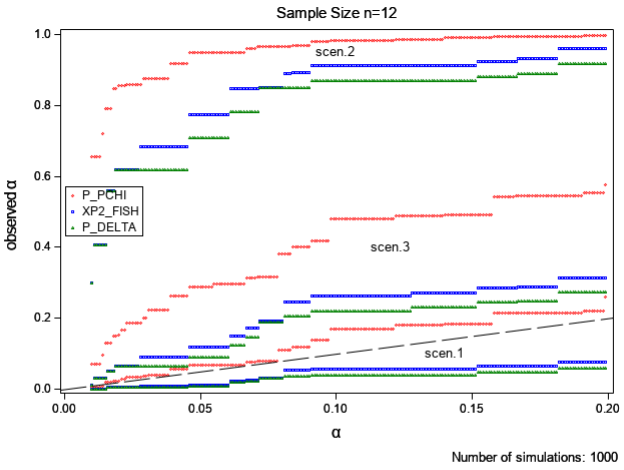


Fig. 2 Comparison between the significance level and the observed significance level. Sample size $n = 20$.

corresponding aberrations of order one. For instance: $a_{X^2Y} = a_{X^2} a_Y$. The procedure proposed here could be extended to the case of binary multi-way tables, where the aberrations are easy to compute and the patterns of dependence or independence among the variables can be difficult to detect. For larger tables, or for rectangular tables, the algebraic expressions of the aberrations become more complicated. One can overcome this difficulty by considering mean aberrations instead of aberrations, as suggested in [3], but this analysis is beyond the scope of the present work.

References

1. Agresti, A.: Categorical Data Analysis, 3rd edn. John Wiley and Sons, New York (2013)
2. Fontana, R., Rapallo, F.: Design of experiments, aberration and market basket analysis. In: SIS 2019-Smart Statistics for Smart Applications, pp. 873–878. Pearson (2019)
3. Fontana, R., Rapallo, F., Rogantin, M.P.: Aberration in qualitative multilevel designs. J. Statist. Plann. Inference **174**, 1–10 (2016)
4. Fries, A., Hunter, W.G.: Minimum aberration 2^{k-p} designs. Technometrics **22**(4), 601–608 (1980)
5. SAS Institute Inc: Sas/iml® 14.1 user's guide (2015)
6. Sullivant, S.: Algebraic Statistics. No. 194 in Graduate Studies in Mathematics. American Mathematical Society, Providence, RI (2018)
7. Xu, H., Wu, C.: Generalized minimum aberration for asymmetrical fractional factorial designs. Ann. Statist. **29**(2), 549–560 (2001)

An investigation of the paradoxical behaviour of κ-type inter-rater agreement coefficients for nominal data

Studio sul comportamento paradossale dei coefficienti Kappa di accordo inter-valutatore per dati nominali

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract A main criticism arising in agreement studies is the lack of robustness for some κ-type coefficients, that is their dependency on the frequency distribution of the ratings over classification categories. In order to investigate the statistical behaviour of three well-known κ-type coefficients, an extensive Monte Carlo simulation study has been conducted. The simulation study analyzed several scenarios, differing for sample size, rating scale dimension, number of raters, frequency distribution of ratings and pattern of agreement across raters.

Abstract *Una criticita' tipica degli studi di accordo riguarda la mancanza di robustezza di alcuni coefficienti Kappa, che deriva dalla dipendenza dei coefficienti dalle distribuzioni marginali di frequenza delle valutazioni fornite. Uno studio in simulazione Monte Carlo e' stato condotto per analizzare la robustezza di tre noti coefficienti Kappa. Lo studio in simulazione ha investigato la robustezza degli indici in scenari diversi tra loro per dimensione campionaria, numero di livelli della scala di valutazione, numero di valutatori, distribuzione di frequenza delle valutazioni e grado di accordo tra i valutatori.*

Key words: Inter-rater agreement, κ-type coefficients, Paradoxical behaviour

1 Introduction

In many fields, researchers and practitioners measure characteristics of interest by having two or more raters (e.g. clinicians, judges, psychologists, field experts) assigning scores to observed subjects, objects, or events. In these cases, a common issue is the reliability of the subjective classification process.

[1]Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; email: amalia.vanacore@unina.it

[2]Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; e-mail: mariasole.pellegrino@unina.it

To deal with this issue, a number of approaches have been proposed over the years and several coefficients have been introduced [15, 20, 22, 23]. For example, just to cite a few of them, the Intraclass Correlation Coefficient (ICC); the Kendall's W, a measure of rank correlation that is a generalization of Spearman's ρ correlation coefficient [16]; the Goodman and Kruskal's γ, based on the notion of concordance [10]; and the Krippendorff's α reliability coefficient [17].

A widespread approach relates the quality of subjective measurement procedures to the concept of inter-rater agreement, that is the agreement observed among two or more independent raters. The easiest way of measuring inter-rater agreement is to calculate the overall percentage of agreement; nevertheless, this measure does not take into account the agreement that would be expected by chance alone. A common solution is to adopt a coefficient belonging to the well-known family of the κ-type agreement coefficients, rescaled measures of the probability of observed agreement corrected with the probability of agreement expected by chance alone. As a matter of fact, different κ-type agreement coefficients have been proposed in the literature over the years changing for scale (nominal, ordinal and interval) and number of raters (two or more).

Despite their popularity, some κ-type agreement coefficients suffer from *paradoxical behaviour* due to two criticisms coming from the dependency on the marginal frequency distributions of the ratings over classification categories [6, 12]. The first paradox is the strong dependency on the *trait prevalence in the item population* which affects the observed marginal distribution of items across rating categories and thus the calculation of the chance agreement term. Specifically, fixing the observed agreement component, "unbalanced marginal totals produce higher values of κ than more balanced marginal totals" [8].
The second paradox is the dependency on the *marginal distribution of ratings over categories*: fixing the observed agreement component, the chance agreement obtained with symmetric marginal distribution is higher than the value obtained with asymmetric distribution and this makes the κ value considerably decrease [8]. These criticisms were firstly observed by Brennan and Prediger in 1981 [2] although they are widely known as *kappa paradoxes* as referred to by Feinstein and Cicchetti [4, 8].

The literature discussions on the paradoxical behaviour of κ-type agreement coefficients are limited to the simplest case of two raters and dichotomous (or at least nominal) data [3, 5, 11, 12, 21] whereas few research efforts have been devoted to κ-type agreement coefficients for inter-rater agreement [7, 13, 18, 19].

To the best of our knowledge, the most recent work is the study conducted by Quarfoot and Levine [19] who defined the robustness of an agreement coefficient as its "ability of giving roughly the same result for a fixed level of agreement irrespective of the frequency distribution used when correcting for chance agreement". Their simulation study aimed at investigating the robustness of five inter-rater agreement indices under a specific experimental condition (i.e. 8 raters classifying 100 items on a k-point ($4 \leq k \leq 11$) ordinal rating scale). In order to get more insight on this topic, this paper exploits the approach proposed in [19] to investigate and compare the robustness of three κ-type coefficients for inter-rater agreement with nominal data,

that is Fleiss' kappa K_F [9], *BP* coefficient [2] and Gwet's AC_1 [13]. Our simulation study analyzes several scenarios differing for sample size, rating scale dimension, number of raters, frequency distribution of ratings and pattern of agreement across raters.

The paper is organized as follows: in Section 2 the κ-type agreement coefficients are introduced; in Section 3 the simulation design is described and the main results are discussed; finally, conclusions are summarized in Section 4.

2 κ-type coefficients for inter-rater agreement

The κ-type agreement coefficients are rescaled measures of the observed proportion of agreement (p_a) corrected with the proportion of agreement expected by chance alone ($p_{a|c}$). All κ-type agreement coefficients, differing in scale type and number of raters, are formulated as follows:

$$\kappa = \frac{p_a - p_{a|c}}{1 - p_{a|c}} \tag{1}$$

The κ-type agreement coefficients range from -1 to +1: they are null when p_a equals $p_{a|c}$; when p_a is greater than $p_{a|c}$ the coefficients are positive; vice-versa, the coefficients are negative and can be interpreted as disagreement.
Different κ-type agreement coefficients have been proposed in the literature, sharing the same formulation of observed agreement but differing from each other in the notion of chance agreement and thus in the formulation of the $p_{a|c}$ term.

Let us consider the case of R raters classifying the same set of n items on a nominal k-point rating scale, with $k > 2$. Being r_{li} the number of raters classifying item l into category i, the proportion of observed agreement among multiple raters [14] is computed as follows:

$$p_a = \frac{1}{n} \sum_{l=1}^{n} \sum_{i=1}^{k} \frac{r_{li}\,(r_{li} - 1)}{R(R-1)} \tag{2}$$

The inter-rater agreement coefficient proposed by Fleiss [9] in 1971 (K_F) formulates the agreement expected by chance assuming that the items are classified into categories independently to each other:

$$p_{a|c}^{K_F} = \sum_{i=1}^{k} r_i^2 \tag{3}$$

where $r_i = 1/n \cdot \left[\sum_{i=1}^{k} r_{li}/R\right]$ is the propensity of classifying an item into category i. The $p_{a|c}^{K_F}$ depends on the propensity of classifying an item into each category so that the categories are not all equally probable but those with higher marginals are preferred over the others.

A solution to the above drawback is to assume that the probability of chance classification is uniform across categories:

$$p_{a|c}^{BP} = 1/k \tag{4}$$

The obtained coefficient is commonly known as Brennan-Prediger coefficient, hereafter *BP*, although it has been firstly suggested by Bennett et al. [1] when assessing the degree of inter-rater agreement on binary classifications.

The Agreement Coefficient (AC_1), proposed by Gwet in 2008 [13] formulates $p_{a|c}$ as the probability of the simultaneous occurrence of random rating in one replication (R) and agreement across replications (G):

$$p_{a|c}^{AC_1} = P(G \cap R) = P(G|R) \cdot P(R) = \frac{1}{k} \cdot \frac{\sum_{i=1}^{k} r_i(1-r_i)}{(k-1)/k} = \frac{1}{k-1} \sum_{i=1}^{k} r_i(1-r_i) \tag{5}$$

The conditional probability, $P(G|R)$, that the raters agree given that at least one of them has performed random ratings is formulated under the assumption of uniform distribution for chance measurements.

3 Simulation study

The simulated data sets are the classifications simultaneously provided by R raters on the same set of n items adopting a k-point nominal scale, considering that the generic i^{th} rater provides her/his classifications according to a given Frequency Distribution (FD) of ratings over classification categories and all the other $R-1$ raters agree with her/him according to a given pattern of agreement among raters (i.e. Agreement Distribution, AD).

The simulation study has been designed taking into account five multi-level factors: k, n, R, FD and AD. The factor k has 3 levels: $k = 3,4,5$ classification categories, the factor n has 3 levels: $n = 10,30,100$ items, the factor R has 3 levels: $R = 2,5,9$ raters, whereas the factors FD and AD have 6 and 1 level, respectively. FDs are all special cases of the beta-binomial distribution with different values of the shape parameters (a,b): FD 1 (0.25, 0.25), FD 2 (1, 1), FD 3 (2, 2), FD 4 (50, 50), FD 5 (25, 50), FD 6 (5, 50). The AD is a binomial distribution scaled on k and centred on R_i. For each combination of k, n, R, FD and AD, $r = 2000$ Monte Carlo data sets have been generated, for a total of $3 \cdot 3 \cdot 3 \cdot 6 \cdot 1 = 162$ simulated scenarios for each coefficient. The robustness of the κ-type agreement coefficients under study has been evaluated in terms of the mean agreement range $\overline{\text{AR}}$ over the 6 different FDs. The simulation algorithm has been implemented using Mathematica (Version 11.0, Wolfram Research, Inc., Champaign, IL, USA).

Simulation results, obtained for every combination of rating scale dimension, number of raters, sample size and agreement coefficient, are represented in Figure 1.

The simulation results show that, for each scenario (i.e. combination of rating scale size, number of rater and sample size), the robustness of the three κ-type coefficients against changes in the FD is generally comparable; the only exception is for $n = 10$ and $R = 2$, when K_F is slightly less robust than AC_1 and BP. Moreover, increasing the group dimension and the sample size makes the robustness improve; instead, with as few as $k = 5$ or less rating categories, the dimension less impacting on robustness seems to be the rating scale dimension.

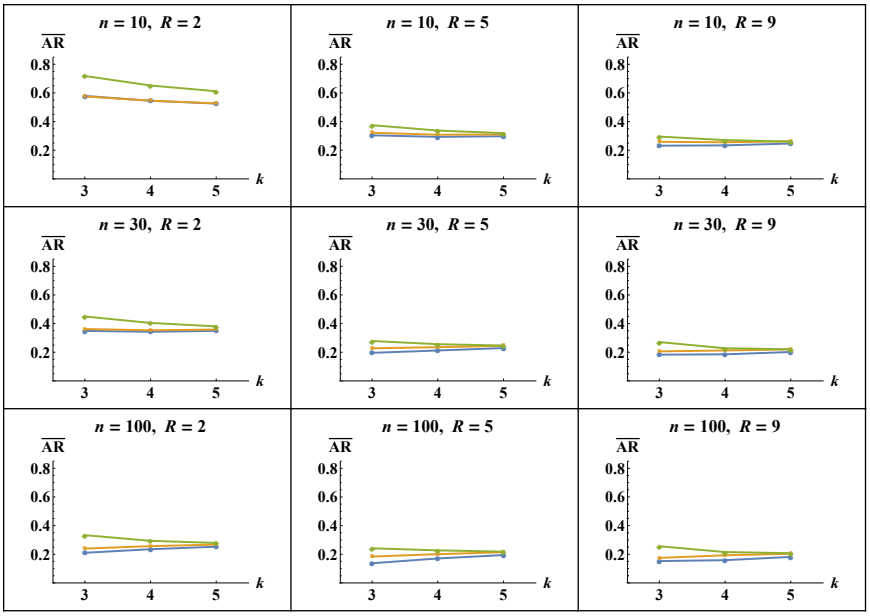


Fig. 1 Mean Agreement Range obtained for K_F (in green), BP (in blue) and AC_1 (in orange) as FD varies

4 Conclusions

The robustness of three κ-type inter-rater agreement coefficients for nominal data has been investigated via Monte Carlo simulation. Our findings suggest that the most robust κ-type coefficient is BP, assuming uniform chance agreement, followed by AC_1, formulating chance agreement assuming that at least one rater provides random ratings. Moreover, our results reveal that, beyond the approach adopted to correct for chance agreement, the statistical behaviour of inter-rater agreement coefficients is affected by the sample size of rated items and by the number of raters involved in the experiment.

A further development of forthcoming research aims at investigating the robustness of an inferential benchmarking procedure for the characterization of the extent

of agreement: whether and how the lack of robustness of the κ-type coefficients under study could affect the final agreement characterization.

References

1. Bennett, E.M., Alpert, R., Goldstein, A.: Communications through limited-response questioning. Public Opinion Quarterly **18**(3), 303–308 (1954)
2. Brennan, R.L., Prediger, D.J.: Coefficient kappa: some uses, misuses, and alternatives. Educational and psychological measurement **41**(3), 687–699 (1981)
3. Byrt, T., Bishop, J., Carlin, J.B.: Bias, prevalence and kappa. Journal of clinical epidemiology **46**(5), 423–429 (1993)
4. Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. Journal of clinical epidemiology **43**(6), 551–558 (1990)
5. Erdmann, T.P., De Mast, J., Warrens, M.J.: Some common errors of experimental design, interpretation and inference in agreement studies. Statistical methods in medical research **24**(6), 920–935 (2015)
6. Eugenio, B.D., Glass, M.: The kappa statistic: A second look. Computational linguistics **30**(1), 95–101 (2004)
7. Falotico, R., Quatto, P.: Fleiss' kappa statistic without paradoxes. Quality & Quantity **49**(2), 463–470 (2015)
8. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of clinical epidemiology **43**(6), 543–549 (1990)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378–382 (1971)
10. Goodman, L.A., Kruskal, W.H.: Measures of association for cross cliassification. Journal of the American Statistical Association **49**, 1732–1769 (1954)
11. Gwet, K.: Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Statistical Methods for Inter-Rater Reliability Assessment Series **2**(1), 1–9 (2002)
12. Gwet, K.: Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical methods for inter-rater reliability assessment **1**(6), 1–6 (2002)
13. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology **61**(1), 29–48 (2008)
14. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
15. Hallgren, K.A.: Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology **8**(1), 23–34 (2012)
16. Kendall, M., Gibbons, J.: Rank correlation methods, trans. JD Gibbons (5th edn ed.). Edward Arnold: London (1990)
17. Klaus, K.: Content analysis: An introduction to its methodology (1980)
18. Marasini, D., Quatto, P., Ripamonti, E.: Assessing the inter-rater agreement for ordinal data through weighted indexes. Statistical methods in medical research **25**(6), 2611–2633 (2016)
19. Quarfoot, D., Levine, R.A.: How robust are multirater interrater reliability indices to changes in frequency distribution? The American Statistician **70**(4), 373–384 (2016)
20. Vanbelle, S.: Agreement between raters and groups of raters. Ph.D. thesis, Universitede Liege, Belgique (2009)
21. Warrens, M.J.: A formal proof of a paradox associated with cohen's kappa. Journal of Classification **27**(3), 322–332 (2010)
22. Watson, P., Petrie, A.: Method agreement analysis: a review of correct methodology. Theriogenology **73**(9), 1167–1179 (2010)
23. Zapf, A., Castell, S., Morawietz, L., Karch, A.: Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC medical research methodology **16**, 93 (2016)

Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach

Analisi di dati soggetti a processi tipo faking-good mediante approcci di mistura

Luigi Lombardi, Antonio Calcagnì

Abstract In this short paper, we describe a novel approach to model and analyse ordinal data in the presence of faking behavior, namely the tendency of survey's participants to falsify their responses in order to achieve a particular purpose. The proposal relies on the use of two statistical approaches commonly used to analyse faking and preference data: the Sampling Generation by Replacement (SGR) and Combination of Uniform and Binomial distributions (CUBE). By combining both SGR and CUBE, we propose CRB (Combination of Replacement and Binomial distributions), where the response ordinal measure is modeled as a convex combination of the shifted-Binomial distribution and the Replacement distribution. Thus, the first component aims to represent the response measure unaffected by faking behavior whereas the second element of the linear model represents the result of a faking strategy. As for the CUBE models, CRB parameters are estimated via Maximum likelihood by means of the EM algorithm. Finally, an application to ordinal data is proposed to show how the CRB model can be used to analyse self-reported data potentially affected by faking behavior.

Abstract *Abstract in Italian*. Questo lavoro presenta alcuni risultati preliminari per la definizione di un modello di analisi dei dati in presenza di *faking* o *malingering*. Dato un campione di misure ordinali - come quelle ottenute nel contesto delle *surveys* o questionari *self-report* - si definisce *faking* quel processo per il quale parte (o la totalità) dei rispondenti modifica la propria risposta in modo deliberato con l'obiettivo di ottenerne un vantaggio. Per l'analisi di tale tipologia di dati, si propone un nuovo approccio, denominato CRB (*Combination of a Replacement and Binomial distributions*), derivante dall'integrazione di due approcci statistici indipendenti, ossia SGR (*Sample Generation by Replacement*) and CUBE (Combination of Uniform and Binomial distributions). CRB modella la risposta ordinale mediante una combinazione lineare convessa di due componenti, una distribuzione Binomiale traslata per la componente ordinale della risposta ed una distribuzione di *Replace-*

Luigi Lombardi, University of Trento, e-mail: luigi.lombardi@unitn.it · Antonio Calcagnì, University of Padova, e-mail: antonio.calcagni@unipd.it

ment per la componente faking. Come per i modelli CUBE, la stima dei parametri del modello è effettuata per massima verosimiglianza. Infine, un breve caso studio è utilizzato per mostrare il funzionamento del modello CRB per l'analisi e valutazione di dati soggetti potenzialmente a faking.

Key words: Fake-good data, CRB approach, Ordinal data, Generalized Mixture distribution, CUBE approach

1 Introduction

Faking behavior in self-report measures, a type of response set, is a tendency to falsify item responses in order to meet strategic goals (e.g., avoiding being charged with a crime, see [1]). This behavior may be observed in some sensitive contexts such as, for example, risky sexual behaviors and drug addictions (e.g., [2, 3]) where individuals may react by hiding their real opinions or honest responses.

SGR (Sample Generation by Replacement) is a probabilistic resampling procedure [4, 5] that can be used to study and evaluate uncertainty in inferences based on possible fake responses as well as to study the implications of fake data for empirical results. In general, a SGR analysis takes an interpretation perspective which incorporates in a global model all the available information (empirical or hypothetical) about the process of faking and the underlying true model representation. In particular, SGR has a statistical descriptive nature which tries to capture the phenomenological effect of faking according to an informational, data-oriented perspective based on a data replacement (information replacement) paradigm. SGR has been normally used as a methodology to study, using Monte Carlo simulation designs, the impact of fake data on parameter estimations and model fit evaluations.

Unlike SGR, CUBE models (Combination of a Uniform and a Binomial distribution) is a class of statistical models that is grounded on the data generating process of the discrete response choice [6, 7] which allows the modeling of rating data expressing preferences and evaluations. The CUBE approach considers the final discrete response as the combination of two components: *feeling* and *uncertainty*. The shifted Binomial component regards the expression of feeling and takes into account for the fraction of responses associated with a precise opinion on the rating. By contrast, the uniform component concerns to uncertainty in rating and mimics aspects not directly associated to the content of the item. This representation allows finer model specifications which include refuge options, response styles and possible overdispersion. Moreover, unlike SGR models, CUBE models are also supported by Maximum likelihood (ML) estimation procedures based on EM algorithms.

In this contribution, we introduce a novel model representation, called CRB (Combination of a Replacement and a Binomial distribution), which combines the two approaches by integrating into a common framework some nice features of the two perspectives to provide an effective data analysis strategy for faking behavior

in self-report measures. In particular, the new representation substitutes the second component (*uncertainty*) of CUBE with a replacement distribution mimicking the faking process in self-report measures.

2 Model

In this section, we will first highlight some connections between the two approaches according to a general probabilistic representation. Next, we will formally describe our proposal.

2.1 CUBE and SGR: similarities and differences

Let Y be a discrete (observed) random variable with a finite support $\{1,2,\ldots,m\}$ (e.g., a rating-type variable). In its general terms, the CUBE representation can be defined as follows:

$$P(Y=y) = \pi P(Z=y) + (1-\pi)P(V=y) \tag{1}$$

here Z and V are two hidden variables with the same support of Y. Note that, Eq. 1 constitutes a mixture representation for the observed variable Y. In particular, let $C \in \{0,1\}$ be a Bernoulli variable with parameter $\pi \in]0,1]$, then Eq. 1 can be rewritten as follows:

$$P(Y=y|C=1) = P(Z=y) \quad \text{and} \quad P(Y=y|C=0) = P(V=y) \tag{2}$$

Therefore, the mixture distribution reduces to a two step process where we first draw a coin C (with probability π of observing the target event), and next we sample the value of Y according to the previous dichotomous result observed on C. Note that, Eq. 1 implies a hidden joint distribution $P(C,Z,V)$ which in its general form *does not require* to satisfy the independence condition for the pair (Z,V). Therefore, $P(Y)$ represents the probability distribution of the transformed random variable

$$Y = CZ + (1-C)V. \tag{3}$$

Unlike CUBE, the SGR representation is defined as follows:

$$P(W=w) = \sum_x P(W=w|X=x)P(X=x) \tag{4}$$

where W and X are hidden variables with the same support of Y. In this context, the conditional distribution $P(W|X)$ is called the *replacement distribution*, whereas $P(X)$ is named the *prior distribution* for the true variable. In the SGR perspective, the random variable W is called the fake response, whereas X represents the true

hidden (and unknown) response. Note that $P(X)$ identifies the prior distribution of the true value X before any direct inspection of the observed data.

Now, we are in the position to link CUBE and SGR by setting the new transformed variable:

$$Y = CZ + (1-C)W \tag{5}$$

where, in this context, W denotes the fake random variable defined in Eq. 4. Note that a similar representation has been adopted in a recent SGR contribution called *mixture* SGR (see Eq. 11 in [8]). Recollecting all the terms we finally have:

$$P(Y = y) = \pi P(Z = y) + (1-\pi)P(W = y) \tag{6}$$

which, at a general level, directly connects SGR with the CUBE representation.

2.2 The CRB model

We now define the model instances for the CRB distribution. The first component is

$$b_y(\xi) = P(Z = y) = \binom{m-1}{y-1}(1-\xi)^{y-1}\xi^{m-y}, \qquad y = 1,2,\ldots,m, \tag{7}$$

the so-called shifted Binomial distribution with parameter $\xi \in [0,1]$. The second component of the CRB distribution is

$$p_y = P(W = y) = \frac{1}{m}\sum_{x=1}^{m} p_{y|x}, \qquad y = 1,2,\ldots,m, \tag{8}$$

with replacement distribution

$$p_{y|x} = \begin{cases} 1, & x = y = m \\ \frac{1}{m-x}, & 1 \le x < y \le m \\ 0, & 1 \le y \le \min\{x, m-1\} \end{cases} \tag{9}$$

The latter component corresponds to a fake good distribution (e.g., see Table 1 and Figure 1). Here we assume an uninformative prior for $P(X)$, that is to say, $P(X = x) = \frac{1}{m}$ for all $x = 1,2,\ldots,m$. Therefore, the CRB distribution takes the following form:

$$g_y \equiv P(Y = y) = \pi b_y(\xi) + (1-\pi)p_y. \tag{10}$$

Clearly, the distribution in Eq. 10 is a discrete one with a well defined two-dimensional parameter space $\Omega = \{(\pi,\xi) : 0 < \pi \le 1, 0 \le \xi \le 1\}$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be a random sample of n i.i.d. rating responses on the finite support $\{1,2,\ldots,m\}$. Then, the CRB log-likelihood function is expressed by:

$$L(\theta) = \sum_{i=1}^{n} \log\{\pi b_{y_i}(\xi) + (1-\pi) p_{y_i}\}. \tag{11}$$

Note that Maximum likelihood (ML) estimates can be obtained by EM algorithm according to the procedure outlined, for instance, by [9] for standard CUBE models.

3 Application

In this application we analyzed an hypothetical set of ordinal data about illicit drug use (cannabis consumption) among young people (see Table 1). In particular, the response variable uses a four-point ordinal scale ranging from 1 = *never* to 4 = *often*, with intermediate levels being 2 = *once* and 3 = *sometimes*. The observed frequencies for the four values were 27, 8, 5, and 3, respectively. To apply the faking-good model, as defined in Eq. 10, we reversed the rating scale. We finally ran the CRB model to the data sample with $n = 43$ independent rating observations. The estimated parameters were as follows:

$$\widehat{\pi} = 0.12090; \qquad \widehat{\xi} = 0.83336;$$

with a log-likelihood function $L(\widehat{\theta}) = -44.76983$ and a very low dissimilarity index $\delta = 0.01139$, which was calculated as the normalized difference between observed relative frequencies, $\widehat{p}_h$, and fitted probabilities, $\widehat{g}_h$:

$$\delta = \sum_{h=1}^{m} |\widehat{p}_h - \widehat{g}_h|.$$

The results suggested the prominent role of the replacement distribution component p_{y_i}, which modeled the faking-good style of response, in modulating the response process as reflected by the mixture parameter $(1-\pi)$.

Table 1 Replacement distribution $p_{y|x}$ and its marginal representation p_y. This corresponds to a fake good scenario such that $Y > X$ with a discrete uniform kernel.

	$X = x$					
$y\|x$	1	2	3	4	5	p_y
1	0.0	0.0	0.0	0.0	0.0	0.0
2	1/4	0.0	0.0	0.0	0.0	0.05
3	1/4	1/3	0.0	0.0	0.0	0.11667
4	1/4	1/3	1/2	0.0	0.0	0.21667
5	1/4	1/3	1/2	1.0	1.0	0.61667

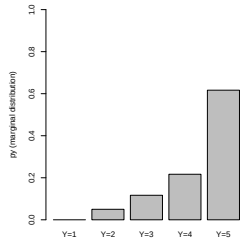


Fig. 1 Marginal distribution p_y.

4 Results and conclusion

In this short contribution we described a new model to analyse self-reported measures which could potentially be affected by faking behavior. Indeed, as many research have previously shown (e.g., see [2]), the latter plays an important role in social research and surveys based on self-reported questionnaires. We proposed a *combination of Replacement and Binomial* (CRB) distributions approach which takes the advantages of two statistical methodologies, namely SGR [4] and CUBE [7] that were independently proposed to analyse faking and preference data respectively. The new CRB approach uses a statistical rationale based on a mixture distributions approach where ordinal measures are represented as convex linear combination of a shifted-Binomial distribution, modeling the component unaffected by faking, and a Replacement distribution, which models instead the faking component. Model parameters were estimated via Maximum likelihood as offered by the EM algorithm in the general CUBE framework [9]. We showed the novel CRB approach on a simple application involving ordinal data from a hypothetical case study. Results suggested how faking response styles should deserve more attention, especially in those research involving analyses based on self-reported surveys and questionnaires.

References

1. M. Ziegler, C. MacCann, R. Roberts, *New perspectives on faking in personality assessment* (Oxford University Press, 2011)
2. M.J. Zickar, C. Robie, Journal of Applied Psychology **84**(4), 551 (1999)
3. L.A. McFarland, A.M. Ryan, Journal of Applied Psychology **85**(5), 812 (2000)
4. L. Lombardi, M. Pastore, Multivariate behavioral research **47**(4), 519 (2012)
5. M. Pastore, L. Lombardi, Quality & Quantity **48**(3), 1191 (2014)
6. D. Piccolo, Quaderni di Statistica **5**(1), 85 (2003)
7. A. D'Elia, D. Piccolo, Computational Statistics & Data Analysis **49**(3), 917 (2005)
8. M. Bressan, Y. Rosseel, L. Lombardi, Frontiers in psychology **9**, 1876 (2018)
9. M. Iannario, Metron **68**(1), 87 (2010)

BOD – min range: A Robustness Analysis Method for Composite Indicators

Intervallo BOD – minimo: un metodo per l'analisi di robustezza degli indicatori compositi

Emiliano Seri, Leonardo Salvatore Alaimo, Vittoria Carolina Malpassuti

Abstract Composite indicators are a useful instrument to represent in a synthetic and easy to read way a phenomenon or a specific reality. However, synthesizing indicators using aggregative approaches inevitably involves a loss of information due to the reduction in size of the original matrix. So, in composite indicators construction, the robustness analysis phase is very important to validate the synthetic construct and to attempt to recover some of the lost information. This paper aims to present an interval, useful for the robustness analysis of composite indicators, between the best and the worst performance obtainable from the synthetic construct. The two margins of the range are obtained: for the upper bound, using the Benefit Of The Doubt approach, that is an application for composite indicators construction of the linear programming technique, Data Envelopment Analysis; and, for the lower bound, it has been used the minimum between the considered elementary indicators. The obtained range will comprehend almost every other synthetic measure that can be calculated with other aggregative methods for the considered matrix of indicators, so, our methodological proposal is useful to see the distance from the best and the worst obtainable cases.

Abstract *Gli indicatori compositi sono uno strumento utile per rappresentare in modo sintetico e di facile lettura un fenomeno o una realtà specifica. Tuttavia, la sintesi di indicatori mediante approcci aggregativi comporta inevitabilmente una perdita di informazioni a causa della riduzione delle dimensioni della matrice originale. Pertanto, nella costruzione di indicatori compositi, la fase di analisi della robustezza è molto importante per validare il costrutto sintetico e tentare di recuperare alcune delle informazioni perse. In questo articolo presentiamo un intervallo utile per l'analisi di robustezza degli indicatori compositi, compreso tra la performance*

Emiliano Seri
Sapienza Università di Roma, e-mail: emiliano.seri@uniroma1.it

Leonardo Salvatore Alaimo
Italian National Institute of Statistics - Istat, e-mail: leonardo.alaimo@istat.it

Vittoria Carolina Malpassuti
Sapienza Università di Roma, e-mail: vittoriacarolina.malpassuti@uniroma1.it

migliore e quella peggiore ottenibili dal costrutto sintetico. L'estremo superiore del range è calcolato usando l'approccio del Benefit Of The Doubt, che è un applicazione per la costruzione di indicatori compositi della tecnica di programmazione lineare Data Envelopment Analysis. Per l'estremo inferiore è stato usato il minimo tra gli indicatori elementari considerati. Il range ottenuto comprenderà quasi ogni altra misura sintetica che che può essere calcolata con altri metodi aggregativi per la matrice di indicatori. Di conseguenza, questa proposta metodologica permette di osservare la distanza dal caso migliore e peggiore.

Key words: Composite indicators, Robustness analysis, BOD, Performance interval

1 Introduction

Constructing a composite indicator involves different subjective choices [4]. Subjectivity is an indispensable element in any measurement process; its presence does not make the process arbitrary [1]. But, different choices lead to different results; furthermore, a synthetic measure is useful to give an easy-to-read information about the phenomenon, but implicitly involves a loss of information due to the transition from a multi-dimensional to a uni-dimensional representation. For those reasons, the robustness analysis is a very important step in the composites construction. Its aim is to verify how the synthetic construct reacts to the different choices made to measure it. So, it is an instrument to validate the results and a way to find out some of the information loss during the previous phases of construction [5].

In this paper, we propose a new method for the robustness analysis of composite indicators based on the Benefit of the Doubt (BOD) approach. This method generates an upper bound and a lower bound to find the best and the worst performance that could be obtained aggregating a system of indicators. The space between the two bounds, it is going to include almost all the other constructs made with other mean-based aggregative methods, like: minimum, harmonic mean, geometric mean, arithmetic mean, quadratic mean, cubic mean and maximum. Thus, the proposed method could be useful to compare each of the cited techniques, observing their distance from the best and the worst performances. It is also useful to shrink the range of variation of the construct to observe the position of each unit in a real range of variation, in which each position can be reached by each unit.

2 Data envelopment analysis

DEA is a linear programming technique, useful to measure the relative efficiency of decision making units (DMU) on the basis of multiple inputs and outputs [7]. The efficiency of a set of variables can be adapted to construct a synthetic indicator using

an input-oriented DEA.[3]. The model assumes N inputs and M outputs for each of the I units. For the i^{th} unit, the inputs are represented by an array x_i and the outputs are represented by an array q_i. A first problem's formulation is the following: for each unit i the ratio of all the outputs over all the inputs is defined as

$$f(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}^{'}\mathbf{q}_i}{\mathbf{v}^{'}\mathbf{x}_i} \tag{1}$$

where $\mathbf{u}$ is an array of output weights and $\mathbf{v}$ is an array of input weights. The model seeks to maximize f, which represents the efficiency of the unit i^{th}, under the constraints that all the efficiency measures must be less than or equal to 1 and that the weights must be positive. This linear program is solved by assigning to each unit the most favourable weight. However, this formulation has infinite solutions of the form $(\delta\mathbf{u}, \delta\mathbf{v})$ for $\delta > 0$. To avoid it, the following constraint is introduced:

$$\mathbf{v}^{'}\mathbf{x}_i = 1$$

Thus the problem can be written as:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^{'}\mathbf{q}_i \\ &\text{s.t. } \mathbf{v}^{'}\mathbf{x}_i = 1 \\ &\quad \mathbf{u}^{'}\mathbf{q}_1 - \mathbf{v}^{'}\mathbf{x}_1 \leq 0 \\ &\qquad \vdots \\ &\quad \mathbf{u}^{'}\mathbf{q}_\mathrm{I} - \mathbf{v}^{'}\mathbf{x}_\mathrm{I} \leq 0 \\ &\quad \mathbf{u},\mathbf{v} \geq 0 \end{aligned} \tag{2}$$

Lastly we call the model DEA with no input if, for the array $\mathbf{x}_i$, we have $\mathbf{x}_i = (1,...,1)^{'}$. The model computes the weights so that the unit under investigation is valued as best as possible. The weights can differ from unit to unit and range between 0 and 1.

3 Benefit of doubt approach

The BOD method is an application of DEA that can be used to composite indicators construction[2][6]. The synthetic measure is expressed as the weighted sum of the elementary indicators relatively to a benchmark; more precisely, is defined as the performance of the single unit divided for the performance of the benchmark:

$$CI_c = \frac{\sum_{q=1}^{Q} I_{cq} w_{cq}}{I_{cq}^{*}} \tag{3}$$

where I_{cq} is the normalized score of the q$^{\text{th}}$ basic indicator ($q = 1, \ldots, Q$) for the unit c$^{\text{th}}$ ($c = 1, \ldots, C$) and w_{cq} is the corresponding weight. The benchmark I^*_{cq} is defined as:

$$I^*_{cq} = \max_{I_{c\in[1\ldots C]}} \sum_{q=1}^{Q} I_{cq} w_{cq} \tag{4}$$

The optimal set of weights (if exist) guarantees that each unit is associated to the best possible position compared to all the others. The optimal weights are obtained by solving equation 3 where the weights are non-negative. The result will be between zero and 1.

4 BOD – min range of performance: a robustness analysis method for composite indicators

In this paper, we propose a new interval of performance for composite indicators using the BOD approach for the upper bound and the minimum for the lower bound. With the obtained range, it is possible to compare the performance of the synthetic construct, calculated using most of the mean-based aggregative methods, for each considered unit, with the best and the worst obtainable performance. The proposed interval can be calculated on a set of indicators standardized using the Min-Max method, that bring each indicator on a range between 0 and 1. Our interval, shrink this range to lead to a real *photograph* of the units for the considered phenomenon and solve the common problem in composite indicators construction, that the results of each unit can have an evaluative function only relatively to those of the others.

We have chosen the BOD approach to make the upper bound because, as explained in Section 3, with this method it is possible to calculate the best reachable performance, associating to each variable the optimal weight for each unit. The lower bound corresponds to the minimum, because it is not reasonable that a unit should want to drop below its worst level.

Let's see now an application on a set of seven variables about the quality of work in the Italian regions in 2017 (Table 1). The selected indicators are:

- share of employed persons with temporary jobs for at least 5 years (V1);
- share of employees with below 2/3 of median hourly earnings (V2);
- share of employed people aged 15-64 years working over 60 hours per week (V3);
- share of employed persons not in regular occupation (V4);
- involuntary part-time (V5);
- share of employed persons who feel satisfied with their work (V6);
- Share of employed persons who feel their work insecure (V7).

Reg.	$V1$	$V2$	$V3$	$V4$	$V5$	$V6$	$V7$	Min	BOD
PIEMONTE	11.6	7.1	23.1	10.8	5.4	7.5	4.6	0.57	1.00
VALLE D'AOSTA	16.9	5.0	21.2	10.4	4.4	7.7	6.3	0.61	1.00
LIGURIA	17.4	5.8	23.4	12.1	6.0	7.3	6.1	0.29	0.91
LOMBARDIA	10.7	4.9	21.5	10.3	5.4	7.4	5.1	0.43	1.00
TRENTINO	19.8	4.3	18.4	9.6	3.4	7.8	4.2	0.64	1.00
VENETO	11.9	5.3	23.6	8.9	3.4	7.5	5.6	0.57	1.00
FRIULI	14.9	5.1	24.6	10.6	4.3	7.5	6.4	0.53	0.98
EMILIA ROMAGNA	16.7	5.0	25.3	10.0	4.9	7.5	6.4	0.48	0.97
TOSCANA	14.8	6.1	25.9	10.9	6.6	7.4	6.4	0.43	0.91
UMBRIA	12.1	6.4	31.7	12.9	6.2	7.4	7.3	0.00	0.95
MARCHE	14.1	6.3	27.5	10.3	5.1	7.4	6.8	0.32	0.91
LAZIO	21.2	8.6	28.4	15.6	7.5	7.3	6.7	0.25	0.70
ABRUZZO	17.7	9.5	30.0	15.9	5.6	7.2	8.4	0.13	0.75
MOLISE	20.9	9.6	25.5	15.6	8.0	7.5	5.9	0.38	0.73
CAMPANIA	21.0	16.4	23.7	20.1	8.2	7.1	8.4	0.00	0.68
PUGLIA	21.8	16.5	23.3	16.7	8.3	7.3	9.2	0.07	0.69
BASILICATA	23.4	10.1	27.9	14.4	6.7	7.2	9.4	0.04	0.60
CALABRIA	31.3	17.7	26.6	22.3	10.6	7.2	9.6	0.00	0.38
SICILIA	35.7	17.4	22.2	19.8	10.8	7.1	9.4	0.00	0.71
SARDEGNA	11.3	12.5	20.7	15.2	8.9	7.4	8.9	0.13	1.00

Table 1 Application of BOD – min range of performance: quality of work indicators; lower bound (Min); upper bound (BOD); Italian regions; year 2017.

First, we standardized the basic indicators and changed the polarity to those with present *negative polarity*[1]. The normalization method must be the relative indices with respect to the variation range, commonly called Min-Max:

$$r_{ij} = \frac{x_{ij} - \min\limits_{i}\left\{x_{ij}\right\}}{\max\limits_{i}\left\{x_{ij}\right\} - \min\limits_{i}\left\{x_{ij}\right\}} \tag{5}$$

Where x_{ij} is the value of ith unit for the jth indicator. We calculated the upper bound with the BOD approach (see equation 3); the lower bound was set equal to the minimum of each basic indicator normalized. The bounds obtained are reported in Table 1. We aggregated the standardized indicators r_{ij} using harmonic mean, geometric mean, arithmetic mean, quadratic mean, cubic mean; we report all the results in Figure 1[2].

[1] Polarity is the sign of the relation between the indicator itself and the phenomenon. All the indicators must have positive polarity, i.e. an increase in the normalized indicators corresponds to an increase in the composite index. To invert polarity, we use a linear transformation [1].

[2] Note that in geometric mean, 0 values have been changed with 0,001 in order to have explanatory results of the considered phenomena.

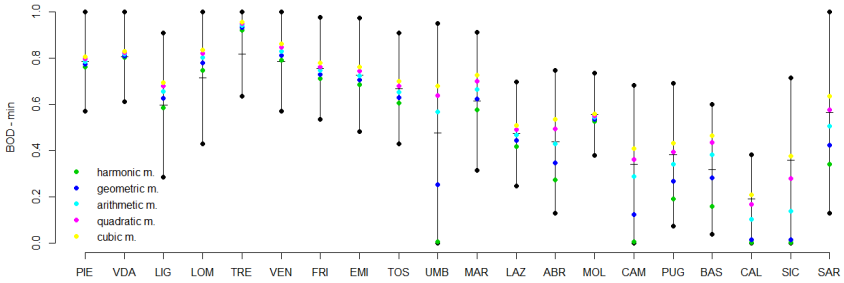


Fig. 1 BOD–min range of performance

5 Conclusions

Looking at Figure 1, it is clear that using the proposed method the range of variation of each unit has been widely narrowed or anyway made more truthful of the reality. It is also easier to compare the results obtained from different aggregation methods and to highlight where each of them is collocated in respect of the obtainable minimum and maximum.

References

1. Alaimo, L.S.: Complexity of Social Phenomena: Measurements, Analysis, Representations and Synthesis. Unpublished Doctoral Dissertation, University of Rome" La Sapienza", Rome, Italy (2020)
2. Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T.: An Introduction to 'Benefit of the Doubt'Composite Indicators. Social Indicators Research **82**(1), 111–145 (2007)
3. Coelli, T.: A Guide to DEAP version 2.1: A Data Envelopment Analysis (Computer) Program. Centre for Efficiency and Productivity Analysis, University of New England, Australia **96**(08) (1996)
4. Commission, J.R.C.E., et al.: Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD publishing (2008)
5. Maggino, F.: Complexity in Society: From Indicators Construction to Their Synthesis, vol. 70. Springer (2017)
6. Rogge, N., Cherchye, L., Moesen, W., Van Puyenbroeck, T.: 'Benefit of the Doubt'Composite Indicators. In: European Conference on Quality in Survey Statistics (2006)
7. Terzi, S., Pierini, A., et al.: Data Envelopment Analysis (DEA) Assessment of Composite Indicators of Infrastructure Endowment. Rivista di statistica ufficiale **17**(1), 5–18 (2015)

Comparing classifiers for ordinal variables

Confronti tra classificatori per variabili ordinali

Silvia Golia and Maurizio Carpita

Abstract To choose a single category of a qualitative variable using its predicted probability distribution is the final task to solve a classification problem. In this study, five predictive criteria are proposed and compared with the modal one, which is the standard criterion. The predictive performances are evaluated considering a set of indicators built from the resulting 3×3 confusion matrix. The data used are the decimal betting odds on the matches, transformed in probabilities of loss, draw and win of the home team, coming from the Kaggle European Soccer database, for seasons from 2008/2009 to 2015/2016 of the Italian League Serie A.

Abstract *Il compito finale per risolvere un problema di classificazione è la scelta di una sola categoria di una variabile qualitativa utilizzando la sua distribuzione di probabilità. In questo studio, si propongono cinque criteri predittivi che vengono confrontati con il criterio usuale che impiega la moda. Le prestazioni predittive sono valutate attraverso un insieme di indicatori costruiti sulla matrice di confusione 3×3. I dati utilizzati sono le quote decimali delle scommesse effettuate sulle partite, trasformate in probabilità di perdita, pareggio e vittoria della squadra di casa, provenienti dal database Kaggle European Soccer, per le stagioni dal 2008/2009 al 2015/2016 giocate in Serie A.*

Key words: Classification, Bipolar Mean, Decimal betting odds, Prediction of the result

Silvia Golia
University of Brescia, Department of Economics and Management, C.da S.Chiara, 50 - 25122 Brescia, Italy, e-mail: silvia.golia@unibs.it

Maurizio Carpita
University of Brescia, Department of Economics and Management, C.da S.Chiara, 50 - 25122 Brescia, Italy, e-mail: maurizio.carpita@unibs.it

1 Some classification criteria for ordinal variables

To choose a single category of a qualitative variable using its predicted probability distribution is the final task to solve a classification problem. Once the predicted probability distribution is obtained by applying a statistical model or other methods, the classification depends on the criterion used. The standard classification criterion is the simple majority rule, that means to choose the mode (the category with higher predicted probability), but it is not the only one that can be applied. In particular, for an ordinal variable A with $k > 2$ categories, $a_1, a_2, \ldots, a_k$, one can take advantage of the ordered categories and use other measures of center such as the median or, after a numerical coding step, the expected value.

In this paper six classification criteria, reported in Table 1, have been considered. M1 is the standard Mode criterion, whereas M2 involves the median. M3 evaluates the difference between the predicted probabilities pr_i of the k categories of the variable A and the corresponding frequencies fr_i computed from previous data, and takes the category corresponding to the maximum difference, that is

Table 1: The proposed classifiers

Classifiers	Criterion
M1	Mode
M2	Median
M3	Maximum Distance
M4	Bipolar Mean + Maximum Distance
M5	Expected Value
M6	Mode + Expected Value

$$\arg \max_{i \in (a_1, a_2, \ldots, a_k)} (pr_i - fr_i). \tag{1}$$

The fourth method M4 involves the computation of the Bipolar Mean (BM) and the criterion M3. The BM for ordinal variables, proposed by [4], is a "synthetic" variable with no more than two adjacent categories with non zero frequencies. The BM is obtained using the Expected Value (EV) of the considered variable, where EV is the arithmetic mean of the numerical recoded categories, weighted with its frequencies, as follows:

1. if EV is an integer number $i \in (1,2,,k)$, then BM has one category a_i with frequency 1 and the others categories with zero frequency;
2. if EV is a non-integer number between i and $i+1$ with factional part equal to $frac(EV)$, then BM has two categories a_i and a_{i+1} with frequency $1 - frac(EV)$ and $frac(EV)$ respectively, and the others categories with zero frequency.

Some applications of the BM are described in [1, 3]. M4 is computed applying the equation (1) where pr_i and fr_i are substituted by the BM distributions pr_i^{BM} and fr_i^{BM}.

The criteria M5 and M6 request to code the categories of the variable A as $a_1 = 1, a_2 = 2, \ldots, a_k = k$. M5 takes the expected value of the recoded variable, whereas M6 resorts to the expected value only if the mode has a frequency less than a given threshold, otherwise it coincides with M1. In general the expected value is not exactly equal to one of the coded categories, so it necessary to apply a rounding procedure following the ceiling (c), round (r) or floor (f) rule.

The predictive performance of a classifier can be evaluated using some indicators computed from the confusion matrix, with generic frequency n_{OP} corresponding to the number of units belonging to the observed category O and predicted to belong to category P.

The first indicator is the *Sensitivity* for category C, that is

$$Sens_C = \frac{n_{CC}}{n_{C\bullet}},$$

which expresses how well the classifier recognizes a unit belonging to the category C. It can be of interest to compare the sensitivities of the categories computing their difference in absolute value; a useful indicator from these differences is the maximum (*Maximum Distance Between Sensitivities* - MDBS), that is:

$$MDBS = \max_{i \neq j} |Sens_{C_i} - Sens_{C_j}|.$$

The lower the MDBS, the better the classification.

Considering only the cases of correct classification for category C, it is possible to compute the *Precision* for the classifier as

$$Prec = \frac{\sum_C n_{CC}}{n},$$

where n is the sample size.

Rearranging the confusion matrix as a 2×2 matrix with reference of each category, it is possible to compute the Accuracy for the category C as

$$ACC_C = \frac{n_{CC} + n_{\overline{C}\,\overline{C}}}{n},$$

where $\overline{C}$ is the complement of category C, and then their average obtaining the *Average Accuracy* (AA), which is the average per-category effectiveness of a classifier.

2 An application to the Italian Soccer League Serie A

The applicative fields of the theory described in Section 1 are various, and include the prediction of the results of the soccer matches. In fact, in this context, the variable *result* with its three categories, loss (L), draw (D) and win (W) of the home team, has an ordinal nature (loss $\prec$ draw $\prec$ win). Moreover, it can be codified as 1-2-3 and considered as a numerical variable with the predicted probabilities as its frequency distribution.

The classification criteria described in the previous section were applied to the decimal betting odds on the matches coming from the Kaggle European Soccer database (KES) [2]. The interest was focused on the matches played in the Italian League Serie A during the seasons from 2008/2009 to 2015/2016. The KES database reports, for each match, the decimal betting odds provided by 10 betting companies and the final result. These odds were averaged and transformed into probabilities of L, D and W [5, 6].

From the set of the resulting 3014 triplets of probabilities, 1000 samples of 1500 matches were randomly sampled without replacement; the remaining 1514 triplets were used to estimate fr_i and fr_i^{BM} for the M3 and M4 criteria.

Criterion M6 was implemented considering four thresholds, that is 0.4 (M6.1), 0.5 (M6.2), 0.6 (M6.3) and 0.7 (M6.4). Moreover, for both M5 and M6 the three rounding rules were applied.

For each of these samples the results of the matches were predicted applying the six criteria under study and evaluated thanks to the indicators introduced in the previous section. Moreover, due to the peculiarity of this application, an extra indicator of performance has been considered. It was called *Mean Weighted Win* (MWW), and it is the average win that a gambler obtains betting on a result of the match. It is computed attributing a win of 3 to the right prediction, a null win if the prediction is win (loss) and the actual result is loss (win) and a win of 1 otherwise. MWW is a weighted average of the previous three scores, $S_j = 3,1,0$, that is

$$MWW = \sum_{j=1}^{3} S_j \cdot V_j,$$

with weights respectively equal to:

$$\begin{aligned} V_1 &= (n_{LL} + n_{DD} + n_{WW})/n \\ V_2 &= (n_{DW} + n_{WD} + n_{DL} + n_{LD})/n \\ V_3 &= (n_{LW} + n_{WL})/n. \end{aligned}$$

The higher the MWW, the better the classification.

Table 2 reports the mean values of all the predictive performance indices with standard errors in parenthesis.

M1 has the highest precision, even if is not different from the one of other classifiers such as M6.1, a high average accuracy but also a high MDBS due to the fact

Table 2: Mean values of the predictive performance indices of the classifiers based on the prediction of 1500 matches results randomly sampled 1,000 times (standard errors in parenthesis)

Classifier	$Sens_L$	$Sens_D$	$Sens_W$	MDBS	*Prec*	AA	MWW
M1	0.491 (0.02)	0.019 (0.01)	0.852 (0.01)	0.834 (0.01)	0.535	0.690	1.868
M2	0.195 (0.01)	0.623 (0.02)	0.498 (0.01)	0.428 (0.02)	0.449	0.633	1.844
M3	0.604 (0.04)	0.281 (0.01)	0.531 (0.03)	0.324 (0.10)	0.486	0.657	1.810
M4	0.376 (0.02)	0.384 (0.04)	0.629 (0.03)	0.269 (0.04)	0.496	0.664	1.885
M5.c	0.000 (0.00)	0.263 (0.02)	0.860 (0.01)	0.860 (0.01)	0.471	0.647	1.803
M5.f	0.485 (0.02)	0.743 (0.02)	0.000 (0.00)	0.743 (0.02)	0.327	0.551	1.590
M5.r	0.048 (0.01)	0.874 (0.01)	0.216 (0.01)	0.826 (0.01)	0.344	0.563	1.674
M6.1.c	0.373 (0.02)	0.092 (0.01)	0.857 (0.01)	0.765 (0.01)	0.524	0.683	1.874
M6.1.f	0.482 (0.02)	0.153 (0.01)	0.745 (0.01)	0.593 (0.02)	0.518	0.679	1.873
M6.1.r	0.373 (0.02)	0.221 (0.02)	0.745 (0.01)	0.525 (0.02)	0.507	0.671	1.875
M6.2.c	0.190 (0.01)	0.209 (0.01)	0.860 (0.01)	0.672 (0.02)	0.508	0.672	1.867
M6.2.f	0.483 (0.02)	0.432 (0.02)	0.497 (0.01)	0.069 (0.02)	0.476	0.651	1.844
M6.2.r	0.190 (0.01)	0.624 (0.02)	0.497 (0.01)	0.434 (0.02)	0.448	0.632	1.841
M6.3.c	0.068 (0.01)	0.253 (0.02)	0.860 (0.01)	0.793 (0.01)	0.486	0.657	1.832
M6.3.f	0.485 (0.02)	0.596 (0.02)	0.274 (0.01)	0.322 (0.02)	0.416	0.611	1.750
M6.3.r	0.068 (0.01)	0.841 (0.01)	0.274 (0.01)	0.773 (0.02)	0.368	0.578	1.715
M6.4.c	0.015 (0.00)	0.261 (0.02)	0.860 (0.01)	0.846 (0.01)	0.474	0.649	1.810
M6.4.f	0.485 (0.02)	0.692 (0.02)	0.114 (0.01)	0.577 (0.02)	0.367	0.578	1.667
M6.4.r	0.048 (0.01)	0.874 (0.01)	0.216 (0.01)	0.826 (0.01)	0.344	0.563	1.674

The standard errors of *Prec* and *AA* and *MWW* are respectively equal to 0.01, 0.01 and 0.02 for all the criteria
Rounding rules for classifiers M5 and M6: ceiling (c), floor (f), round (r)

that mainly it is not able to correctly predict D. M2 is able to correctly predict a higher number of draws ($Sen_D = 0.623$) but it loses the capability to correctly predict the other two results; this finding translates into a reduction of the precision. M4 improves the performances of M3, it has the highest MWW and a more balance ability in correctly predicting all the three results. Considering the classifiers M5 and M6, it has to be noted the role played by the rounding procedure. As an example, for the classifier M5, applying the ceiling or round roundings, L is rarely correctly predicted, whereas, if the floor rounding is used, one loses the possibility to correctly predict W.

References

1. Brentari, E., Dancelli, L. and Maffenini, W.: The Bipolar Mean in Sensory Analysis. Electronic Journal of Applied Statistical Analysis *4*, 277–286 (2011)
2. Carpita, M., Ciavolino, E. and Pasca, P.: Exploring and modelling team performances of the Kaggle European Soccer Database. Statistical Modelling *19*, 74–101 (2019)
3. Dancelli, L., Manisera, M. and Vezzoli, M.: Interpreting clusters and their bipolar means: a case study. Statistica & Applicazioni *XI*(1), 49–62 (2013)
4. Maffenini, W. and Zenga, M.: Bipolar mean for ordinal variables. Statistica & Applicazioni *III*(1), 3–18 (2005)

5. Strumbelj, E.: On determining probability forecasts from betting odds. International Journal of Forecasting *30*, 934–943 (2014)
6. Strumbelj, E. and Sikonja, M.R.: Online bookmakers' odds as forecasts: The case of European soccer leagues. International Journal of Forecasting *26*, 482–488 (2010)

Discovering Interaction Effects Between Subject-Specific Covariates: A New Probabilistic Approach For Preference Data

La Scoperta Di Effetti Interazione Tra Variabili: Un Nuovo Approccio Probabilistico per i dati di preferenza

A. Baldassarre, C. Conversano, A. D'Ambrosio, M. De Rooij, E. Dusseldorp

Abstract This research presents a new probabilistic approach for the analysis of preference data when dealing with paired comparisons. The combination of the extended log-linear Bradley-Terry model with the regression trunk methodology generates a new model that allows finding interaction effects between subject-specific covariates. By fitting Poisson regressions to find the best split points and applying the final pruning procedure, the result is a small regression tree (so-called regression trunk). It represents a compromise between an easier interpretation of higher-order interaction effects and an efficient partition of individuals according to their preference scales.

Abstract *Questa ricerca mette in luce un nuovo approccio probabilistico per l'analisi dei dati di preferenza espressi con comparazioni a coppie. La combinazione del modello Bradley-Terry log-lineare con il metodo regression trunk consente la generazione di un albero di regressione. Attraverso una serie di regressioni Poisson è possibile ricercare i migliori punti di split e trovare le principali interazioni tra variabili. Il risultato finale è un albero di regressione di piccole dimensioni, che costituisce il giusto compromesso tra interpretabilità dei risultati ed efficienza di ripartizione degli individui in base alle preferenze espresse*

Key words: Preferences, Bradley-Terry Model, Regression Trunk, Decision tree

---

A. Baldassarre
Università degli Studi di Cagliari, Cagliari, e-mail: al.baldassarre1@gmail.com

C. Conversano
Università degli Studi di Cagliari, Cagliari

A. D'Ambrosio
Università Federico II di Napoli, Napoli

M. De Rooij
Universiteit Leiden, Leiden

E. Dusseldorp
Universiteit Leiden, Leiden

1 Introduction

Recently, several scientific fields, such as behavioral, political, and computational sciences, are showing a growing interest in preference data analysis. Generally, a number m of judges express their preferences for n objects by assigning values to each of them or by ordering the items from the most preferred to the least one. In the first case, they are called rankings; in the second one, we talk about orderings (Marden, 1995).

In literature, there are several statistical models and methodologies to analyze these data. Among these, there are methods based on badness-of-fit adaptation (Carrol, 1972; Coombs, 1950; Heiser and De Leeuw, 1981; D'Ambrosio and Heiser, 2016) and probabilistic models (Marden, 1995; Heiser and D'Ambrosio, 2013).

In some circumstances, different groups of judges may express different preferences scales given their different characteristics. The heterogeneity can be addressed by introducing subject-specific covariates (Dittrich, Hatzinger and Katzenbeisser, 1998; Strobl, Wickelmaier and Zeileis, 2011).

This paper presents a model that follows a probabilistic approach and generates a regression tree into which the individuals are divided according to their preferences and characteristics. It overcomes the problems related to additive models that become complex when there are interactions. Furthermore, it solves the typical difficulty of tree-based models, namely capturing linear effects between variables. One of the strengths of the regression trunk is that it is not necessary to have a priori information about the interactions between the covariates to build the regression tree. Through the extended log-linear Bradley-Terry model, we apply the regression trunk methodology for the analysis of preference data when subject-specific covariates are observable within a dataset.

In section 2, we introduce the Bradley-Terry model and its log-linear extended version. In section 3, our methodology is explained with a brief reference to the regression trunk model.

2 The log-linear Bradley-Terry model

The model that we propose works with preferences expressed through paired comparisons. In literature, the Bradley-Terry model (Bradley and Terry, 1952), is one of the most widely used methods for these kinds of data. It is a logistic model for paired preference data, and it has been applied in psychology and several other disciplines.

For example, when the number of objects is n, then $\frac{n\times(n-1)}{2}$ comparisons can be calculated. For example, if $n = 3$ and the items are A B C, the paired comparisons will be A-B, A-C, and B-C.

Let $\Pi_{(ij)i}$ denote the probability that the item i is preferred in comparison with j. The probability that j is preferred is $\Pi_{(ij)j} = 1 - \Pi_{(ij)i}$ (ties cannot occur). The Bradley-Terry model has item parameters β_i such that (Agresti, 2007, p. 286)

$$logit(\Pi_{(ij)i}) = \log\left(\frac{\Pi_{(ij)i}}{\Pi_{(ij)j}}\right) = \beta_i - \beta_j. \tag{1}$$

The Bradley-Terry model can also be fitted as a log-linear model (Fienberg and Larntz, 1976; Sinclair, 1982; Dittrich, Hatzinger, and Katzenbeisser, 1998), in which the number of preferences for i in comparison with j is assumed to follow a Poisson distribution. Let $e_{(ij)i}$ be the expected number of comparisons in which i is preferred to j. Then $e_{(ij)i} = n_{ij}\Pi_{(ij)i}$ has a log-linear representation

$$\begin{aligned} \ln(e_{(ij)i}) &= \mu_{(ij)i} + \lambda_i^O - \lambda_j^O \\ \ln(e_{(ij)j}) &= \mu_{(ij)j} - \lambda_i^O + \lambda_j^O, \end{aligned} \tag{2}$$

where μ is a nuisance parameter representing the items involved in the comparison, and λ_i^O is the item i related term such that $\lambda_i^O = \frac{1}{2}\log(\pi_i)$, with π_i equal to the probability of preferring the item i. Note that the superscript O refers to the position of the object (object-specific parameter) without considering the characteristics of the judges. It will be removed for easier reading from the equations that follow.

The Bradley-Terry model can be extended to incorporate categorical (e.g., gender) or numerical covariates (e.g., age). The model building becomes more complicated when dealing with continuous subject-specific covariates than when dealing with categorical ones. In fact, in the first case, it has to be extended for each subject and each value of the covariates. The log-linear Bradley-terry model equations for subject m and comparison ij with continuous subject-specific covariates are

$$ln(e(y_{ij;m})) = \mu_{ij;m} + y_{ij;m}(\lambda_{i;m} - \lambda_{j;m}), \tag{3}$$

where $y_{ij;m} \in \{-1, 1\}$ indicates whether the judge m expresses ($y_{ij,m} = 1$) or not ($y_{ij,m} = -1$) a preference in the comparison between i and j. If it is preceded by the letter e, then we deal with the expected value of y.

The parameter $\lambda_{i;m}$ can be expressed through a linear relation

$$\lambda_{i;m} = \lambda_i + \sum_{p=1}^{P} \beta_{ip} x_{p;m}, \tag{4}$$

where $x_{p;m}$ is the pth covariate ($p = 1...P$) associated to judge m. The parameters β express the effect of the subject-specific covariates on item i. λ_i is an intercept and indicates the location of object i in the overall preference scale.

3 The regression trunk applied to preference data

The extension of the log-linear Bradley-Terry model with subject-specific covariates allows the application of a regression model to paired comparisons. Here, we fol-

low the regression trunk methodology, which combines a multiple regression model and a regression tree (Dusseldorp and Meulman, 2004). In specific, it makes predictions by considering linear main effects and interaction effects for the same set of covariates.

The idea underlying the regression trunk is that, in some cases, it is not necessary building a large tree to capture the interactions between variables. The final result of the model is a small regression tree (so-called regression trunk) that represents a compromise between the interpretability of interaction effects and the ability to summarize the available information about the individuals' characteristics. The basic regression trunk is defined by a single linear model

$$\hat{Y} = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p + \sum_{t=1}^{T-1} \hat{\beta}_{P+t} I\{(X_1, ..., X_P) \in t\}, \tag{5}$$

where T is the total number of terminal nodes, and $T-1$ represents the total number of indicator variables, where the region not included works as reference group. $\hat{\beta}_0$ indicates the estimated intercept for the reference group, while for the node t the estimated intercept is represented by $\hat{\beta}_0 + \hat{\beta}_{P+t}$. The first part of the equation represents the linear part of the model estimated using all the observations and covariates for the main effects. The second part indicates the interaction effects obtained by partitioning the m observations.

In this research, we adapt the extended Bradley-Terry model with subject-specific covariates to the regression trunk methodology for the analysis of preference data.

The tree is built by fitting Poisson regressions and combining a regression model with main effects and a small number of higher-order interaction effects. This model allows observing the preference scale in each node of the tree and quantifying the probability of preferring certain items for different groups of individuals.

Usually, the building process of a generic regression tree is composed of different steps: the definition of a splitting criterion, the definition of a stopping rule, and a pruning procedure.

Similarly, the regression trunk for preference data uses as splitting criterion the deviance between one Poisson regression and the subsequent. Given the subject-specific covariate X_p, we consider all the values as a possible split point. Each of these values represents the rule to discretize the candidate variable X_p and to create the associated dichotomous variable Z. This variable is added to the model, and the deviance is calculated and stored. The process is repeated in each node, for each variable, and each value, calculating in total (number of nodes) $\times$ (number of variables) $\times$ (number of judges) different regressions. After that, the split point that minimizes the deviance in node t is chosen as the best split point. This value is used as a rule to create a new dichotomous variable Z_i^*, which gives information about the judges' partition. The building procedure is allowed by using Z_i^* for the subsequent regressions.

The definition of a stopping rule represents a choice that is declined to the users and their aims. The option that we propose is to split a parent node with m observations only if the child node includes a number greater than $\sqrt{m}$. Another option

would be to repeat the process until a predefined number of T terminal nodes is reached.

To apply the final pruning, the V-fold cross-validation is computed for each step of the process in the same way as it is used in CART and STIMA (Dusseldorp, Conversano and Van Os, 2010). The Poisson regression is fitted on the training set, and the predicted values $\hat{Y}_i$ are calculated on the test set to compute the cross-validation deviance. We choose as pruning rule to cut the nodes that induce a deviance increase, also considering the standard error through the $c \times SE$ rule.

Generally, the pruning procedure restitutes a small regression trunk (e.g., $T = 5$) given that the cross-validation deviance usually decreases and, after a certain split, it starts to increase.

References

1. Agresti, A.: An introduction to categorical data analysis, Hoboken, NJ: Wiley-Interscience. (2007)
2. Bradley, R. A., & Terry, M. A.: Rank analysis of incomplete block designs. I. Biometrika, **39**, 324–345. (1952)
3. Carroll, J. D.: Individual differences and multidimensional scaling. In: R. N. Shepard, et al. (Eds.), Multidimensional scaling theory (Vol. I, pp. 105–155). New York: Seminar Press (1972)
4. Coombs, C. H.: Psychological scaling without a unit of measurement. Psychological Review, **57**, 145—158. (1950)
5. D'Ambrosio, A., & Heiser, W.J.: A recursive partitioning method for the prediction of preference rankings based upon Kemeny distances, Psychometrika, **81**(3), 774–794. (2016)
6. Dittrich, R., Hatzinger, R., & Katzenbeisser, W.: Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings, Journal of the Royal Statistical Society C, **47**, 511–525. (1998)
7. Dusseldorp, E., & Meulman, J. J.: The regression trunk approach to discover treatment covariate interaction, Psychometrika, **69**(3), 355–374. (2004)
8. Dusseldorp, E., Conversano, C., & Van Os, B.J.: Combining an additive and tree-based regression model simultaneously: STIMA, Journal of Computational and Graphical Statistics, **19**(3), 514-–530. (2010)
9. Fienberg, S. E. & Larntz, K.: Loglinear representation for paired and multiple comparison models, Biometrika, **63**, 245–254. (1976)
10. Heiser, W. J., & De Leeuw, J.: Multidimensional mapping of preference data. Mathématiques et Sciences Humaines, **19**, 39—96. (1981)
11. Heiser, W.J., & D'Ambrosio, A.: Clustering and prediction of rankings within a Kemeny distance framework, In B, Lausen, D., Van den Poel, Ultsch, A. (Eds.), Algorithms from and for Nature and Life, Springer series in Studies in Classification, Data Analysis, and Knowledge Organization, Springer International Publishing Switzerland, 19–331. (2013)
12. Marden, J. I.: Analyzing and modelling rank data. London: Chapman Hall (1995)
13. Sinclair, C. D.: GLIM for preference, Biometrika, **14**, 164–178. (1982)
14. Strobl, C., Wickelmaier, F., & Zeileis, A.: Accounting for individual differences in Bradley-Terry models by means of recursive partitioning, Journal of Educational and Behavioral Statistics, **36**(2), 135–153. (2011)

Hybrid random forests for ordinal data

Un approccio ibrido alle foreste random per dati ordinali

Rosaria Simone and Gerhard Tutz

Abstract Prediction of ordinal outcomes is a compelling statistical task for which no definitive method is universally acknowledged. A semi-parametric approach to random forests is proposed that exploits the efficiency of ordinal regression models to gain information on the impact of explanatory variables on ordered response categories. In contrast to other proposals, scoring of categories is avoided, and both the partitioning process and the assignment rule use their ordering only. The proposed hybrid approach between model-based structuring and classification trees is illustrated on ratings on probability to vote for German Parties.

Abstract Il contributo propone un approccio semi-parametrico alle foreste random per dati ordinali basato sull'uso di modelli di regressione. Il metodo non richiede la scelta di codifiche numeriche per le categorie né per l'algoritmo di partizionamento ricorsivo né per le regole di assegnazione a fini previsionali. L'approccio è illustrato sulla base di valutazioni ordinali della probabilità di voto per i partiti Tedeschi.

Key words: Decision trees, ordinal data models, random forests

1 The State of the Art

The present study fits in research efforts devoted to developing trustworthy methods for predicting ordinal outcomes. For these data, tree methods mostly rely on the assignment of numeric scores to ordered categories, which is a questionable practice (see, for instance, both the packages rpartOrdinal [1] and its improved version rpartScore [5]). Numeric scoring of categories underlies also some proposals of ran-

Rosaria Simone
Department of Political Sciences, University of Naples Federico II, Italy, e-mail: rosaria.simone@unina.it

Gerhard Tutz
Ludwig-Maximilians-Universität München, Germany e-mail: gerhard.tutz@stat.uni-muenchen

dom forests for ordinal responses [8, 10], as well as conditional inference trees [9]. Since the latter approach allows to grow unbiased trees, hereafter it will be considered as benchmark.

A hybrid approach between model-based structuring and classification trees for ordinal outcomes to grow random forests [3] is introduced. The concept resorts to well-suited regression models for ordinal responses in the partitioning process only. Prediction accuracy of alternative base-learner for hybrid random forests will be checked in terms of the well-known Rank Probability Score (RPS, [7]). On this basis, a new accuracy measure based on a majority of vote rule for the minimum RPS is advanced to identify the best suited base-learner for a given hybrid random forest. Hybrid random forests based on ordinal logit and adjacent category models [12] will be illustrated on ratings for the probability to vote for German parties (taken from the General German Social Survey in 2008, [6]).

2 Hybrid random forests

In the following, first the construction of hybrid model-based trees is introduced and then the concept is used to grow random forests.

Hybrid trees

Parametric ordinal models are efficient tools to evaluate which variables are influential when response categories are ordered and the order is to be exploited. Therefore, we use a baseline model $\mathscr{M}(Y;\boldsymbol{x},\boldsymbol{\theta})$, where Y is the ordered response, which takes values in m categories $c_1 \prec c_2 \prec \cdots \prec c_m$, $\boldsymbol{x}$ is a vector of predictors, and $\boldsymbol{\theta}$ is the estimable parameter vector. Simple and easy-to-fit models hereafter considered are:

- the cumulative model: $P(Y \leq r) = F(\beta_{0r} + \boldsymbol{x}^T\beta), \quad r = 1,\ldots,m,$
- the adjacent categories model: $P(Y_i = r | Y_i \in \{r-1, r\}) = F(\beta_{0r} + \boldsymbol{x}^T\beta), r = 2,\ldots,m$

where $F(\cdot)$ is a suitable link function. Thus, a tree is built by using fits of the baseline model to select the best variable to perform the split along with the split points. Specifically, an ordinal regression model is fitted with predictor $\eta_r = \beta_{0r} + x_{j,c}\beta$, where $x_{j,c} = I(x_j \geq c)$, and $I(.)$ denotes the indicator function ($I(a) = 1$ if a is true, and $I(a) = 0$, otherwise). The impact of variable x_j at split point c is investigated by testing the null hypothesis $H_0 : \beta = 0$. The resulting p-value is an indicator of the relevance of the split, with small values indicating high relevance. Therefore, at node k the model $\mathscr{M}(Y;x_{j,c},(\{\beta_{0r}\},\beta))$ is fitted to the observations in that node for each candidate splitting variable and possible split points. The combination of variable and split point (x_j, c) that yields the smallest p-value is selected to generate the descendant nodes. The procedure is iterated for each node until a stopping criterion is met or when the minimum p-value is above some threshold α, possibly

corrected for multiple testing. Since partitioning of the predictor space is obtained by evaluating splits in specific covariates, typically several splits are possible (especially for continuous ones), which are considered as equally important. In order to obtain unbiased trees in the spirit of [9], one could resort to permutation tests for the test statistics [2]. However, this issue will not be addressed in the following preliminary analysis since it is mitigated by the random selection of predictors run to grow trees in the forest. A first application of the concept of hybrid trees appears in a benchmark study carried out in [4] based on the model-based procedure of [13].

Random forests

Random forests as a combination of trees typically show much better performance in terms of prediction than single trees [3]. They belong to the class of ensemble methods, which means that various predictors are aggregated for prediction of the response. The problem of possibly correlated splits is overcome by choosing, for each tree and at each step, a random selection of candidate splitting variables. Assume that model $\mathscr{M}$ is considered in the partitioning process of each of the B hybrid trees in the forest, say $\mathscr{T}^1,\ldots,\mathscr{T}^b,\ldots,\mathscr{T}^B$. Then:

- each $\mathscr{T}^b$ is grown by using a learning sample $\mathscr{S}_L$ obtained by bootstrapping (with replacement) from the original data $\mathscr{S}$, with size n. For the b-th tree, the test set to evaluate its prediction accuracy will be composed of those *out-of-bag observations* in $\mathscr{S}$ that were not used to grow the tree (oob for short).
- Each tree is grown up to a large depth. Growing descendants from a node stops if the number of observations to attempt a split or if any of the sample sizes of the candidate children nodes is below some threshold.

For each replicate b, the same training set and test set will be used to grow a hybrid random forest when assuming different base learners for the partitioning process. For each tree, a random set of $\sqrt{K}$ splitting covariates among those that were not yet included are selected in each step of the splitting procedure, where K denotes the number of (binary) candidate variables of the predictor space. For these variables, tests are performed to find the next split. In order to reduce the computational time but also account for the varying number of splits in covariates, for continuous covariates a random split point can be selected (yet ensuring that a minimum number of observations fall in both groups).

Prediction accuracy of a random forest can be measured in terms of the rank probability score (RPS, [7]). For each $i=1,\ldots,n$ corresponding to an oob for tree $\mathscr{T}$, let $RPS_i^{(\mathscr{T})}$ be the (empirical) RPS of observation Y_i in the test set computed on the basis of the classification obtained by tree $\mathscr{T}$. If $(f_1^{(\mathscr{T})},\ldots,f_m^{(\mathscr{T})})$ is the relative frequency distribution at the terminal node of $\mathscr{T}$ corresponding to the profile $\boldsymbol{x}_i$ of Y_i, then one computes

$$RPS_i^{(\mathscr{T})} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\sum_{s=1}^{j} f_s^{(\mathscr{T})} - I(Y_i \le j)\right)^2 \tag{1}$$

with lower values indicating a higher accuracy for prediction of Y_i ($I(Y_i \leq j)$ denotes the indicator function of the event $(Y_i \leq j)$). The predicted distribution resulting from a random forest is composed of the predictions from the single trees. Specifically:

- let $\mathscr{F}^{(i)}$ be the set of trees in the forest for which Y_i is an out-of-bag observation. For each $\mathscr{T} \in \mathscr{F}^{(i)}$, let $RPS_i^{(\mathscr{T})}$ be its RPS in the corresponding terminal node (established on the basis of its covariate profile $\boldsymbol{x}_i$).
- For Y_i, consider the average $\widehat{RPS}_i$ of the RPS values $RPS_i^{(\mathscr{T})}$, for $\mathscr{T} \in \mathscr{F}^{(i)}$.
- In the end, the performance of the hybrid random forest based on model $\mathscr{M}$ for the partitioning process will be measured by the index:

$$I = \sum_{i=1}^{n} \widehat{RPS}_i$$

- Lower values of the performance index I will denote higher prediction accuracy.

3 An application to voting probability for political parties

More and more often, opinion polling of voting intentions is accompanied by the collection of ratings for the probability to vote for each candidate in order to foster the understanding of the electoral behaviour. This is the case of some rating variables collected within the General German Social Survey: as shown in Figure 1, in several cases the distributions are U-shaped or J-shaped, and the modal value might not be meaningful when used to summarize the distribution and predict future observations. For illustrative purposes, a predictors matrix of $K = 12$ (binary)

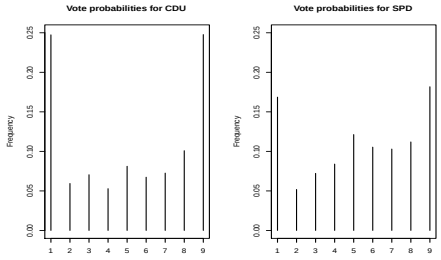


Fig. 1 Observed distribution of ratings for voting probabilities for the main German Parties

covariates is considered to grow a random forest. Table 1 reports some instances of the proposed *majority of vote* rule for prediction based on the distribution of median values. They are obtained in the following way:

- Given Y_i, consider the trees $\mathscr{T} \in \mathscr{F}^{(i)}$ for which Y_i is an out-of-bag observation;

- For each $\mathscr{T} \in \mathscr{F}^{(i)}$, let $(f_1, \ldots, f_m)$ be the frequency distribution in the terminal node corresponding to profile $\boldsymbol{x}_i$, and set

$$\hat{Y}_i^{(\mathscr{T})} = Median(f_1, \ldots, f_m).$$

  This choice corresponds to assigning, as a prediction for Y_i, the category with the minimum RPS value in the associated node. It is worth mentioning that predictions based on the modal value of the distribution of a terminal node would result mostly in the first and last response categories for the data under examine.
- Set $\hat{Y}_i$ the modal values of $\{\hat{Y}_i^{(\mathscr{T})} : \mathscr{T} \in \mathscr{F}^{(i)}\}$.

Observation	Prediction		
	Hybrid with Ordinal logit	Hybrid with adjacent	Ctree
1	3	3	3
9	8	8	8
6	4	3	3
8	8	7	8
6	6	6	6
7	6	6	6
8	6	7	6
2	3	3	3
5	6	7	8
4	5	5	5

Table 1 Some instances of observations and predicted outcomes established on the basis of the proposed *majority of vote* rule

For pointwise prediction, the overall accuracy (when using model $\mathscr{M}$ for the partitioning process of hybrid trees in a random forest) can be assessed in terms of the distance between the (minimum) RPS value associated to the predicted category, say $RPS_{\hat{Y}_i}^{(\mathscr{T})}$, and the RPS value $RPS_{Y_i}^{(\mathscr{T})}$ of the observed one, namely:

$$a^{(\mathscr{M},\mathscr{T})}(Y_i|\boldsymbol{x}_i) = |RPS_{\hat{Y}_i}^{(\mathscr{T})} - RPS_{Y_i}^{(\mathscr{T})}| \tag{2}$$

for each tree $(\mathscr{T})$ in $\mathscr{F}^{(i)}$. Then, the model $\mathscr{M}$ that attains the lowest distance in the sense of (2), after averaging over trees in the forest and over observations, can be considered as the best performing one. Table 2 reports the chosen performance indicators for the proposed hybrid random forests; for comparisons, random forests using the conditional inference approach as base learner (implemented within the R package `partykit`) have been considered.

	Accuracy as in (2)			RPS		
	Ordinal logit	Adjacent Category	Ctree	Ordinal logit	Adjacent Category	Ctree
CDU	0.092	0.091	0.092	620.04	618.46	613.27
SPD	0.098	0.098	0.098	578.48	579.14	578.96

Table 2 Prediction performances of hybrid random forests and forests with `ctree` as base learner

4 Ongoing research

The main contribution of the proposal is the introduction of a hybrid decision tree for ordinal data that exploits model fitting for the partitioning process and that does not need artificial numerical coding of ordered categories. Accordingly, hybrid random forests can be developed for prediction of rating outcomes. The rationale allows to introduce a new accuracy measure based on a majority of vote rule for the minimum RPS to identify the best base-learner for the hybrid random forests. The idea is illustrated on an original application to the understanding of the electoral behaviour. Further research will be addressed to variable importance measures for predictors, designed without the need of numerical scoring (which is considered in [10], for instance). Non-parametric trees tailored to provide flexible interpretation of response patterns yet respecting the ordinal nature of data, can be pursued via quantile trees, by basing the partitioning process on differences in multiple quantiles [11].

References

1. Archer, K.J.: rpartOrdinal: an R package for deriving a classification tree for predicting an ordinal response. Journal of Statistical Software **34**, (2010)
2. Berger, M., Tutz, G., Schmid, M.: Tree-structured modelling of varying coefficients Stat Comput **29**: 217–229 (2019)
3. Breiman, L.: Random Forests. Machine Learning **45**,5–32 (2001)
4. Cappelli, C., Simone, R., Di Iorio, F.: CUBREMOT: A tool for building model-based trees for ordinal responses. Expert Systems with Applications **124**, 39–49 (2019)
5. Galimberti, G., Soffritti, G., Di Maso, M.: Classification trees for ordinal responses in R: the rpartScore package. Journal of Statistical Software **47**, (2012)
6. GESIS Leibniz Institute for the Social Sciences (2016). German General Social Survey (ALLBUS) - Cumulation 1980-2014, GESIS Data Archive, Cologne. ZA4584 Data file version 1.0.0. DOI: 10.4232/1.12574
7. Gneiting, T., Raftery, A.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**, 359–376 (2007)
8. Hornung, R.: Ordinal Forests. Journal of Classification, doi:10.1007/s00357-018-9302-x (2019)
9. Hothor, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics **15**, 651–674 (2006)
10. Janitza, S., Tutz, G., Boulesteix, A.L.: Random Forests for Ordinal Responses: Prediction and Variable Selection. Computational Statistics and Data Analysis **96**,57–73 (2016)
11. Simone, R., Tutz, G., Davino, C., Vistocco, D.: Quantile trees for ordinal responses. (Preprint) (2020)
12. Tutz G.: Regression for Categorical Data. Cambridge: Cambridge University Press (2012)
13. Zeileis, A., Hothorn, T., Hornik, K.: Model-Based Recursive Partitioning. Journal of Computational and Graphical Statistics, **17**,92–514 (2008)

Model-based approach to biclustering ordinal data

Approccio di biclustering basato su modello per dati ordinali

Monia Ranalli and Francesca Martella

Abstract A finite mixture model to simultaneously cluster the rows and columns of two-mode ordinal data matrix is proposed. Due to the numerical intractability of the likelihood function, estimation of model parameters is based on composite likelihood (CL) methods and essentially reduces to a computationally efficient Expectation-Maximization type algorithm. The performance of the proposed approach is discussed on both simulated and real datasets. The results are encouraging and would deserve further discussion.

Abstract *Si propone una mistura finita per classificare sia le righe che le colonne di una matrice di dati ordinali. La stima dei parametri è ottenuta tramite un algoritmo di tipo EM definito sulla base di una verosimiglianza composita. La performance dell'approccio proposto è discussa sia su dati simulati che reali. I risultati ottenuti sono incoraggianti e meritano ulteriori approfondimenti.*

Key words: Mixture models, biclustering, ordinal data, latent variable models

1 Introduction

Cluster analysis has been mainly developed for continuous data. Only in the last decades, an increasing interest in clustering ordinal data has been observed. Such data are encountered very frequently in practice to measure, for example, attitudes, abilities or opinions. However, practitioners often apply on their ranks models and techniques developed for continuous data. To take into account the ordinal nature of

Monia Ranalli
Sapienza University of Rome, Piazzale Aldo Moro 5, Rome e-mail: monia.ranalli@uniroma1.it

Francesca Martella
Sapienza University of Rome, Piazzale Aldo Moro 5, Rome e-mail: francesca.martella@uniroma1.it

the data there exist two approaches developed mainly in the factor analysis framework: the Item Response Theory (IRT; [3, 1]) and the Underlying Response Variable (URV; [10, 11, 19]). In the former, the probabilities of the categories are assumed to be analytic functions of some latent variables having a particular cluster structure. The best known model is the Latent Class Analysis (LCA; [8]) where the latent variable is nominal. Examples where the latent variables are continuous are found in [4, 16, 7]. In the URV approach, the ordinal variables are seen as a discretization of continuous latent variables jointly distributed as a finite mixture model; examples can be found in [6, 13, 21]. In both approaches, the use of latent continuous variables makes the estimation rather complex because it requires the computation of many high dimensional integrals. The problem is usually solved by surrogating the likelihood function.

In some cases, the interest would be to provide a joint clustering of units and variables that is, to partition the data matrix into homogeneous blocks with respect to some observed features. The task, which may be thought of as a extension of standard clustering approaches to group both units (rows) and variables (columns) in a data matrix, is often referred to as biclustering, but it is also known under a broad range of different names, including double clustering, block clustering, bidimensional clustering, co-clustering, grid clustering, simultaneous clustering and block modelling. In case of ordinal data, such approach may be useful in many research fields, as for instance in marketing studies where finding a subset of customers that tends to evaluate similarly a subset of products or services may help the researchers to target some products or services according to customer profiles. A number of distance and model based biclustering methods exists. Among distance-based methods, we mention the double k-means method of [26] and [24], for example. Among model-based approaches, we found proposals developed for continuous data such as in [14], for binary or count data such as in [20].

Specifically looking at ordinal data, to the best of our knowledge, we found two proposals. One approach, proposed by [15], is a generalization of [20] in case of ordinal data. It relies on the proportional odds model and can be seen also as an extension of the one-mode clustering model given by [5]. The other proposal has been introduced by [9] and it relies on a recent distribution for ordinal data (BOS for Binary Ordinal Search model [2]).

The aim of this paper is to propose a new model for biclustering ordinal data. Following the URV approach, the observed variables are considered as a discretization of latent continuous variables distributed as a mixture of Gaussians. To introduce a partition of the P variables within the g-th component of the mixture, we adopt a factorial representation of the data following the modelling approach developed in [14], where a binary row stochastic matrix, representing variable membership, is used to cluster variables. In this way, we associate a component in the finite mixture to a cluster of variables and define a bicluster of units and variables. Notice that the number of clusters of variables (and therefore the partition of variables) may vary with clusters of units. More precisely, we consider the CMAP criteria [22] to define unit clustering, whereas a binary and row stochastic matrix for representing the variable partition.

However, the model specification involves multidimensional integrals that make the maximum likelihood estimation rather cumbersome and in some cases infeasible. To overcome this issue, the model is estimated within the EM framework maximizing a composite likelihood based on m-dimensional marginals. In the current work, we present the model estimation considering $m = 2$, i.e. a pairwise likelihood, that is the product of all possible likelihoods based on the bivariate marginals (see e.g. [21] and references therein). However, as long as sparsity is not a problem and computations are feasible, it is possible to use a higher m, as shown in the simulation study and real data applications. Under some regularity conditions [18], the estimators are consistent, asymptotically unbiased and normally distributed. In general, they are relatively less efficient [12, 25] compared to full maximum likelihood estimators, but much more efficient in terms of computational complexity.

2 Model

Let $x_1,\ldots,x_P$ be ordinal variables and $c_i = 1,\ldots,C_i$ the associated categories for $i = 1,\ldots,P$. There are $R = \prod_{i=1}^{P} C_i$ possible response patterns $\mathbf{x}_r = (x_1 = c_1,\ldots,x_P = c_P)$, with $r = 1,\ldots,R$. The ordinal variables are generated by thresholding $\mathbf{y}$, which is a latent multivariate continuous random variable distributed as a mixture of G Gaussian densities $\phi(\cdot)$ with means $\boldsymbol{\mu}_g$, covariance matrices $\boldsymbol{\Sigma}_g$ and mixture weights p_g $(g = 1,\ldots,G)$. The link between $\mathbf{x}$ and $\mathbf{y}$ is expressed by a threshold model defined as $x_i = c_i \Leftrightarrow \gamma_{c_i-1}^{(i)} \leq y_i < \gamma_{c_i}^{(i)}$. where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \ldots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ are the thresholds defining the C_i categories. Let $\boldsymbol{\psi} = \{p_1,\ldots,p_{G-1},\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_G,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_G,\boldsymbol{\Gamma}\}$ be the set of model parameters, where $\boldsymbol{\Gamma}$ is the set of vectors $\boldsymbol{\gamma}^{(i)}$ $(i = 1,\ldots,P)$. The probability of response pattern $\mathbf{x}_r$ can be expressed as follows

$$\Pr(\mathbf{x}_r;\boldsymbol{\psi}) = \sum_{g=1}^{G} p_g \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_P-1}^{(P)}}^{\gamma_{c_P}^{(P)}} \phi(\mathbf{y};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)d\mathbf{y} = \sum_{g=1}^{G} p_g \pi_r(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Gamma}) \qquad (1)$$

where $\pi_r(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Gamma})$ is the probability of response pattern $\mathbf{x}_r$ in cluster g. Thus, for a random i.i.d. sample of size N, the log-likelihood has the following form

$$\ell(\boldsymbol{\psi};\mathbf{X}) = \sum_{r=1}^{R} n_r \log\left[\sum_{g=1}^{G} p_g \pi_r\left(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\boldsymbol{\Gamma}\right)\right], \qquad (2)$$

where n_r is the observed sample frequency of response pattern $\mathbf{x}_r$ and $\sum_{r=1}^{R} n_r = N$. According to [14], in order to cluster not only the units but also the variables of the data matrix, we suggest a factorial representation of component-specific covariance structure of the latent variable $\mathbf{y}$ involving the binary row stochastic matrix $\mathbf{A}_g = \{a_{giq}\}$ with $i = 1,\ldots,P$; $q = 1,\ldots,Q_g$ $(g = 1,\ldots,G)$ representing variable membership. Specifically, $a_{giq} = 1$ if and only if the $i-$th variable belongs to the

$q-$th cluster in the g-th component, and 0 otherwise. In detail, conditional on the $g-$th component of the mixture with probability p_g $(g = 1,\ldots,G)$, $\mathbf{y}_r$ is specified as follows

$$\mathbf{y}_{rg} = \boldsymbol{\eta}_g + \mathbf{A}_g\mathbf{f}_{rg} + \boldsymbol{\varepsilon}_{rg} \tag{3}$$

where $\boldsymbol{\eta}_g$ is a component-specific mean vector, $\mathbf{f}_{rg}$ is a Q_g-dimensional ($Q_g < P$) vector of component-specific latent variables (factors), which is assumed to be i.i.d. Gaussian variate with vector mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_{Q_g}$, where $\mathbf{I}_{Q_g}$ denotes the $Q_g \times Q_g$ identity matrix. Furthermore, $\boldsymbol{\varepsilon}_{rg}$ are i.i.d. Gaussian component-specific random variables with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}_g = \text{diag}(\sigma^2_{1g},\ldots,\sigma^2_{Pg})$. According to this assumption, a unit in the g-component follows a multivariate density with mean $\boldsymbol{\eta}_g$ and covariance $\mathbf{A}_g\mathbf{A}'_g + \mathbf{D}_g$, that is

$$\mathbf{y}_{rg} \sim N(\boldsymbol{\eta}_g, \mathbf{A}_g\mathbf{A}'_g + \mathbf{D}_g). \tag{4}$$

Note that, the covariance specification implies: (i) a block diagonal correlation structure; (ii) the correlation between variables depends on the variances only. It is worth to note that, the proposed model may assume different specifications whether the $\mathbf{D}_g$ and $\mathbf{A}_g$ matrix are constrained to be equal across row clusters.

2.1 Estimation

The maximization of (2) is quite time consuming and becomes infeasible when P increases due to the presence of multidimensional integrals. For this reason, model parameters are estimated through an EM framework maximizing the pairwise log-likelihood, i.e. the sum of all possible log-likelihoods based on the bivariate marginals. The estimators obtained have been proven to be consistent, asymptotically unbiased and normally distributed. In general, they are less efficient than the full maximum likelihood estimators, but in many cases the loss in efficiency is very small or almost null [25].
The pairwise log-likelihood is

$$\begin{aligned} p\ell(\boldsymbol{\psi};\mathbf{x}) &= \sum_{i=1}^{P-1}\sum_{j=i+1}^{P} \ell(\boldsymbol{\psi};(x_i,x_j)) \\ &= \sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j} n^{(ij)}_{c_ic_j} \log\left[\sum_{g=1}^{G} p_g \pi^{(ij)}_{c_ic_j}(\boldsymbol{\eta}_g, \mathbf{A}_g\mathbf{A}'_g + \mathbf{D}_g, \boldsymbol{\gamma})\right], \end{aligned} \tag{5}$$

where now, after the reparameterization, the set of models parameters is $\boldsymbol{\psi} = \left\{p_1,\ldots,p_{G-1},\boldsymbol{\eta}_{11},\ldots,\boldsymbol{\eta}_{1Q_1},\ldots,\boldsymbol{\eta}_{G1},\ldots,\boldsymbol{\eta}_{GQ_G},\mathbf{A}_1,\ldots,\mathbf{A}_G,\mathbf{D}_1,\ldots,\mathbf{D}_G,\boldsymbol{\Gamma}\right\}$. Furthermore, $n^{(ij)}_{c_ic_j}$ is the observed frequency of a response in category c_i and c_j for variables x_i and x_j respectively, while $\pi^{(ij)}_{c_ic_j}(\boldsymbol{\eta}_g, \mathbf{A}_g\mathbf{A}'_g + \mathbf{D}_g, \boldsymbol{\gamma})$ is the corresponding probability under the model obtained by integrating on dimensions i and j the density of a normal distribution with parameters $\left(\boldsymbol{\eta}_g, \mathbf{A}_g\mathbf{A}'_g + \mathbf{D}_g\right)$ between the corresponding threshold parameters γ. It is clear that the pairwise approach is feasible as it requires

the evaluation of integrals of bivariate normal distributions, regardless of the number of observed or latent variables. The computation of parameter estimates is carried out using a EM-type algorithm.

2.2 Classification, model selection and identifiability

The CMAP criteria [22] is used to define unit clustering, whereas a binary and row stochastic matrix for representing the variable partition. One of the features of model-based clustering methods is that they provide quite a rigorous theoretical basis to choose the number of unit and variable clusters; in this perspective, the choice of G and Q_g $(g = 1, \ldots, G)$ involves several comparisons between different potential models. The best model is chosen by minimizing the composite version of penalized likelihood selection criteria like BIC or CLC (see [23] and references therein). A further important point of the proposed model that is worth to be discussed is parameter identifiability. Adopting a pairwise likelihood estimation approach, the number of essential parameters equals the number of parameters of a log linear model with only two factor interaction terms. Thus it means that we can estimate a lower number of parameters compared to a full maximum likelihood approach. Furthermore, to estimate both thresholds and component parameters if all the observed variables have three categories at least and when groups are know, we set the first two thresholds to 0 and 1, respectively [17]. Finally, the factorial reparameterization of component-specific covariance of the latent variable involving $\mathbf{A}_g$ is unique since the matrix $\mathbf{A}_g$ does not suffer from the identifiability issues, apart from clusters labels permutations, differently from the factor loadings matrix in the factor analysis model.
Further details will be given in the extended version of the paper along with simulation and real data results to show the effectiveness of the proposal.

References

1. Bartholomew, D., Knott, M., Moustaki, I.: Latent Variable Models and Factor Analysis: A Unified Approach, third edn. Wiley Series in Probability and Statistics. Wiley (2011)
2. Biernacki, C., Jacques, J.: Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing **26**(5), 929–943 (2016)
3. Bock, D., Moustaki, I.: Handbook of Statistics on Psychonometrics, chap. Item response theory in a general framework. Elsevier (2007)
4. Cagnone, S., Viroli, C.: A factor mixture analysis model for multivariate binary data. Statistical Modelling **12**, 257–277 (2012)
5. DeSantis, S.M., Houseman, E.A., Coull, B.A., Stemmer-Rachamimov, A., Betensky, R.A.: A penalized latent class model for ordinal data. Biostatistics **9**(2), 249–262 (2008)
6. Everitt, B.: A finite mixture model for the clustering of mixed-mode data. Statistics & Probability Letters **6**(5), 305–309 (1988)
7. Gollini, I., Murphy, T.: Mixture of latent trait analyzers for model-based clustering of categorical data. Statistics and Computing **24**(4), 569–588 (2014)

8. Goodman, L.A.: Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika **61**(2), 215–231 (1974)
9. Jacques, J., Biernacki, C.: Model-based co-clustering for ordinal data. Computational Statistics & Data Analysis **123**, 101–115 (2018)
10. Jöreskog, K.G.: New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares. Quality and Quantity **24**(4), 387–404 (1990)
11. Lee, S.Y., Poon, W.Y., Bentler, P.: Full maximum likelihood analysis of structural equation models with polytomous variables. Statistics & Probability Letters **9**(1), 91–97 (1990)
12. Lindsay, B.: Composite likelihood methods. Contemporary Mathematics **80**, 221–239 (1988)
13. Lubke, G., Neale, M.: Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. Multivariate Behavioral Research **43**(4), 592–620 (2008)
14. Martella, F., Alfo, M., Vichi, M.: Biclustering of gene expression data by an extension of mixtures of factor analyzers. The international journal of biostatistics **4**(1) (2008)
15. Matechou, E., Liu, I., Fernández, D., Farias, M., Gjelsvik, B.: Biclustering models for two-mode ordinal data. Psychometrika **81**(3), 611–624 (2016)
16. McParland, D., Gormley, I.C., McCormick, T.H., Clark, S.J., Kabudula, C.W., Collinson, M.A.: Clustering south african households based on their asset status using latent variable models. Ann. Appl. Stat. **8**(2), 747–776 (2014). DOI 10.1214/14-AOAS726
17. Millsap, R.E., Yun-Tein, J.: Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research **39**(3), 479–515 (2004)
18. Molenberghs, G., Verbeke, G.: Models for discrete longitudinal data. Springer Series in Statistics Series. Springer Science+Business Media, Incorporated New York (2005)
19. Muthén, B.: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika **49**(1), 115–132 (1984)
20. Pledger, S., Arnold, R.: Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Statistics & Data Analysis **71**, 241–261 (2014)
21. Ranalli, M., Rocci, R.: Mixture models for ordinal data: a pairwise likelihood approach. Statistics and Computing pp. 1–19 (2014). DOI 10.1007/s11222-014-9543-4
22. Ranalli, M., Rocci, R.: Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. Analysis of Large and Complex Data. Studies in Classification,Data Analysis and Knowledge Organization. Editors: Adalbert F.X. Wilhelm Hans A. Kestler. (2016). DOI 10.1007/978-3-319-25226-1
23. Ranalli, M., Rocci, R.: Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. In: Analysis of Large and Complex Data, pp. 53–68 (2016)
24. Rocci, R., Vichi, M.: Two-mode multi-partitioning. Computational Statistics & Data Analysis **52**(4), 1984–2003 (2008)
25. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Statistica Sinica **21**(1), 1–41 (2011)
26. Vichi, M., Kiers, H.A.: Factorial *k*-means analysis for two-way data. Computational Statistics & Data Analysis **37**(1), 49–64 (2001)

New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution

Nuovi algoritmi e diagnostiche di bontà di adattamento per la modellizzazione dei dati di ranking con la distribuzione di Plackett-Luce estesa

Cristina Mollica and Luca Tardella

Abstract The *forward order* assumption postulates that the ranking process of the items is carried out by assigning the positions from the top (most-liked) to the bottom (least-liked) alternative. This assumption has been recently relaxed in the *Extended Plackett-Luce model* (EPL) through the introduction of the discrete *reference order* parameter, describing the rank attribution path. By starting from two formal properties of the EPL, we derive novel diagnostic tools for testing appropriateness of the EPL assumption. We also show how one of the two statistics can be exploited to construct a heuristic method, that surrogates the maximum likelihood approach for inferring the underlying reference order. The performance of the proposals was compared with more conventional approaches through an extensive simulation study.

Abstract *L'ipotesi di ordinamento in avanti postula che il processo di ranking venga eseguito assegnando le posizioni dalla prima all'ultima. Questa ipotesi è stata recentemente rilassata nel modello di Plackett-Luce esteso (EPL) attraverso l'introduzione del parametro discreto detto ordine di riferimento, che descrive il percorso di attribuzione del rango. Partendo da due proprietà formali dell'EPL, deriviamo nuovi strumenti diagnostici per testare l'adeguatezza del modello EPL. Mostriamo anche come una delle due statistiche possa essere impiegata per costruire un metodo euristico, che surroga l'approccio di massima verosimiglianza per l'inferenza sull'ordine di riferimento. La performance delle proposte è stata confrontata con approcci più convenzionali attraverso un ampio studio di simulazione.*

Key words: ranking data, Plackett-Luce model, model specification, goodness-of-fit assessment, bootstrap, heuristic methods

---

Cristina Mollica
Dipartimento MEMOTEF, Sapienza Università di Roma
e-mail: cristina.mollica@uniroma1.it

Luca Tardella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: luca.tardella@uniroma1.it

1 Introduction

Let us consider an experiment in which a sample of N judges is asked to rank a set $I = \{1,\dots,K\}$ of K labeled alternatives, namely *items*, according to a certain criterion. A *ranking* is a vector $\pi = (\pi(1),\dots,\pi(K))$ collecting the ranks assigned to each item, specifically the entry $\pi(i)$ indicates the position attributed to the i-th alternative. Equivalently, data can be recorded in the *ordering* format $\pi^{-1} = (\pi^{-1}(1),\dots,\pi^{-1}(K))$, where the generic component $\pi^{-1}(j)$ indicates the item ranked in the j-th position. Therefore, ranking data are multivariate ordinal data taking values in the set of permutations $\mathscr{S}_K$ of the first K integers.

This work concentrates on the parametric family of stagewise ranking models, in particular on the *Extended Plackett-Luce model* (EPL) introduced by [1]. The EPL generalizes the *Plackett-Luce model* (PL) [2, 3] by relaxing the implicit forward order assumption with the introduction of the *reference order parameter* $\rho = (\rho(1),...,\rho(K))$. It indicates the rank assignment order, i.e., $\rho(t)$ is the position attributed at the stage t. The probability of an ordering π^{-1} under the EPL is

$$\mathbf{P}_{\text{EPL}}(\pi^{-1}|\rho,\underline{p}) = \mathbf{P}_{\text{PL}}(\pi^{-1}\rho|\underline{p}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^{K} p_{\pi^{-1}(\rho(v))}} \qquad \pi^{-1} \in \mathscr{S}_K, \qquad (1)$$

where ρ is the discrete parameter, specifically a permutation of the first K integers, and the positive quantities p_i's are referred to as *support parameters*, proportional to the probabilities for each item to be ranked in the position indicated by the first entry of ρ. Hereinafter, we will shortly refer to (1) as EPL$(\rho,\underline{p})$. Inference on the EPL and its generalization into a finite mixture framework was originally addressed from the Maximum Likelihood Estimation (MLE) perspective in [1], via the hybrid Expectation-Maximization-Minorization (EMM) algorithm. Recently, [4] introduced the Bayesian version of the EPL, where a tuned joint Metropolis-within-Gibbs sampling was developed to conduct approximate posterior inference. The MCMC algorithm was also adapted to infer on the EPL with order constrains on ρ [6] and on the Bayesian EPL mixture [7].

2 Novel EPL diagnostics and comparative evaluation

Specific diagnostic tools to evaluate model adequacy of multistage ranking models are very limited in the ranking literature and their effectiveness has not been deeply explored. The present work aims at providing some contributions in this direction.

Let us suppose that we have some data simulated from an EPL$(\rho,\underline{p})$. We expect the marginal frequencies of the items at the first stage to be ranked according to the order of the corresponding support parameter component. On the other hand, we expect the marginal frequencies of the items at the last stage to be ranked according to the reverse order of the corresponding support parameter component. One can then derive that the ranking of the marginal frequencies of the items corresponding

to the first and last stage should sum up to $(K+1)$, no matter what their support is. Of course, this is less likely to happen when the sample size is small or when the support parameters are not so different of each other. In any case, one can define a test statistic by considering, for each couple of integers (j, j') candidate to represent the first and the last stage ranks, namely $\rho(1)$ and $\rho(K)$, a discrepancy measure $T_{jj'}(\pi)$ between $K+1$ and the sum of the observed ranks of the frequencies corresponding to the same item extracted in the first and in the last stage. Formally, let $\underline{r}_j^{[1]} = (r_{j1}^{[1]}, \ldots, r_{jK}^{[1]})$ and $\underline{r}_{j'}^{[K]} = (r_{j'1}^{[K]}, \ldots, r_{j'K}^{[K]})$ be the marginal item frequency distributions for the j-th and j'-th positions, to be assigned respectively at the first [1] and last [K] stage. In other words, the generic entry $r_{ji}^{[s]}$ is the number of times that item i is ranked j-th at the s-th stage. The proposed EPL diagnostic relies on the following discrepancy

$$T_{jj'}(\pi) = \sum_{i=1}^{K} |\text{rank}(\underline{r}_j^{[1]})_i + \text{rank}(\underline{r}_{j'}^{[K]})_i - (K+1)|, \tag{2}$$

implying that the smaller the value $T_{jj'}(\pi)$, the larger the plausibility that the two integers (j, j') represent the first and the last components of the reference order. In this sense, $T_{jj'}(\pi)$ is a measure of the closeness of the positions j and j' in ρ. To globally assess the conformity of the sample with the EPL, we consider the statistic

$$T_m(\pi) = \min_{j<j'} T_{jj'}(\pi). \tag{3}$$

With the aim at further enlarging the collection of diagnostics for the EPL class, we focus our attention also on another specific property of the EPL, knows as *Luce's choice axiom* or *independence of irrelevant alternatives* (IIA). The IIA hypothesis implies that the probability ratio of selecting item i over item i' is constant over the stages of the ranking process (constant ratio rule), as long as the two items are both still available. Formally, it implies that the expected paired comparison frequency at stage t of choosing item i over item i' is $\tau^*_{ii't} = T_{ii't} p_i/(p_i + p_{i'})$, given by the product between the total number $T_{ii't}$ of pairwise comparisons between i and i' at stage t and the theoretical pairwise comparison probability under the EPL. Hence, a chi-squared statistic for the IIA assumption can be defined as follows

$$X^2_{\text{IIA}} = \sum_{t=1}^{K-1} \sum_{i<i'} \frac{(\tau_{ii't} - \tau^*_{ii't})^2}{\tau^*_{ii't}},$$

where $\tau_{ii't}$ is the observed paired comparison frequency at stage t.

After introducing novel test statistics, one should enquire into their power, for instance through a bootstrap approach, and preferably compare it with that of some standard goodness-of-fit tools for ranking models. To this aim, we conducted a simulation study under alternative model specifications, involving in the comparison the chi-squared discrepancies based on the top frequencies, the pairwise comparisons and the first-order marginals, given respectively by

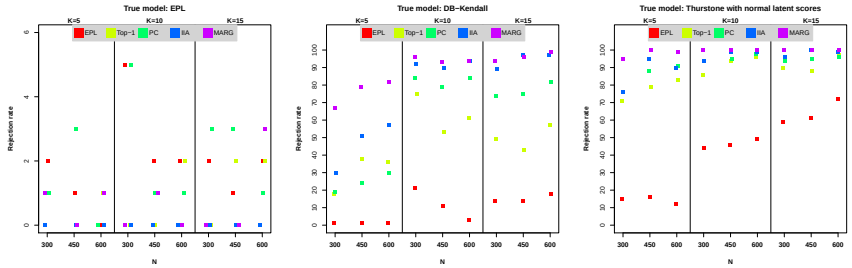


Fig. 1: Rejection rates of the EPL assumption for alternative goodness-of-fit diagnostics computed on simulated data from different model scenarios.

$$X^2_{\text{TOP}} = \sum_{i=1}^{K} \frac{(m_{1i} - m^*_{1i})^2}{m^*_{1i}} \qquad X^2_{\text{PC}} = \sum_{i<i'} \frac{(\tau_{ii'} - \tau^*_{ii'})^2}{\tau^*_{ii'}} \qquad X^2_{\text{M}} = \sum_{j=1}^{K} \sum_{i=1}^{K} \frac{(m_{ji} - m^*_{ji})^2}{m^*_{ji}}$$

where the expected frequencies are $m^*_{1i} = Np_i$ and $\tau^*_{ii'} = N\frac{p_i}{p_i + p_{i'}}$, whereas m^*_{ji} were estimated with a Monte Carlo simulation.

A comparative evaluation of the competing model specification tools was carried out by means of an extensive simulation study. For each possible combination (K,N), with values varying respectively in the grids $K \in \{5,10,15\}$ and $N \in \{300,450,600\}$, we drew 100 datasets with N orderings of K items from the following ranking distributions: i) EPL, ii) distance-based model (DB) with the Kendall metric (DB-Kend), iii) DB with the Hamming distance (DB-Ham) and iv) TH with normal latent scores (TH-norm). where the true parameter values were uniformly generated. To approximate the reference distribution of the test statistics under the EPL assumption, the bootstrap method was applied. Finally, for each model adequacy criterion, we estimated the mis- and correct rejection rates with the relative frequency of the times that the p-value was smaller than or equal to 0.05.

The simulation study revealed a satisfactory performance of all the considered diagnostics regarding the rates of mis-rejections, testified by estimated Type I error probabilities below 0.05 (Figure 1, left). On the other hand, noteworthy differences emerged in terms of the power. Firstly, measure (3) exhibited a consistent poor behavior of the estimated power under each considered model scenario, see for instance the plots in the middle and on the right of Figure 1. At least two motivations can be put forward to argue this evidence. The former is related to the formal definition of $T_m(\pi)$; in fact, (3) is a parameter-free measure based on the ranks of the expected marginal frequencies, rather than on the computation of the parameter-dependent first- and last-stage theoretical probabilities. This makes $T_m(\pi)$ by construction a rougher diagnostic in the comparison with the other statistics. Secondly, the remarkably low power of (3) under the DB with the Kendall metric (Figure 1, center) suggested that the monotonicity property of the first- and last-stage item probabilities is not specific of the EPL, but it is shared by other rankings models too. Another stable evidence highlighted by the comparative analysis concerns the

best-performing diagnostic, which turned to be the one relying on the marginal item distributions. However, it is no less apparent that, for higher values of K and N, the performance of the new IIA statistic is pretty much equivalent to that of the chi-squared based on the marginal distributions and, in general, always better than the remaining competing statistics typically used in the real-data applications.

3 Likelihood-free estimation of the reference order

Rather than from a model specification perspective, in this section we explored the utility of the statistic (3) from the inferential point of view.

Let $\mathbf{T}(\pi) = (T_{jj'}(\pi))$ be the $K \times K$ matrix with entries defined in (2). For each component $T_{jj'}(\pi)$, the inequality $T_{jj'}(\pi) \leq u_K$ holds, where the upper bound u_K corresponds to the constant value in the main diagonal of the matrix, i.e., $u_K = T_{jj}(\pi) = \sum_{l=1}^{K} |2l-(K+1)|$ depending on data only through K. Our heuristic method to estimate the unknown parameter ρ is composed of the following steps:

1. compute $\mathbf{D}(\pi) = |\mathbf{T}(\pi) - u_k \mathbf{J}_K|$, where $\mathbf{J}_K$ is $K \times K$ all-ones matrix, so that each component $D_{jj'}(\pi)$ can be interpreted as a measure of the distance between positions j and j' in the sequential rank assignment process;
2. use the matrix $\mathbf{D}(\pi)$ as the input of a Principal Component Analysis (PCA);
3. estimate ρ by taking the non-decreasing ordering of the scores $(s_1, \ldots, s_K)$ of the K positions on the first PC.

The inferential effectiveness of the proposal to recover the true discrete parameter was explored by means of a simulation study. For each possible combination (K,N), where $K \in \{5,10,15\}$ and $N \in \{50,200,1000,10000\}$, we drew 100 datasets $\underline{\pi}_{(R)}^{-1}$ with $R = 1, \ldots, 100$ from the EPL with uniformly generated parameters. For comparison purposes, we inferred the reference order of each simulated sample $\underline{\pi}_{(R)}^{-1}$ with the heuristic strategy described above, with the one replacing the PCA with the Multidimensional Scaling (MDS) and with the MLE approach [1], which is considered as the reference method for the present estimation task. Finally, the estimation performance of the competing strategies were compared in terms of: i) percentage of matching between $\dot{\rho}^{(R)}$ and $\hat{\rho}^{(R)}$ (% recoveries) and ii) average cograduation $\bar{r}_{\text{Spear}}(\dot{\rho}, \hat{\rho})$ between the estimated and the actual reference order computed with Spearman's rank correlation.

As apparent in Table 1, PCA and MDS exhibited essentially the same ability. Compared with the MLE, one can appreciate very good results for the heuristic methods. The percentage of matching consistently grows with N and, by checking also the cases where there is not an exact correspondence, on average an analogous trend is highlighted for the relative Spearman correlation. Additionally, if we look at a fixed N, the percentage of recoveries shows a worse tendency for larger values of K. In this regard, the cases $K \in \{10,15\}$ combined with a relatively very low ($N = 50$) and very high ($N = 10000$) sample size deserve some considerations to stress typical issues which can be encountered in a ranking data analysis. First, in a

Table 1: Inferential performance of the heuristic methods via PCA and MDS to estimate the reference order on simulated data compared to the MLE.

	% recoveries			$\bar{r}_{\text{Spear}}(\hat{\rho},\hat{\rho})$		
(K,N)	PCA	MDS	MLE	PCA	MDS	MLE
$(5,50)$	58	60	45	0.86	0.86	0.69
$(5,200)$	79	80	77	0.89	0.89	0.95
$(5,1000)$	91	90	100	0.92	0.93	1.00
$(5,10000)$	97	97		0.94	0.94	
$(10,50)$	3	5	3	0.90	0.89	0.87
$(10,200)$	14	16	23	0.93	0.93	0.98
$(10,1000)$	55	54	68	0.96	0.95	0.99
$(10,10000)$	79	78		0.96	0.96	
$(15,50)$	0	0	0	0.92	0.92	0.91
$(15,200)$	0	1	3	0.94	0.94	0.97
$(15,1000)$	13	15	26	0.97	0.97	0.99
$(15,10000)$	48	51		0.97	0.97	

sparse data situation, all of the estimation techniques exhibit a great uncertainty in recovering the actual ρ, testified by the negligible values of the recovery percentage. On the other hand, although a better behavior of the MLE is expected for $N = 10000$, this has not been implemented since, without a specialized program, fitting the EPL to a large sample can be deeply computational demanding if not actually unfeasible. Moreover, the computational burden is further aggravated by the multiple initialization needed to address the issue of local maxima. In the light of these remarks, the likelihood-free approach can be motivated as a straightforward method that can be combined with the MLE or with an MCMC method in the Bayesian framework. In fact, without computational costs, it can implemented as a preliminary step to obtain a promising initialization, that can guide the parameter space exploration towards the global optimum and substantially reduce the elaboration time.

References

1. Mollica, C., Tardella, L.: Epitope profiling via mixture modeling of ranked data. Stat Med. **33**(21), 3738–3758 (2014)
2. Luce, R.D.: Individual choice behavior: A theoretical analysis. John Wiley & Sons Inc (1959)
3. Plackett, R.L.: The analysis of permutations. J Royal Stat Soc C. **24**(2), 193–202 (1975)
4. Mollica, C., Tardella L.: Bayesian analysis of ranking data with the Extended Plackett-Luce model. Statistical Methods and Applications. doi: 10.1007/s10260-020-00519-5 (2020)
5. Mollica, C., Tardella L.: Bayesian mixture of Plackett-Luce models for partially ranked data. Psychometrika. **82**(2), 442–458 (2017)
6. Mollica, C., Tardella, L.: Constrained Extended Plackett-Luce model for the analysis of preference rankings. In: Book of Short Papers SIS 2018, (eds.) Springer Italia, pp 480-486, ISBN: 9788891910233 (2018)
7. Mollica, C., Tardella, L. Modelling unobserved heterogeneity of ranking data with the Bayesian mixture of Extended Plackett-Luce models. In: Book of Short Papers CLADAG 2019, pp. 346-349, ISBN: 978-88-8317-108-6 (2019)

Non-metric unfolding on augmented data matrix: a copula-based approach

Unfolding non-metrico basato sull'aumento di matrici: un approccio basato sulla funzione copula

Marta Nai Ruscone and Antonio D'Ambrosio

Abstract In this contribution an effective procedure to avoid degeneracies in multidimensional unfolding for preference rank data is proposed. We adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, by using copula-based association measures among rankings (individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). Our proposal is able to both recover the order of the preferences and reproduce the position of both rankings and objects in a geometrical space. Application on real datasets show that our procedure returns non-degenerate unfolding solutions.

Abstract *In questo contributo viene proposta un'efficace procedura per evitare soluzioni degeneri nell'unfolding multidimensionale per dati di tipo rank. La strategia utilizzata é quella di aumentare la matrice dei dati, cercando di costruire una matrice di dissimilaritá completa, utilizzando misure di associazione basate su copule tra rank (individui) e tra rank e oggetti (ovvero, una rappresentazione dell'ordine dei rank e degli oggetti attraverso rank appaiati). La nostra proposta é quindi in grado di catturare sia l'ordine dei rank sia la posizione di rank e oggetti in uno spazio geometrico. Applicazioni su dataset reali mostrano che la nostra procedura restituisce soluzioni non degeneri.*

Key words: copula, unfolding, multidimensional scaling

Marta Nai Ruscone
Universitá di Genova, Via Dodecaneso, 35 - 16146 Genova, e-mail: marta.nairuscone@unige.it

Antonio D'Ambrosio
Universitá di Napoli Federico II, Via Cinthia, M.te S. Angelo - 80125 Napoli, e-mail: antdambr@unina.it

1 The copula function

Copula are functions that join multivariate distribution functions to their marginal distribution functions [8]. They describe the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate normal distribution.

A bivariate copula $C : I^2 \to I$, with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$, is the cumulative bivariate distribution function of a random variable (U_1, U_2) with uniform marginal random variables in [0,1]

$$C(u_1, u_2; \theta) = P(U_1 \leq u_1, U_2 \leq u_2; \theta), \quad 0 \leq u_1 \leq 1 \quad 0 \leq u_2 \leq 1 \tag{1}$$

where θ is a parameter measuring the dependence between U_1 and U_2.

The following theorem by Sklar [8] explains the use of the copula in the characterization of a joint distribution. Let (Y_1, Y_2) be a bivariate random variable with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cdf $F_{Y_1,Y_2}(y_1, y_2; \theta)$, then there always exists a copula function $C(\cdot, \cdot; \theta)$ with $C : I^2 \to I$ such that

$$F_{Y_1,Y_2}(y_1, y_2; \theta) = C\big(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta\big), \quad y_1, y_2 \in \mathbb{R}. \tag{2}$$

Conversely, if $C(\cdot, \cdot; \theta)$ is a copula function and $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are marginal cdfs, then $F_{Y_1,Y_2}(y_1, y_2; \theta)$ is a joint cdf.
If $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous functions then the copula $C(\cdot, \cdot; \theta)$ is unique. Moreover, if $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous the copula can be found by the inverse of (2):

$$C(u_1, u_2) = F_{Y_1,Y_2}(F_{Y_1}^{-1}(u_1), F_{Y_2}^{-1}(u_2)) \tag{3}$$

with $u_1 = F_{Y_1}(y_1)$ and $u_2 = F_{Y_2}(y_2)$. This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The dependence structure is explained by the copula function $C(\cdot, \cdot; \theta)$. Moreover the (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the normal distribution.

Each copula is related to the most important measures of dependency: the Pearson correlation coefficient and the Spearman grade correlation coefficient. The Spearman grade correlation coefficient (see [8] pp. 169-170 for the definition of the grade correlation coefficient for continuous random variables) measures the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula C with parameter θ, then the Spearman grade correlation is

$$\rho_s(C) = 12 \int_{I^2} C_\theta(u_1, u_2) du_1 du_2 - 3. \quad (4)$$

For continuous random variables it is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient [5].

2 Unfolding as a special case of multidimensional scaling on copula-based association between rankings

Unfolding, originally formulated by Coombs [3] for the analysis of the two-mode preference choice data, is a technique that allows the estimation of two configurations usually representing the coordinates for a set of m individuals and a set of n objects on the basis of proximity values between them, typically expressing preferences of each individual over each object.

Therefore unfolding applies multidimensional scaling [4] to an off-diagonal $n \times m$ matrix, usually representing the scores (or the rank) assigned to a set of m items by n individuals or judges [1]. Using of either scores or rankings traditionally discriminates between metric and non-metric unfolding.

The goal is to obtain two configuration of points representing the position of the judges (X) and the items (Y) in a reduced geometrical space. Each point representing the individuals is considered as an ideal point so that its distances to the object points correspond to the preference scores [3].

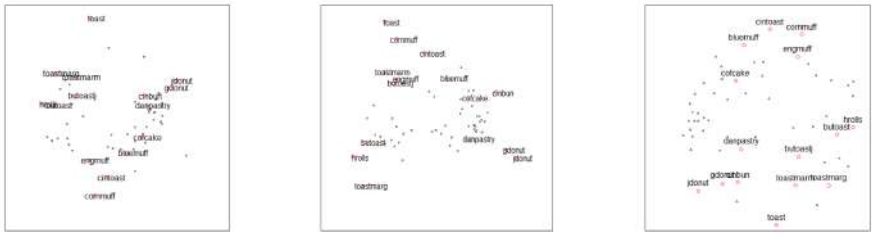
Unfolding can be seen as a special case of multidimensional scaling because the off-diagonal matrix is considered as a block of an ideal distance matrix in which both the within judges and the within items dissimilarities are missing. The presence of blocks of missing data causes the phenomenon of the so-called degenerate solutions, i.e., solutions that return excellent badness of fit measures but not graphically interpretable at all.

To tackle the problem of degenerate solutions, several proposals have been presented in the literature [1]. By following the approach introduced by [9], we adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, and then applying any MDS algorithms.

Let $\boldsymbol{\Gamma}$ be the original $m \times n$ original preference data matrix. In order to augment the data matrix we add to this n additional rows, one for each of the n objects, that correspond to tied rankings representing the jth item, $j = 1, \ldots, n$. As a result, a new $N \times n$ $\boldsymbol{\Gamma}^*$ matrix is obtained, with $N = n + m$. Then we use copula-based association measures among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings), obtaining in fact a $N \times N$ dissimilarity matrix to be analyzed with any MDS algorithm.

3 An application on a real data set

Fig. 1 shows a comparison between the Unfolding solutions of PRESCAL [2], which actually is the most used algorithm for Unfolding analysis, and our proposal by using the Spearman grade correlation coefficient via copula on the breakfast data set. Green and Rao [6] collected 42 rankings of 15 objects by asking 21 students and their wives to order 15 breakfast items in terms of their preference.



(a) Row-conditional Prefscal solution (b) Unconditional Prefscal solution (c) Spearman augmented unfolding

Fig. 1: Unfolding solutions for breakfast data. Breakfast items are labeled as follow: Toast pop-up; Buttered toast; English muffin and margarine; Jelly donut; Cinnamon toast; Blueberry muffin and margarine; Hard rolls and butter; Toast and marmalade; Buttered toast and jelly; Toast and margarine; Cinnamon bun; Danish pastry; Glazed donut; Coffee cake; Corn muffin and butter.

PREFSCAL works by setting two penalties on a modified loss function in such a way to guarantee non degenerate solutions. A possible drawback of this algorithm is that it is not always clear how set the penalty terms. In fact the user must make attempts in order to find the right solution.
The figure emphasizes that the solution of our procedure is not degenerate and it is comparable with the one of PREFSCAL, especially with its unconditional output. It is normal that our output looks like the unconditional PREFSCAL solution because we propose a solution that, depending on how we defined the dissimilarity matrix, is unconditional as well.

4 Concluding remarks

We propose an unfolding algorithm based on the augmentation of the data matrix and a copula-based association between rankings. The shown result highlights that our proposal produces non-degenerate unfolding solutions that are comparable with the ones obtained with PREFSCAL. With respect to PREFSCAL, any parameter

must be a priori chosen by the user. On the other hand PREFSCAL always guarantees non degenerate solutions. A robust simulation study will be discussed.

References

1. Borg, I., Groenen, P.: Modern multidimensional scaling. Theory and applications. Springer-Verlag, New York (1997)
2. Busing, Frank M. T. A. and Groenen, Patrick J. K and Heiser, Willem J. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika, **70(1)** 71–98, (2005).
3. Coombs,C.H.: Psychological scaling without a unit measurement. Psychological review, **57** 145–158 (1998)
4. Cox, T.F., Cox, M. A. A.: Multidimensional scaling. Chapman & Hall, London (1994)
5. Embrechts, P., McNeil, A. J., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. Risk management: value at risk and beyond, 176–223 (1998)
6. Green, P E and Rao, V R: Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. Holt. Rinheart and Winston, New York (1972).
7. Kaufman, L., Rousseeuw, P. J.: Finding groups in data. An introduction to cluster analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York (1990)
8. Nelsen, R.B.: An introduction to copulas. Springer Series in Statistics. Springer, New York (2013)
9. Van Deun, K., Heiser W., Delbeke, L.: Multidimensional unfolding by nonmetric multidimensional scaling of spearman distances in the extended permutation polytope. Multivariate Behavioral Research (2007) doi: 10.1080/00273170701341167

Ordinal probability effect measures for dyadic analysis in cumulative models

Misure di probabilità ordinali per l'analisi diadica in modelli cumulativi

Maria Iannario and Domenico Vistocco

Abstract Dyadic data analysis (DDA) is increasingly being used to better understand, analyze and model intra- and inter-personal mechanisms in various types of dyads such as husband-wife, caregiver-patient, doctor-patient, parent-child or athlete-coach as in our example. A key strength of the DDA is its flexibility to take the (non)independence available in the dyads into account. In this article, we illustrate the value of using DDA to examine how sports performance is perceived by an athlete and if it is consistent with the declared performance by his/her coach. A probability summary for ordered comparison of groups referred to a measure of stochastic superiority is used to indicate the consistency of perceived assessments.
Abstract *L'analisi dei dati diadici (DDA) è sempre più utilizzata per comprendere, analizzare e modellare i meccanismi intra e interpersonali in vari tipi di diade come marito-moglie, assistente-paziente, medico-paziente, genitore–bambino o atleta–allenatore. Quest'ultimo caso viene analizzato nel presente contributo. In particolare, la flessibilitá della DDA nel tener conto della (non)indipendenza presente nelle diadi viene qui considerata per esaminare in che modo un atleta percepisce la prestazione sportiva e se questa percezione è coerente con la valutazione della prestazione dichiarata dal suo allenatore. Al fine di considerare la coerenza delle valutazioni espresse dai due attori del processo considerato, si utilizza una misura di probabilità stocastica che permette il confronto ordinato dei gruppi.*

Key words: Dyadic analysis, Ordinal data models, Ordinal superiority measures

Maria Iannario
University of Naples Federico II, Via L. Rodinò - Naples, e-mail: maria.iannario@unina.it

Domenico Vistocco
University of Naples Federico II, Via L. Rodinò - Naples, e-mail: domenico.vistocco@unina.it

1 Introduction

The study deals with modelling the actor/partner interdependence in case of categorical dyadic data by presenting an alternative approach with respect to the current used methods [7]. The research aims at evaluating the consistency of perceived assessments on some topics related to the performance in sport dyads (athletes/coaches). The proposal may adjust the score perceived by the athletes in the definition of their performance by means of coach evaluation. The analysis is about a cross–sectional study on 100 couples of athlete-coach from Italian Swimming Federation–Campania region section, collected between November and December 2019. Satisfaction and other psychological domains were measured by rating scales on a $k = 7$ point scale. Frequency distribution of some items are displayed in Figure 1.

Figure 2 is about two items, *Challenging* and *Talent*, whose response distribution of the two groups of dyads is almost the same for the first item and very different for the second one, as remarked in the next analysis. The observer agreement plot ([4], [5]) provides a graphical representation of agreements of the assessment perceived by athletes and coaches. Each chart depicts a $n \times n$ square, where n denotes the total sample size and the black squares, each of size $n_{jj} \times n_{jj}$, $j = 1,2,\ldots,k$, denoting the observed agreement. The larger boxed rectangles depict the maximum possible agreement, since they are obtained starting from the observed marginals. Therefore, a visual impression of the strength of the agreement is obtained comparing the areas of dark squares with the area of white rectangles. A further information is provided by the lighter shaded rectangles, that depict the weighted contribution from off-diagonal cells. This allows to interpret the partial agreement among athletes and coaches comparing the areas of lighter shaded rectangles to the area of external rectangles. Finally, the positions of the dark squares with respect to the diagonal line provide information about "observer bias", i.e. the case where the observers consistently tend to classify the objects into higher or lower categories than the other.

Figure 3 depicts the path of the levels of the considered two items exploiting a correspondence analysis map [6], *Challenging* being on the left–side and *Talent* on the right–side. The two paths inside each plot represent the Athletes' levels (solid line) and the Coaches' levels (dotted line) of the two items. Multiple correspondence analysis has been carried out on the whole set of items, even if only the two illustrative items are represented for the sake of illustration. The plots confirm the different patterns of the two items, already suggested by the previous agreement plot.

2 Ordinal Superiority Measures for dyadic data analysis

When Y is a k-category ordinal response variable, one of the candidate model to analyse the rating is the cumulative link model [8]:

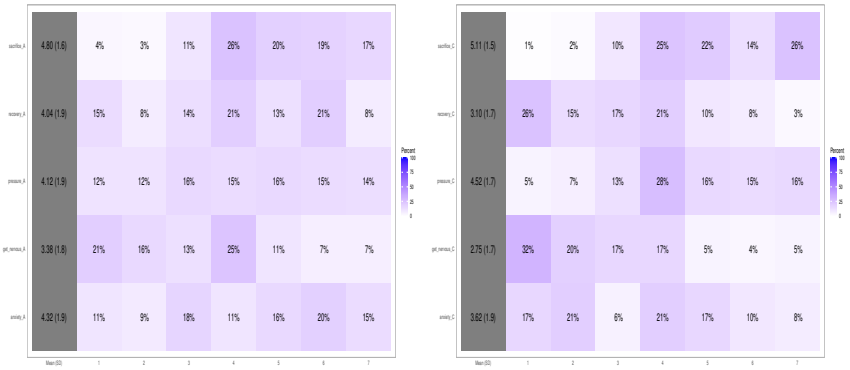


Fig. 1 Frequency distribution of subjective perceptions concerning 5 items collected on a 7 point scale (1=total disagreement; 7=total agreement). First column reports means and standard deviation. Left: Athletes' responses. Right: Coaches' responses

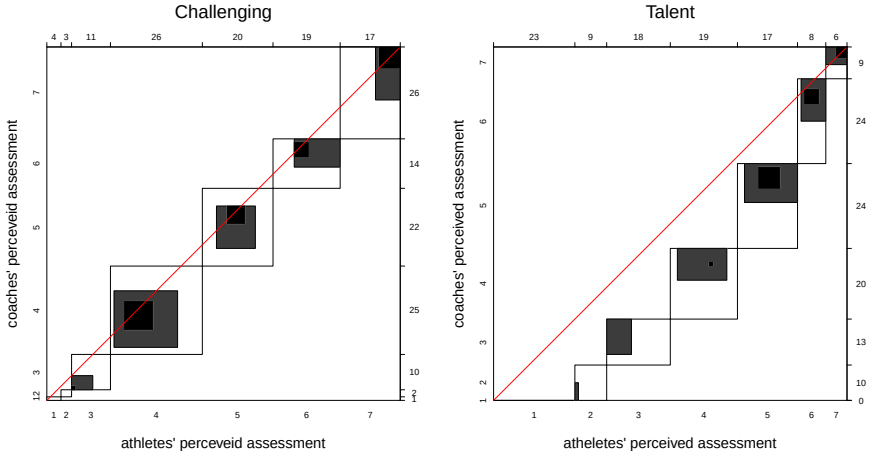


Fig. 2 Agreement plots of subjective perceptions concerning 2 items collected on a 7 point scale (1 = total disagreement; 7 = total agreement). First one is related to Athletes' challenging. Second one concerns Athletes' talent. The Athletes' responses are depicted on the horizontal axis, the Coaches' responses on the vertical axis.

$$P(Y_i = j \mid \boldsymbol{X}_i;\boldsymbol{\theta}) = P\left(\alpha_{j-1} < Y_i^* \leq \alpha_j | \boldsymbol{X}_i\right) = \\ = F_\varepsilon(\alpha_j - \boldsymbol{X}_i\boldsymbol{\beta}) - F_\varepsilon(\alpha_{j-1} - \boldsymbol{X}_i\boldsymbol{\beta}), \tag{1}$$

where

$$P(Y_i \leq j | \boldsymbol{X}_i;\boldsymbol{\theta}) = F_\varepsilon(\alpha_j - \boldsymbol{X}_i\boldsymbol{\beta}), \qquad i = 1,2,\ldots,n;\ j = 1,2,\ldots,k. \tag{2}$$

The parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ contains the intercept values $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_{k-1})'$, where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_k = +\infty$ are the thresholds of the scale of a latent vari-

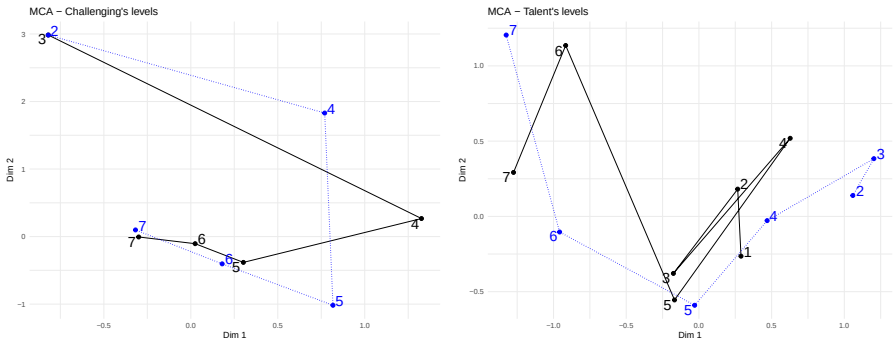


Fig. 3 Multiple correspondence maps of subjective perceptions concerning 2 items collected on a 7 point scale (1 = total disagreement; 7 = total agreement). The analysis is carried out on the whole set of items but only the levels of *Challenging* (left–side) and *Talent* (right–side) are represented. The two lines in each panel refer athletes' levels (solid black line) and coaches' levels (dotted blue line), respectively.

able Y^* surrounding the response Y, and the covariates coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. The vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ contains instead the covariates. Here, $\boldsymbol{\theta} \in \Omega(\boldsymbol{\theta})$, where $\Omega(\boldsymbol{\theta})$ is an open subset of $\mathbb{R}^{p+k-1}$.

Some common choices for $F_\varepsilon(.)$ are the Gaussian, the logistic or the standard Gumbel distribution and the related models are named probit, logit, and extreme value, respectively. The latter, related to log-log and the complementary log-log link function, is adopted when a skewed distribution is assumed. The ordinal superiority measures apply directly to the latent variable model (see [2, 3]). These measures represent our selected approach to compare groups by studying the probability that an observation from one group (=coaches) is scored above an independent observation from the alternative group (=athletes) of dyad.

We consider a dichotomous variable to identify the two clusters of coaches and athletes. Let Y_C^* and Y_A^* be independent underlying latent response variables at $\boldsymbol{X}$. Denoting by g_C (coaches) and g_A (athletes) two groups of statistical units related to the observed Y. An "*ordinal superiority measure*" or "*measure of stochastic superiority*" is

$$\gamma = P(Y_{g_C}^* > Y_{g_A}^*; \boldsymbol{X}) = P\left[\frac{(Y_{g_C}^* - Y_{g_A}^* - \beta)}{\sqrt{2}} > \frac{-\beta}{\sqrt{2}}\right],$$

regardless of $\boldsymbol{X}$ values, which for the probit, logit and complementary loglog link [2] is, respectively:

$$\gamma = \Phi(\hat{\beta}/\sqrt{(2)}), \qquad \gamma \approx \frac{exp(\hat{\beta}/\sqrt{2})}{1+exp(\hat{\beta}/\sqrt{2})}, \qquad \gamma = \frac{exp(\hat{\beta})}{1+exp(\hat{\beta})}.$$

In the three expressions β is the group parameter. From γ index is possible to obtain the other measure

$$\Delta = 2\gamma - 1.$$

Both γ and Δ indexes are normalized ($\gamma \in [0,1]$ and $\Delta \in [-1,1]$); when $\gamma > 1/2$ or $\Delta > 0$, the ratings from g_C tend to be larger than the ratings from g_A assessing an ordinal superiority of g_C over g_A.

For γ and Δ, simple confidence intervals result directly from ordinary confidence intervals for β for the corresponding ordinal cumulative link model.

3 Results

Results point out the role played by coach adjustment enhancing the feeling and the relationship of the dyads. Table 1 reports the ordinal superiority measures along with the corresponding 95% confidence intervals for the probit model. The same results for the logit model are shown in Table 2 and for the extreme value model (complementary log–log link function) in Table 3. The value of $\hat{\gamma}$ lower than $1/2$, with exception of *pressure*, *sacrifice*, *talent*, and *work out*, indicates an ordinal inferiority of the ratings of coaches with respect to athletes. It means that coaches are less critical than their athletes. Actually, for the items *improvement* and *challenging* in the case of logit and probit models, $\hat{\Delta} \simeq 0$ providing similar responses in the two groups. Results of the latter model for an underlying extreme value distribution, which is plausible for the evaluations expressed by coaches and athletes, are quite different from the logit or probit models but lead to the same conclusions.

The ordinal superiority measures extend directly to summary comparisons of multiple groups, based on more general models that have multiple indicator variables for the groups. Thus, a possible extension may consider a triad analysis where athletes who are in teams may be analysed. In this case captain's rating may be introduced in the analysis reporting the evaluation of a peer.

Table 1 Ordinal superiority measures for probit model

	Probit model			
	$\hat{\gamma}$	$CI(\gamma)$	$\hat{\Delta}$	$CI(\Delta)$
anxiety	0.394	(0.319, 0.474)	-0.212	(-0.362, -0.052)
recovery	0.354	(0.281, 0.434)	-0.292	(-0.438, -0.132)
pressure	0.564	(0.483, 0.642)	0.128	(-0.034, 0.284)
get nervous	0.396	(0.319, 0.478)	-0.208	(-0.362, -0.044)
sacrifice	0.562	(0.481, 0.641)	0.124	(-0.038, 0.282)
challenging	0.479	(0.392, 0.568)	-0.042	(-0.216, 0.136)
talent	0.697	(0.621, 0.765)	0.394	(0.242, 0.530)
work out	0.593	(0.509, 0.673)	0.186	(0.018, 0.346)
improvement	0.477	(0.394, 0.560)	-0.046	(-0.212, 0.120)

Acknowledgments: This research was carried out in the context of the project "Statistical Modelling and Data Analytics for Sports. Psychosocial aspects to assess the performance: the case of swimmers" (University of Naples Federico II-Federazione Italiana Nuoto Campania) and partially supported by Osservatorio Re-

Table 2 Ordinal superiority measures for logit model

	Logit model			
	$\hat{\gamma}$	$CI(\gamma)$	$\hat{\Delta}$	$CI(\Delta)$
anxiety	0.388	(0.309, 0.473)	-0.224	(-0.382, -0.054)
recovery	0.345	(0.269, 0.428)	-0.310	(-0.462, -0.144)
pressure	0.563	(0.477, 0.646)	0.126	(-0.046, 0.292)
get nervous	0.386	(0.306, 0.471)	-0.228	(-0.388, -0.058)
sacrifice	0.556	(0.470, 0.640)	0.112	(-0.060, 0.280)
challenging	0.469	(0.377, 0.562)	-0.062	(-0.246, 0.124)
talent	0.703	(0.623, 0.773)	0.406	(0.246, 0.546)
work out	0.603	(0.513, 0.688)	0.206	(0.026, 0.376)
improvement	0.470	(0.381, 0.561)	-0.060	(-0.238, 0.122)

Table 3 Ordinal superiority measures for extreme value model

	Extreme value model			
	$\hat{\gamma}$	$CI(\gamma)$	$\hat{\Delta}$	$CI(\Delta)$
anxiety	0.405	(0.335, 0.478)	-0.190	(-0.330, -0.044)
recovery	0.371	(0.305, 0.442)	-0.257	(-0.390, -0.116)
pressure	0.534	(0.458, 0.608)	0.068	(-0.084, 0.216)
get nervous	0.419	(0.350, 0.491)	-0.162	(-0.300, -0.018)
sacrifice	0.557	(0.478, 0.633)	0.113	(-0.044, 0.266)
challenging	0.443	(0.347, 0.542)	-0.114	(-0.306, 0.084)
talent	0.640	(0.560, 0.696)	0.262	(0.120, 0.392)
work out	0.516	(0.433, 0.597)	0.032	(-0.134, 0.194)
improvement	0.408	(0.330, 0.490)	-0.184	(-0.340, -0.020)

gionale delle Politiche Giovanili 2- POR CAMPANIA FSE 2014-2020-Cup project: E64I19002390005.

References

1. Agresti, A.: *Analysis of Ordinal Categorical Data*, 2^{nd} ed. Hoboken: Wiley (2010)
2. Agresti, A. and Kateri, M.: Ordinal Probability Effect Measures for Group Comparsons in Multinomial Cumulative Link Models. *Biometrics*, **73**, 214–219 (2017)
3. Agresti, A., and Tarantola, C.: Simple Ways to Interpret Effects in Modeling Ordinal Categorical Data, *Statistica Neerlandica*, **72**, 210–223 (2018)
4. Bangdiwala, K.: Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS Users Group International Conference*, 12:1083–1088 (1987)
5. Friendly, M.: *Visualizing Categorical Data*. Cary, NC: SAS Institute Inc. (2000)
6. Greenacre, M.: *Correspondence Analysis in Practice*. Third Edition. New York: Chapman and Hall/CRC (2017)
7. Kenny, D. A., Kashy, D. A., Cook, W.: *Dyadic data analysis*. New York: Guilford (2006)
8. McCullagh, P.: Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980)

Simulated annealing for maximum rater agreement

Simulated annealing per il calcolo della massima concordanza

Fabio Rapallo and Maria Piera Rogantin

Abstract In this work we consider the weighted kappa as a measure of rater agreement for ordinal variables, and we use a simulated annealing algorithm to find the maximum agreement configuration. The proposed algorithm allow us to show, through some examples, that the maximum agreement depends strongly on the choice of the weighting scheme.

Abstract *In questo lavoro consideriamo l'indice kappa pesato come misura di rater agreement per variabili ordinali e utilizziamo un algoritmo di simulated annealing per determinare la massima concordanza. L'algoritmo permette di mostrare, tramite esempi, come la massima concordanza dipenda dalla scelta dei pesi.*

Key words: Algebraic Statistics, Markov bases, MCMC maximization

1 Introduction

Rater agreement analysis is one of the most active research areas in the field of categorical data analysis. Let us consider two (or more) raters which independently classify a set of N items based on a rating scale with m categories. For two raters, the data are therefore collected in a square $m \times m$ contingency table, where the diagonal cells correspond to agreement, while the off-diagonal cells correspond to disagreement. When multiple raters are involved and each of them classify all the N items, the data are collected in a multi-way contingency table.

Fabio Rapallo
Dipartimento di Economia, Università di Genova, via Vivaldi 5, 16126 Genova, e-mail: fabio.rapallo@unige.it

Maria Piera Rogantin
Dipartimento di Matematica, Università di Genova, via Dodecaneso 35, 16146 Genova, e-mail: rogantin@dima.unige.it

For two raters, the Cohen's kappa introduced in [4] and the weighted Cohen's kappa introduced in [5] have become the most popular measures of agreement. For a discussion on the Cohen's kappas, its generalizations to the multi-observer case, and other related indices of agreement, refer to [7].

The un-weighted version of the Cohen's kappa only distinguishes between agreement cells and disagreement cells, and thus it is used in case of ratings on a nominal scale. When the rating scale is ordinal, or in general when there are some disagreements to be considered more serious than others, then the weighted kappa should be preferred. In this framework, the choice of the weights is a delicate issue, see for instance [10] and the references therein for a recent discussion on the use and interpretation of different weighting schemes.

Another classical issue concerning the Cohen's kappa is the normalization. The kappa belongs to the family of "chance corrected measures of agreement", and its expression can be summarized in

$$\kappa = \frac{O-E}{1-E} \tag{1}$$

where O is the proportion of observed agreement and E is the proportion of expected agreement under independence of the ratings. In Eq. (1), κ can be considered a normalized measure, since the maximum value of κ is 1. But the value $\kappa = 1$ can not be actually reached when the marginal distributions are different. This fact leads to some difficulties in the interpretation of the measured κ, and even to classical counterexamples, see for instance [9].

In this work, we introduce a MCMC algorithm (simulated annealing) based on the notion of Markov bases to find the maximum agreement. Due to space limitation, we introduce the algorithm only in the case of two raters, but it can be extended also to the multi-rater case. As an application, we also discuss some examples to show that different weighting schemes lead to different maximum agreement tables.

2 Cohen's kappa and weighted Cohen's kappa

In the two-rater setting, let us denote with p_{ij} the probability of the cell (i,j), and with p_{i+} and p_{+j} the marginal distributions ($i,j = 1,\dots,m$). The Cohen's kappa is:

$$\kappa = \frac{\sum_{i=1}^m p_{ii} - \sum_{i=1}^m p_{i+}p_{+i}}{1-\sum_{i=1}^m p_{i+}p_{+i}} = 1 - \frac{\sum_{(i,j)\in D} p_{ij}}{\sum_{(i,j)\in D} p_{i+}p_{+j}}, \tag{2}$$

where $D = \{(i,j) \ : \ i \neq j\}$ is the set of the disagreement cells. Given a matrix of (symmetric) weights of agreement $W = (w_{ij})$ with $0 \leq w_{ij} \leq 1$ for all (i,j) and $w_{ii} = 1$ for all i, the weighted kappa is:

$$\kappa_w = \frac{\sum_{i,j=1}^m w_{ij}p_{ij} - \sum_{i,j=1}^m w_{ij}p_{i+}p_{+j}}{1-\sum_{i,j=1}^m w_{ij}p_{i+}p_{+j}} = 1 - \frac{\sum_{(i,j)\in D} u_{ij}p_{ij}}{\sum_{(i,j)\in D} u_{i,j}p_{i+}p_{+j}}, \tag{3}$$

where in the second expression $u_{ij} = 1 - w_{ij}$. In practice, the u_{ij} are weights of disagreement, and it is easily seen that $u_{ij} = 0$ on the main diagonal. Note that when $w_{ij} = 0$ for $i \neq j$ in Eq. (3) one recovers the un-weighted kappa in Eq. (2). When a sample is available, the indices κ and κ_w are estimated through the sample proportions.

Among the most commonly used weights there are:

1. the quadratic weights (see [8]):

$$w_{ij} = 1 - \frac{(i-j)^2}{(m-1)^2}$$

2. the linear weights (see [3]):

$$w_{ij} = 1 - \frac{|i-j|}{m-1}$$

In order to illustrate the theory, we also consider a square-root version of the weights, namely:

$$w_{ij} = 1 - \frac{\sqrt{|i-j|}}{\sqrt{m-1}}$$

To avoid confusion, we denote with $\kappa_{w,2}$, $\kappa_{w,1}$, and $\kappa_{w,0.5}$ the weighted kappa computed with the quadratic, linear, and square-root weights, respectively.

Remark 1. Note that in the case of the quadratic weights, the disagreement weights u_{ij} do not define a distance on the set $\{1,\ldots,m\}$, since the triangular inequality is not satisfied.

3 The algorithm

While the computation of the maximum agreement is simple for the un-weighted kappa (at least in the two-rater setting), the problem is not trivial when the weighted kappa is considered. In order to determine the configuration with maximum agreement, we use here the simulated annealing within the context of Algebraic Statistics, and therefore we make use of a Markov basis in order to navigate the fiber (or reference set) of an observed contingency table. For an introduction to Markov bases and their use in MCMC techniques for categorical data see [2].

The fiber of an observed table n_{obs} with respect to a linear sufficient statistic $T : \mathbb{N}^{m\times m} \longrightarrow \mathbb{N}^s$ is the set

$$\mathscr{F}_T(n_{\text{obs}}) = \left\{ n \in \mathbb{N}^{m\times m} \ , \ T(n) = T(n_{\text{obs}}) \right\} . \tag{4}$$

To ease the notation we drop the dependence on n_{obs} and we write simply $\mathscr{F}_T$ instead of $\mathscr{F}_T(n_{\text{obs}})$ when there is no ambiguity.

In our case, the function T is the sufficient statistics corresponding to the margins, i.e., $T(n) = (n_{1+}, \ldots, n_{m+}, n_{+1}, \ldots, n_{+m})$. To define a connected Markov chain on the fiber $\mathscr{F}_T$, we use an approach based on the Markov moves and Markov bases.

A *Markov move* is any table f with integer entries that preserves the linear function T, i.e. $T(n \pm f) = T(n)$ for all $n \in \mathscr{F}_T$. A *Markov basis* for a fiber $\mathscr{F}_T$ is a finite set of moves $\mathscr{M} = \{f^{(1)}, \ldots, f^{(r)}\}$ such that for all n_1 and n_2 in $\mathscr{F}_T$, there exist a sequence of moves $f^{(i_1)}, \ldots, f^{(i_A)}$ such that

$$n_2 = n_1 + \sum_{a=1}^{A} f^{(i_a)} \tag{5}$$

and

$$n_1 + \sum_{j=1}^{a} f^{(i_j)} \geq 0 \quad \text{for all } \; a = 1, \ldots, A\,. \tag{6}$$

Eqs. (5) and (6) state that with a Markov basis it is possible to connect any two tables of $\mathscr{F}_T$ staying non-negative.

When T is the linear statistics preserving the margins, it is known that a Markov basis is formed by basic moves, i.e., moves of the form

$$\begin{array}{cc} -1 & +1 \\ +1 & -1 \end{array} \qquad \text{or} \qquad \begin{array}{cc} +1 & -1 \\ -1 & +1 \end{array}$$

on each 2×2 minor of the table. More formally (see [6] and [11]):

$$\mathscr{M} = \left\{ f \in \mathbb{Z}^{m \times m} \; : \; f_{i_1 j_1} = f_{i_2 j_2} = 1, f_{i_1 j_2} = f_{i_2 j_1} = -1, \text{ and 0 otherwise} \right\}.$$

For brevity, a move of $\mathscr{M}$ is denoted with $f_{i_1 i_2 j_1 j_2}$ where the pedices specify the row and column indices of the relevant 2×2 minor. With such a Markov basis we can use the simulated annealing to maximize the observed agreement (or equivalently to minimize the disagreement) among all tables with fixed margins. Define the observed disagreement of a table n as

$$D(n) = \frac{1}{N} \sum_{i,j=1}^{m} u_{ij} n_{ij}\,.$$

Using the Markov moves to define the neighbors, the MCMC simulated annealing starts from the observed table and runs at each step b ($b = 1, \ldots, B$) as follows:

1. Choose a move $f \in \mathscr{M}$ and define $n' = n + f$;
2. If n' is non-negative, then move the chain from n to n' with transition probability $\min\{\exp((D(n) - D(n'))/\tau_b), 1\}$, where τ_b is the temperature at time b.

The reader can refer to [12] for a general introduction to simulated annealing in the discrete case and for a discussion on the computational details of the algorithm, as for instance the choice of the temperature function τ_b.

4 Examples and discussion

In this section we use our algorithm to show, through two numerical examples, that the maximum agreement depends strongly on the choice of the weighting scheme and that the solution is not unique in the case of linear weights. The first example is based on artificial data, while the second one is a real-world example.

The first example concerns a 3×3 table. In Table 1 on the left, there is the table of the observed counts. For this table, we have $\kappa_{w,2} = 0.5087$ with quadratic weights, $\kappa_{w,1} = 0.4328$ with linear weights, and $\kappa_{w,0.5} = 0.3934$ with square-root weights. Then, we can compute the maximum values of the weighted kappa by means of the MCMC algorithm introduced above: using quadratic weights, we obtain a weighted kappa $\kappa_{w,2} = 0.7399$ (Table 1, center-left); using linear weights we have $\kappa_{w,1} = 0.6073$ (Table 1, center-right); using square-root weights, we have $\kappa_{w,0.5} = 0.6586$ (Table 1, right).

Table 1 A 3×3 example. The observed table (left), the table with maximum $\kappa_{w,2}$ with quadratic weights (center-left), a table with maximum $\kappa_{w,1}$ with linear weights (center-right), the table with maximum $\kappa_{w,0.5}$ with square-root weights (right).

	1	2	3	
1	5	4	2	11
2	1	4	3	8
3	0	1	5	6
	6	9	10	25

	1	2	3	
1	6	5	0	11
2	0	4	4	8
3	0	0	6	6
	6	9	10	25

	1	2	3	
1	6	4	1	11
2	0	5	3	8
3	0	0	6	6
	6	9	10	25

	1	2	3	
1	6	1	4	11
2	0	8	0	8
3	0	0	6	6
	6	9	10	25

Remark 2. The configuration with maximum $\kappa_{w,1}$ is in general not unique. For instance, in the 3×3 example above, there are 5 different tables reaching the maximum value of $\kappa_{w,1} = 0.6073$. Among such tables, the value of $\kappa_{w,2}$ ranges from 0.5087 to 0.7399.

As a second example, we analyze the data from [1], page summarizing the diagnoses of multiple sclerosis for two neurologists on an ordinal scale. In Table 2 on the left, there is the table of the observed counts. For this table, we have $\kappa_{w,2} = 0.5246$ and $\kappa_{w,1} = 0.3797$. Here the $\kappa_{w,0.5}$ is not considered. In the configuration of maximum agreement using quadratic weights we have $\kappa_{w,2} = 0.7849$, while using linear weights we have $\kappa_{w,1} = 0.5715$.

Through the examples above we conclude that, in the weighted case, the solution of maximum agreement is not merely the table with the highest possible counts on the main diagonal, and this can be easily seen by comparing the solutions for the quadratic and the square-root weights. The differences between the optimal configurations under different weighting schemes decrease when the marginal distributions are nearly homogeneous.

Further work on this topic includes the extension to the multi-rater case, and an analytical study of the agreement in terms of the Markov moves. In fact, the

Table 2 The example from [1]. The observed table (left), the table with maximum $\kappa_{w,2}$ with quadratic weights (center), a table with maximum $\kappa_{w,1}$ with linear weights (right).

	1	2	3	4	
1	38	5	0	1	44
2	33	11	3	0	47
3	10	14	5	6	35
4	3	7	3	10	23
	84	37	11	17	149

	1	2	3	4	
1	44	0	0	0	44
2	40	7	0	0	47
3	0	30	5	0	35
4	0	0	6	17	23
	84	37	11	17	149

	1	2	3	4	
1	44	0	0	0	44
2	24	23	0	0	47
3	15	10	10	0	35
4	1	4	1	17	23
	84	37	11	17	149

non-uniqueness of the optimal configuration with linear weights suggests that the combinatorial structure of the fiber $\mathscr{F}_T$ has a relevant role in the rater agreement analysis. Finally, a simulation study should complement the analysis, especially for large tables or multi-rater problems, where the curse of dimensionality may affect the performance of the MCMC algorithm.

References

1. Agresti, A.: An Introduction to Categorical Data Analysis. John Wiley and Sons, New York (2007)
2. Aoki, S., Hara, H., Takemura, A.: Markov Bases in Algebraic Statistics. Springer-Verlag, New York (2012)
3. Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. American Journal of EEG Technology **11**(3), 101–110 (1971)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
5. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**(4), 213–220 (1968)
6. Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Statist. **26**(1), 363–397 (1998)
7. von Eye, A., Mun, E.Y.: Analyzing Rater Agreement: Manifest Variable Methods. Lawrence Erlbaum Associates, Mahwah, NJ (2004)
8. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ. Psychol. Meas. **33**(3), 613–619 (1973)
9. Flight, L., Julious, S.A.: The disagreeable behaviour of the kappa statistic. Pharm. Stat. **14**(1), 74–78 (2015)
10. Kvålseth, T.O.: An alternative interpretation of the linearly weighted kappa coefficients for ordinal data. Psychometrika **83**(3), 618–627 (2018)
11. Rapallo, F.: Algebraic markov bases and MCMC for two-way contingency tables. Scand. J. Statist. **30**(2), 385–397 (2003)
12. Suman, B., Kumar, P.: A survey of simulated annealing as a tool for single and multiobjective optimization. J. Oper. Res. Soc. **57**(10), 1143–1160 (2006)

Models and methods – Regression

A Clusterwise regression method for Distributional-valued Data

Un metodo di regressione cluster-wise per dati distribuzionali

Rosanna Verde, Francisco de A. T. de Carvalho, Antonio Balzanella

Abstract In this paper we propose a cluster-wise regression strategy for aggregated data represented by distributions. The basic idea is that the set of observed data are related by local causal relationships for different clusters. Some cluster-wise regression methods are based on K-means clustering algorithms, where the representatives of clusters are expressed by linear models and the assignment of an element to a cluster is performed according to the minimum distance to the model, in the sense of ordinary least squares (OLS). The proposed method extends this strategy to data expressed by empirical distributions or histograms. The present work refers to one of the regression models proposed in the analysis of the dependence of histogram type variables and to the k-means algorithm developed for such data. The metric used is the L_2 Wasserstein distance. An application on real distributional data corroborate the method.

Abstract *In questo lavoro si propone una strategia di regressione cluster-wise per dati in forma di distribuzioni. L'idea é che l'insieme dei dati osservati presentino una differente relazione di causualitá tra una variabile dipendente e un insieme di variabili indipendenti, per ciascuna classe. In letteratura diversi metodi di regressione cluster-wise si basano su un algoritmo di tipo K-means, dove i rappresentanti dei cluster sono espressi da modello di regressione lineare e l'assegnazione avviene sulla base della minima distanza dal modello, nel senso dei minimi quadrati. Il metodo proposto estende tale strategia a dati espressi da distribuzioni empiriche o istogrammi. Il presente lavoro fa riferimento ad uno dei modelli di regressione*

Rosanna Verde
Departement of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Viale A. Lincoln, 5, 81100 Caserta, Italie e-mail: rosanna.verde@unicampania.it

Francisco T. de A. de Carvalho
CIN-UFPE, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitria 50.740-560, Recife, PE, Brasil e-mail: fatc@cin.ufpe.br

Antonio Balzanella
Departement of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Viale A. Lincoln, 5, 81100 Caserta, Italie e-mail: antonio.balzanella@unicampania.it

proposti nell'ambito dell'analisi della dipendenza di variabili di tipo istogramma e all'algoritmo k-means, esteso a tali dati. In particolare, la metrica utilizzata é la distanza L_2 di Wasserstein. Un'applicazione su distribuzioni di dati reali permette di validare la metodologia proposta.

Key words: Cluster-wise regression, Distributional data, Wasserstein distance

1 Introduction

In this paper we propose a cluster-wise regression strategy for distributional-valued data, in the framework of distributional data analysis [1], [8].

Cluster-wise linear regression (CLR) comprises a collection of methods developed to find the best fitting models according to the clustering of data [14], therefore, they aim at finding simultaneously the clusters and the models, by optimizing suitable criteria.

A pioneering paper in the search of local models and a clustering of data is the Typological Principal Component Analysis [7]. It is based on the simultaneous search, by alternating steps, of k subspaces of maximal inertia and on the assignment of the elements to the cluster according to the minimum distances to the local plane, until the convergence to a stable partition and K final sub-spaces. Then, Spath ([14], [15], [16]) focuses on the partitioning of a set of objects into K classes and establishing a regression model within each class. Thus, CLR can be characterized as a combination of cluster analysis (partitioning) and regression, where the representatives of the clusters are expressed by linear models and the assignment of an element to a cluster is performed according to the minimum distance to the model, in the sense of ordinary least squares. Preda et al. [13] proposed a clusterwise method based on PLS regression. Mixture-model formulations of CLR have been also proposed in [4], [12], [18] that assume the response variable estimations to be obtained from a mixture of K conditional density distributions.

In the framework of Symbolic Data Analysis [1], [3] presented a cluster-wise generalization for interval data based on a double regression on the centers and radii of the intervals, according to a suitable strategy in the analysis of this kind of multivalued data. Here, we propose a generalization of the K-means based cluster-wise regression for another type of symbolic data, the distributional-valued data, which are expressed by empirical distributions or histograms.

Distributional-valued data are a way to summarize large amounts of data, as they allow to treat information already in aggregate form. Recently many methods have been developed based on a suitable metric, the L_2 Wasserstein distance; they have allowed to extend classical data analysis methods to distribution-valued data. The cluster-wise regression method, here proposed, refers to one of the regression models for histogram type variables [8] and to the k-means algorithm developed for such data [9].

The main contribution of the proposed algorithm consists in analysing linear dependence structures for clusters of distributional data, which take into account the different characteristics (location, variability and shape) of the distributions in each cluster, also related to the decomposition property of the L_2 Wasserstein distance. An application on distribution of aggregated real data corroborates the procedure.

2 Distributional-valued data and L_2 Wasserstein metric

Let $Q = \{o_1, \ldots, o_N\}$ be a set of N objects described by $p+1$ distributional-valued variables. We assume that one of the $p+1$ distributional-valued variables is related to a dependence relationship by the other p ones. Then, we denote with Y the dependent variable and with X_j $(j = 1, \ldots, p)$ the independent variables, observed on the N individuals. Each i-th object o_i $(1 \leq i \leq N)$ is represented by $p+1$ distributions (or distributional-valued data): f_i^y, f_{ij}^x $(j = 1, \ldots, P)$.

Several distances to compare density or frequency distributions have been proposed in the literature. Among them, we consider the L_2 Wasserstein distance [17] due to its properties.

Given two *pdf*'s $f(x)$ and $g(x)$, with means $\bar{x}_f$ and $\bar{x}_g$ and finite standard deviations s_f and s_g. From the associated cumulate distribution functions $F(x)$ and $G(x)$ and the respective *quantile functions* $x_f(t) = F^{-1}(t)$ and $x_g(t) = G^{-1}(t)$, the L_2 Wasserstein distance [17] is defined as follows:

$$d_W(f,g) = \sqrt{\int_0^1 \left[x_f(t) - x_g(t)\right]^2 dt}. \tag{1}$$

In [2] it is shown that the squared L_2 Wasserstein distance can be rewritten as:

$$d_W^2(f,g) = \left(\bar{x}_f - \bar{x}_g\right)^2 + \int_0^1 \left[x_f^c(t) - x_g^c(t)\right]^2 dt. \tag{2}$$

Where $\bar{x}_f$ and $\bar{x}_g$ are the averages of the distributions and $x_f^c(t)$ and $x_g^c(t)$ are the respective centered quantiles functions.

3 Clusterwise Regression for Distributional-valued data (CRD)

The proposed method looks for a partitioning of the set Ω of distributional data into K homogeneous clusters and the best fitting regression equation for each cluster.

The regression model used to fit distributional data was introduced by [10], as follows:

$$y_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j x_{ij}^c(t) + e_i(t), \quad \forall t \in [0,1] \tag{3}$$

where: β_o is the constant, β_j's are the coefficients associated to the vectors of the averages $\bar{x}_{ij}$ of the distributions f_{ij}, and γ_j are the coefficients of the centered quantile functions x_{ij} ($j = 1, \ldots, p$).

This model decomposition, due to the Wasserstein distance properties, allows to consider separately the two components related to the location and the variability-shape of the distributions. In order to estimate these parameters, we define the Sum of Square Errors function (SSE) like in the LS method, using the squared Wasserstein L_2 metric:

$$\begin{aligned} SSE(\beta_0, \beta_j, \gamma_j) &= \sum_{i=1}^{n} \int_0^1 e_i^2(t)dt = \\ &= \sum_{i=1}^{n} \int_0^1 \left[y_i(t) - \left(\beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j x_{ij}^c(t) \right) \right]^2 \end{aligned} \tag{4}$$

According to the aim of the cluster-wise regression model, fixed the number K of clusters, the better partition $P_k = C_1, \ldots, C_K$ and the best fitting models $\hat{y}^k$ for each cluster C_k, are achieved by minimising the $SSE(\beta_0^k, \beta_j^k, \gamma_j^k | P_k)$:

$$\begin{aligned} SSE(\beta_0^k, \beta_j^k, \gamma_j^k | P_k) &= \sum_{k=1}^{K} \sum_{i \in C_k} \int_0^1 e_{ik}^2(t)dt = \\ &= \sum_{k=1}^{K} \sum_{i \in C_k} \int_0^1 \left[y_i^k(t) - \left(\beta_0^k + \sum_{j=1}^{p} \beta_j^k \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j^k x_{ij}^c(t) \right) \right]^2 dt \end{aligned} \tag{5}$$

The assignment of the element o_i to a cluster C_k is performed according to the minimum error e_{ik}.

The convergence of the algorithm is guaranteed by the decreasing of the criterion relating to the improvement of the fitting of the regression models of clusters.

We consider two indexes to evaluate the goodness of fit of the cluster-wise regressions: the Ω index proposed by [6], and the $RMSE_W$ (Root Mean Square Error, according to the L_2 Wasserstein distance), computed for each cluster (here denoted as Ω^k and $RMSE_W(C_k)$), and globally ($RMSE_W(P_k)$) for the partition:

$$\Omega^k = \frac{\sum_{i \in C_k} d_W^2(\hat{y}_i^k(t), \bar{y}^k)}{\sum_{i \in C_k} d_W^2(y_i^k(t), \bar{y}^k)}; \; RMSE_W(C_k) = \sqrt{\frac{SSE(\beta_0^k, \beta_j^k, \gamma_j^k | C_k)}{n_k}} \tag{6}$$

$$RMSE_W(P_k) = \sqrt{\frac{\sum_{k=1}^{K} SSE(\beta_0^k, \beta_j^k, \gamma_j^k | C_k) \cdot n_k}{N}} \tag{7}$$

where: $\bar{y}^k$ (for $k = 1, \ldots, K$) and $\bar{y}$ are the averages of the distributional-values of the y_i's belonging to the cluster C_k, for each $C_k \in P_k$, and the average of the all

observed distributional-values y_i (for $i = 1, \ldots, N$), respectively; n_k is the cardinality of the cluster C_k (for $k = 1, \ldots, K$).

4 Application

The proposed method has been corroborated on the air quality dataset from the Clean Air Status and Trends Network (CASTNET), also used in [10]. The data are refereed to the Ozone concentration in 78 USA sites and to some related variables. Especially, we have selected Temperature (X_1) ($^\circ C$), Solar Radiation (X_2) (W/m^2) and of Wind Speed (X_3) (m/s), as explicative variables to study the linear effect on the Ozone concentration (Y) (ppb). The period of reference to summarize the hourly data is the summer season of 2010 and the central hours of the days (10 a.m. - 5 p.m.). For each one of the 78 site, Ozone concentration, Temperature, Solar Radiation and Wind Speed are expressed by their "summer-daylight hours" distributions (histograms). Setting the number of clusters to $K = 5$, the results of CDR are shown in Tab. 4 compared to the global Regression model of Distributional data (RD):

Method	N	Ω	$RMSE_W$
RD	78	0.81	4.54
	n_k	Ω^k	$RMSE_W(C_k)$
CRD(C_1)	18	0.93	2.71
CRD(C_2)	34	0.92	3.02
CRD(C_3)	4	0.90	4.76
CRD(C_4)	10	0.87	3.35
CRD(C_5)	12	0.86	3.47
			$RMSE_W(P_k)$
CRD			**3,15**

The clusterwise regression performs a better fitting of the regression models, for each cluster data, as shown by the higher values of the indexes Ω^k and $RMSE_W(C_k)$ compared with the $\Omega = 0.81$ and $RMSE_W = 4.54$ of the global model (RD). Only the $RMSE_W(C_3)$ is a little higher than the $RMSE_W$ compensate for the Ω^3 lower than Ω. Based on these results, we are confident that the CRD can better catch the linear dependence of the regressor from the exploratory distributional variables when data arise from different sub-populations or the relationship between Y and the X_j variables is not linear.

5 Conclusion

In this paper we have introduced a suitable clusterwise regression model for distributional data. The proposed model combines the K-means clustering algorithm with a regression model for distributional-valued data in order to identify both a partition of the data and the best fitting regression models (one for each cluster). The use of the L_2 Wasserstein distance has allowed to analyse the linear relationships of the independent variables on the response one, with respect to the averages of the distributions and the centred quantile functions which express the variability and shape characteristics of the distribution of data. The application of the proposed strategy on environmental data sets has shown the interest of this approach.

References

1. Bock H.-H., Diday E., Analysis of symbolic data: exploratory methods for extracting statistical information from complex data, Springer Verlag, Heidelberg, 2000.
2. , Matrán C., Tuero-Díaz A., Optimal transportation plans and convergence in distribution, J. Multivar. Anal., vol. 60, 1997, pages 72–83.
3. De Carvalho F. de A. , Saporta G., Queiroz D. N., A Clusterwise Center and Range Regression Model for Interval-Valued Data, Proceeding of COMPSTAT 2010, Lechevallier Y., Saporta G. (eds), Springer Physica-Verlag, 2010, pp. 461-468.
4. DeSarbo, W.S., Cron, W.L., A maximum likelihood methodology for clusterwise linear regression. Journal of Classification, 5, pp.249-282.
5. DeSarbo, W.S , Kamakura W.A. , Wedel M. (2005): Latent Structure Regression, In: Handbook of Marketing Research , R. Grover, M. Vriens, (eds), London, Sage, 394-417.
6. Dias, S. Brito, P. (2005): Linear regression model with histogram-valued variables, Statistical Analysis and Data Mining, Wiley, 8, 2, pp. 75-113.
7. Diday, E., Introduction lanalyse factorielle typologique, Revue de Statistique Applique, 22, 4, pp.29-38
8. Irpino A., Verde R., Basic statistics for distributional symbolic variables: A new metric-based approach, Adv. Data Anal. Classif., vol. 9, issue 2, 2015, pp. 143–175.
9. Verde R., Irpino A., Dynamic Clustering of Histogram Data: Using the Right Metric, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 123–134.
10. Irpino A., Verde R., Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance, Adv. Data Anal. Classif., vol. 9, issue 1, 2015, 81-106
11. Hennig, C., Models and methods for clusterwise linear regression. In: Classification in the Information Age, Springer, pp.179-187.
12. Hennig, C., Identifiability of models for Clusterwise linear regression. Journal of Classification, 17, pp.273-296.
13. Preda, C., Saporta, G., Clusterwise PLS regression on a stochastic process. Computational Statistics and Data Analysis, 49, 2005, pp.99108.
14. Spaeth, H., Clusterwise linear regression, Computing, 22, 1979, pp.367-373
15. Spaeth, H., Algorithm 48: A fast algorithm for clusterwise linear regression. Computing, 29, 1982, 175181.
16. Spaeth, H., Mathematical algorithms for linear regression. San Diego, CA: Academic, 1991.
17. Wasserstein L. Markov processes over denumerable products of spaces describing large systems of automata, Prob. Inf. Transmission, vol. 5, 1969, pp. 47–52.
18. Wedel M., DeSarbo W.S., A Mixture Likelihood Approach for Generalized Linear Models, Journal of Classification, vol. 12, 1995 pp. 2155.

A nonparametric approach for nonlinear variable screening in high-dimensions

Un approccio non parametrico per la selezione di variabili non lineari in alta dimensionalità

Giordano Francesco, Milito Sara and Parrella Maria Lucia

Abstract We present a fully-nonparametric method for screening selection called Derivative Empirical Likelihood Independent Screening (D-ELSIS) for regression analysis in high-dimensions. Since our method is model-free, it is able to identify explanatory variables that contribute to the explanation of the response variable in nonparametric and non-additive contexts. The proposed method is a two-step approach that combines the estimation of marginal derivatives by local polynomials together with the empirical likelihood technique. In the first step D-ELSIS selects the relevant variables, while in the second step it identifies the nonlinear ones. The simulation results give evidence of a good performance for D-ELSIS.

Abstract *Presentiamo un metodo non parametrico per la selezione di variabili chiamato Derivative Empirical Likelihood Independent Screening (D-ELSIS) per la regressione in alta dimensionalità. Poichè il nostro metodo è model-free, è capace di identificare le variabili esplicative che contribuiscono alla spiegazione della variabile di risposta in constesti non parametrici e non additivi. Il nostro approccio prevede due step e combina la stima della derivata marginale effettuata con i polinomi locali con la tecnica dell'empirical likelihood. Nel primo step D-ELSIS seleziona le variabili rilevanti, mentre nel secondo identifica le variabili non lineari. Le simulazioni presentate mostrano come il metodo di screening D-ELSIS funzioni in modo soddisfacente.*

Key words: screening selection, high-dimensions, nonparametric regression.

Giordano Francesco
University of Salerno, Via Giovanni Paolo II 84084 Fisciano (SA), Italy e-mail: giordano@unisa.it

Milito Sara
University of Salerno, Via Giovanni Paolo II 84084 Fisciano (SA), Italy e-mail: smilito@unisa.it

Parrella Maria Lucia
University of Salerno, Via Giovanni Paolo II 84084 Fisciano (SA), Italy e-mail: mparrell@unisa.it

1 Introduction

Statistical analysis with high-dimensional data has the main purpose of identifying a set of relevant variables that contribute to the explanation of a given response variable. Often, however, we are also interested in the type of relationship between the explanatory variables and the response. Suppose that we have a random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the data model

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i \tag{1}$$

where Y is the response variable, $\mathbf{X}$ is the vector of p candidate variables and ε is the error, with $E(\varepsilon|\mathbf{X}) = 0$. Let X_j denote a generic univariate covariate. Furthermore, $m(\cdot)$ is the unknown regression function. In our method, we do not impose any particular form to the function $m(\cdot)$, so we consider a general class of regression problems, including additive and non-additive models. Regarding the dimensionality p of the variable vector $\mathbf{X}$, this can grow exponentially with the sample size n, and, without loss of generality, we assume that $E(Y) = 0$ implying that $E\{m(\mathbf{X})\} = 0$. Let us denote with

$$M_* = \{1 \leq j \leq p : \text{the } j\text{-th variable in } \mathbf{X} \text{ is relevant for explanation of } Y\}$$

the set of s true relevant covariates in model (1). Moreover, we consider a very sparse model: only a small fraction of the candidate variables contributes to the response ($s \ll p$). Among these s covariates, we assume that l have a linear effect on the response and $s - l$ have a nonlinear effect.

In order to identify the s relevant covariates and the nonlinear ones in M_*, in a high-dimensional nonparametric framework with p very large, we propose a two-step independence model-free feature screening technique that combines two different elements: the local polynomial regression and the empirical likelihood technique. The procedure is called Derivative Empirical Likelihood Independence Screening (D-ELSIS). In the first step, we apply the local polynomial regression to estimate a marginal derivative with respect to the covariate X_j, for $j = 1, \ldots, p$ (so, p derivatives in total). By using the derivatives, we investigate the marginal contribution from each variable in explaining Y to justify whether it is relevant or not. Then, the procedure checks by the empirical likelihood (a nonparametric inference method, see Owen (2001)) if these derivatives are uniformly zero (or not) on the support of each variable. Those variables for which the test is passed are chosen as relevant covariates. In the second step, we choose the top ranked variables from the first step, and we consider, for each of these, the difference between the marginal derivative and its mean. If this difference is uniformly zero on the support, then we conclude that the covariate taken into consideration is linear. In order to test this difference we use, again, the empirical likelihood. Those variables for which the test is passed are chosen as nonlinear covariates. Until now, based on our knowledge, no other screening method uses the local polynomial estimation of the first marginal

derivative with a twofold purpose: screening of relevant variables and selection of nonlinear ones (see Chang et al. (2016), Lian et al. (2014) and the reference therein).

In Section 2 we review the local polynomials for derivative estimation. In Section 3 we introduce the D-ELSIS procedure to carry out screening of relevant variables and we show how to use it for nonlinear screening. A simulation study in Section 4 assesses the performance of the proposed D-ELSIS method.

2 Derivative estimation by local polynomials

Denoting with $f(x)$ the marginal relation between Y and X, where $X = X_1, \ldots, X_n$ is a one-dimensional explanatory variable, its local polynomial estimation is derived by means of a weighted least squares regression fitted at each point of interest, x, using data from its neighbourhood. Suppose that the $(d+1)^{th}$ derivative of $f(x)$ at the point x exists and is continuous. We approximate locally the marginal function $f(x)$ using Taylor's Expansion by a polynomial of order d. Then, we can estimate the expansion terms using weighted least squares by minimizing the following equation:

$$\sum_{i=1}^{n} \left[Y_i - \sum_{v=0}^{d} \beta_v(x)(X_i - x)^v \right]^2 K_h(X_i - x) \tag{2}$$

where h, called the bandwidth, controls the size of the neighbourhood around x, $K_h(\cdot)$ controls the weights and $K_h(x) \equiv K(x/h)/h$ with K a kernel density function. Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$ be the solution of (2), then $\hat{f}^{(v)}(x) = v!\hat{\beta}_v$ is an estimator for $f^{(v)}(x)$ with $v = 0, \ldots, d$. In matrix notation, let $\mathbb{X}$ be the design matrix centred at x:

$$\mathbb{X} = \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^d \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \ldots & (X_n - x)^d \end{pmatrix}, \tag{3}$$

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\mathbf{W}$ a diagonal matrix of weights with diagonal elements $K_h(X_i - x)$, for $i = 1, \ldots, n$. Then, the local estimate of $f^{(v)}(x)$ with a dth degree polynomial is

$$\hat{f}^{(v)}(x; d, h) = v!\mathbf{e}_{v+1}^T (\mathbb{X}^T \mathbf{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbf{W} \mathbf{Y} = v! \sum_{i=1}^{n} W_{i,d,h}(x) Y_i \tag{4}$$

for $v = 0, \ldots, d$, where $W_{i,d,h}(x) = \mathbf{e}_{v+1}^T (\mathbb{X}^T \mathbf{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbf{W} \mathbf{e}_i$. Here $\mathbf{e}_r$ is the $(d+1) \times 1$ vector having 1 in the rth entry and zeros elsewhere. In general, we need a polynomial of order $d = v + 1$ (see Fan and Gijbels (1996)), so we use the local quadratic estimator, *i.e.* $d = 2$, in order to estimate the first derivative ($v = 1$).

3 The proposed procedure

In the first step of the procedure, we estimate the first marginal derivative of our nonparametric model (1) using (4) with $d = 2$. In order to use an univariate local estimator, we consider $f_j(x) = E(Y|X_j = x)$, that is the marginal contribution of X_j locally at x. Although this significantly alters the regression function, it does not entail significant consequences in terms of variable selection, since both the original (non-additive) model and the one we work with (its marginalized additive approximation) depend on the same set of relevant variables.

In order to implement the proposed method for each variable X_j we select n_{X_j} equally spaced points between the first and the third quartile of the support $\mathscr{X}_j$, obtaining the set $C_j = (x_{1j}, \ldots, x_{n_{X_j}j})$. For assessing $f_j'(x) := f^{(1)}(x) = 0$ at given x without distributional assumptions, we construct the following empirical likelihood ratio (*e.l.r.*) statistic, following the same steps of Chang et al. (2016):

$$EL_j(x,0) = \sup_w \left\{ \prod_{i=1}^{n} w_i : w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i W_{i,2,h}(x) Y_i = 0 \right\}. \tag{5}$$

Applying the Lagrange multiplier method, we obtain the logarithm of the *e.l.r.*

$$l_j(x,0) = -2\log\{EL_j(x,0)\} - 2n\log n = 2\sum_{i=1}^{n} \log\{1 + \lambda W_{i,2,h}(x) Y_i\} \tag{6}$$

where λ is the univariate Lagrange multiplier solving $\sum_{i=1}^{n} \frac{W_{i,2,h}(x)Y_i}{1+\lambda W_{i,2,h}(x)Y_i} = 0$.
Since $\sum_{i=1}^{n} W_{i,2,h}(x)Y_i$ converges in probability to the marginal derivative of X_j evaluated at x, a large value of $l_j(x,0)$ is taken as evidence against $f_j'(x) = 0$. Then, $l_j(x,0)$ is a statistic for testing whether or not (4) with $d = 2$ has zero mean locally at x. For assessing $f_j'(x) \equiv 0$ uniformly on $\mathscr{X}_j$, we use

$$l_j(0) = \sup_{x \in \mathscr{X}_j} l_j(x,0) \quad \forall\, j = 1, \ldots, p.$$

We evaluate the statistic l_j using $l_j(0) = \max_{1 \leq i \leq n_{X_j}} l_j(x_{ij}, 0)$. With this expedient, we can use the univariate optimisation to solve (6) by the Lagrange multiplier method. For feature screening purposes, we sort l_j for all $j = 1, \ldots, p$ in decreasing order. In this way, we identify, for a given threshold γ_n, the set of relevant covariates

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : l_j \geq \gamma_n\}.$$

Since γ_n is difficult to estimate, we only consider the top ranked covariates.

In the second step of the procedure, we consider the null hypothesis that the covariate j is linear, which means $f_j'(x) - E(f_j'(\cdot)) = 0$ for each $x \in C_j$. Moreover, let $\widehat{\mu}_j = 1/n_{X_j} \sum_{u=1}^{n_{X_j}} \widehat{f}_j'(\mathrm{x}_{uj})$ be an estimator of the expected value of $f_j'(\cdot)$. Then, we construct the following empirical likelihood statistic for non linear covariates

$$ELN_j(x,0) = \sup_w \left\{ \prod_{i=1}^{n} w_i : w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i [W_{i,2}(x) Y_i - \widehat{\mu}_j] = 0 \right\}, \quad (7)$$

and the respective logarithm of *e.l.r.*

$$ln_j(x,0) = -2\log\{ELN_j(x,0)\} - 2n\log n = 2\sum_{i=1}^{n} \log\left\{1 + \lambda\left[W_{i,2}(x)Y_i - \widehat{\mu}_j\right]\right\} \quad (8)$$

where λ is the univariate Lagrange multiplier solving $\sum_{i=1}^{n} \frac{W_{i,2}(x)Y_i - \widehat{\mu}_j}{1+\lambda\left[W_{i,2}(x)Y_i - \widehat{\mu}_j\right]} = 0$. Since $\sum_{i=1}^{n}\left[W_{i,2}(x)Y_i - \widehat{\mu}_j)\right]$ converges in probability to the difference between the marginal derivative of X_j evaluated at x and its mean, a large value of $ln_j(x,0)$ is taken as evidence that the covariate j is nonlinear, which means $f_j'(x) - E(f_j'(\cdot)) \not\equiv 0$. For assessing $f_j'(x) - E(f_j'(\cdot)) \equiv 0$ uniformly on $\mathscr{X}_j$, we use

$$ln_j(0) = \sup_{x \in \mathscr{X}_j} ln_j(x,0) \quad \forall\ j = 1, \ldots, p.$$

Again, we evaluate the sup by the max of the test-statistic over C_j. For nonlinear feature screening purposes, we sort ln_j for all $j = 1, \ldots, p$ in decreasing order and we take the top ranked covariates as the nonlinear covariates of model (1).

4 Simulations

Several simulation experiments are conducted to investigate the performance of the proposed D-ELSIS method. Concerning the empirical likelihood, we used the R package **emplik**, while for the estimation of the bandwidth we used the R package **NonpModelCheck**. We set $n = (200, 500)$ and $p = (50, n/2, 2n)$. We report three criteria: (i) the median of the Minimum Model Size (MMS), *i.e.* the minimum number of predictors needed to keep all the relevant ones, for 100 repetitions; (ii) the IQR divided by 1.34 (SD), *i.e.* a robust measure of the standard error of MMS; (iii) the True Positive Rate in percentage (TPR) that controls the precision measuring the proportion that all active predictors are selected. To calculate the TPR we consider that the predicted relevant variables are included in the first 20, while the nonlinear ones are in the first 10. We consider a model with both linear and non linear covariates. This case is taken from Example 2 of Chang et al. (2016). Data are generated from the model

$$Y = 5X_1 + 3(2X_2 - 1)^2 + 4\frac{\sin(2\pi X_3)}{2 - \sin(2\pi X_3)} + 6\big[0.1\sin(2\pi X_4) +$$
$$0.2\cos(2\pi X_4) + 0.3(\sin(2\pi X_4))^2 + 0.4(\cos(2\pi X_4))^3 + 0.5(\sin(2\pi X_4))^3\big] + \sigma\varepsilon.$$

Here the predictors X_j are *i.i.d* random variables of Uniform distribution $U(0,1)$, and $\varepsilon \sim N(0,1)$ is independent of all X_j. In this case we have 3 relevant nonlinear

Table 1 Simulation results

n	p		relevant variables				linear vs nonlinear			
			$\sigma^2=1$	$\sigma^2=1.74$	$\sigma^2=2$	$\sigma^2=3$	$\sigma^2=1$	$\sigma^2=1.74$	$\sigma^2=2$	$\sigma^2=3$
200	50	MMS(SD)	4(0.00)	4(0.00)	4(0.00)	4(0.00)	3(0.75)	3(0.75)	3(0.75)	3(0.75)
		TPR	100.00	100.00	100.00	100.00	99.00	100.00	100.00	99.33
	100	MMS(SD)	4(0.00)	4(0.00)	4(0.00)	4(0.00)	3(0.75)	3(0.75)	3(0.75)	3(0.75)
		TPR	100.00	100.00	100.00	100.00	99.33	100.00	99.33	100.00
	400	MMS(SD)	4(0.00)	4(0.00)	4(0.75)	4(0.19)	3(0.75)	3(0.75)	3(0.75)	3(0.75)
		TPR	100.00	100.00	100.00	100.00	99.67	99.33	98.67	99.33
500	50	MMS(SD)	4(0.00)	4(0.00)	4(0.00)	4(0.00)	3(0.00)	3(0.00)	3(0.00)	3(0.00)
		TPR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	250	MMS(SD)	4(0.00)	4(0.00)	4(0.00)	4(0.00)	3(0.00)	3(0.00)	3(0.00)	3(0.00)
		TPR	100.00	100.00	100.00	99.75	100.00	100.00	99.67	99.67
	1000	MMS(SD)	4(0.00)	4(0.00)	4(0.75)	4(0.19)	3(0.00)	3(0.00)	3(0.75)	3(0.00)
		TPR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

covariates and 1 relevant linear covariate. We consider four different signal to noise ratios by varying σ^2, as in Chang et al. (2016).

From Table 1 we are able to correctly identify the set of the 4 true relevant variables in the first 20 top ranked covariates and the 3 true nonlinear variables in the first 10 position for all signal to noise ratio considered, with a precision at least of 98%. In the first step we have a value 4 for the median of MMS, while in the second step this value is 3. This means that the only linear covariate is identified among the relevant ones in the first step, while in the second step it is positioned in the ranking after the nonlinear ones. The (average) computational times in seconds for one iteration of D-ELSIS procedure, using a Windows 10 PC with Intel Core i7, Dual-Core 2.70 GHz, are reported in the following table:

n	200			500		
p	50	100	400	50	250	1000
	5.85	8.79	40.71	7.24	27.71	120.07

References

Chang, J., Tang, C. Y., and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Annals of statistics*, 44(2):515.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press.

Lian, H., Du, P., Li, Y., and Liang, H. (2014). Partially linear structure identification in generalized additive models with np-dimensionality. *Computational Statistics & Data Analysis*, 80:197–208.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Adjusted scores for inference in negative binomial regression

Funzioni di punteggio aggiustate per l'inferenza nella regressione binomiale negativa

Euloge C. Kenne Pagui, Alessandra Salvan, Nicola Sartori

Abstract Overdispersion is often encountered when fitting a Poisson model. An alternative model to handle the issue can be the negative binomial characterized by the mean with an additional dispersion parameter. With small sample sizes, the maximum likelihood estimator of the dispersion parameter may be subject to a significant bias. This paper studies negative binomial regression using the adjusted scores of Firth (1993) and Kenne Pagui et al. (2017) for mean and median bias reduction, respectively. A simulation study based on an application compares the proposed methods with maximum likelihood.

Abstract La sovradispersione si riscontra spesso quando si adatta un modello di Poisson. Un modello alternativo puo` essere la binomiale negativa carat-terizzata dalla media e con un parametro di dispersione aggiuntivo. Con campioni di piccole dimensioni, lo stimatore di massima verosimiglianza del parametro di dis-persione puo` essere soggetto ad una distorsione significativa. Questo articolo studia l'inferenza nel modello di regressione binomiale negativa usando le funzione di punteggio aggiustate di Firth (1993) e Kenne Pagui et al. (2017) per la riduzione della distorsione in media e in mediana, rispettivamente. Lo studio di simulazione basato su un'applicazione confronta i metodi proposti con quello della massima verosimiglianza.

Key words: Overdispersion, bias reduction, likelihood, count data.

Euloge C. Kenne Pagui
University of Padova, Department of Statistical Sciences, e-mail: kenne@stat.unipd.it

Alessandra Salvan
University of Padova, Department of Statistical Sciences, e-mail: salvan@stat.unipd.it,

Nicola Sartori
University of Padova, Department of Statistical Sciences, e-mail: sartori@stat.unipd.it

1 Introduction

The negative binomial model (e.g. Fisher, 1941; Engel, 1984; Lawless, 1987; Saha and Paul, 2005) is often used in the analysis of count data when the latter display overdispersion. With moderate sample sizes, the maximum likelihood estimator of the dispersion parameter may be subject to a substantial bias that can influenced the inferential conclusions.

Saha and Paul (2005) proposed a bias corrected maximum likelihood (ML) estimator for the mean and dispersion parameters of the negative binomial. Their paper does consider model with regression structure on the mean.

This paper extends the work of Saha and Paul (2005) by developing mean bias reduction (meanBR) for negative binomial regression following Firth (1993), whose adjusted score does not depends on the ML estimate. The estimator, solution of the bias reducing estimating equation has smaller mean bias with respect to the ML estimator. A second alternative adjusted score is proposed in Kenne Pagui et al. (2017), aiming at median bias reduction (medianBR). The resulting estimator is componentwise third-order median unbiased, and is equivariant under componentwise monotone reparameterization. We analyse the properties of the ML, meanBR and medianBR estimators in a simulation study based on real dataset.

2 Negative binomial regression model

Let y_i, $i = 1,\ldots,n$, be independent realizations of negative binomial random variables with mean $\mu_i = g^{-1}(\eta_i)$ and dispersion parameter κ, where $g^{-1}(\cdot)$ is the inverse of the link function, $\eta_i = x_i\beta$, $\beta = (\beta_1,\cdots,\beta_p)^\top$ and $x_i = (x_{i1},\ldots,x_{ip})$ is a row vector of covariates. The probability mass function of Y_i is

$$f_{Y_i}(y_i;\mu_i,\kappa) = \frac{\Gamma(y_i+\kappa^{-1})}{y_i!\Gamma(\kappa^{-1})}\left(\frac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\left(\frac{1}{1+\kappa\mu_i}\right)^{\kappa^{-1}}, \quad y_i = 0,1,\cdots, \qquad (1)$$

with $\kappa > 0$ and $\beta \in \mathbb{R}$. Noting that for any $\kappa > 0$, $\Gamma(y+\kappa^{-1})/\Gamma(\kappa^{-1}) = \kappa(\kappa+1)\cdots(\kappa+y-1)$, the log likelihood is

$$\ell(\beta,\kappa) = \sum_{i=1}^{n}\left\{\sum_{j=0}^{y_i^*}\log(1+\kappa j) + y_i\log\frac{\mu_i}{1+\kappa\mu_i} - \frac{1}{a}\log(1+\kappa\mu_i)\right\},$$

where $y_i^* = y_i - 1$ and $\sum_{j=0}^{y_i^*}$ is zero when $y_i^* < 0$. Let us denote by $\theta = (\beta^\top,\kappa)^\top$ and by $\hat{\theta}$ the maximizer of $\ell(\theta)$.

3 Bias reduction

For a general parametric model with p-dimensional parameter θ and log likelihood $\ell(\theta)$, based on a sample of size n, let $U_r = U_r(\theta) = \partial\ell(\theta)/\partial\theta_r$ be the r-th component of the score function $U(\theta)$, $r = 1,\dots,p$. Let $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^\top$ be the observed information and $i(\theta) = E_\theta\{j(\theta)\}$ the expected information.

Firth (1993) proposes a suitable adjusted score for meanBR, of the form

$$\tilde{U}(\theta) = U(\theta) + A^*(\theta),$$

where the vector $A^*(\theta)$ has components $A^*_r = \frac{1}{2}\mathrm{tr}\{i(\theta)^{-1}[P_r + Q_r]\}$, with $P_r = E_\theta\{U(\theta)U(\theta)^\top U_r\}$ and $Q_r = E_\theta\{-j(\theta)U_r\}$, $r = 1,\dots,p$. The resulting estimator, $\hat{\theta}^*$, has a mean bias of order $O(n^{-2})$, compared with $O(n^{-1})$ of the ML estimator.

The medianBR estimator, $\tilde{\theta}$, is obtained as solution of the estimating equation based on the adjusted score (Kenne Pagui et al., 2019)

$$\tilde{U}(\theta) = U(\theta) + \tilde{A}(\theta),$$

with $\tilde{A}(\theta) = A^*(\theta) - i(\theta)F(\theta)$. The vector $F(\theta)$ has components $F_r = [i(\theta)^{-1}]_r^\top \tilde{F}_r$, where $\tilde{F}_r$ has elements $\tilde{F}_{r,t} = \mathrm{tr}\{h_r[(1/3)P_t + (1/2)Q_t]\}$, $r,t = 1,\dots,p$, with the matrix h_r obtained as $h_r = \{[i(\theta)^{-1}]_r[i(\theta)^{-1}]_r^\top\}/i^{rr}(\theta)$, $r = 1,\dots,p$. Above, we denoted by $[i(\theta)^{-1}]_r$ the r-th column of $i(\theta)^{-1}$ and by $i^{rr}(\theta)$ the (r,r) element of $i(\theta)^{-1}$.

In the continuous case, each component of $\tilde{\theta}$, $\tilde{\theta}_r$, $r = 1,\dots,p$, is median unbiased with error of order $O(n^{-3/2})$, i.e. $\Pr_\theta(\tilde{\theta}_r \le \theta_r) = \frac{1}{2} + O(n^{-3/2})$, compared with $O(n^{-1/2})$ of ML estimator. Moreover, the asymptotic distribution of $\hat{\theta}^*$ and $\tilde{\theta}$ is the same as that of the ML estimator, that is $\hat{\theta} \dot{\sim} \mathcal{N}_p(\theta, i(\theta)^{-1})$.

4 Applications: ants and favourite sandwich

Consider the Ants dataset, obtained through an experiment conducted by students of a degree course in Applied Sciences of an Australian university (Mackisack, 2017). The goal of the experiment is to evaluate the preference of ants of the species Iridomyrmex purpureus, with respect to different types of sandwich. For each of 4 types of bread, Bread (1, rye; 2, wholemeal; 3, multigrain; 4, white), 3 types of filling, Filling (1, Vegemite; 2, peanut butter; 3 ham and pickles), and presence or not butter, Butter (1, present; -1, absent), two sandwich samples have been positioned, at random times near the entrance of an anthill. After 5 minutes, on each a glass was placed and the number of captured ants was counted, Ant_count. The interest is in evaluating how the number of ants depends on the characteristics of the sandwich. In this analyses, we consider only two covariates, Filling and Butter. We assume model (1), with

$$\log \mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i2} x_{i4} + \beta_6 x_{i3} x_{i4},$$

where x_{i2} is coded 1 if Filling is peanut butter o 0 otherwise, x_{i3} is 1 if Filling is ham and pickles o 0 otherwise while x_{i4} is 1 if Butter is present o 0 otherwise.

The parameter estimates of model (1) and their corresponding standard errors are displayed in Table 1. The regression coefficient estimates are very close for the three methods, ML, meanBR and medianBR. On the other hand there is difference in the dispersion parameter estimate for ML as compared with the meanBR and medianBR.

Table 1 Ants and favourite sandwich. Estimates of parameters of (1) using ML, meanBR and medianBR. The quantities in parentheses are the corresponding estimated standard errors.

	ML		meanBR		medianBR	
β_1	3.426	(0.121)	3.434	(0.130)	3.431	(0.131)
β_2	0.097	(0.170)	0.097	(0.183)	0.097	(0.185)
β_3	0.468	(0.166)	0.468	(0.180)	0.468	(0.181)
β_4	0.063	(0.170)	0.063	(0.183)	0.063	(0.185)
β_5	0.248	(0.237)	0.248	(0.256)	0.248	(0.259)
β_6	0.277	(0.233)	0.276	(0.252)	0.277	(0.255)
ϕ	0.084	(0.022)	0.103	(0.026)	0.105	(0.026)

To assess the properties of ML, meanBR and medianBR estimators, we performed a simulation study. We considered 10000 replications with parameter values equal to the ML fit of model (1) displayed in Table 1 and covariates held fixed at the observed values. Estimators performance are evaluated in terms of empirical relative bias (RB%), $100\times$ B/$|\beta|$; empirical relative increase in mean squared error from its absolute minimum due to bias (IBMSE%), $100\times$ B^2/SD^2; empirical percentage of underestimation (PU%) and coverage of nominally 95% Wald-type confidence intervals (WALD%). Here, B denotes the absolute bias estimation and SD, its standard deviation. The quantity PU is calculated as the proportion of times that the estimate is smaller than the corresponding true parameter value. Simulation results are shown in Table 2. Both meanBR and medianBR estimators achieve the desired goals and are preferable to ML estimator, especially in terms of coverage of confidence intervals.

References

1. Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statistica Neerlandica*, **38**, 159 – 167.
2. Fisher, R. A. (1941). The negative binomial distribution. *Annals of Eugenics*, **11**, 182–187.
3. Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27 – 38.
4. Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, **104**, 923 – 938.

Table 2 Ants and favorite sandwich. Simulation of size 10000 of the parameters of (1) under ML fit.

		β_1	β_2	β_3	β_4	β_5	β_6	ϕ
RB%	ML	-0.19	-1.36	0.45	-2.30	1.45	-0.24	-15.93
	meanBR	0.02	-1.56	0.29	-2.50	1.29	-0.33	0.14
	medianBR	-0.06	-1.45	0.38	-2.39	1.33	-0.31	2.17
PU%	ML	50.85	50.47	49.81	50.48	49.44	50.02	75.65
	meanBR	48.81	50.51	49.95	50.55	49.51	50.05	52.34
	medianBR	49.75	50.50	49.85	50.53	49.51	50.04	49.58
WALD%	ML	92.77	92.66	92.46	92.80	92.82	92.03	79.52
	meanBR	94.43	94.52	94.18	94.24	94.38	93.81	91.16
	medianBR	94.62	94.65	94.35	94.39	94.47	93.98	92.05
IBMSE%	ML	0.29	0.01	0.02	0.01	0.02	0.00	41.94
	meanBR	0.00	0.01	0.01	0.01	0.02	0.00	0.00
	medianBR	0.03	0.01	0.01	0.01	0.02	0.00	0.58

5. Mackisack, M. (2017). What is the use of experiments conducted by Statistics students? *Journal of Statistics Education*, **2**, 12 – 15.
6. Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, **15**, 209 – 225.
7. Saha, K., and Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, **61**, 179 – 185.

Estimation of the treatment effect variance in a difference-in-differences framework

Stima della varianza dell'effetto del trattamento nel metodo della differenza nelle differenze

Marco Doretti and Giorgio E. Montanari

Abstract We compare two regression-based estimators of treatment/policy effects in a standard difference-in-differences setting without additional covariates. These two estimators, deriving from what we term the *piled* and *unpiled* regression models, produce identical finite-sample estimates, thus sharing the same theoretical variance. However, they are not equivalent with regard to the estimation of such a variance. We show that under the assumed model the piled approach should be preferred since it results in a more efficient variance estimator, thereby returning more reliable standard errors for the treatment effect. We further show that this result holds even when the normality assumption for the regression error terms is violated.

Abstract *Nel metodo della differenza nelle differenze, l'effetto del trattamento può essere stimato secondo due diversi modelli di regressione lineare, detti* impilato *e* non impilato*. I due approcci producono le medesime stime nel campione finito, condividendo di fatto la stessa varianza teorica. Tuttavia, essi non sono equivalenti per quanto concerne la stima di questa varianza. In questo lavoro si mostra come, sotto le assunzioni fatte, l'approccio impilato vada preferito, in quanto caratterizzato da uno stimatore della varianza più efficiente. Questo risultato vale anche in caso di violazione dell'ipotesi di normalità degli errori nella regressione.*

Key words: difference-in-differences, kurtosis, piled model, policy evaluation, unpiled model, variance estimation

Marco Doretti
University of Perugia, Department of Political Science, via A. Pascoli 20, 06123 Perugia (Italy), e-mail: marco.doretti@unipg.it

Giorgio E. Montanari
University of Perugia, Department of Political Science, via A. Pascoli 20, 06123 Perugia (Italy), e-mail: giorgio.montanari@unipg.it

1 Introduction

Difference-in-differences (DiD) is a well-known technique to estimate a treatment/policy effect in the presence of repeatedly measured data [4]. In the most basic setting, an outcome of interest is measured on two groups of units over two temporal occasions, in between of which one group receives the treatment while the other - also called control group - does not. The name of the method stems from the fact that the target effect is identified by subtracting the difference between the post- and pre-treatment outcome expectations in the control group from the same difference in the treated group. In a causal inference perspective, the DiD method identifies an average treatment effect on the treated [1].

Identification of DiD effects relies on the so called *parallel trends* assumption, stating that the expected difference in the outcome between the two time points in the treated group, had its units not received the treatment, would have been equal to that of the control group. In other words, the spontaneous (*i.e.*, not due to the treatment) dynamic governing the temporal evolution of the outcome, possibly driven by unobserved factors, has to be the same for the two groups. When some factors make the spontaneous dynamic differ between the two groups - and data for them are available - a covariate adjustment is in principle possible, though with some provisos concerning endogeneity.

In the basic setting without additional covariates, which is the one considered here, there are many equivalent techniques to estimate the target effect. The most intuitive one is based on a trivial replacement of expectations with sample averages in the DiD identification formula. However, several alternatives including linear regression exist; see for example [2]. All the resulting estimators are known to be algebraically equivalent, yielding exactly the same results in finite samples. As a consequence, they all share the same theoretical variance. However, the way this variance is estimated changes with the undertaken approach. In detail, the considered techniques give rise to two regression-based variance estimators with different performances, thus justifying a choice between them. In this paper, we address this comparison, which - to the best of our knowledge - has not been discussed in the literature yet.

The paper is arranged as follows. In Section 2, we formalize the DiD framework introducing the target effect as well as its equivalent estimators together with their common theoretical variance. The focus on variance estimation is contained in Section 3, where we describe the two competing regression-based variance estimators and state which estimation method should be favoured.

2 Difference-in-differences effect estimators

As typical in a standard DiD framework, we use the index $t = \{0,1\}$ to label the pre- ($t = 0$) and post-treatment ($t = 1$) temporal occasions, while $g = \{0,1\}$ indicates the control ($g = 0$) and treatment ($g = 1$) unit groups. In this setting, y_{gti} denotes

the outcome at time t for the i-th unit of the g-th group, with $i = 1,\ldots,n_0$ for the control group and $i = 1,\ldots,n_1$ for the treatment group. The data generating process is assumed to follow the linear model

$$y_{gti} = \mu + \alpha t + \beta g + \delta gt + e_{gti}, \tag{1}$$

where e_{gti} are independent and identically distributed error terms with $E(e_{gti}) = 0$ and $V(e_{gti}) = \sigma_e^2$, β captures the pre-treatment average difference between the two groups and α represents the trend effect, that is, the aforementioned spontaneous dynamic of the outcome. As a consequence, the interaction parameter

$$\delta = \big\{E(y_{11i}) - E(y_{10i})\big\} - \big\{E(y_{01i}) - E(y_{00i})\big\}$$

isolates the additional effect due to the treatment/policy only, identifying the DiD effect under investigation. Such a parameter can estimated by

$$\hat{\delta} = (\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00}),$$

where $\bar{y}_{gt}$ is the sample average of the outcome for units of group g at time t ($g,t = \{0,1\}$). Due to the independence of the error terms, the variance of the $\hat{\delta}$ estimator is given by

$$V(\hat{\delta}) = \sum_{g,t=0}^{1} V\left(\frac{1}{n_g}\sum_{i=1}^{n_g} y_{gti}\right) = 2\sigma_e^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right) = 2\sigma_e^2\frac{n}{n_0 n_1}, \tag{2}$$

where $n = n_0 + n_1$ is the total sample size. Another formulation originates from differentiating Equation (1) with respect to time. Specifically, we have

$$d_{gi} = \alpha + \delta g + u_{gi}, \tag{3}$$

where $d_{gi} = y_{g1i} - y_{g0i}$ and $u_{gi} = e_{g1i} - e_{g0i}$ is a compound error term with $V(u_{gi}) = \sigma_u^2 = 2\sigma_e^2$. The $\hat{\delta}$ estimator can be analogously obtained as $\hat{\delta} = \bar{d}_1 - \bar{d}_0$, with $\bar{d}_g$ representing the sample average of the post- and pre-treatment differences for units in group $g = \{0,1\}$. Clearly, the estimator variance remains as in (2), a fact that can be straightforwardly verified considering that $V(\bar{d}_g) = \sigma_u^2/n_g$.

As mentioned in the Introduction, a regression approach can also be adopted. In detail, two regression models based on Equations (1) and (3) can be fitted to obtain exactly the same effect estimate as above, that is, $\hat{\delta}$. In the first one, the $2n$ observations resulting by piling the pre- and post-treatment outcome values are regressed against an intercept, the two binary indicators of time and group and their interaction. In the second one, the n temporal differences in the outcome d_{gi} are first computed and then regressed against an intercept term and the group indicator only. Henceforth, we will refer to these models as to the *piled* and the *unpiled* model, respectively. In this regression-based perspective, it is easy to analytically show that the variance of the estimated effect $\hat{\delta}$ remains as in (2) in both regression models.

3 Variance estimation

In a classical linear regression framework with homoskedastic, independent and normally distributed error terms, the residual sum of squares divided by the number of residual degrees of freedom of the model is the usual estimator of the error variance. In our simplified context without additional covariates, the regression fitted values reduce to group sample averages. Therefore,

$$S_e^2 = \frac{1}{2n-4} \sum_{g,t=0}^{1} \sum_{i=1}^{n_g} (y_{gti} - \bar{y}_{gt})^2 \tag{4}$$

turns out to be the residual-based estimator of σ_e^2 in the piled regression (1) while

$$S_u^2 = \frac{1}{n-2} \sum_{g=0}^{1} \sum_{i=1}^{n_g} (d_{gi} - \bar{d}_g)^2 \tag{5}$$

is the analogous estimator of $\sigma_u^2 = 2\sigma_e^2$ in the unpiled model (3). It is important to note that the mathematical relationship between the two theoretical variances is not preserved when estimation comes about, meaning that in finite samples we generally have $S_u^2 \neq 2S_e^2$. In other words, the two regression models originate two distinct estimators - namely, S_e^2 and $S_u^2/2$ - of the same variance parameter σ_e^2. As already mentioned, a comparison of these two alternative estimators is therefore sensible. Since both estimators are known to be unbiased, the whole problem reduces to the evaluation of their variances.

In the classical setting outlined above, standard results about the variance of the residual-based estimator of the error variance apply. Indeed, denoting by σ^2 the true error variance and by $\hat{\sigma}^2$ its estimator, it is known that

$$\frac{(N-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-p},$$

with $N-p$ being the number of residual degrees of freedom of the model. Consequently, $V(\hat{\sigma}^2) = 2\sigma^4/(N-p)$. An application to our two variance estimators gives

$$V(S_e^2) = \frac{2\sigma_e^4}{2n-4} = \frac{\sigma_e^4}{n-2} \tag{6}$$

and

$$V(S_u^2/2) = \frac{2\sigma_u^4/(n-2)}{4} = \frac{2\sigma_e^4}{n-2}, \tag{7}$$

showing that the ratio $V(S_u^2/2)/V(S_e^2)$ is equal to 2. Therefore, the variance estimator resulting from the piled model is twice as efficient as the unpiled variance estimator, in line with the intuition suggesting to use the former because it is based on a double amount of information. Thus, the piled approach should be preferred

since it provides a less variable estimate of σ_e^2 and, consequently, a more reliable standard error for the DiD effect $\hat{\delta}$.

The standard theory leading to (6) and (7) provides exact results if the error terms e_{gti} follow a normal distribution, which implies also the u_{gi} terms do. Alternatively, a large sample size is required to rely on central-limit-theorem approximations. If the sample size is small and deviations from the normality assumption occur, $V(\hat{\delta})$ is still given by (2) and S_e^2 and $S_u^2/2$ remain unbiased estimators of σ_e^2. Nevertheless, the variance expressions in (6) and (7) are no longer appropriate. Distribution-free formulas for them can still be obtained under the assumption that the error terms e_{gti} are independent and identically distributed.

To derive the requested generalizations of (6) and (7), it is worth to bear in mind two further probability results. The first one states that, given N independent and identically distributed observations $x_1,\ldots,x_N$ from a (not necessarily normal) population with variance σ^2 and kurtosis K (that is, the standardized fourth moment), the variance of the sample variance estimator $s^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$, with $\bar{x}$ denoting the sample mean, is

$$V(s^2) = \frac{\sigma^4}{N}\left(K-1+\frac{2}{N-1}\right); \tag{8}$$

see [3]. The second result states that the kurtosis of the difference of two independent random variables having a common variance and a common kurtosis K is equal to $(K+3)/2$; see Appendix 5 for a proof. On the one hand, (8) is of use since the numerators of both (4) and (5) can be thought of as linear combinations of independent sample variances within the respective unit groups; notice that this is not the case with more complex models involving additional covariates. On the other hand, the second result allows to express the kurtosis of the u_{gi} error terms as a simple function of the kurtosis of the error terms e_{gti}, say K_e.

Building on the two probability results above, some algebraic derivations lead to update the variance formulas (6) and (7) when normality of the errors is not invoked. Specifically, we have

$$V(S_e^2) = \frac{\sigma_e^4}{2(n-2)^2}\left\{(K_e-1)\left(n-4+\frac{n}{n_0n_1}\right)+2\left(2-\frac{n}{n_0n_1}\right)\right\} \tag{9}$$

and

$$V(S_u^2/2) = \frac{\sigma_e^4}{(n-2)^2}\left\{\frac{K_e+1}{2}\left(n-4+\frac{n}{n_0n_1}\right)+2\left(2-\frac{n}{n_0n_1}\right)\right\}. \tag{10}$$

In the normal case, it is immediate to verify that (9) and (10) reduce to (6) and (7) respectively because $K_e = 3$. In general, it is easy to show that the ratio between (10) and (9) is always greater than 1, thereby confirming that the piled regression approach should be preferred also when non-normality in the error terms is spotted. Since such a variance ratio is essentially driven by $(K_e+1)/(K_e-1)$, the efficiency gain is higher for smaller values of K_e, and tends to vanish as K_e grows.

4 Conclusions

This paper suggests the use of the piled rather than the unpiled approach for the estimation of treatment/policy effects in a standard DiD framework, even when normality of the regression errors is questionable. However, all the derivations rely on the assumptions of the basic DiD setting, including the absence of additional covariates as well as of a temporal correlation between the error terms e_{g0i} and e_{g1i}. Removing these assumptions might lead to novel variance estimators or, more importantly, to alternative effect estimators with different variances. These topics are worth investigating in future lines of work.

5 Appendix

Let X and Y be two independent random variables with a common variance σ^2 and a common kurtosis K. For simplicity we assume $E(X) = 0 = E(Y)$ so that $K = E(X^4)/\sigma^4 = E(Y^4)/\sigma^4$, though generality is by no means lost. The kurtosis of the difference random variable $Z = X - Y$ is

$$K(Z) = E\left\{\left(\frac{X-Y}{\sqrt{2}\sigma}\right)^4\right\} = E\left(\frac{X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4}{4\sigma^4}\right)$$
$$= \frac{1}{4}\left\{\frac{E(X^4)}{\sigma^4} + 6\frac{E(X^2)E(Y^2)}{\sigma^4} + \frac{E(Y^4)}{\sigma^4}\right\} = \frac{2K+6}{4} = \frac{K+3}{2},$$

which hols true since all the odd-power terms in the right-hand side of the second equality have a null expectation because of the independence assumption.

Acknowledgements We thank Fondazione Cassa di Risparmio di Perugia for financial support.

References

[1] A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 01 2005.

[2] A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.

[3] E. Cho and M. J. Cho. Variance of sample variance with replacement. *International Journal of Pure and Applied Mathematics*, 52(1):43–47, 2009.

[4] Myoung-Jae Lee. *Matching, regression discontinuity, difference in differences and beyond*. Oxford University Press, 2016.

Exploring multicollinearity in quantile regression

Analisi della multicollinearità nella regressione quantile

Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco

Abstract The aim of the paper is to propose a simulation study to explore the multicollinearity problem in quantile regression, as compared to the classical linear regression. The simulation exploits the concept of a relevant subspace and relevant predictors, considering different degrees of collinearity among the involved predictors. The approach is based on principal components and consists in evaluating the degree of dependence between the predictors on the basis of the eigenvalue structure of their covariance matrix. It is well known that in case of highly intercorrelated predictors the least squares coefficients, although determinate, posses large standard errors, causing precision problems for their estimation. For this reason, results of the simulation study focus on standard errors estimated according to different collinearity levels. A possible solution based on regression on principal components is briefly presented.

Abstract *L'articolo presenta uno studio di simulazione che mira ad esplorare il problema della multicollinearità nella regressione quantile, offrendo al contempo un parallelo con la regressione lineare classica. Lo studio sfrutta il concetto di sottospazio rilevante e predittori rilevanti, prendendo in considerazione diversi gradi di collinearità tra i predittori. L'approccio, basato sulle componenti principali, consiste nel valutare il grado di dipendenza tra i predittori rispetto alla struttura degli autovalori della loro matrice di covarianza. In caso di predittori altamente intercorrelati, è infatti noto che i coefficienti dei minimi quadrati, sebbene determinati, presentano errori standard elevati, causando problemi di precisione delle stime. I*

Cristina Davino
University of Naples Federico II, Italy, e-mail: cristina.davino@unina.it

Tormod Næs
NOFIMA AS, Norway, e-mail: tormod.naes@nofima.no

Rosaria Romano
University of Naples Federico II, Italy, e-mail: rosaroma@unina.it

Domenico Vistocco
University of Naples Federico II, Italy, e-mail: domenico.vistocco@unina.it

risultati presentati si concentrano pertanto sugli errori standard dei coefficienti, stimati in base a diversi livelli di collinearità. Una possibile soluzione al problema è mostrata sfruttando la regressione sulle componenti principali, in luogo dei regressori originari, al fine di eliminare il problema di collinearità.

Key words: multicollinearity, least squares regression, quantile regression, principal component regression

1 Introduction

Regression is widely used in the analysis of socio-economic phenomena to study the dependence of a response variable on a set of predictors. One typical problem in such fields concerns the structure of relations among the predictors, engendering the well-known problem of collinearity [**?**]. Multicollinearity has been extensively explored for least square regression (LS). This is not the case for quantile regression (QR), a well established statistical method aimed to explore the whole conditional distribution of a response variable without posing any parametric assumption for the error (and hence response) distribution ([**?**]; [**?**]). QR estimates separate models for different quantiles $\tau \in [0,1]$, namely QR provides a regression model for each conditional quantile of interest [**?**]. Even if an infinite number of conditional quantiles can be estimated (the quantile process [**?**]), in practice the researcher defines few quantiles of interest, in most cases, the three quartiles, $\tau = [0.25, 0.50, 0.75]$, along with two extreme quantiles, typically $\tau = [0.10, 0.90]$. For each considered quantile, a regression model is estimated, providing a set of coefficients and a fitted response vector. This paper presents a simulation study to explore the collinearity in quantile regression, as compared to the classical linear regression model. Results of a possible solution based on principal component regression are presented. The simulation scheme considers classical normal i.i.d. errors. Further developments will include normal heteroscedastic errors, costant skewness in the response (error) and increasing skewness in the response (error).

2 Simulation study

2.1 Experimental design

The concept of a relevant subspace has been exploited to simulate different degrees of correlation among predictors. It essentially consists of the subspace of the predictor space that is relevant for the variation in the response variable. Principal components analysis allows to take into account of different degrees of correlation. We carried out the analysis using the software R [**?**] and the *simrel* package [**?**]

for linear model data simulations. In particular we set the following values for the simulations:

- number of observations: 100
- number of predictors: 3
- number of relevant principal components: 1
- theoretical R^2 for generating data: 0.7
- coefficient controlling the degree of collinearity: γ
- number of iterations: 1000

According to the concept of relevant subspace, and given the small number of selected predictors, we assume that only one component is relevant for prediction. With respect to the γ coefficient, it regulates the speed of decline in eigenvalues (variances) of the principal components. In particular, the eigenvalues are assumed to decline according to an exponential model and the first eigenvalue is set equal to 1. We considered a grid of values for γ ranging from 0 to 5, using an increment of 0.5. In case of low values of γ we expect no or very low collinearity among predictors, while high collinearity should be present by incrementing γ. As an example, Table **??** reports the percentage of cumulated explained variance for the three components ($Comp_1$, $Comp_2$ and $Comp_3$ on the columns) using one random sample for each level of γ (rows).

Table 1 Percentage of cumulated explained variance on the three principal components (columns) for the considered scenarios (different values of γ on the rows)).

	$Comp_1$	$Comp_2$	$Comp_3$
$\gamma = 0.0$	36.33	70.20	100.00
$\gamma = 0.5$	45.44	79.60	100.00
$\gamma = 1.0$	55.19	91.56	100.00
$\gamma = 1.5$	65.17	95.88	100.00
$\gamma = 2.0$	66.17	98.10	100.00
$\gamma = 2.5$	72.16	99.19	100.00
$\gamma = 3.0$	85.03	99.77	100.00
$\gamma = 3.5$	91.79	99.91	100.00
$\gamma = 4.0$	95.43	99.97	100.00
$\gamma = 4.5$	97.47	99.99	100.00
$\gamma = 5.0$	97.87	100.00	100.00

For each value of the γ grid, the standard errors of the classical linear regression and QR models were computed. We decided to compute LS standard errors using the bootstrap procedure in order to have a fair comparison with QR, where bootstrap is typically used to this end.

2.2 Main results

Simulation results for LS and the three quartiles of QR are shown in Figure **??**. In particular, the different models (LS bootstrap, QR bootstrap for $\tau \in (0.25, 0.5, 0.75)$, where τ denotes the conditional quantile) are depicted on the columns, while rows refer to the regression coefficient ($X1$, $X2$, $X3$). The different boxplot in each panel represent the distribution of the standard error (vertical axis) for the different values of the γ coefficient (horizontal axis). includes several panels. Each panel corresponds to one of the estimated parameters by row ($X1$, $X2$, $X3$).

Results highlight how increasing the degree of collinearity, the distribution of the standard errors increases both in size and variability for each predictor. This trend can be found both in the LS and in the QR. In case of QR the inaccuracy of the estimates is even larger and no particular pattern stands out with respect to the considered quantile. This effect is higher for extreme quantiles (results not shown).

Principal component regression (PC–R) exploits principal component analysis to overcome multicollinearity [**?**]. It exploits the principal components of regressors as explanatory variables. Typically, only a subset of the principal components is used, and in particular the ones with higher variances. Therefore PC–R is considered also a shrinkage method [**?**]. Since in this first simulation study only three regressors have been considered, we present PC–R results on all the three principal components. Figure **??** presents the same structure of the previous Figure **??**, the two columns representing results for the two extreme conditional quantiles ($\tau = 0.1$ and $\tau = 0.9$) being the only exception. It is safe to say that rows in Figure **??** refers to the principal components (and not to the origiginal regressors). It is evident how PC–R represents a possible solution also for QR in case of multicollinearity: the first two regression coefficients become really stable, the only variability remains on the third coefficients.

2.3 Concluding remarks

The simple simulation study shows that QR coefficients suffer the same problem of LS coefficients in case of strong collinearity among regressors. In fact, they become unstable, and this in particular for quantiles far from the median. PC–R is a possible solution to deal with such problem, providing more stable solutions.

This early study will be expanded considering different error (response) distribution, that is normal heteroscedastic response, constant skewness in the response and increasing skewness in the response. Moreover, the influence of prediction performances will be studied in the various scenarios, both for prediction inside the sample, i.e. inside the typical range of the regressor(s), and for prediction outside the sample. We expect that prediction performances will not be necessarily influenced from collinearity when the input is inside the range of the data used for model fitting. As soon as one moves outside, collinearity can have a huge effect on the performance. The small eigenvalue directions are the most susceptible to this because

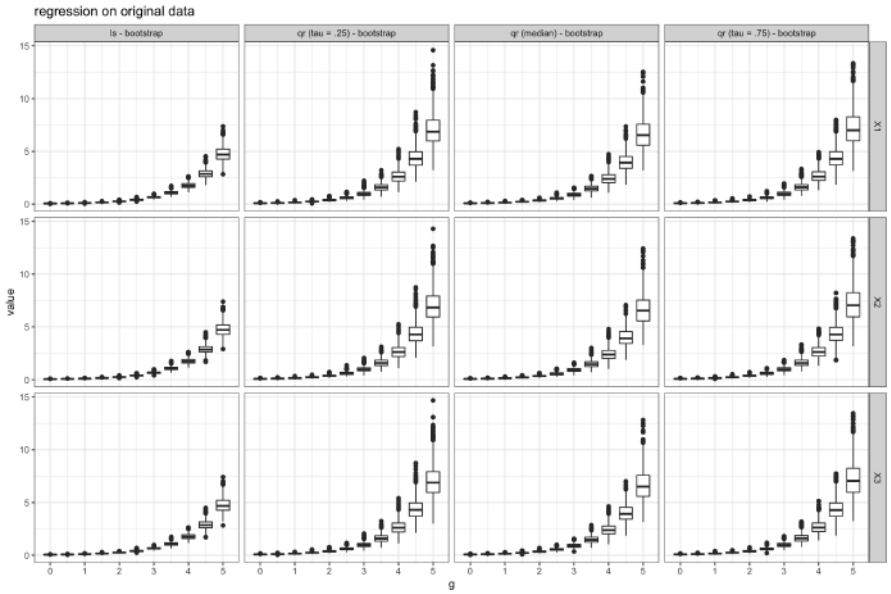


Fig. 1 Results from the simulation study: standard errors of the involved coefficients (different boxplots) for LS and QR models (columns) for three selected values of the γ coefficient (rows), regulating the degree of collinearity.

they are more unstable. Also in this case, we expect that PC–R may help to overcome the problem.

References

1. Davino, C., Furno, M., Vistocco, D.: *Quantile Regression. Theory and applications*. Wiley Series in Probability and Statistics, John Wiley & Sons (2013)
2. Furno, M., Vistocco, D.: *Quantile Regression. Estimation and Simulation*. Wiley Series in Probability and Statistics, John Wiley & Sons (2018)
3. Hastie, T., Tibshirani, R., Friedman, J: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd edition. Springer Series in Statistics, Springer (2009)
4. Koenker, R., Bassett, G.: Regression quantiles. *Econometrica. Journal of the Econometric Society*. 33–50 (1978)
5. Koenker, R.: *Quantile Regression* (Econometric Society monographs; no. 38). Cambridge university press (2005)
6. Næs, T., Isaksson, T.: Data Compression by PLS/PCR. *NIR News*. **3**(1), 10–11 (1992)
7. Næs, T., Mevik, B.H.: Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*. **15**(4), 413–426 (2001)
8. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
9. Sæbø, S., Almøy, T., Helland, I.S.: simrel—A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. Chemometrics and Intelligent Laboratory Systems. **146**, 128–135 (2015)

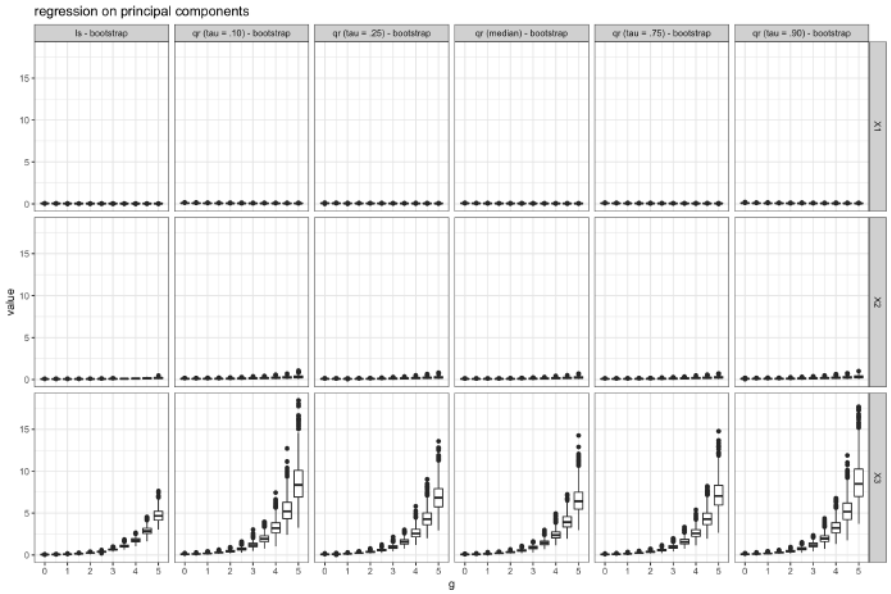


Fig. 2 Results from the simulation study: standard errors of the involved coefficients (different boxplots) for LS and QR models (columns) for three selected values of the γ coefficient (rows), regulating the degree of collinearity, when regressions are carried out on the principal components of the regressors.

Generalized M-quantile random effects model

Modello di regressione M-quantile generalizzato ad effetti misti

Francesco Schirripa Spagnolo and Vincenzo Mauro

Abstract In this paper we extend the linear M-quantile random intercept model (MQRE) to discrete data and use the proposed model to evaluate the effect of selected covariates on two count responses: the number of generic medical examinations and the number of specialized examinations for Health Districts in three regions of central Italy. The proposed approach represents an outlier-robust alternative to the Poisson generalized linear mixed model with Gaussian random effects and allows estimating the effect of the covariates at various quantiles of the conditional distribution of the target variable.

Abstract *L'obiettivo del lavoro è quello di estendere il modello di regressione M-quantile ad effetti misti al fine di modellare dati discreti. Il modello proposto è stato applicato per studiare l'effetto di alcune covariate sul numero di visite generiche e sul numero di visite specialistiche nei distretti sanitari di tre regioni del Centro Italia. Tale metodo rappresenta un'alternativa robusta ai modelli lineari generalizzati ad effetti misti e permette di stimare l'effetto delle covariate a diversi quantili della distrubuzione condizionata della variabile risposta.*

Key words: Robust methods; Poisson distribution; Health Conditions and Appeal to Medicare Survey (HCAMS)

Francesco Schirripa Spagnolo
Department of Economics and Management University of Pisa, Via Cosimo Ridolfi, 10, Pisa, Italy, e-mail: francesco.schirripa@ec.unipi.it

Vincenzo Mauro
Department of Economics and Management University of Pisa, Via Cosimo Ridolfi, 10, Pisa, Italy, e-mail: vincenzo.mauro@unipi.it

1 Introduction

Generalized linear mixed models (GLMM) represent the most common method to analyse discrete, binary, or count outcomes in presence of clustered data including random effects to account for between-group variations. Standard GLMM approaches have two main drawbacks. Firstly, the estimation of the parameters is implemented by means of parametric assumptions like the normal distribution of random effects. Secondly, they target the expected value of the conditional distribution of the outcome given a set of covariates. However, when the entire conditional distribution is concerned, it is not recommended to analyse only the central tendency (i.e. the conditional mean) and it is better to use regression models that allow to summarise the behaviour at different percentage points (quantiles) of the target variable. Moreover, when the distribution of the outcome is asymmetric and data contain outliers, the assumptions of the standard GLMM approaches do not hold and classical estimators present very poor performance, producing misleading results. In this work we propose an extension of the linear M-quantile random intercept model (MQRE) to discrete data. Compared to quantile regression, M-quantile (MQ) approach represents a more flexible alternative in modelling different parts of the conditional distribution of the outcome variable.

2 Model

M-quantile regression is usually defined as a 'quantile-like' generalization of regression based on influence functions (M-regression) able to estimate the differential effect of a set of covariates $\mathbf{X}$ at different levels of the conditional distribution of the response variable. The M-quantile of order q, $q \in (0,1)$, for the conditional distribution of y given the set of covariates $\mathbf{X}$, $f(y|\mathbf{X})$, is defined as the solution $MQ_y(q|\mathbf{X},\psi)$ of the estimating equation $\int \psi_q[y - MQ_y(q|\mathbf{X};\psi)]f(y|\mathbf{X})dy = 0$, where $\psi_q(u) = 2\psi(u)\{qI(u>0) + (1-q)I(u\leq 0)\}$ and ψ is a influence function.

Given $\mathbf{X}$, the linear M-quantile regression model for the qth conditional M-quantile of y is defined by

$$MQ_y(q|\mathbf{X};\psi) = \mathbf{X}\beta_{\psi q}, \tag{1}$$

where $\beta_{\psi q}$ represents the vector of the regression parameters.

To model clustered data, [7] extend M-quantile regression in order to include random effects to take into account a two-level hierarchical structure of the data. The M-quantile random intercept model (MQRE) is defined as follows:

$$MQ_y(q|\mathbf{X},\mathbf{u}_q;\psi) = \mathbf{X}\beta_{\psi q} + \mathbf{Z}\mathbf{u}_q, \tag{2}$$

where $\mathbf{u}_q$ is a vector of group random effects at the qth M-quantile and $\mathbf{Z}$ represents and incidence matrix to identify the groups.

In order to model discrete outcomes an MQ regression for discrete data has been proposed in [2, 6]:

$$MQ_y(q|\mathbf{X};\psi) = \eta_{1q} = \mathbf{X}\beta_{\psi q}. \tag{3}$$

Following the idea of [7] we can include a group-specific random intercept in the Eq. (3) in order to define a Generalized MQRE (G-MQRE) suitable to model a discrete outcome in the presence of clustered data:

$$MQ_y(q|\mathbf{X};\mathbf{u}_q;\psi) = \eta_q + \mathbf{e}_q = \mathbf{X}\beta_{\psi q} + \mathbf{Z}\mathbf{u}_q + \mathbf{e}_q, \tag{4}$$

where $\mathbf{u}_q$ is a vector of group random effects at the qth M-quantile (as in Eq. (2)) and $\mathbf{e}_q$ is a vector of level 1 (unit) residuals at the qth M-quantile with no distributional assumption imposed.

Using the same idea of [4], we define a new variable $\mathbf{y}_q^\star = \mathbf{X}\beta_{\psi q} + \mathbf{Z}\mathbf{u}_q + \mathbf{e}_q^\star$ and then it is possible to obtain the estimating equation of the model in Eq. (4) where $\mathbf{e}_q^\star = \mathbf{e}_q\left(\partial\eta_q/\partial\mu_q\right)$ is an adjusted error term.

In particular, following [7] and [5] it is possible to define the following modified estimating equations for estimating the regression coefficients and the variance parameters in the G-MQRE:

$$\mathbf{X}^T\mathbf{V}_q^{\star-1}\mathbf{U}_q^{\star 1/2}\psi_q\left\{\mathbf{U}_q^{\star-1/2}(\mathbf{r}_q^\star)\right\} = \mathbf{0}, \tag{5}$$

$$\psi_q\left\{\mathbf{U}_q^{\star-1/2}(\mathbf{r}_q^\star)\right\}\mathbf{U}_q^{\star 1/2}\mathbf{V}_q^{\star-1}(\partial_{\theta_k}\mathbf{V}_q^\star)\mathbf{V}_q^{\star-1}\mathbf{U}_q^{\star 1/2}\psi_q\left\{\mathbf{U}_q^{\star-1/2}(\mathbf{r}_q^\star)\right\}^T - \mathrm{tr}(K_{2q}(\partial_{\theta_k}\mathbf{V}^\star)\mathbf{V}_q^{\star-1}) = 0, \tag{6}$$

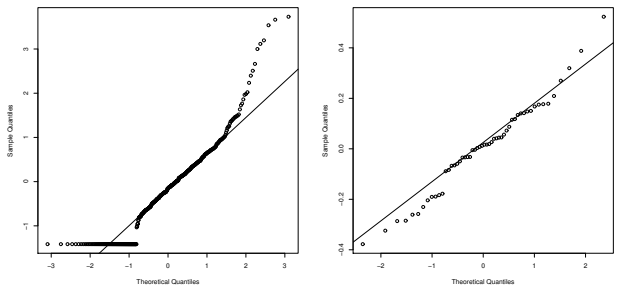
where $\mathbf{V}_q^\star = \delta\Sigma_{e_q}^\star + \mathbf{Z}\Sigma_{u_q}^\star\mathbf{Z}^T$; $\mathbf{r}_q^\star = \mathbf{y}^\star - \mathbf{X}\beta_{\psi q}$; $\mathbf{U}_q^\star$ is a diagonal matrix with diagonal elements equal to the diagonal elements of the variance-covariance matrix $\mathbf{V}_q^\star$.

3 Modelling the number of visits to physicians for Health Districts in Liguria, Toscana and Umbria (Italy).

The proposed methodology is applied to model the number of generic medical examinations and the number of specialized examinations for Health Districts (HD) in three administrative regions of Italy, namely Liguria, Toscana, and Umbria. Data are obtained using the 1999/2000 wave of Health Conditions and Appeal to Medicare Survey (HCAMS), a national, multistage sample survey periodically conducted by the National Institute of Statistics in Italy. The two target variables are the generic and specialised visits that took place in the last four weeks for the elderly (aged 65 and above) in the 60 HDs of Toscana, Liguria and Umbria. The total sample size for the three regions is $n = 4021$.

3.1 Preliminary analysis

We here fit a Poisson GLMM and present some diagnostics in order to motivate the use of the alternative approach proposed in this paper. The covariates included in the model are: five-year age groups (65-69, baseline; 70-74; 75-79; 80-84; 85 and above), gender (baseline: female) and region (baseline: Liguria). Figures 1 and 2 present normal probability plots of level 1 and level 2 Pearson residuals derived by fitting GLMM on Generic visits and GLMM on Specialized visits, respectively. These plots suggest severe departures from the Gaussian assumptions of the level 1 residuals for GLMM for both outcomes. Conversely, the distribution of the random effects can be considered approximately normal as confirmed by the Shapiro-Wilk test ($W = 0.978$ with $p - value = 0.430$ for random effects obtain fitted a GLMM on generic visits and $W = 0.968$ with $p - value = 0.161$ for GLMM on specialized visits). The skewed distribution of level 1 residuals and the presence of large values is confirmed by the plots in Figure 3, plotting the distribution of level 1 residuals by health districts. For both outcomes, some health districts contain some positive outliers, i.e. residuals exceeding 2.



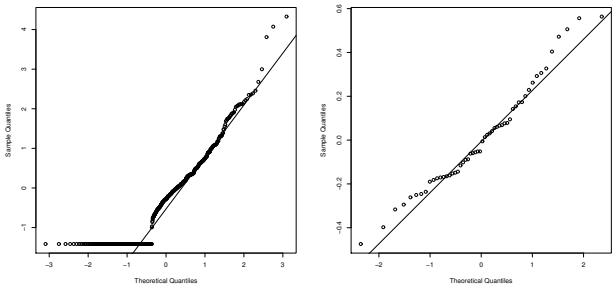
(a) (b)

Fig. 1 Normal probability plots of level 1 (a) and level 2 (b) Pearson residuals derived by fitting the Poisson GLMM on Generic visits

As the diagnostics reported in this subsection provide some evidence towards the adoption of an approach alternative to the standard Poisson GLMM, we decided to fit a G-MQRE model, whose results are presented in the next section.

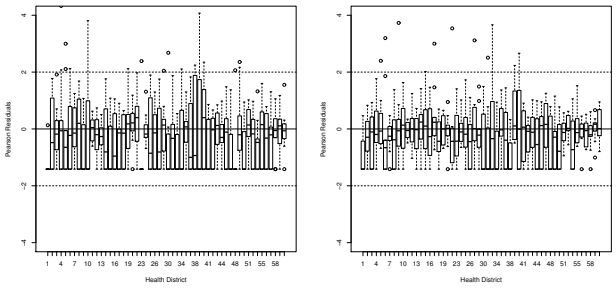
3.2 G-MQRE on HCAMS data

The influence function ψ in the G-MQRE is set to be the Huber Proposal 2 [3]. Given the presence of level 1 outliers, we set the tuning constant on fixed effects $c = 1.6$ [1]. On the other hand, given the approximately normal distribution of the



(a) (b)

Fig. 2 Normal probability plots of level 1 (a) and level 2 (b) Pearson residuals derived by fitting the Poisson GLMM on Specialized visits



(a) (b)

Fig. 3 Box-plots of level 1 Pearson residuals by health districts for the Poisson GLMM on Generic visits (a) and on Specialized visits (b)

level 2 residuals obtained after fitting GLMMs, we set a large c for the estimation of random effects.

For both outcomes, the effects measured by the regression coefficients tend to be different at different quantiles. The number of generic visits increases as people grow older, even if at a slower pace as the quantiles increase. After controlling for the effects of variables "age" and "region", women seems to report more generic visits than men, but also for this variable the effect seems to be less sharp at higher quantiles. People tend to visit generic physicians more often in Tuscany and Umbria than in Liguria, but as quantiles increase, the effect is stable for Tuscany while it decreases for Umbria. The model for specialized visits reports slightly different results. The number of specialized visits seems to be lower for elder people, as shown by the negative coefficient for people over 85 for all the quantiles analysed. The hypothesis of an overall decreasing trend in the number of specialised visits as people get older is consistent with an observed decrease of the number of visits for the 80-84 range (it must be noted that the lack of significance of the latter effect

is only due to the choice of the baseline, so that the difference between the 80-84 age class and the previous class is highly significant). As observed in the model for generic visits, women seem to schedule more specialized than men, but the effect tends to be sharper at $q = 0.90$. The differences for the variable "Region" does not appear significant.

Table 1 Results of G-MQRE with tuning constants $c = 1.6$ on fixed effects and $c = 100$ on random effects†

	Number of Generic Visits			**Number of Specialized Visits**		
Variables	$\beta_{q=0.50}$	$\beta_{q=0.75}$	$\beta_{q=0.90}$	$\beta_{q=0.50}$	$\beta_{q=0.75}$	$\beta_{q=0.90}$
Intercept	-1.507^{***}	-0.427^{***}	0.036	-2.327^{***}	-0.768^{***}	0.111
Age 70-74	0.025	0.062	0.093	0.065^{***}	0.073	-0.014
Age 75-79	0.482^{***}	0.299^{***}	0.299^{***}	0.246^{***}	0.183^{***}	0.023
Age 80-84	0.490^{***}	0.286^{***}	0.313^{***}	0.002	-0.038	-0.018
Age >84	0.670^{***}	0.475^{***}	0.445^{***}	-0.236^{***}	-0.168^{**}	-0.227^{**}
Gender	-0.204^{***}	-0.105^{***}	-0.082^{*}	-0.049^{***}	-0.097^{**}	-0.154^{***}
Region Toscana	0.255^{**}	0.186^{**}	0.170^{**}	0.091	0.105	0.010
Region Umbria	0.345^{**}	0.213^{*}	0.186^{*}	0.001	-0.011	-0.102

†*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

References

[1] Cantoni, E., Ronchetti, E.: Robust inference for generalized linear models. J. Am. Stat. Assoc. **96(445)**, 1022–1030 (2001)

[2] Chambers, R., Dreassi, E., Salvati, N.: Disease mapping via negative binomial regression M-quantiles. Stat. Med. **33(27)**, 4805–4824 (2014)

[3] Huber, P. J.: Robust Statistics. John Wiley Sons, New York (1981)

[4] Schall, R.: Estimation in generalized linear models with random effects. Biometrika **78(4)**, 719–727 (1991)

[5] Sinha, S. K., Rao, J.: Robust small area estimation. Can. J. Stat. **37(3)**, 381–399 (2009)

[6] Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., Chambers, R.: Robust small area prediction for counts. Stat. Methods Med. Res. **24(3)**, 373–395 (2015)

[7] Tzavidis, N., Salvati, N., Schmid, T., Flouri, E., Midouhas, E.: Longitudinal analysis of the strengths and difficulties questionnaire scores of the millennium cohort study children in England using M-quantile random-effects regression. J. Roy. Stat. Soc. A Sta. **179(2)**, 427–452 (2016)

Goodness-of-fit assessment in linear quantile regression

Bontà di adattamento delle regressioni quantili lineari

Ilaria Lucrezia Amerise and Agostino Tarsitano

Abstract In this paper we examine three summary statistics used to evaluate the reduction in variation due to fitting a quantile regression to data. The goal is to identify characteristics and behaviors that enable a coefficient of determination to validly assess the agreement between observed and estimated responses.
Abstract *Obiettivo del presente lavoro è di esaminare caratteristiche e comportamenti di tre statistiche utilizzabili per misurare la bontà di adattamento nella regressione quantile.*

Key words: Goodness-of-fit, Coefficient of determination, Model validation.

1 Goodness-of-fit for quantile regression

Let y be the response variable of interest and $\mathbf{x}=(x_1,x_2,\cdots,x_m)$ be a m-dimensional vector of regressors. Suppose we observe a random sample $(y_i,\mathbf{x}_i), i=1,2,\cdots,n$. Let $Q_{y|\mathbf{x}_i}$ denote the τ-th conditional quantile of y given $\mathbf{x}(\tau)$. In recent years, there has been a growing interest in quantile regression, meaning that rather than focus on $E(y|\mathbf{x}_i)$, the goal is to model $Q_{y|\mathbf{x}_i}(\tau)$, that is,

$$y_i=\mathbf{x}_i\beta_\tau+e_i, \qquad Q_{y|x_i}(\tau)=\mathbf{x}_i\beta_\tau, \quad i=1,2,\cdots,n\,, \tag{1}$$

where β is a $(m\times 1)$ vector of unknown parameters specific of the τ-th quantile regression (QR). The $\mathbf{x}_i$ are rows of a $n\times m$ design matrix $\mathbf{X}$ with full rank m and

Ilaria Lucrezia Amerise
Universitá della Calabria – Dipartimento di Economia, Statistica e Finanza, Via Pietro Bucci, Cubo 1c, 87036 Rende (CS) - Italy e-mail: ilaria.amerise@unical.it

Agostino Tarsitano
Universitá della Calabria – Dipartimento di Economia, Statistica e Finanza, Via Pietro Bucci, Cubo 1c, 87036 Rende (CS) - Italy e-mail: agostino.tarsitano@unical.it

$m<n$. We assume that $x_{i,1}=1\ \forall\ i$ so that the design matrix $\mathbf{X}$ contains a column of ones. We further assume that the e_i are unobservable independent (not necessarily identically distributed) random variables, each with $Q_\tau=0$.

QR is considered a methodological improvement with respect to multiple linear regression (MLR) because it can depict a more detailed picture of the relationship between response and regressors by estimating multiple slopes along the entire response distribution. However, MLR can rely on the variance decomposition formula that subdivides the total variance of into two components: the variance explained by the model and the one associated with the experimental error. The idea is condensed in the classical ratio between explained and total variance, R^2, which is in common use. Unfortunately, the variance decomposition formula does not generally hold for a metric different from the Euclidean metric. Therefore, one must look to other measures for goodness-of-fit based on the absolute deviations of the observed responses from their median or other quantiles.

To judge the relative effectiveness of model (1), it is natural to resort to the coefficient of determination, which is well-known in the field of linear mean regression. Some proposals in this direction are reported in the following sections.

1.1 Koenker-Machado criterion

[3] generalizes the proposal of [4] for median regression to an arbitrary quantile,

$$\begin{aligned} V_\tau &= 1-\frac{\tau\sum_{y_i\geq\hat{y}_{i,\tau}}|y_i-\hat{y}_{i,\tau}|+(1-\tau)\sum_{y_i<\hat{y}_{i,\tau}}|y_i-\hat{y}_{i,\tau}|}{\tau\sum_{y_i\geq\widehat{q}_\tau}|y_i-\widehat{q}_\tau|+(1-\tau)\sum_{y_i<\widehat{q}_\tau}|y_i-\widehat{q}_\tau|}, \\ &= 1-\frac{C(y_i-\hat{y}_{i,\tau})}{C(y_i-\widehat{q}_\tau)} \end{aligned} \tag{2}$$

where $C(n_i)=\sum_{i=1}^{n} n_i[\tau-I(n_i<0)]$, $I()$ is the indicator function that takes 1 if the argument is true and 0 otherwise, $\widehat{q}_\tau$ is the estimated intercept in an intercept-only model and $\hat{y}_{i,\tau}$ is the i-th estimated response at the τ-th quantile. Notice that $\widehat{q}_\tau$ coincides with the τ-th unconditional quantile of the response. The difference $y_i-\hat{y}_{i,\tau}$ are regression residuals while $y_i-\widehat{q}_\tau$ are residuals from a model which only contains a constant term.

[4] lists some desirable properties for a coefficient of determination and V_τ satisfies many but not all of the properties. For example, it is dimensionless, ranges between zero ($\widehat{y}_{i,\tau}=\widehat{q}_\tau,\ \forall\ i$) and one ($\widehat{y}_{i,\tau}=y_i,\ \forall\ i$), with a larger value indicating better model fit. However, the quality of the fit does not change in the expected direction if a regressor is added to or deleted from the model. [3] claim that V_τ constitutes a local measure of goodness of fit for a particular quantile rather than a global measure of goodness of fit over the entire conditional distribution. Therefore, it should not be surprising that a regressor might exert a significant effect on one tail of the conditional distribution of the response but might have no effect or even an opposite effect in the other tail, with the final balance uncertain. However, coef-

ficient (2) maintains a strong analogy with the R^2 in least squares regression, which we consider not appropriate in the setting of QR.

1.2 Amerise criterion

To develop an index alternative to (2), [1] considered the triangle inequality

$$\tau \sum_{y_i \geq \widehat{y}_{i,\tau}} |y_i - \widehat{y}_{i,\tau}| + (1-\tau) \sum_{y_i < \widehat{y}_{i,\tau}} |y_i - \widehat{y}_{i,\tau}| \leq \tau \sum_{y_i \geq \widehat{y}_{i,\tau}} |y_i - \widetilde{q}_\tau| + (1-\tau) \sum_{y_i < \widehat{y}_{i,\tau}} |y_i - \widetilde{q}_\tau| + \tau \sum_{y_i \geq \widehat{y}_{i,\tau}} |\widehat{y}_{i,\tau} - \widetilde{q}_\tau| + (1-\tau) \sum_{y_i < \widehat{y}_{i,\tau}} |\widehat{y}_{i,\tau} - \widetilde{q}_\tau|. \quad (3)$$

The coefficient of determination proposed is:

$$R_\tau \quad = 1 - \frac{C_\tau (y_i - \widehat{y}_{i,\tau})}{C_\tau (y_i - \widetilde{q}_\tau) + C_\tau (\widehat{y}_{i,\tau} - \widetilde{q}_\tau)}. \quad (4)$$

The numerator is a weighted sum of absolute deviations between observed y_i and estimated $\widehat{y}_{i,\tau}$ responses. The denominator provides a standardization factor so that the coefficient varies in a limited range. More specifically, the first and second addends, respectively, are proportional to the weighted sum of absolute deviations of the $y_{i,\tau}$ and $\widehat{y}_{i,\tau}$ from $\widetilde{q}_\tau$ with no contribution from regressors. In terms of the classical variance decomposition, the denominator quantifies the total variability in the response (as expressed by the weighted sum of absolute deviations) and $C_\tau (y_i - \widehat{y}_{i,\tau})$ represents the variability of residuals. Therefore, the ratio R_τ is a measure of what percentage of the total variability in the response is explained by the regressors at the given quantile.

Coefficient (4) is, by construction, invariant to units of measurement, takes values in the $[0,1]$ interval and becomes larger as the fit tends to be perfect. It turns out that $R_\tau = 1$ if $y_i = \hat{y}_{i,\tau}, i = 1,2,\ldots,n$, that is, when the observed responses coincide with the τ-th conditional quantile regression. $R_\tau = 0$ if $\widehat{y}_{i,\tau} = \widetilde{q}_\tau \ \forall i$ implying that a model with only a constant as a regressor provides an explanation for the responses. Of course, $R_\tau > V_\tau$ simply because $\widehat{q}_\tau$ minimizes the asymmetrical liner loss function so that $C_\tau (y_i - \widehat{q}_\tau) < C_\tau (y_i - \widetilde{q}_\tau)$ for each τ.

The choice of $\widetilde{q}_\tau$ in (3) over $\widehat{q}_\tau$ can be illustrated as follows. Suppose that the variations of the estimated values were entirely due to random errors, that is, $\widehat{y}_{i,\tau} = \theta + e_i$. In this case, $\widehat{y}_{i,\tau} = \widetilde{q}_\tau$ with $R_\tau = 0$. If deviations were computed from $\widehat{q}_\tau$, then $C_\tau (\widehat{y}_{i,\tau} - \widehat{q}_\tau) > 0$ and, consequently, $R_\tau > 0$, suggesting that random regressors have some predictive power, which is the wrong conclusion. On the other hand, if $y_i = \theta + e_i$ then $R_\tau = 0$ if and only if $\widehat{q}_\tau = \widetilde{q}_\tau$. Obviously, R_τ and V_τ constitute local measures of fit because they are specific to the τ-th quantile and have little, if any, meaning for the other.

1.3 A new criterion

Let $\widetilde{q}_\tau$ be the τ-th quantile of the estimated responses $\widehat{y}_\tau$, which, in general, is different from the corresponding quantile $\widehat{q}_\tau$ of the observed response y. To judge the overall goodness of fit of a QR we can generalize the proposal of ([2], p. 223) for median regression and adapt it to an arbitrary quantile. The suggested statistic is

$$B_\tau = \frac{C_\tau(\widehat{y}_{i,\tau} - \widetilde{q}_\tau)}{C_\tau(\widehat{y}_{i,\tau} - \widetilde{q}_\tau) + C_\tau(\widehat{y}_{i,\tau} - y_i)}. \tag{5}$$

The index B_τ is a true proportion of two quantities of which the numerator is part of the denominator. Thus, the criterion is dimensionless and varies between 0 and 1. In particular, $B_\tau = 0$ denotes a non-informative model $\widehat{y}_i = \widetilde{q}_\tau \ \forall\, i$. The value of B_τ increases as $\widehat{y}_{i,\tau} \to y_i \ \forall\, i$. Finally, $B_\tau = 1$ if $y_i = \widehat{y}_i, \ \forall\, i$, that is, the model perfectly fits the observed data. The B_τ criterion can be interpreted as the relative reduction (but not necessarily monotonically) in the sum of weighted absolute deviations due to fitting the complete model with respect to fitting an intercept-only model.

2 Applications

We offer two examples; the first illustrates how the coefficients of determination are sensitive to the presence of outliers and heterosckedasticity in the data set. The second example demonstrates the usefulness of various criteria in interpreting regression fits when the number of regressors varies. Both examples are based on well-known data sets freely available on [6].

2.1 Skeena River Sockeye salmon data

Spawning stock and resulting recruitment of sockeye salmon were recorded over the years 1940–1967 in Skeena River in British Columbia (see [7]). The data set consists of $n = 28$ yearly values. The most prominent aspect of the data is their heterosckedasticity. The year 1951 (Observation 12) is an outlier resulting from a rock slide that interfered with spawning. The year 1955 (Observation 16) has a low value of S since the spawning population that year came from 1951.

Table 1 reports the coefficient of determination of ordinary least squares together with those computed for linear quantile regression at $0.25, 0.50, 0.75$. The first row concerns the original data set. Successive rows show analogous computations after excluding the design points indicated in the first column of the row.
Observation 12 is highly influential and, in fact, its exclusion alters the fit of the quantile hyperplanes considerably. Additional deletion of data point 16 has a negative impact on the agreement between observed and estimated responses because,

Table 1 Goodness-of-fit statistics for different design matrices

		$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
Deleted	R^2	V_τ	B_τ	R_τ	V_τ	B_τ	R_τ	V_τ	B_τ	R_τ
none	0.308	0.488	0.373	0.331	0.251	0.506	0.581	0.460	0.608	0.568
12	0.262	0.496	0.369	0.320	0.230	0.500	0.568	0.455	0.599	0.554
12, 16	0.211	0.485	0.296	0.244	0.191	0.473	0.531	0.443	0.582	0.528

in these case, V_τ, B_τ and R_τ decrease. In general, the values of the QR goodness-of-fit indices look similar in all the model fits, but only apparently. On a closer examination, it is pretty evident that the QR criteria are more stable than R^2, thus demonstrating their robustness. From Table 1, it emerges also that R_τ not only is the index least affected by deletion of observations but also ranges in a wider interval.

It is important to consider the magnitude of the differences between the criteria across the quantiles. This result is in line with what is expected on the basis of the local character of the criteria, which can be satisfactory in the upper tail of the response, but gradually attenuated to the point that it appears negligible below the median. This is the case for B_τ and R_τ. Index V_τ is low around the median, but it is more acceptable at the first and third quartiles. Such a behavior may be explained by noting that V_τ includes the term $C(y_i - \widehat{q}_\tau)$, which deviates widely in presence of outliers and heterosckedasticity.

2.2 Hald cement data

The data set has been widely used in the statistics literature to illustrate collinearity and variable selection methods. The data resulted from a five-component mixture experiment that was concerned with the effect of the composition of cement on heat evolved during hardening. See [5], for more details. The data frame includes $n = 13$ observations and $m = 4$ regressors. The response variable y is the heat evolved in a cement mix. The four explanatory variables are ingredients of the mix, x_1 (tricalcium aluminate), x_2 (tricalcium silicate), x_3 (tetracalcium alumino ferrite), x_4 (dicalcium silicate). An important feature of these data is that the variables x_1 and x_3 are highly correlated, as well as the variables x_2 and x_4. Thus we should expect any subset of (x_1, x_2, x_3, x_4) that includes one variable from highly correlated pair to do as any subset that also includes the other member.

Table 2 displays the results of the quantile regression analysis for the best fit of each order for three different quantiles: $0.25, 0.50, 0.75$.
Since the data are well behaved in this example, it is not surprising that the four coefficients are in close agreement and, furthermore, conclusions concerning this data set based on them cannot be too dissimilar from those based on least squares. It is worth noting that the quantile criteria V_τ, B_τ and R_τ do not always increase when a regressor is added to the model. This happens, for example, for B_τ when $\tau = 0.75$ and x_3 joins the other three regressors.

Table 2 Goodness-of-fit statistics for various models

		$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
m	R^2	V_τ	B_τ	R_τ	V_τ	B_τ	R_τ	V_τ	B_τ	R_τ
2	0.666	0.721	0.759	0.755	0.496	0.628	0.728	0.642	0.642	0.654
4	0.675	0.677	0.689	0.701	0.491	0.575	0.703	0.686	0.686	0.691
1,2	0.979	0.932	0.941	0.946	0.860	0.881	0.931	0.908	0.907	0.918
1,2,3	0.982	0.932	0.942	0.946	0.885	0.896	0.942	0.932	0.928	0.940
1,2,4	0.982	0.933	0.943	0.946	0.881	0.893	0.940	0.932	0.929	0.939
1,2,3,4	0.982	0.933	0.943	0.946	0.886	0.897	0.943	0.932	0.928	0.940

3 Concluding remarks

As a premise we note that the debate on how important goodness of fit is as a tool in quantile regression analysis is not well developed in the literature. To our knowledge only [3] has dealt with this issue. Our goal in this paper is to compare their index with other two indices originally proposed to measure the closeness between observed ad the estimated response in the case of the median regression. The comparison is carried out by means of two well-known data sets, namely, Brownlee's stack loss plant data and Hald cement data. Our findings reveal that QR criteria are sufficiently robust to suspect outliers and hence especially useful in situation of long tailed residuals. A critical characteristic of all the three indices is that they are quantile-specific and can vary significantly depending on the tail of the conditional distribution where regressors exert their effects. This prevents using them as measures for assessing how well the dependent variable can be predicted from knowledge of the independent variables.

The conclusions can be expressed by a simple recommendation: rather than searching for an optimal local measure of goodness of fit, it is more useful to combine the individual goodness-of-fit indices to form a global synthetic index of goodness-of-fit concerning all the conditional distributions of interest.

References

1. Amerise, I. L.: Iteratively reweighted constrained quantile regressions. Adv. Appl. Stat. **49**, 417–441 (2016)
2. Bonferroni, C. E.: Elementi di statistica generale. Litografia Gili, (1940-41)
3. Koenker, R. and Machado, J. A. F.: Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. **94**, 1296–1310 (1999)
4. McKean, J. W. and Sievers, G. L.: Coefficient of determination for least absolute deviation analysis. Stat. Probab. Lett. **5**, 49–54 (1987)
5. Piepel, G. and Trish Redgate, T.: A mixture experiment analysis of the Hald cement data. Am. Stat. **52**, 23–30 (1998)
6. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2018)
7. Ruppert, D. and Cressie, N. ans Carroll, R. J.: A Transformation/Weighting model for estimating Michaelis-Menten parameters. Biometrics, **45**, 637–656 (1989)

Joint Redundancy Analysis by a multivariate linear predictor

Analisi di ridondanza condivisa sulla base di un predittore lineare multivariato

Laura Marcis, Renato Salvatore

Abstract A common multi-group Redundancy Analysis is introduced, when the reduced space is given by a singular value decomposition of a multivariate best linear predictor. The algorithm finds a nearby *OLS* fixed-effects estimates by a least squares closed-form solution, provided by the standardized predictor. The empirical predictor is given by an extension of the distribution-free variance least squares method to an iterative multivariate response algorithm.

Abstract *Il lavoro introduce una Analisi di Ridondanza sulla base di gruppi indipendenti, utilizzando la decomposizione ai valori singolari di un predittore lineare multivariato. L'algoritmo fornisce stime di effetti fissi vicini alle stime OLS, attraverso una soluzione esatta sulla base del predittore standardizzato. La stima del predittore empirico è basata sull'estensione del metodo ai minimi quadrati della varianza del modello al caso multivariato, seguendo un approccio iterativo.*

Key words: Redundancy analysis, linear mixed model, empirical best linear unbiased predictor, variance least squares

Laura Marcis
Department of Economics and Law, University of Cassino and Southern Latium, e-mail: laura.marcis@unicas.it

Renato Salvatore
Department of Economics and Law, University of Cassino and Southern Latium e-mail: rsalvatore@unicas.it

1 Introduction

Redundancy Analysis (*RDA*) was originally introduced [7] in order to capture the effects on a reduced space of the linear dependence by a set of criterion variables $\mathbf{Y}$ from a set of predictors $\mathbf{X}$. Partial *RDA*, constrained *RDA* and ridge-type regularized *RDA* were also introduced, where the goal is substantially of two types: firstly, to highlight the effects of a subset of some conditioning predictors [2], to remove, and, secondly, to assess a ridge estimator to reduce the mean squared error of the multivariate regression by some nearby collinear predictors [6]. Even though the *RDA* provides a constrained analysis of the whole linear relations between the two sets of variables, and an unconstrained analysis given by the set of multivariate regression residuals, it is straightforward to relate *RDA* with principal component analysis (*PCA*), see for example [4]. One of research issues in the field of the *PCA* is the simultaneous *PC*-reduction of a set of independent groups of observations, collected by a multivariate random vector $\mathbf{Y}$ ([5],[1]). A general linear mixed model [3] is usually employed to represent the relationship between the sets of criterion and predictor variables, when the goal is to predict a specific group (subject) contribution to the linear dependence. For this reason, a *RDA* of the predicted criterion variables by the best linear unbiased predictor at group level may be quite representative into this contribution. Further, it is also useful to perform *RDA* of the modeled predicted data to investigate the "common" groups redundancy on the criterion variables, and on the multivariate mixed model conditioned residuals. This paper introduces a joint *RDA* by a least-squares solution for an optimal fixed-effects estimate from the data collected by the linear mixed model predictors of the dependent variables. The singular value decomposition of the resulting linear regression model estimates gives the best projection in the common reduced subspace of the best unbiased predictor by the whole set of random effects. The application uses an extension to the multivariate case of the variance least squares algorithm to estimate a variance components MANOVA model for data repeated over time.

2 Joint Redundancy Analysis. Estimation of model parameters

Given a q-variate random vector $\mathbf{Y}$, consider the case when $\mathbf{Y}$ is partitioned in n subjects (groups), each of them with n_i individuals. We assume that the population model for the n subjects is $\mathbf{y}_{i|q\times 1} = \mathbf{B}'_{q\times p}\mathbf{x}_{i|p\times 1} + \mathbf{a}_{i|q\times 1}$, where $\mathbf{B}$ is the

matrix of fixed regression coefficients, and $\mathbf{a}_i \sim N(0,\Sigma_a)$ is a q-variate random effect. Given a sample of N units (repeated measurements), then the multivariate random effects model assumes the general structure $\mathbf{Y}_{N\times q} = \mathbf{X}^{+}_{N\times p}\mathbf{B}_{p\times q} + \mathbf{Z}^{+}_{N\times n}\mathbf{A}_{n\times q} + \mathbf{E}_{N\times q}$, with $\mathbf{X}^{+}$ the matrix of data covariates, $\mathbf{Z}^{+}$ the design matrix of random effects, and $\mathbf{E}$ the matrix of regression within-subject errors. Further both $\mathbf{Y}$ and $\mathbf{X}^{+}$ are assumed as columnwise centered and standardized. Rewriting the last model in the vector form, with $\mathbf{y}^{*} = vec(\mathbf{Y})$, $\mathbf{X} = (\mathbf{I}\otimes\mathbf{X}^{+})$, $\beta = vec(\mathbf{B})$, $\mathbf{Za} = vec(\mathbf{ZA}) = (\mathbf{I}\otimes\mathbf{Z})vec(\mathbf{A})$, and given for the sake of simplicity a balanced design ($n_i = k$), the multivariate linear best predictor is given by $\widetilde{\mathbf{y}}^{*} = vec(\widetilde{\mathbf{Y}}) = \mathbf{X}\widehat{\beta} + \mathbf{Z}\widetilde{\mathbf{a}} = \left\{\mathbf{I} - (\Sigma_a\otimes\mathbf{ZZ}')cov(\mathbf{y}^{*})^{-1}\right\}\mathbf{X}\widehat{\beta} + (\Sigma_a\otimes\mathbf{ZZ}')cov(\mathbf{y}^{*})^{-1}\mathbf{y}^{*} = \Gamma\mathbf{y}^{*} + (\mathbf{I}-\Gamma)\mathbf{X}\widehat{\beta}$.

Here $\widehat{\beta} = \widehat{\beta}_{gls}$, $\mathbf{Z}\widetilde{\mathbf{a}} = (\Sigma_a\otimes\mathbf{ZZ}')cov(\mathbf{y}^{*})^{-1}(\mathbf{y}^{*} - \mathbf{X}\widehat{\beta})$, $cov(\mathbf{y}^{*}) = (\mathbf{I}\otimes\mathbf{Z})\ (\Sigma_a\otimes\mathbf{I}_k)(\mathbf{I}\otimes\mathbf{Z}') + cov(vec(\mathbf{E})) = (\Sigma_a\otimes\mathbf{ZZ}') + (\Sigma_e\otimes\mathbf{I}_n)\otimes\mathbf{I}_k$, $\Gamma = (\Sigma_a\otimes\mathbf{ZZ}')cov(\mathbf{y}^{*})^{-1}$.

By standard *Redundancy Analysis* (*RDA*), a reduced-rank subspace is given by a *singular value decomposition* (*SVD*) of the multivariate regression predicted values $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}_{ols}$. To get a reduced subspace by the multivariate linear best predictor $\widetilde{\mathbf{Y}}$, thus we set $\widetilde{\mathbf{Y}} = \mathbf{X}^{+}\widehat{\mathbf{B}} + \mathbf{Z}\widetilde{\mathbf{A}} = \mathbf{U}_{\widetilde{Y}}\Lambda_{\widetilde{Y}}\mathbf{V}'_{\widetilde{Y}}$ as a possible *SVD* representation of the redundant information in the criterion variables, captured by the dispersion matrix $\Sigma_{\mathbf{y}^{*}|\mathbf{X}} = var(\mathbf{y}^{*}) - cov(\mathbf{y}^{*},\mathbf{X})var(\mathbf{X})^{-1}cov(\mathbf{X},\mathbf{y}^{*})$. Even though a joint-subject reduction subspace is given by the fixed-effects model estimates $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}_{gls}$, we are interested in the simultaneous representation of all the predicted $\widetilde{\mathbf{a}}_i$'s, given by a common projection subspace. To do this, we find the matrix of fixed effects $\mathbf{B}$ closest to the fixed $\widehat{\mathbf{B}}_{ols}$, by setting the minimum Froebenius norm by the multivariate predictor $\widetilde{\mathbf{Y}}$, of the difference $F = \widetilde{\mathbf{Y}}^{**}var(\widetilde{\mathbf{y}})^{-\frac{1}{2}} - \mathbf{X}^{+}(\mathbf{B}), \widetilde{\mathbf{Y}}^{**} = \widetilde{\mathbf{Y}} - E(\widetilde{\mathbf{Y}}) = \widetilde{\mathbf{Y}} - \mathbf{1}_N E[(\widetilde{\mathbf{y}}_{|q\times 1})']: \|F\|^2 = tr(\mathbf{f}'\mathbf{f}) = \left\|\widetilde{\mathbf{Y}}^{**}\Sigma^{-\frac{1}{2}} - \mathbf{X}^{+}(\mathbf{B})\right\|^2 = \min$.

Note that $var(\widetilde{\mathbf{y}}) = \Sigma = E\left\{(\widetilde{\mathbf{Y}}-\mathbf{Y})'(\widetilde{\mathbf{Y}}-\mathbf{Y})\right\}$, with $(\widetilde{\mathbf{Y}}-\mathbf{Y})'(\widetilde{\mathbf{Y}}-\mathbf{Y})$ the random matrix of $q\times q$ cross products $\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}$ given on the basis of the subjects covariances $cov(\widetilde{\mathbf{y}}^{*}_i - \mathbf{y}^{*}_i) = \mathbf{X}_i(cov\widehat{\beta}_{gls})\mathbf{X}'_i + \mathbf{Z}_i cov(\widetilde{\mathbf{a}}_i - \mathbf{a}_i)\mathbf{Z}'_i + cov(vec(\mathbf{E}))$.

Now, setting $\mathbf{f} = vec(F) = (\Sigma^{-\frac{1}{2}}\otimes\mathbf{I}_N)\widetilde{\mathbf{y}}^{**} - (\mathbf{I}_q\otimes\mathbf{X}^{+})\overline{\beta} = \overline{\Sigma}^{-\frac{1}{2}}\widetilde{\mathbf{y}}^{**} - \mathbf{X}\overline{\beta}$, $\overline{\beta} = vec(\mathbf{B})$, $\overline{\Sigma}^{-\frac{1}{2}} = (\Sigma^{-\frac{1}{2}}\otimes\mathbf{I}_N)$, we come the following properties of $\overline{\beta}$: $tr(\mathbf{f}'\mathbf{f}) = tr\left\{(\overline{\Sigma}^{-\frac{1}{2}}\widetilde{\mathbf{y}}^{**} - \mathbf{X}\overline{\beta})'(\overline{\Sigma}^{-\frac{1}{2}}\widetilde{\mathbf{y}}^{**} - \mathbf{X}\overline{\beta})\right\} = (\widetilde{\mathbf{y}}^{**} - \overline{\mathbf{X}}\overline{\beta})'\overline{\Sigma}^{-1}(\widetilde{\mathbf{y}}^{**} - \overline{\mathbf{X}}\overline{\beta})$, where $\overline{\mathbf{X}} = \overline{\Sigma}^{\frac{1}{2}}\mathbf{X}$. Thus, $\widehat{\overline{\beta}} = (\overline{\mathbf{X}}'\overline{\Sigma}^{-1}\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}'\overline{\Sigma}^{-1}\widetilde{\mathbf{y}}^{**}$ is the q-variate vector in the subspace spanned by the columns of the matrix $\overline{\mathbf{X}}$, with $\widetilde{\mathbf{y}}^{**}$ orthogonal to the columns of $\overline{\mathbf{X}}$ in the metric of $\overline{\Sigma}^{-1}$, $\widetilde{\mathbf{y}}^{**\prime}\overline{\Sigma}^{-1}\overline{\mathbf{x}} = 0$. Then, $P_{\overline{X}} = \overline{\mathbf{X}}(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}'$ is the projection matrix of the predictor $\widetilde{\mathbf{y}}^{**}$ onto the joint subspace by $\overline{\mathbf{X}}$. The *SVD* of $\widehat{\overline{\mathbf{Y}}}^{**} = \mathbf{X}^{+}\widehat{\overline{\beta}}$

gives the common rescaled predictor's coordinates, $\mathbf{U}_{\widetilde{Y}}\Lambda_{\widetilde{Y}}\mathbf{V}'_{\widetilde{Y}}$, further noticing that $\mathbf{U}^*_{\widetilde{Y}} = \widetilde{\mathbf{Y}}\mathbf{V}^{-1}\Lambda_{\widetilde{Y}}$ contains the row joint reduced coordinates in the space of $\widetilde{\mathbf{Y}}$.

In order to avoid distributional assumptions for the multivariate data vector $\mathbf{Y}$, we introduce an Iterative multivariate Variance Least Squares (*IVLS*) estimation. The objective function to minimize is $VLS = trace(\Xi - \mathbf{U} - \mathbf{D})^2$, with $\Xi_{|N\times N}$ the empirical model covariance matrix. The algorithm is based on alternating least squares in a two-step iterative optimization process. At every iteration IVLS first fixes $\mathbf{U}$ and solves for $\mathbf{D}$, and following that it fixes $\mathbf{D}$ and solves for $\mathbf{U}$. Since *LS* solution is unique, in each step the *VLS* function can either decrease or stay unchanged, but never increase. Alternating between the two steps iteratively guarantees convergence only to a local minima, because it ultimately depends on the initial values for $\mathbf{U}$. The iterations are related to the following steps: a) from the group covariance matrices $\mathbf{U}$, first minimize VLS to obtain the estimates of $\mathbf{D}$, where Ξ is given by the multivariate *OLS* cross-products of residuals; b) after the estimation of the matrix $\widehat{\mathbf{B}}_{GLS}$, minimize VLS, setting the same error covariance matrix among groups, and c), Iterate a) and b), until convergence to the minimum. The number of iterations may vary by the choice of the specific model random effects and error covariance matrices. Applications of the *Joint RDA* may be related to different types of available data, and then accommodate a variety of patterned covariance matrices. Further, groups can be dependent or independent, even in space, time, and space-time correlated data. The *IVLS* estimator at each step is unbiased, by the following Lemma.

Lemma *(Unbiasedness of the IVLS estimator)*. Under the balanced p -variate variance components model $\mathbf{Y} = \mathbf{XB} + \mathbf{ZA} + \mathbf{E}$, with covariance matrix $\mathbf{D} + \mathbf{U}$, $\mathbf{D} = (\mathbf{I} \otimes \mathbf{Z})cov(vec(\mathbf{A}))(\mathbf{I} \otimes \mathbf{Z}')$, $\mathbf{U} = cov(vec(\mathbf{E}))$, and known matrix $\mathbf{U}$, for the *IVLS* estimator of the parameters θ in $\mathbf{D}$ we have $E\left\{\mathbf{D} = D(\widehat{\theta}_{IVLS})\right\} = D(\theta)$.

3 Application and concluding remarks

Recent national laws reformed the Italian Budget law, provided that the "Benessere Equo e Sostenibile (BES) - Fair and Sustainable Well-being (FSW)" [8] indicators should contribute to define those economic policies which largely affect some fundamental dimensions for the quality of life. The Italian Statistical Institute provide these indicators annually. The Ministry of Finance and Economics most recent publication is the Budget Law 2019 where it is possible to find the trend and programmatic forecasts relating to the 12 FSW indicators and the anal-

ysis of most recent trends, at the levels NUTS2 and NUTS3. We analyze 5 of the 12 FSW indicators available in the years 2013-2016 (4 time istants), at the level of NUTS2. The random multivariate vector is partitioned in repeated observations of the same administrative Region of Italy in the 4 time instants. We take in consideration a balanced multivariate Mixed *MANOVA* Model (*MMM*), with an *AR(1)* error stucture: $\mathbf{Y}_{|mk\times p} = \mathbf{X}_{|mk\times l}\mathbf{B}_{|l\times p} + \mathbf{Z}_{|mk\times pm}\mathbf{A}_{|pm\times p} + \mathbf{E}_{|mk\times p}$, where $p = 5$, $m = 20$, $k = t = 4$, that is a balanced model, with a random intercept and an *AR(1)* error. Then: $vec(\mathbf{Y}) = (\mathbf{I}_p \otimes 1_{mt})vec(\mathbf{B}) + (\mathbf{I}_p \otimes Z)vec(\mathbf{A}) + vec(\mathbf{E})$; $\mathbf{y}^* = vec(\mathbf{Y}), \mathbf{X}^* = (\mathbf{I}_p \otimes \mathbf{X}) = (\mathbf{I}_p \otimes \mathbf{1}_{mt})$, $\beta^* = vec(\mathbf{B})$, $\mathbf{Z}^*\mathbf{a}^* = (\mathbf{I}_p \otimes \mathbf{Z})vec(\mathbf{A})$.

Further we have $cov(\mathbf{y}^*) = (\mathbf{I}_p \otimes \mathbf{I}_m \otimes \mathbf{1}_k)(\Sigma_a \otimes \mathbf{I}_m)(\mathbf{I}_p \otimes \mathbf{I}_m \otimes \mathbf{1}_k') + cov(vec(\mathbf{E})) = \Sigma_a \otimes (\mathbf{I}_m \otimes \mathbf{1}_k\mathbf{1}_k') + (\Sigma_e \otimes \mathbf{I}_n) \otimes \Omega$. Finally, after the iterative *VLS* estimation, the predictor is given by $\widetilde{\mathbf{y}}^* = \mathbf{X}^*\widehat{\beta}^*_{GLS} + \mathbf{Z}^*\widetilde{\mathbf{a}}^* = \Gamma\mathbf{y}^* + (\mathbf{I} - \Gamma)\mathbf{X}^*\widehat{\beta}^*_{GLS}$, $\Gamma = (\Sigma_a \otimes \mathbf{Z}\mathbf{Z}')cov(\mathbf{y}^*)^{-1}$. Note that the matrix Γ specifies both the contribution of the regression model and the observed data to the prediction.

We assume equicorrelation both of the multivariate random effects and the residual covariance, together with the *AR(1)* structure of the error:

$$\Sigma_a = \sigma_a^2 \times \begin{bmatrix} 1 & \rho_a & \cdots & \rho_a \\ \rho_a & 1 & \cdots & \rho_a \\ \vdots & \cdots & \ddots & \vdots \\ \rho_a & \rho_a & \cdots & 1 \end{bmatrix}_{5\times 5} \Sigma_e = \sigma_e^2 \times \begin{bmatrix} 1 & \rho_e & \cdots & \rho_e \\ \rho_e & 1 & \cdots & \rho_e \\ \vdots & \cdots & \ddots & \vdots \\ \rho_e & \rho_e & \cdots & 1 \end{bmatrix}_{5\times 5} \Omega = \frac{1}{1-\rho_t^2} \begin{pmatrix} 1 & \rho_t & \rho_t^2 & \rho_t^3 \\ \rho_t & 1 & \rho_t & \rho_t^2 \\ \rho_t^2 & \rho_t & 1 & \rho_t \\ \rho_t^3 & \rho_t^2 & \rho_t & 1 \end{pmatrix}_{4\times 4}$$

Figure 1 reports a comparison between observed and predicted data. Bold lines refer to predicted loadings and black dots are the predicted scores. Dashed lines and grey dots come from the standard *PCA*.

In conclusion, the paper introduces *RDA* of a multivariate predictor to perform a common survey of the predicted data, a joint *RDA* analysis. Given a multivariate vector with independent groups, and a random effects population model, the joint *RDA* relies on the assumption that the linear model itself is able to predict accurately specific subjects or group representatives, even in time and spatial dependent data. After using a linear mixed model, the joint *RDA* explores data that originates in part from regressive process and in part from the observed, to understand the contribution to the linear dependence of the observed and of predictions. We suggest the use of this approach when the research issues are related to the use of model covariates and

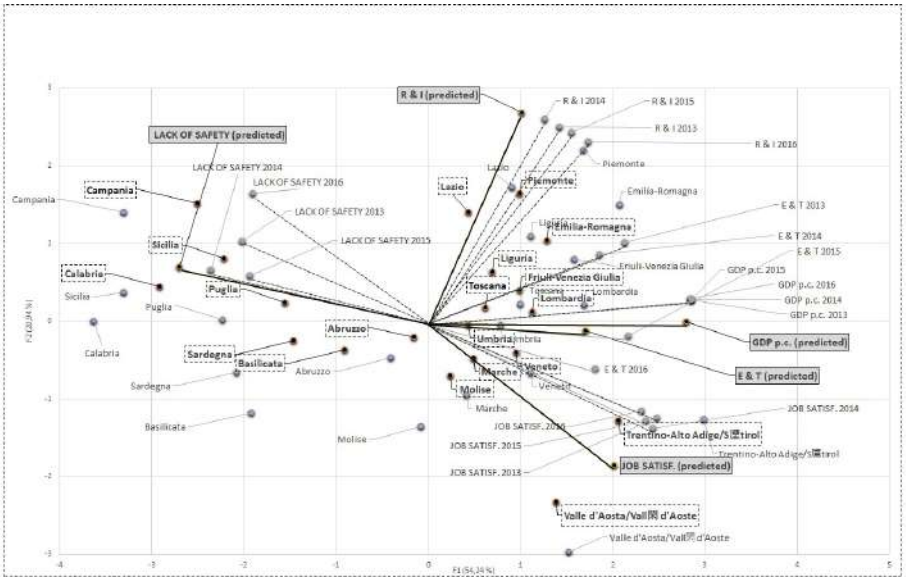


Fig. 1 Multiple Factor Analysis (MFA), observed factor loadings and scores per year (in grey); predicted loadings and scores (in black).

specific patterned covariance matrices. Further, the impact of choosing the model structure is easily recognizable when we investigate changes in the data description by the common factors.

References

1. Bechger, T.M., Blanca, M.J., Maris, G.: The analysis of multivariate group differences using common principal components. Structural Equation Modelling **21**, 577-587 (2014)
2. Borcard, D., Legendre, P., Drapeau, P.: Partialling out the spatial component of ecological variation. Ecology **84**, 511-525 (1992)
3. Demidenko: Mixed Models: theory and applications. Wiley (2004)
4. Härdle, W.K., Simar, L.: Principal Components Analysis. In: Applied Multivariate Statistical Analysis, pp. 319-358. Springer Berlin Heidelberg (2015)
5. Neuenschwander, B.E., Flury, B.D.: Common Principal Components for Dependent Random Vectors. J. of Multivar. Analysis **75**, 163-183 (2000)
6. Takane, Y., Hwang, H.: Regularized linear and kernel redundancy analysis. Computational Statistics and Data Analysis. **52(1)** 394-405 (2007)
7. van den Wollenberg, A.L.: Redundancy analysis an alternative for canonical correlation analysis. Psychometrika **42**, 207–219 (1977)
8. https://www.istat.it/en/

M-quantile regression shrinkage and selection via the lasso

Regressione M-quantilica con regolarizzazione lasso

M. Giovanna Ranalli and Lea Petrella and Francesco Pantalone

Abstract Standard regression analysis investigates the average behavior of a response variable, y, given a vector of predictors, $\mathbf{x}$. However, in some cases, the mean does not give a complete picture of a distribution. Therefore, quantile regression analyzes how the q-th quantile of the conditional distribution of y given $\mathbf{x}$ varies with $\mathbf{x}$, and M-quantile regression generalizes this idea through the use of influence functions. When dealing with a large number of predictors, selection of a subset of them improves the interpretability of the model. Towards this end, in this paper, we introduce M-quantile regression with lasso regularization. This allows us to investigate the extreme behavior of y conditional on $\mathbf{x}$ and to shrink the predictors in order to perform model selection.

Abstract *I classici metodi di regressione studiano il comportamento medio di una risposta, y, dati alcuni predittori,* $\mathbf{x}$*. In alcuni casi, il comportamento medio non fornisce un'informazione adeguata di un fenomeno. Per questo motivo, la regressione quantilica analizza come il q-esimo quantile della distribuzione condizionata di y dato* $\mathbf{x}$ *varia a seconda di* $\mathbf{x}$*, e la regressione M-quantilica la generalizza utilizzando funzioni di influenza. Inoltre, quando il numero di predittori è elevato, selezionarne un sottoinsieme migliora l'interpretibilità del modello. Per questo motivo, in questo lavoro introduciamo la regressione M-quantilica con la regolarizzazione lasso. Questo permette di investigare il comportamento a diversi livelli della distribuzione di y condizionata ad* $\mathbf{x}$ *e di effettuare, allo stesso tempo, la selezione del modello.*

Key words: Robust regression, variable selection, prostate cancer data.

M. Giovanna Ranalli
Dept. of Political Science, University of Perugia, Italy, e-mail: giovanna.ranalli@unipg.it

Lea Petrella
Dept. MEMOTEF, Sapienza University of Rome, Italy, e-mail: lea.petrella@uniroma1.it

Francesco Pantalone
Dept. of Economics, University of Perugia, Italy, e-mail: francesco.pantalone@studenti.unipg.it

1 Introduction

When investigating the relationship between a response variable y and some predictors $\mathbf{x}$, it may be useful to employ *quantile regression* (Koenker, 1978; Koenker and D' Orey, 1987), that is to analyze how the q-th quantile of the conditional distribution of y given $\mathbf{x}$ varies with $\mathbf{x}$. This is particularly useful when the average behavior of y given $\mathbf{x}$, investigated by means of the classical regression analysis, does not give a complete picture of the distribution. In particular, quantile regression assumes that the q-th quantile of y is a linear function of the predictors, in other words $Q_q(y|\mathbf{x}) = \mathbf{x}^T \mathbf{b}(q)$, and it leads to a family of hyperplanes indexed by the value of the corresponding quantile coefficient $q \in (0,1)$. Given a sample of n observations, the vector $\mathbf{b}(q)$ is estimated by minimizing

$$\sum_{i=1}^{n} |r_i[\mathbf{b}(q)]| \{(1-q) I(r_i[\mathbf{b}(q)] \leq 0) + qI(r_i[\mathbf{b}(q)] > 0)\},$$

with respect to $\mathbf{b}(q)$ through linear programming methods, where $r_i[\mathbf{b}(q)] = y_i - \mathbf{x}_i^T \mathbf{b}(q)$.

M-quantile regression (Breckling and Chambers, 1988) generalizes the quantile regression idea through the use of influence functions. Specifically, the M-quantile regression line of order q is defined as the solution $Q_q(y|\mathbf{x}, \psi_q) = \mathbf{x}^T \mathbf{b}_\psi(q)$ to $\int \psi_q(y - Q_q)\, dF(y|\mathbf{x}) = 0$, where F denotes the distribution of y given $\mathbf{x}$ underlying the data and ψ_q denotes the influence function associated to q-th M-quantile. The general M-estimator of $\mathbf{b}_\psi(q)$ is obtained by solving the set of estimating equations $\sum_{i=1}^{n} \psi_q\left(y_i - \mathbf{x}_i^T \mathbf{b}_\psi(q)\right) \mathbf{x}_i = \mathbf{0}$, with respect to $\mathbf{b}_\psi(q)$. It is assumed that

$$\psi_q\left(r_{iq\psi}\right) = 2\psi\{s^{-1} r_{iq\psi}\}\{(1-q) I\left(r_{iq\psi} \leq 0\right) + qI\left(r_{iq\psi} > 0\right)\},$$

where $r_{iq\psi} = y_i - \mathbf{x}_i^T \mathbf{b}_\psi(q)$ and s is a robust estimate of the scale, such as the MAD estimate for which $s = \text{median}|r_{iq\psi}|/0.6745$. We consider the Huber (Huber, 1981) influence function, given by $\psi(u) = u$, if $-c \leq u \leq c$, and $\psi(u) = c\ \text{sign}(u)$, otherwise, where c is a cutoff constant. In particular, robustness is increased as c decreases, while efficiency is increased as c increases.

Generally, we can improve the interpretability of a model selecting a subset of predictors, and we may improve the efficiency shrinking the coefficients towards zero. One useful tool is the lasso (Tibshirani, 1996), which is a penalized least square method that imposes an L_1-penalty on the coefficients, i.e. the sum of the absolute value of the coefficients must be less than a given constant. In this regard, quantile regression with lasso regularization has been introduced by Li and Zhu (2008).

In this work, we propose an M-quantile regression that takes advantage of a lasso regularization, so that we can analyze the behavior of y given $\mathbf{x}$ and do selection and shrinkage at the same time. The paper is organized as follows. In Section 2, we introduce the proposed method while, in Section 3, we illustrate the method through the use of a classical data set and provide some conclusions and future research.

2 M-quantile regression and the lasso

Suppose we have a regression data set $(\mathbf{X},\mathbf{y})$ with $\mathbf{X}\in\mathbb{R}^{n\times p}$ and $\mathbf{y}\in\mathbb{R}^n$. The lasso-estimator for M-quantile regression coefficients can be defined as

$$\hat{\mathbf{b}}=\left(\hat{b}_0,\hat{\mathbf{b}}_1^T\right)^T=\text{argmin}\{L(\mathbf{X},\mathbf{y},\mathbf{b}):b_0\in\mathbb{R},\mathbf{b}_1\in\mathbb{R}^p\}\text{ subject to }\sum_{j=1}^{p}|b_j|\leq t,$$

where the loss function L is given by

$$L(\mathbf{X},\mathbf{y},\mathbf{b})=\sum_{i=1}^{n}\psi_q\left(r_i(\mathbf{b})/\sigma_q\right), \tag{1}$$

with $r_i(\mathbf{b})=y_i-b_0-\mathbf{x}_i^T\mathbf{b}_1$, and σ_q is the scale of the residuals and it can be estimated by MAD. Here $t\geq 0$ is the tuning parameter and it controls the amount of shrinkage that is applied to the estimates. Setting the derivatives of (1) with respect to $\mathbf{b}$ to zero yields the following system of equations,

$$\mathbf{W}\left(\mathbf{y}-\hat{b}_0\mathbb{1}_n-\mathbf{X}\hat{\mathbf{b}}_1\right)=\mathbf{0}, \tag{2}$$

$$\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)\hat{\mathbf{b}}_1=\mathbf{X}^T\mathbf{W}\left(\mathbf{y}-\hat{b}_0\mathbb{1}_n\right), \tag{3}$$

where $\mathbf{W}=\text{diag}\,(w_i)$ and $w_i(t_i)=\psi_q(t_i/2t_i)$. This is a weighted version of normal equations for the classical lasso estimator for the regression coefficients. As usual in M-quantile regression, these weighted normal equations require an iterative procedure. We describe a stable algorithm to solve this system of equations.

Let $\delta_i, i=1,2,\ldots,2^p$ be the p-tuples of the form $(\pm1,\pm1,\ldots,\pm1)$. The condition $\sum_{j=1}^{p}|b_j|\leq t$ is equivalent to $\delta_i^T\mathbf{b}_1\leq t$, $\forall i$. For a given $\mathbf{b}_1$, let $E=\{i:\delta_i^T\mathbf{b}_1=t\}$ and $S=\{i:\delta_i^T\mathbf{b}_1<t\}$. Denote by $\mathbf{G}_E$ the matrix whose rows are δ_i for $i\in E$. The algorithm starts with $E=\{i_0\}$ where $\delta_{i_0}=sign\left(\hat{\mathbf{b}}_1\right)$, $\hat{\mathbf{b}}_1$ being the overall M-quantile estimate where $\hat{b}_0$ and $\hat{\mathbf{b}}_1$ are computed iteratively solving (2) and (3). Then, find $\hat{\mathbf{b}}_1$ to minimize $L(\mathbf{X},\mathbf{y},\mathbf{b})$ subject to $\mathbf{G}_E\hat{\mathbf{b}}_1\leq t\mathbb{1}_p$ by using a constrained M-quantile procedure. Given $\hat{\mathbf{b}}_1$, we can estimate b_0 by using (2). While $\{\sum_{j=1}^{p}|b_{1j}|>t\}$, we add i to the set E where $\delta_i=sign\left(\hat{\mathbf{b}}_1\right)$. Find $\hat{\mathbf{b}}_1$ to minimize $L(\mathbf{X},\mathbf{y},\mathbf{b})$ subject to $\mathbf{G}_E\hat{\mathbf{b}}_1\leq t\mathbb{1}_p$. Given $\hat{\mathbf{b}}_1$, we can estimate b_0 by using (2). The procedure must always converge in a finite number of steps since one element is added to the set E at each step, and there is a total of 2^p elements.

The lasso parameter t can be estimated by cross-validation or generalized cross-validation.

3 Illustration and final remarks

Data used for this illustration come from a study of prostate cancer (Stamey et al., 1989) and are used in Tibshirani (1996) to illustrate the standard lasso. The response variable is the logarithm of prostate-specific antigen (lpsa), while the predictors are given by eight clinical measures: the logarithm of cancer volume (lcavol), the logarithm of prostate weight (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), the seminal vesicle invasion (svi), the logarithm of the capsular penetration (lcp), the Gleason score (gleason), and the percentage of the Gleason score 4 or 5 (pgg45).

We perform the M-quantile regression with lasso selection at different values of q and compare the results with standard lasso regression and M-quantile regression. The predictors have been standardized. In the case of standard lasso regression and M-quantile regression with lasso the optimal t is computed by means of five-fold cross-validation. The tuning constant c used in the M-quantile regression and M-quantile regression with lasso is kept fixed at 1.345. This value provides fairly high efficiency in the normal case and still offers protection against outliers (Huber, 1981).

In Table 1 we report the estimated coefficients and corresponding standard errors. The latter, in the case of M-quantile regression and M-quantile regression with lasso, are obtained by $M = 1000$ bootstrap replicates, in which the optimal t is computed at each replicate. We can see that OLS and M-quantile provide similar results, and that Lasso and M-quantile lasso provide similar results. This can be explained by the fact that the response variable is on the log-scale and that there are not influential observations. When comparing M-quantile regression with M-quantile lasso, we can see that the shrinkage effectively performs model selection. If we fit OLS, M-quantile regression and M-quantile with lasso for the response variable on the original scale, differences among the methods are more pronounced as the response variable has a distinctive skew distribution (see the second half of Table 1). In fact, OLS and M-quantile provide quite different estimates, by this providing evidence of the need for robustness. In addition, M-quantile lasso provides model selection and standard errors that are generally lower than M-quantile, further improving on efficiency in this case.

In Table 2 we compare the M-quantile and the M-quantile lasso at $q = \{0.25, 0.50, 0.75\}$ for the log-scale response. In Figure 1 we show the value of the estimates for M-quantile lasso as a function of t, and in Figure 2 we show the value of the estimates as a function of q. Again, model selection is performed at different levels of the conditional distribution of the response given the predictors. In addition, we can notice that for most of the predictors, the effect is the same for different values of q, while for others it increases with q (svi).

We plan to make this approach more flexible by extending it to Elastic Net and to allow for non-linear relationships using penalized spline regression.

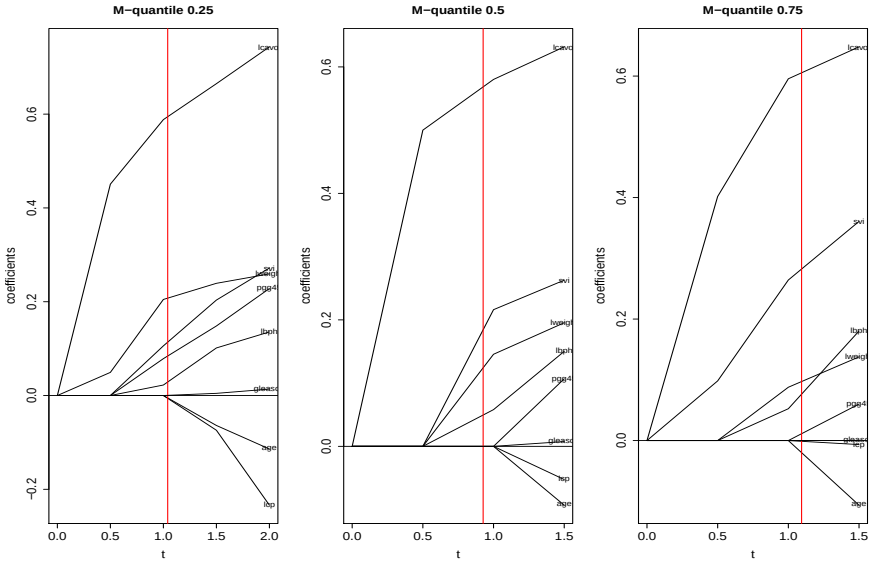


Fig. 1 Solution path for M-quantile regression with lasso at $q = \{0.25, 0.50, 0.75\}$. Each curve represents the value of the estimated coefficient at the corresponding value of t. The red line is at the value of t selected by fivefold cross-validation.

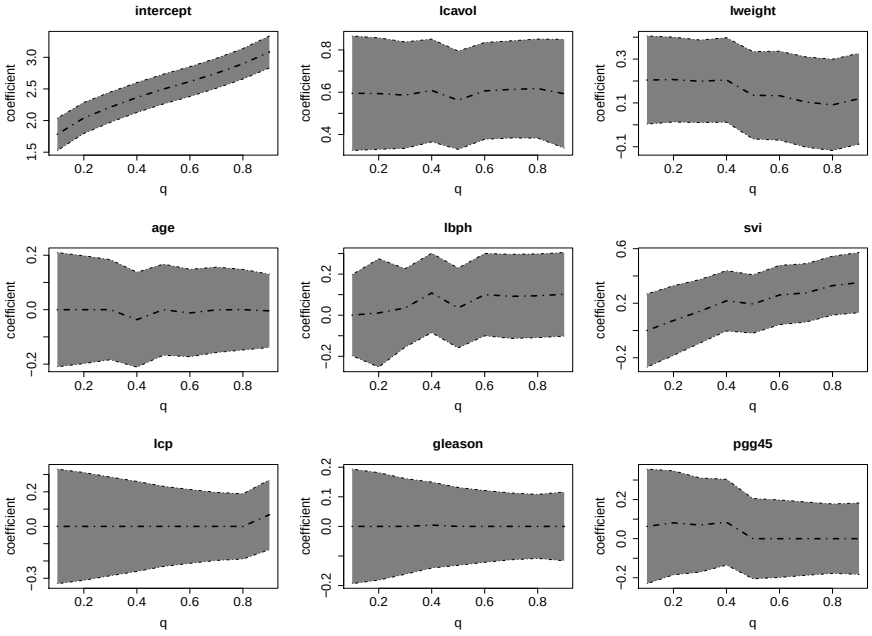


Fig. 2 Parameter estimates for M-quantile regression with lasso and 95% confidence intervals for varying q.

	Response on log-scale								Response on original scale							
	OLS		Lasso		M-quantile		M-q lasso		OLS		Lasso		M-quantile		M-q lasso	
Predictor	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.
Intercept	2.48	0.07	2.48	0.07	2.48	1.32	2.50	0.12	23.74	3.44	23.74	4.20	17.40	23.21	15.84	2.23
lcavol	0.69	0.10	0.56	0.10	0.69	0.09	0.56	0.12	12.87	4.95	6.95	3.77	7.58	2.10	5.12	1.59
lweight	0.23	0.08	0.10	0.08	0.23	0.25	0.13	0.10	1.53	4.04	0.00	1.40	2.61	3.30	0.63	1.23
age	-0.15	0.08	0.00	0.05	-0.15	0.01	0.00	0.09	-4.71	3.98	0.00	2.91	-2.14	0.24	0.00	1.00
lbph	0.16	0.08	0.00	0.07	0.19	0.07	0.04	0.10	3.32	4.05	0.00	2.51	0.88	1.14	0.00	0.91
svi	0.32	0.10	0.16	0.09	0.33	0.24	0.19	0.11	14.93	4.83	9.78	5.54	7.35	8.30	3.43	3.30
lcp	-0.15	0.13	0.00	0.03	-0.20	0.10	0.00	0.12	3.12	6.08	0.00	4.65	-2.44	1.95	0.00	1.21
gleason	0.03	0.11	0.00	0.02	0.02	0.14	0.00	0.07	-3.52	5.43	0.00	4.34	-0.73	2.85	0.00	0.75
pgg45	0.13	0.12	0.00	0.03	0.18	0.00	0.00	0.10	-1.28	5.96	0.00	5.84	2.46	0.09	0.00	1.24

Table 1 Parameter estimates and corresponding estimated standard errors for the prostate cancer data for ordinary least squares (OLS), Lasso, M-quantile regression and M-quantile regression with lasso at $q = 0.5$.

	$q = 0.25$				$q = 0.50$				$q = 0.75$			
	M-quantile		M-q lasso		M-quantile		M-q lasso		M-quantile		M-q lasso	
Predictor	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.	*b*	s.e.
Intercept	2.16	1.74	2.14	0.12	2.48	1.32	2.50	0.12	2.81	1.36	2.82	0.12
lcavol	0.76	0.09	0.60	0.13	0.69	0.09	0.56	0.12	0.67	0.10	0.61	0.12
lweight	0.26	0.25	0.21	0.10	0.23	0.25	0.13	0.10	0.14	0.28	0.10	0.11
age	-0.13	0.01	0.00	0.10	-0.15	0.01	-0.00	0.09	-0.13	0.01	-0.00	0.08
lbph	0.15	0.07	0.03	0.10	0.19	0.07	0.04	0.10	0.20	0.08	0.09	0.10
svi	0.29	0.26	0.12	0.12	0.33	0.24	0.19	0.11	0.38	0.26	0.30	0.11
lcp	-0.28	0.10	-0.00	0.15	-0.20	0.10	0.00	0.12	-0.05	0.10	-0.00	0.10
gleason	0.01	0.17	-0.00	0.09	0.02	0.14	-0.00	0.07	-0.00	0.15	0.00	0.06
pgg45	0.25	0.00	0.08	0.13	0.18	0.00	0.00	0.10	0.08	0.00	0.00	0.09

Table 2 Parameter estimates and corresponding estimated standard errors for the prostate cancer data for M-quantile regression and M-quantile regression with lasso at $q = \{0.25, 0.50, 0.75\}$.

References

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

Huber, P. J. (1981). *Robust statistics / Peter J. Huber*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley & Sons, New York.

Koenker, R. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. W. and D' Orey, V. (1987). Computing regression quantiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):383–393.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185.

Stamey, T. A., Kabalin, J. N., Ferrari, M., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. iv. anti-androgen treated patients. *The Journal of urology*, 141(5):1088–1090.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

New insights into the Conditioning and Gain Score approaches in multilevel analysis

Alcune riflessioni su approccio condizionato e approccio alle differenze nell'analisi multilivello .

Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini

Abstract We consider the issue of estimating the effect of a treatment variable on student achievement when a pre-test is available, taking into account the hierarchical structure of the data, with students nested into schools. The treatment variable can be either at student level or at school level. This effect can be estimated alternatively by adjusting for the pre-test score, i.e. conditioning, or by using the difference between post-test and pre-test scores, namely the gain score. The performance of the two approaches depends on pre-test reliability and validity of the common trend assumption. We derive approximated analytical results and we compare the two approaches via a simulation study.

Abstract *Questo lavoro affronta il problema della stima dell'effetto di uno specifico intervento (trattamento) sull'apprendimento degli studenti, nel caso in cui si disponga di una misura dell'abilità sia prima che dopo il trattamento, tenendo in considerazione la natura gerarchica dei dati, con gli studenti raggruppati in scuole. Il trattamento può essere sia a livello di singolo studente che a livello di scuola. Considerando il caso in cui l'assegnazione al trattamento non è casuale, esistono due approcci alternativi per la stima dell'effetto di interesse. Un primo approccio si basa sul condizionamento rispetto al test di abilità effettuato prima del trattamento, mentre il secondo approccio considera il guadagno, cioè la differenza tra i punteggi al test prima e dopo il trattamento. La validità dei due approcci dipende dalla affidabilità del test prima del trattamento come misura dell'abilità dello studente e dall'ipotesi di effetto comune dell'abilità nel determinare il punteggio ai due test. In questo lavoro, mostriamo alcuni risultati teorici e i risultati di uno studio di simulazione per il confronto tra questi due approcci.*

Key words: Achievement tests, Random effects model, Treatment effect.

Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence e-mail: bruno.arpino@unifi.it, silvia.bacci@unifi.it, leonardo.grilli@unifi.it, raffaele.guetto@unifi.it, carla.rampichinii@unifi.it

1 Introduction

In the education literature, student achievement is typically measured using multilevel models [1], with students nested into schools. In this contribution we aim to assess the effect of a specific treatment at student level (e.g. an individual support program) or at the school level (e.g. a given school policy). We consider a setting where student achievement is measured by means of a standardized test in two occasions, one before the treatment (pre-test), and the other one after the treatment (post-test). Two main methodological approaches have been proposed to estimate the treatment effect in this setting. The first approach consists in estimating the effect of the treatment on the post-test score, conditionally on the pre-test score (*conditioning* approach). The second approach considers the difference between the post-test score and the pre-test score as response variable, the so called *gain score* approach.

These two approaches give unbiased estimates under different assumptions, which are difficult to evaluate in observational studies. The debate on which of the two methods has to be preferred is still ongoing. Recently, Kim and Steiner [2] reconsidered the choice between the two approaches using graphs to illustrate the conditions under which a method has to be preferred. Despite the important contribution of this and several other studies that we review next, this literature has overlooked the fact that often test scores are collected in data sets with a multilevel structure, like in the educational setting.

We extend the results of Kim and Steiner [2] in two directions. First, we consider the most frequent case of a binary treatment variable rather than a continuous one. Second, we carry out the comparison between the two approaches when data have a multilevel structure, like in education setting, relying on a two level linear model. The multilevel setting has additional characteristics, playing a role in the comparison of the two approaches. In particular, the treatment can be either at student level or at school level. Moreover, the model includes random effects and it may also include cluster means.

2 Conditioning and gain score approaches

We consider individuals ($i = 1,\dots,n_j$) nested into clusters ($j = 1,\dots,J$), for example students nested into schools. Let Y_{1ij} and Y_{2ij} be continuous variables describing the observable scores on the pre-test and the post-test. These scores are error prone measures of a latent ability A_{ij}. The difference $G_{ij} = Y_{2ij} - Y_{1ij}$ is the gain score.

We are interested in assessing the effect of a treatment Z_{ij} on the post-test score Y_{2ij}, taking into account that the unobservable ability A_{ij} acts as a confounder affecting both Z_{ij} and Y_{2ij}. We assume that the two scores are generated by the following random intercept models:

$$Y_{1ij} = \mu_1 + \beta_1 A_{ij} + \psi_1 \bar{A}_j + u_{1j} + \lambda_1 e_{ij} \tag{1}$$

$$Y_{2ij} = \mu_2 + \beta_2 A_{ij} + \psi_2 \bar{A}_j + \tau Z_{ij} + u_{2j} + \lambda_2 e_{ij} + v_{ij}, \tag{2}$$

where $\bar{A}_j$ is the cluster-mean ability, μ_1 and μ_2 are intercepts, β_1 and β_2 are the within effects of the ability on the pre- and post-test scores, respectively, whereas ψ_1 and ψ_2 are the corresponding contextual effects. The treatment effect of interest is τ. The random variables u_{1j} and u_{2j} are the level 2 errors, while e_{ij} is a common level 1 error, and v_{ij} is a level 1 error specific to the post-test. All errors are assumed to be normally distributed with 0 mean and constant variances. Without loss of generality, we assume that both the individual ability and the error terms are normally distributed with mean 0 and unit variance, that is, $E(A_{ij}) = E(e_{ij}) = E(u_j) = 0$, and $Var(A_{ij}) = Var(e_{ij}) = 1$. Then, $Var(Y_{1ij}) = \beta_1^2 + \sigma_u^2 + \lambda_1^2$. When $\lambda_1 = 0$ the ability A_{ij} is measured without error by the pre-test Y_{1ij}. The gain score is defined as $G_{ij} = Y_{2ij} - Y_{1ij}$. The data generating path defined by equations (1) and (2) is represented in Figure 1. In order to estimate the treatment effect τ it is necessary to rely on some assumptions on the unobserved confounding due to the ability A_{ij}.

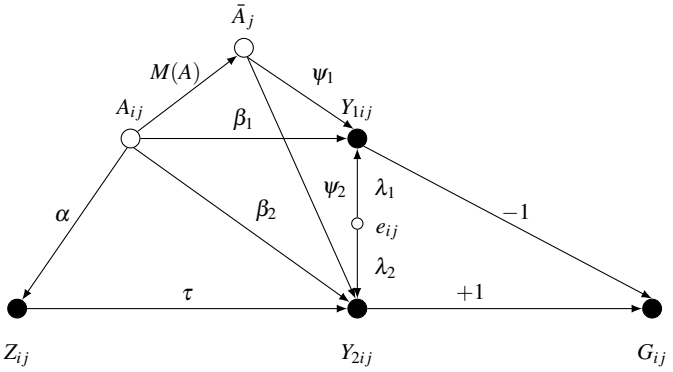


Fig. 1 *Path diagram for conditioning and gain score approaches; different effect of ability at within- and between-levels.*

In particular, according to the conditioning approach, we use the pre-test score Y_{1ij} as a proxy of the latent ability A_{ij}. Thus the conditioning model is specified as follows:

$$Y_{2ij} = \mu_2 + \beta_2 Y_{1ij} + \psi_2 \bar{Y}_{1j} + \tau Z_{ij} + u_{2j} + \varepsilon_{ij}. \tag{3}$$

The treatment effect τ is correctly estimated from model (3) if A_{1ij} is measured without error by Y_{1ij}, i.e. when λ_1 in equation (2) is equal to zero.

On the other hand, considering equations (1) and (2), we obtain $E(G_{ij}) = Y_{2ij} - Y_{1ij} = (\mu_2 - \mu_1) + (\beta_2 - \beta_1)A_{ij} + (\psi_2 - \psi_1)\bar{A}_j + \tau Z_{ij}$. Thus, under the common trend assumption, i.e. $\beta_2 = \beta_1$ and $\psi_2 = \psi_1$, the gain score is unaffected by the ability and the treatment effect τ can be estimated without bias from the following model:

$$G_{ij} = \tilde{\mu} + \tau Z_{ij} + \tilde{u}_j + \tilde{\varepsilon}_{ij}. \tag{4}$$

3 Main results

Considering a binary treatment, we derived the bias formula for the conditioning approach using model (3) without random effects, e.g. for the OLS estimator of τ, thus extending the results of Kim and Steiner [2]. We checked by a simulation experiment that this formula is a good approximation of the bias for the GLS estimator of τ in the random effects model (3).

The simulation study is based on 1000 data sets, each of them being composed of $10,000$ individuals uniformly distributed in 100 groups. Values of parameters used to generate data mimic the structure of Invalsi data [3]. In particular, the true value of τ (treatment effect) is equal to 2.

Table 1 displays the main results of the simulation study. Nine different configurations are taken into account, which distinguish for: (i) absence ($\lambda_1 = \lambda_2 = 0$) or presence of the measurement error, only on the pre-test ($\lambda_1 \neq 0$) and also on the post-test ($\lambda_2 \neq 0$), and (ii) validity of the common trend assumption, at level 1 ($\beta_2 = \beta_1$), at level 2 ($\psi_1 = \psi_2$), or at both levels. For each configuration, we show the mean of the estimated treatment effects and the corresponding relative error.

Table 1 Simulation study: Means of estimated treatment effect ($\hat{\tau}$) and corresponding relative error ($\%err$), individual-level treatment: conditioning and gain score approaches.

Conf.	Measurement error	Common trend level 1	Common trend level 2	Conditional $\hat{\tau}$	Conditional %err	Gain $\hat{\tau}$	Gain %err
1	no ($\lambda_1 = \lambda_2 = 0$)	yes	yes	2.0	0.0	2.0	0.0
2	no ($\lambda_1 = \lambda_2 = 0$)	no	yes	2.0	0.0	9.0	348.1
3	no ($\lambda_1 = \lambda_2 = 0$)	no	no	2.0	0.0	9.0	348.5
4	yes ($\lambda_1 = 6$; $\lambda_2 = 0$)	yes	yes	3.9	95.5	2.0	0.0
5	yes ($\lambda_1 = 6$; $\lambda_2 = 0$)	yes	no	3.9	95.5	2.0	0.0
6	yes ($\lambda_1 \neq \lambda_2$)	yes	yes	3.0	47.5	2.0	0.0
7	yes ($\lambda_1 = \lambda_2 = 6$)	yes	yes	2.0	0.0	2.0	0.0

Our simulations show that the results of Kim and Steiner generalize to a multilevel setting with some adjustments and further assumptions. Indeed, the treatment effect is correctly estimated using the conditioning approach when the pre-test is measured without error (configurations 1, 2, 3), whereas the gain score approach gives unbiased estimates when the common trend assumption is satisfied (configurations 4, 6, 7). Thus, under the common trend assumption, the gain score approach has to be preferred to the conditioning one in presence of measurement error on the pre-test, even if the reliability is quite high (say around 0.85).

As a further peculiarity, the conditioning approach provides satisfactorily results when the measurement error acts both on the pre-test and the post-test, but at the same extent (configuration 7). In addition, when the treatment is at level 1 (i.e., the target of the treatment are the single students), as in the simulation study here presented, its effect is correctly estimated using the gain score approach if the common trend assumption holds at level 1, even if it is violated at level 2 (configuration 5).

For the future development of this work, we intend to extend the simulation study to consider when the treatment acts at level 2 (i.e., the target of the treatment are the schools). We expect that in such a situation the common trend assumption at level 2 is crucial.

References

1. Goldstein, H.: Multilevel Statistical Models, 4th ed., Wiley (2010)
2. Kim, Y., Steiner, P.M.: Gain Scores Revisited: A Graphical Models Perspective. Sociological Methods & Research. (2019) doi: 10.1177/0049124119826155
3. Martini, A.: L'effetto scuola (valore aggiunto) nelle prove Invalsi 2018. Tech. Rep., Invalsi (2018)

Simultaneous confidence regions and curvature measures in nonlinear models

Regioni di confidenza simultanee e misure di curvatura nei modelli nonlineari

Claudia Furlan and Cinzia Mortarino

Abstract Accuracy measures for parameter estimates require specific procedures for nonlinear models. For low parameter dimensions, routines for evaluating *approximate* simultaneous confidence regions (sCRs) are available in the most common software programs; however, the degree of accuracy may be very low. In this paper, we investigate on the relationship between coverage probability and degree of nonlinearity for approximate and bootstrap sCRs, using simulation results. Given our results, we provide a rule based on the value of the maximum parameter effects curvature to decide whether relying on the approximate sCR or on the bootstrap sCR, even if the latter usually gives coverage probability slightly lower than the nominal level.

Abstract *L'analisi dei modelli nonlineari prevede procedure specifiche per determinare l'accuratezza delle stime dei parametri. Per i casi in cui la numerosità parametrica è limitata, sono disponibili nei più comuni software routine di calcolo della regione di confidenza simultanea approssimata. Tuttavia la sua accuratezza può essere molto bassa. In questo lavoro, sulla base di simulazioni, esaminiamo la relazione fra la probabilità di copertura e il grado di nonlinearità per la regione approssimata e per la regione bootstrap. Anche se quest'ultima ha probabilità di copertura leggermente inferiori al livello nominale, a partire dai risultati ottenuti, proponiamo una regola per decidere quale delle due regioni utilizzare sulla base della curvatura dovuta alla parametrizzazione.*

Key words: Estimation accuracy, Bootstrap, Bass Model

Claudia Furlan
Department of Statistical Sciences, Padova, Italy, e-mail: furlan@stat.unipd.it

Cinzia Mortarino
Department of Statistical Sciences, Padova, Italy, e-mail: mortarino@stat.unipd.it

1 Introduction

Nonlinear models are an essential tool to describe many real-world phenomena. Unlike linear models, accuracy measures for parameter estimates, such as confidence intervals or confidence regions, may represent a difficult task due to nonlinearity. The issue of constructing a simultaneous confidence region (sCR) was mostly developed in the 1960s-1980s, but this research was limited by the computational difficulties of that time, especially for exact regions, [7]. A few approximations have been derived [8], given the complexity of obtaining an exact sCR. For instance, the so-called *approximate* sCR, $\mathfrak{I}(\Theta)$, is derived by approximating the nonlinear model via a linear Taylor expansion of the first order, taking advantage of the asymptotic normal distribution of the estimator. Thus, the approximate sCR's confidence levels are valid asymptotically. The approximate sCR is computationally attractive, since it corresponds to hyperellipsoids, and its use is facilitated by default routines available in the most common software programs, such as R or Mathematica, at least for low parameter dimensions.

Recently, Furlan and Mortarino [6] compared two exact sCRs and the approximate sCR with a parametric and a non parametric bootstrap sCR, for the specific class of nonlinear diffusion models, via simulation studies and real data. The authors focused on two of the most widespread S-shaped diffusion models for the lifecycle of products: the Bass model (BM) and the Generalized Bass model (GBM), with three and six parameters, respectively [1, 2], fitted to cumulative data. The GBM is defined by:

$$y(t) = m\frac{1-e^{-(p+q)\int_0^t w(\tau)d\tau}}{1+\frac{q}{p}e^{-(p+q)\int_0^t w(\tau)d\tau}} + \varepsilon(t), \tag{1}$$

where $y(t)$ are the cumulative sales of a product at time t, m is the market potential, p is the innovation coefficient, q is the imitation coefficient, and $w(t)$ can be any integrable function (in [6], $w(t)$ is a function of 3 parameters). The BM corresponds to Eq. (1) with $w(t) = 1$.

Since new observations at the end of the finite lifecycle do not provide useful information, some parameters of the BM cannot be estimated consistently and asymptotic normality does not apply [5]. However, estimation variance decreases with increasing lifecycle stage, that is, when observations cover a longer portion of the finite lifecycle. Furlan and Mortarino [6], then, examined situations observed at different lifecycle stages and, for the approximate sCR, found that the coverage probability is seriously affected by the parameter dimension, the lifecycle stage and the error variability. The parametric bootstrap sCR, $\mathfrak{B}_p(\Theta)$, proposed in [6], conversely reaches coverage probability values less affected by lifecycle stage, parameter dimension, and error variability. The coverage level of $\mathfrak{B}_p(\Theta)$, however, is systematically lower than the nominal level.

In assessing the appropriateness of the linear approximation, what is important is the overall degree of nonlinearity, irrespective of the causes contributing to it (i.e., parameter dimension, error variability, and lifecycle stage for nonlinear diffusion models). In this work, we make a further step in the assessment of the applicability

of the approximate sCR, by measuring the overall degree of nonlinearity of situations studied in [6]. Moreover, we explore the relationship between coverage probability and the degree of nonlinearity. At the same time, we explore the reliability of $\mathfrak{B}_p(\Theta)$, especially for situations when $\mathfrak{I}(\Theta)$ cannot be trusted.

2 Intrinsic and parameter-effects curvature

Nonlinearity is based on the notion of curvature and is of two different types: the intrinsic nonlinearity, which is due to the bending and twisting of the expectation surface; and the parameter-effects nonlinearity, which is due to the parametrization [4, 8]. Bates and Watts [3] proposed some measures of nonlinearity based on the notion of curvature, which indicate the adequacy of the linear approximation, and are independent of scale changes in both the data and the parameters.

To use the approximate sCR, it is necessary that both planar and uniform coordinate assumptions are satisfied, that is if "expectation surface is sufficiently flat to be replaced by the tangent plane" and if "straight, parallel equispaced lines in the parameter space map into nearly straight, parallel equispaced lines on the expectation surface", respectively [4]. It means that both the intrinsic and the parameter-effects nonlinearity need to be small.

In this work we use, as measures of nonlinearity, the two indicators γ^N_{max} and γ^T_{max} proposed by Bates and Watts [3], which are the maximum intrinsic and parameter-effects curvature, respectively, over the direction h in the parameter set (see Seber and Wild [8] for details). To determine the impact of nonlinearity upon a sCR, the two maximum curvature measures γ^N_{max} and γ^T_{max} must be compared to $1/\sqrt{F(k,n-k;1-\alpha)}$, which can be seen as the radius of the curvature of the 100(1-α)% confidence region. In the rest of the paper, we will omit the degrees of freedom and confidence level, denoting this threshold simply as $1/\sqrt{F}$. If the ratio $\gamma^T_{max}/\sqrt{F}$ is small compared to 1, than the expectation surface is sufficiently flat over the confidence region, and if the ratio $\gamma^N_{max}/\sqrt{F}$ is small compared to 1, the uniform coordinate assumption is good. If both assumptions are satisfied, then the approximate sCR can be substantially trusted.

Bates and Watts [3] found that the parameter-effects curvature is much higher than the intrinsic curvature. In their study, the author found that in 13 out of 16 examples, the maximum of the intrinsic curvature is fine and the ratio $\gamma^N_{max}/\sqrt{F}$ is even smaller than 0.5, while in 13 examples the maximum of the parameter-effects curvature, γ^T_{max}, is unacceptable.

3 Simulation results

In [6], three real cases were considered, different in terms of both the lifecycle stage and pattern, with a moderate sample size. The real cases were represented by the annual sales (in millions of units) of music cassettes in the United States from 1973 to 2005, the Austrian thermal solar capacity, in MW_{th}, with annual data from 1982 to

Table 1 Simulation settings: name of the setting, model used, parameter dimension (k), simulation length (n), lifecycle stage (%) (see [6], for further details).

Setting	Model	k	n	lifecycle stage (%)
Gas.2	restricted BM	2	46	61
Sol.2	restricted BM	2	27	54
Tgas.2	restricted BM	2	35	52
Cass.3	BM	3	33	100
Gas.3	BM	3	46	61
Sol.3	BM	3	27	54
Tgas.3	BM	3	35	52
Sol.6	GBM	6	27	68
Gas.6	GBM	6	46	58
Tgas.6	GBM	6	35	44

2008, and the Algerian natural gas production, in billion cubic metres (BCM), with annual data from 1970 to 2015. The latter series was also analyzed with truncated data until 2004. Furlan and Mortarino [6] investigated the coverage probability of the sCRs, with $1-\alpha = 0.95$ in 10 settings, by simulating from the restricted BM[1], BM, and GBM, and using the estimates found in the real cases as true values for the parameters (see Table 1, for the list of the settings analyzed). Using the three nested models allows to explore different parameter sizes, besides different sample sizes and different lifecycle lenghts. For each setting, $N = 1,000$ time series have been generated, and independent errors ε_t, $t = 1,2,\ldots,n$, have been sampled from a normal distribution with a zero mean and variance σ^2. $B = 1,000$ replications are used for the parametric bootstrap. For analyzing the effect of the variability of the data around the mean trajectory, different values of the error variance σ^2 were used. Since the market potential is different across the settings, $(\sigma/m)^2 \times 10^6$ represents a comparable measure of variability: it ranges from 0.1 to 150, allowing the processes to vary from essentially the trajectory of the model, to an extremely distressed pattern[2].

The results in [6] show that $\mathfrak{I}(\Theta)$, although it is the most used sCR, is often unreliable, but the degree of inaccuracy cannot be directly linked to a specific parameter size or lifecycle stage. In other words, it is difficult to suggest a strategy to decide whether $\mathfrak{I}(\Theta)$ can be trusted. For this reason, in this paper we try to examine the relationship between accuracy (measured by coverage probability) and maximum curvature measures.

With regard to the maximum parameter-effects curvature, Figure 1 shows the box-plots of the ratios $\gamma^T_{max}/\sqrt{F}$ for the simulated datasets[3] of the four settings with $k = 3$, for the different values of $(\sigma/m)^2 \times 10^6$. Values above 1 (red line) represent situations for which the parameter-effects curvature value advises against using

[1] The restricted BM corresponds to the BM with m fixed (parameter dimension: $k = 2$).

[2] The 11 values used for $(\sigma/m)^2 \times 10^6$ were: $0.1, 0.5, 1, 5, 7.5, 10, 12.5, 15, 20, 100, 150$.

[3] For space reasons, we cannot present here the results for $k = 2$ and $k = 6$.

$\mathfrak{I}(\Theta)$. We notice that curvature is related to lifecycle stage: the four settings are ordered according to this quantity, with decreasing levels of lifecycle stage. We also underline, that, given lifecycle stage, increasing variance corresponds to increasing curvature, but the variance level at which curvature exceeds 1 changes for different settings. Notice that we do not discuss here results for γ^N_{max}, because the intrinsic curvature came out to be very small in our simulations (the largest value for $\gamma^N_{max}/\sqrt{F}$ was equal to 0.131), confirming the outcomes obtained in [3] with real cases.

Let us examine, now, the relationship between coverage probability for $\mathfrak{I}(\Theta)$ (as assessed by simulations in [6]) and maximum parameter-effects curvature, by using, for each setting, the median of the N ratios $\gamma^T_{max}/\sqrt{F}$ (Fig. 2). In the plot, the two vertical lines correspond to thresholds 0.5 and 1. We observe that the coverage probability is essentially set on the nominal level when the ratio $\gamma^T_{max}/\sqrt{F}$ is smaller than 0.5, irrespectively for lifecycle stage or variance level. Conversely, curvature ratios between 0.5 and 1 lead to coverage probabilities always below the nominal level, but to a different extent according to different settings. For values above 1, we observe that the decrease of coverage probability is almost linear with the maximum parameter-effects curvature for all the settings. This turns out to produce a simple rule to predict coverage probability of $\mathfrak{I}(\Theta)$, given evaluated $\gamma^T_{max}/\sqrt{F}$.

Fig. 3 shows the same kind of plot for $\mathfrak{B}_p(\Theta)$ coverage probability. The relationship between coverage probability and the maximum parameter-effects curvature is here much less definite than we observed for $\mathfrak{I}(\Theta)$. Coverage probabilities are almost always below the nominal level, but we do not observe large discrepancies that, conversely, affected $\mathfrak{I}(\Theta)$. Moreover, Fig. 3 suggests that a poor performance for $\mathfrak{B}_p(\Theta)$ does not correspond to higher nonlinearity measures, making bootstrap a good alternative to $\mathfrak{I}(\Theta)$ whenever the maximum parameter-effects curvature exceeds $1/\sqrt{F}$.

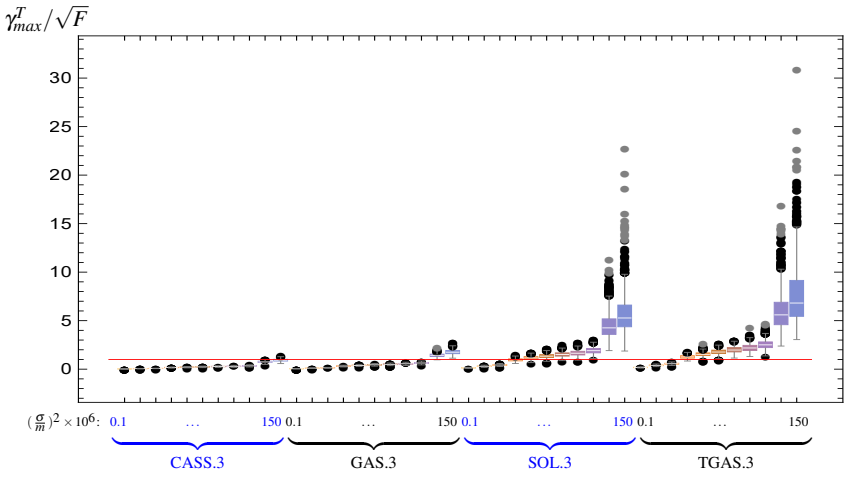


Fig. 1 Box-plots of the ratios $\gamma^T_{max}/\sqrt{F}$ for the 1000 simulated datasets of the four settings with $k = 3$, for the 11 different values of $(\sigma/m)^2 \times 10^6$. The red line corresponds to 1.

References

1. Bass, F.M.: A new product growth for model consumer durables. Manag. Sci. **15**(5), 215–227 (1969)
2. Bass, F.M., Krishnan, T.V., Jain, D.C.: Why the Bass model fits without decision variables. Mark. Sci. **13**(3), 203–223 (1994)
3. Bates, D.M., Watts, D.G.: Relative curvature measures of nonlinearity. J. R. Stat. Soc. Ser. B (Methodol) **42**, 1–25 (1980)
4. Bates, D., Watts, G.: Nonlinear Regression Analysis and Its Applications. John Wiley and Sons, New York, NY (1988)
5. Boswijk, H.P., Franses, P.H.: On the econometrics of the Bass diffusion model. J. Bus. Econ. Stat. **23**(3), 255–268 (2005)
6. Furlan, C., Mortarino, C.: Comparison among simultaneous confidence regions for nonlinear diffusion models. Comput. Stat. (2020) doi: 10.1007/s00180-019-00949-0
7. Lee, A., Nyangoma, S., Seber G.: Confidence regions for multinomial parameters. Comput. Stat. Data Anal. **39**(3), 329–342 (2002)
8. Seber, G., Wild, C.: Nonlinear Regression. Wiley, New York (1989)

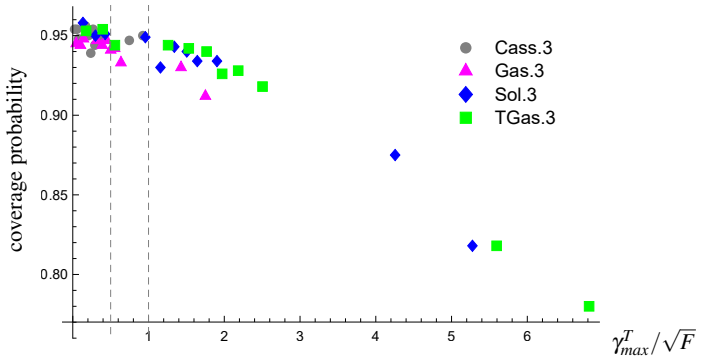


Fig. 2 $\mathfrak{I}(\Theta)$: coverage probability and median of the ratios $\gamma^T_{max}/\sqrt{F}$, for settings with $k = 3$ at different variance levels. The 2 vertical dashed lines correspond to 0.5 and 1.

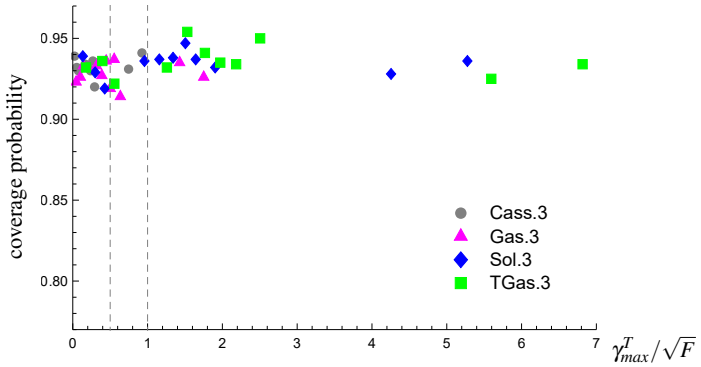


Fig. 3 $\mathfrak{B}_p(\Theta)$: coverage probability and median of the ratios $\gamma^T_{max}/\sqrt{F}$, for settings with $k = 3$ at different variance levels. The 2 vertical dashed lines correspond to 0.5 and 1.

Models and methods – Sampling

Design-based consistency of the Horvitz-Thompson estimator for spatial populations

Consistenza da disegno dello stimatore di Horvitz-Thompson per popolazioni spaziali

L. Fattorini, M. Marcheselli, C. Pisani and L.Pratelli

Abstract Spatial populations are usually located on a continuous support and can be distinguished into surfaces giving the value of the survey variable at any point of the support, finite collections of units, scattered onto the support, with attached values of the survey variables, or finite collections of areal units partitioning the support where the survey variable is the total amount of an attribute within each areal unit. Fattorini et al. (2020) prove design-consistency of the Horvitz-Thompson estimator of totals by focusing on the properties of the sampling design sequences under the minimal requirement that the survey variable is bounded.

Abstract *Le popolazioni spaziali sono di solito situate su un supporto continuo e si possono distinguere in popolazioni continue, identificabili con una superficie che rappresenta in ogni punto il valore della variabile di interesse, insiemi finiti di unità dislocate sul supporto, a ciascuna delle quali è associato un valore della variabile, o insiemi finiti di unità areali che costituiscono una partizione del supporto, in cui la variabile di interesse è il totale di un attributo all'interno di ciascuna unità areale. Fattorini et al. (2020) dimostrano la consistenza basata sul disegno dello stimatore di Horvitz-Thompson del totale sulla base delle caratteristiche della successione di disegni di campionamento sotto l'ipotesi minimale che la variabile di interesse sia limitata.*

Lorenzo Fattorini
Department of Economics and Statistics, University of Siena, P.zza S. Francesco 8, Siena, e-mail: lorenzo.fattorini@unisi.it

Marzia Marcheselli
Department of Economics and Statistics, University of Siena, P.zza S. Francesco 8, Siena, e-mail: marzia.marcheselli@unisi.it

Caterina Pisani
Department of Economics and Statistics, University of Siena, P.zza S. Francesco 8, Siena, e-mail: caterina.pisani@unisi.it

Luca Pratelli
Naval Academy, Viale Italia, 72, Livorno, e-mail: luca_pratelli@marina.difesa.it

Key words: continuous populations, finite populations, Horvitz-Thompson estimation

1 Introduction

In the framework of design-based inference, spatial populations are constituted by fixed sets of locations on a continuous support with fixed values of the survey variable attached to each location. In particular, spatial populations can be distinguished into continuous populations, finite populations of units and finite populations of areal units. Continuous populations arise when a spatial phenomenon on a continuous support is conceptualized as a population constituted by a continuous set of locations with a surface giving the value of the survey variable at each location. Finite populations of units are constituted by a finite set of regularly or irregularly units scattered over the support with the value of the survey variable attached to each unit, usually referred to as the mark of the unit. Finally, partitioning the support into spatial subsets gives rise to a finite population of areal units and the survey variable turns out to be the total amount of an attribute of interest within each areal unit, where the total can refer either to a continuous or a finite population of units scattered over the support.
Design-based inference has been widely adopted for estimating spatial populations totals (e.g. Thompson, 2002). When a finite population of units or areal units is considered, unbiased estimation is performed by using the Horvitz-Thompson (HT) estimator (e.g. Särndal et al., 1992) and when continuous populations are of interest it is usually performed extending the HT criterion (e.g. Cordy, 1993).
In this paper we present and discuss the results by Fattorini et al. (2020), which, in the spirit of Isaki and Fuller (1982), give conditions for the design-based consistency of the HT estimators which involve the characteristics of the design sequence and hold under minimal and realistic assumptions regarding populations.

2 Consistency results

2.1 Consistency for continuous populations

Let y be a Borelian and bounded function on the compact support $\mathscr{A}$ with values on $[0,L]$, where $y(p)$ is the value of the survey variable Y at the location $p \in \mathscr{A}$. The population total is $T = \int_{\mathscr{A}} y(p)\lambda(dp)$, where λ is the Lebesgue measure on $\mathbf{R}^2$.

Following Cordy (1993), suppose a sequence of designs $\{d_k\}$, each of them selecting an increasing number n_k of points onto $\mathscr{A}$, say $P_{k,1},...,P_{k,n_k}$ in such a way that the n_k-tuple $(P_{k,1},\ldots,P_{k,n_k})$ is a random vector. Let $g_i^{(k)}$ be a version of the marginal probability density of $P_{k,i}$ with respect to λ and $g_{ih}^{(k)}$ be a ver-

sion of the marginal probability density of $(P_{k,i}, P_{k,h})$ with respect to $\lambda \otimes \lambda$, with $i \neq h = 1, ..., n_k$.

Denoting by $\pi_k(p) = \sum_{i=1}^{n_k} g_i^{(k)}(p)$ the inclusion function and by $\pi_k(p,q) = \sum_{i \neq h=1}^{n_k} g_{ih}^{(k)}(p,q)$ the pairwise inclusion function, if $\pi_k(p) > 0$ for each $p \in \mathscr{A}$, then the extension of the HT estimator

$$\widehat{T}_k = \sum_{i=1}^{n_k} \frac{y(P_{k,i})}{\pi_k(P_{k,i})}$$

is an unbiased estimator of T.

Fattorini et al. (2020) prove that if the design sequence is such that

$$\lim_{k \to \infty} \sup_p \frac{1}{\pi_k(p)} = 0$$

$$\lim_{k \to \infty} \sup_{p \neq q} \left\{ \frac{\pi_k(p,q)}{\pi_k(p)\pi_k(q)} - 1 \right\}^+ = 0$$

then $\lim_{k \to \infty} \operatorname{var}(\widehat{T}_k) = 0$ and $\widehat{T}_k$ converges in probability to T.

2.2 Consistency for finite populations of units

Following the approach by Isaki and Fuller (1992), let $\{U_k\}$ be a nested sequence of populations of units of increasing size N_k scattered throughout the compact support $\mathscr{A}$ and Y be a survey variable with values on $[0, L]$, in such a way that y_j is the mark of unit $j \in U_k$. It is at once apparent that the population sequence determines a corresponding sequence of totals $\{T_k\}$.

Supposing a sequence of designs $\{d_k\}$, each of them selecting a sample S_k from U_k of increasing size n_k, the HT estimator is given by

$$\widehat{T}_k = \sum_{j \in S_k} \frac{y_j}{\pi_j^{(k)}}$$

where $\pi_j^{(k)}$ denotes the first-order inclusion probability of the j-th unit.

Denoting by $\pi_{jh}^{(k)}$ the second-order inclusion probability of units j and h ($h > j \in U_k$), if the design sequence ensures

$$\lim_{k \to \infty} \max_{h > j} \left\{ \frac{\pi_{jh}^{(k)}}{\pi_j^{(k)} \pi_h^{(k)}} - 1 \right\}^+ = 0$$

and there exists $\pi_0 > 0$ such that $\min_j \pi_j^{(k)} \geq \pi_0$, then $\lim_{k \to \infty} \operatorname{var}(\widehat{T}_k / T_k) = 0$ and $\widehat{T}_k / T_k$ converges in probability to 1.

2.3 Consistency for finite populations of areal units

Consider a sequence of partitions $\{\mathscr{P}_k\}$ of the compact support $\mathscr{A}$ constituted by an increasing number M_k of areal units $\mathscr{A}_1^{(k)},\ldots,\mathscr{A}_{M_k}^{(k)}$ of decreasing size, so that $\lim_{k\to\infty} \max_l \lambda(\mathscr{A}_l^{(k)}) = 0$. Suppose a sequence of designs $\{d_k\}$, each of them selecting a sample Q_k of increasing size m_k from $\mathscr{P}_k$ and let $\pi_l^{(k)}$ and $\pi_{lh}^{(k)}$ for $h > l \in \mathscr{P}_k$ be the first- and second-order inclusion probabilities, respectively.

If a continuous population is considered and y is a Borelian function on $\mathscr{A}$, with values on $[0,L]$, let $T = \int_{\mathscr{A}} y(p)\lambda(dp)$ be the population total and let $T_l^{(k)} = \int_{\mathscr{A}_l^{(k)}} y(p)\lambda(dp)$ be the total amount of the survey variable within $\mathscr{A}_l^{(k)}$. Since $T = \sum_{l=1}^{M_k} T_l^{(k)}$ for each k, estimating the total of a continuous population can be approached as the total estimation in a finite population of units, whose list is always available. In this setting, the HT estimator is given by

$$\widehat{T}_k = \sum_{l\in Q_k} \frac{T_l^{(k)}}{\pi_l^{(k)}}. \tag{1}$$

and if the design sequence is such that

$$\lim_{k\to\infty} \max_l \frac{\lambda(\mathscr{A}_l^{(k)})}{\pi_l^{(k)}} = 0 \tag{2}$$

and

$$\lim_{k\to\infty} \max_{h>l} \Big(\frac{\pi_{lh}^{(k)}}{\pi_l^{(k)}\pi_h^{(k)}} - 1\Big)^+ = 0 \tag{3}$$

then $\lim_{k\to\infty} \mathrm{var}(\widehat{T}_k) = 0$ and $\widehat{T}_k$ converges in probability to T. Note that a sufficient condition for (2) to hold is $\min_l \pi_l^{(k)} \geq \pi_0 > 0$ for any k.

On the other hand, if a finite population of units is scattered throughout $\mathscr{A}$, consistency cannot be proven taking the population fixed, but it is mandatory to consider a sequence of nested populations U_k of increasing size N_k. Let y_j be the value of the survey variable on unit $j \in U_k$, so that the total for k-th population is $T_k = \sum_{j\in U_k} y_j$.

Similarly to the continuous case, estimation of the total of finite populations of units can be switched into estimation in finite populations of areal units, whose lists are always available.

Let $T_l^{(k)} = \sum_{j\in U_l^{(k)}} y_j$ be the total amount of the survey variable within the areal unit $\mathscr{A}_l^{(k)}$, where $U_l^{(k)}$ denotes the sub-population of the $N_l^{(k)}$ units located within $\mathscr{A}_l^{(k)}$. As $T_k = \sum_{l\in\mathscr{P}_k} T_l^{(k)}$, the HT estimator $\widehat{T}_k$ is given by (1) .

For finite populations of units, if (3) holds and

$$\lim_{k\to\infty} \max_{l} \frac{T_l^{(k)}}{\pi_l^{(k)} T_k} = 0$$

then $\lim_k \mathrm{var}(\widehat{T}_k/T_k) = 0$ and $\widehat{T}_k/T_k$ converges in probability to 1.

There is a perfect analogy with the continuous case. In the case of finite populations of units, it suffices to replace $\lambda(\mathscr{A}_l^{(k)})$ with $T_l^{(k)}/T_k$. However, while in the continuous case only partitions are involved, in the discrete case it is necessary to presume a sort of eveness in the enlargements of the nested populations, in such a way that population units do not aggregate into some areal units.

3 Discussion

Consistency is obtained, under minimal assumptions regarding populations, by considering the design sequences, whose characteristics are related to the sampling scheme actually adopted to select the sample. Indeed, for continuous populations, it suffices to hold the support and the surface as fixed and simply considering a design sequence selecting an increasing number of sample points in the support.

For with-list finite populations of units scattered onto a support, the asymptotic framework by Isaki and Fuller (1982) is exploited taking the support fixed and considering a sequence of nested populations increasing within and a sequence of designs selecting sample of increasing size. On the other hand, when the list of the population units is not available, as frequently happens in environmental surveys, and hence it is necessary to sample them by points, eventually identifying plots or transects, the total estimator can be rephrased as a Monte Carlo estimator. In this case, the population can be held fixed and consistency can be achieved from the scheme adopted to locate an increasing number of points on the support.

Finally, when populations of areal units are considered, consistency is obtained both if a continuous population and a finite population of units are located on the support. In the first case, the support and the surface are held fixed and consistency is achieved from the scheme adopted to select areal units from a sequence of partitions constituted by an increasing number of areas of decreasing extents. In the second case, in addition, consistency requires the introduction of a sequence of nested populations and conditions on their enlargement.

Consistency can be easily proven for the simplest sampling schemes and for some widely applied schemes ensuring spatial balance, which is usually pursued in spatial surveys in order to obtain sample locations evenly spread over the support. As a matter of fact, for continuous populations consistency holds under uniform random sampling and tessellation stratified sampling, for with-list populations of units under simple random sampling without replacement and stratified sampling with proportional allocation and for population of areal units under simple random sampling without-replacement and one-per-stratum stratified sampling.

Finally, it is worth noting that spatial balance can be achieved by means of more complex schemes (see e.g. Grafström et al., 2012, Grafström and Tillé, 2013). Unfortunately, owing to their complexity, we cannot prove consistency for these schemes. However, owing to their effectiveness in providing spatial balance and their empirical performance (Fattorini et al., 2015), consistency presumably holds also for these schemes.

References

[1] Cordy, C. B., 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Stat. Probab. Lett.* 18, 353–362.

[2] Fattorini, L., Corona, P., Chirici, G., Pagliarella, M. C., 2015. Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use and land cover estimation. *Environmetrics* 26, 216–228.

[3] Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L., 2020. Design-based consistency of the Horvitz-Thompson estimator under spatial sampling with applications to environmental surveys *Spat. Stat.* 35, in press.

[4] Isaki, C. T., Fuller, W. A., 1982. Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.* 77, 89–96.

[5] Särndal, C. E., Swensson, B., Wretman, J., 1992. *Model Assisted Survey Sampling*, Springer, New York.

[6] Grafström, A., Lundström, N. L. P., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. *Biometrics* 68, 514–520.

[7] Grafström, A., Tillé, Y., 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 24, 120–131.

[8] Thompson, S. K., 2002. *Sampling*, 2nd ed., Wiley, New York.

Empirical likelihood in the statistical matching for informative samples

Verosimiglianza empirica nel matching statistico per campioni informativi

Daniela Marella and Danny Pfeffermann

Abstract Statistical matching aims at combining information available in distinct sample surveys referred to the same target population, but composed of nonoverlapping sets of units. The aim of the present paper is to analyze the statistical matching problem under informative sampling designs, when applying the empirical likelihood approach.

Abstract *Lo scopo del presente lavoro è analizzare il problema del matching statistico in caso di campioni informativi utilizzando l'approccio basato sulla verosimiglianza empirica*.

Key words: empirical likelihood, informative sampling.

1 Introduction

The statistical matching problem is becoming nowdays popular. In practice, information usually comes from different micro databases, each of them contains some of the variables of interest. The main target of statistical matching consists in the estimation of the joint distribution of variables separately observed in independent samples. Formally, the problem can be described as follows. Let A and B be two independent samples of sizes n_A and n_B respectively, selected from a

[1]Daniela Marella, Dipartimento di Scienze della Formazione, Università Roma Tre; email: daniela.marella@uniroma3.it.

Danny Pfeffermann, Central Bureau of Statistics and Hebrew University of Jerusalem, University of Southampton ; email: D.Pfeffermann@soton.ac.uk.

population of N identically distributed (*i.i.d.*) records, generated from some joint probability distribution function (*pdf*, for short). We assume that the population is large such that the samples A and B have no units in common. The statistical matching problem is that (X,Y,Z) are not completely observed in the two samples: only (X,Y) are observed for the units in sample A and only (X,Z) are observed for the units in sample B, see D'Orazio *et al.* (2006). Due to the lack of joint information on Z and Y given X, the joint distribution of (X,Y,Z) is not identifiable since the available sample information is not able to discriminate among a set of plausible models for (X,Y,Z) leading to uncertainty about the data generating model. See Conti *et al.* (2016) and reference therein for a discussion of the statistical matching uncertainty. Alternative techniques have been proposed in literature to overcome the identification problem: (i) techniques based on the conditional independence assumption between Y and Z given X (CIA); (ii) techniques using external auxiliary information regarding the statistical relationship between Y and Z.

The aim of the present paper is to analyze the statistical matching problem for the case where the sampling processes used to select the samples A and B are informative. The sample selection in survey sampling involves complex sampling designs based on different levels of clustering and differential inclusion probabilities. When the inclusion probabilities are related to the value of the target outcome variable even after conditioning on the model covariates, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population. This, quite common phenomenon is known as *informative sampling*, see Pfeffermann and Sverchkov (2009). Returning to statistical matching, knowledge of the sampling designs underlying the selection of the samples A and B and accounting for them is crucial for successful matching. The case of informative sampling designs in the statistical matching problem in a parametric setting is analysed in Marella and Pfeffermann (2019).

The paper is organized as follows. In Section 2 the empirical likelihood (EL) approach under informative sampling designs is discussed. Finally, in Section 3 a simulation study has been performed.

2 Statistical framework under informative sampling designs

In this section an empirical likelihood method based on the sample likelihood as proposed in Pfeffermann (2011) is developed to deal with the statistical matching problem. The empirical likelihood (Owen (1988)) is essentially the likelihood of the multinomial distribution used in Hartley and Rao (1968), where the parameters are the point masses assigned to the distinct sample values. The main advantages of empirical likelihood approach are: (i) it is much more flexible and easy to implement

since it does not require to specify the population model; (ii) it facilities the use of calibration constraints.

Associated with the *i*th population unit are values of (X,Y,Z), where X is a categorical variable taking K values with probability $P(X=k)=p_k^X$ (for $k=1,...,K$) while (Y,Z) are continuous random variables. We assume that the finite population values are generated from a multinomial distribution with parameter $p_i^{XYZ}=P(x_i,y_i,z_i)$. Under the CIA, p_i^{XYZ} can be factorized as follows

$$p_i^{XYZ}=P(x_i=k,y_i,z_i)=P(x_i=k)P(y_i\mid x_i)P(z_i\mid x_i)=p_k^X\,p_i^{Y|X}\,p_i^{Z|X} \qquad (1)$$

for $k=1,...,K$. It is assumed that X has its support both in sample A and in sample B, $p_i^{Y|X}$ has its support in sample A and $p_i^{Z|X}$ has its support in sample B. Suppose that the value k arises $n_{k,A}^X$ ($n_{k,B}^X$) in sample A (B), for $k=1,...,K$. If Z was observed in A, then following Pfeffermann (1998), the marginal *pdf* of (x_i,y_i,z_i) for $i\in A$ is defined as

$$p_{i,A}^{XYZ}=P(x_i=k,y_i,z_i\mid I_i^A=1)=p_{k,A}^X p_{i,A}^{Y|X}\,p_{i,A}^{Z|XY} \qquad (2)$$

where I_i^A is the sampling indicator taking the value 1 if the ith population unit is drawn to the sample A and 0 otherwise. Then, the sample distribution in A is also multinomial, with cell probabilities given by (2), where

$$p_{k,A}^X=\frac{\tau_{k,A}^X p_k^X}{\sum_{j=1}^K \tau_{j,A}^X p_j^X}\ ,\ p_{i,A}^{Y|X}=\frac{\tau_{i,A}^{XY} p_i^{Y|X}}{\sum_{i\in A_k}\tau_{i,A}^{XY} p_i^{Y|X}}\ ,\ p_{i,A}^{Z|XY}=\frac{\tau_{i,A}^{XYZ} p_i^{Z|X}}{\sum_{i\in A_k}\tau_{i,A}^{XYZ} p_i^{Z|X}} \qquad (3)$$

with $A_k=\{i\in A: x_i=k\}$ and $\tau_{i,A}^{XYZ}=E_p(I_i^A=1\mid x_i=k,y_i,z_i)$,

$\tau_{i,A}^{XY}=\sum_{i\in A_k}\tau_{i,A}^{XYZ}\,p_i^{Z|X}$, $\tau_{k,A}^X=\sum_{i\in A_k}\tau_{i,A}^{XY}\,p_i^{Y|X}$.

An expression analogous to (2) and (3) can be obtained for sample B. Under independence between observations corresponding to different sampling units (Pfeffermann (1998)), the *observed sample empirical likelihood* of the sample $A\cup B$ is given by

$$EL_{Obs}^{A\cup B}\propto\prod_{k=1}^K (p_{k,A}^X)^{n_{k,A}^X}\prod_{i\in A_k} p_{i,A}^{Y|X}\prod_{k=1}^K (p_{k,B}^X)^{n_{k,B}^X}\prod_{i\in B_k} p_{i,B}^{Z|X} \qquad (4)$$

and the EL estimators of the probabilities p_k^X, $p_i^{Y|X}$, $p_i^{Z|X}$ are obtained maximizing the likelihood (4) under the constraints

$$p_k^X \geq 0, p_i^{Y|X} \geq 0, p_i^{Z|X} \geq 0, \sum_{k=1}^{K} p_k^X = 1, \sum_{i \in A_k} p_i^{Y|X} = 1, \sum_{i \in B_k} p_i^{Z|X} = 1 \quad (5)$$

for $k = 1,...,K$. The solution is

$$\hat{p}_k^{X(A)} = [n_{k,A}^X (\tau_{k,A}^X)^{-1}] / \sum_{j=1}^{K} [n_{j,A}^X (\tau_{j,A}^X)^{-1}], \quad \hat{p}_k^{X(B)} = [n_{k,B}^X (\tau_{k,B}^X)^{-1}] / \sum_{j=1}^{K} [n_{j,B}^X (\tau_{j,B}^X)^{-1}],$$

$$\hat{p}_i^{Y|X} = (\tau_{i,A}^{XY})^{-1} / \sum_{j \in A_k} (\tau_{j,A}^{XY})^{-1}, \quad (6)$$

$$\hat{p}_i^{Z|X} = (\tau_{i,B}^{XZ})^{-1} / \sum_{j \in B_k}^{n_B} (\tau_{j,B}^{XZ})^{-1}$$

where $\hat{p}_k^{X(A)}$ and $\hat{p}_k^{X(B)}$ are the estimates of p_k^X obtained from sample A and B, respectively. Such estimates can be harmonized as $\hat{p}_k^X = \lambda \hat{p}_k^{X(A)} + (1-\lambda) \hat{p}_k^{X(B)}$ with $\lambda = n_A / (n_A + n_B)$. Sample estimates of the expectations $\tau_{i,A}^{XY}$, $\tau_{i,B}^{XZ}$ appearing in (6) are needed. The probabilities $\tau_{i,A}^{XY} = 1 / E_A(w_{i,A} \mid x_i, y_i)$ ($\tau_{i,B}^{XZ} = 1 / E_B(w_{i,B} \mid x_i, z_i)$) can be estimated by regressing the sample weights $w_{i,A}$ ($w_{i,B}$) on (x_i, y_i) ((x_i, z_i)) using the observed data in A (B). See Pfeffermann and Sverchkov (2009) for different approaches of modeling and estimating such expectations.

Once the parameters governing the multinomial population model have been estimated, a fused dataset can be constructed where each record includes measurements of (X, Y, Z), which users may treat as a "completely" observed dataset, with a similar distribution to the population distribution.

Finally, when population means of variables measured in the sample A and/or in sample B are known, constraints can be added to the maximization problem of sample empirical likelihood.

3 Simulation

In order to illustrate the effects of ignoring the sampling process in statistical matching and to assess the performance of the sample empirical likelihood approach, a simulation study has been performed according to the following steps:

Step 1. Generate a population of $N = 2000$ measurements (x_i, y_i, z_i) as follows: (i) X is a categorical variable taking the value $k = 10,11,12,13$ with probabilities $(p_1^X, p_2^X, p_3^X, p_4^X) = (0.2,0.3,0.4,0.1)$; (ii) $y_i \mid x_i$ is normal with parameters

$\theta_{Y|X} = (\beta_0 + \beta_1 x_i, \sigma^2_{Y|X})$; $\beta_0 = 2$, $\beta_1 = 1$, $\sigma^2_{Y|X} = 2$; (iii) $z_i \mid x_i$ is normal with parameters $\theta_{Z|X} = (\alpha_0 + \alpha_1 x_i, \sigma^2_{Z|X})$; $\alpha_0 = 1$, $\alpha_1 = 0.5$, $\sigma^2_{Z|X} = 1$.

Step 2. Draw samples A and B of size $n_A = n_B = 400$ from the population by Poisson sampling, with selection probabilities

$$\pi_{i,S} = n_S \frac{\exp(\gamma_X^S x_i + \gamma_Y^S y_i + \gamma_Z^S z_i)}{\sum_{j=1}^{N} \exp(\gamma_X^S x_j + \gamma_Y^S y_j + \gamma_Z^S z_j)}$$

where $\gamma^S = (\gamma_X^S, \gamma_Y^S, \gamma_Z^S)$ (for $S = A, B$) represent the sampling parameters.

Step 3. The population parameters p_k^X, $p_i^{Y|X}$, $p_i^{Z|X}$ are estimated: (i) by ignoring the sample selection effects; (ii) by maximizing the sample empirical likelihood (4). The expectations $E_A[w_{i,A} \mid x_i, y_i; \gamma^A]$ ($E_B[w_{i,B} \mid x_i, z_i; \gamma^B]$) are estimated by an exponential regression model of $w_{i,A}$ ($w_{i,B}$) against (x_i, y_i) ((x_i, z_i)).

Step 4. Repeat Step 2-3 500 times for different sampling parameters γ^S.

Figure 1 shows the population *pdf of* $Y \mid X = 10$ and the kernel density estimate of the sample *pdf* when ignoring the selection effects, for $\gamma^A = (0.5, 0.5, 0)$. As clearly seen, the sample *pdf* is very different from the population *pdf* due to the use of informative sampling .

In order to study the effect of ignoring informative sampling mechanisms we proceed comparing the estimates of the population parameters p_k^X, $p_i^{Y|X}$, $p_i^{Z|X}$. The results are shown in Table 1 where the mean of Kolmogorov distance (Kd) over the 500 samples between the X true population *pdf* and the X estimated *pdf*s corresponding to point (i) and (ii) in Step 3 is evaluated. Such distances are denoted by $Kd^{X,I}, Kd^X$, respectively. For $\gamma^A = \gamma^B = (0,0,0)$, Kd^X coincides with $Kd^{X,I}$ since the sampling processes acting in the two samples are not informative. For the remaining values of γ^A, γ^B $Kd^{X,I}$ is always larger than Kd^X. Then, ignoring the sample selection process affects negatively the quality of the estimates of p_k^X.

Analogous results have been obtained for $Y \mid X$ and $Z \mid X$ *pdf*s. Furthermore, if constraints are introduced in the EL maximization a reduction of Kolmogorov distances is obtained. For instance, for $\gamma^A = (0.5, 0.5, 0), \gamma^B = (0.5, 0, 0.5)$ if constraints regarding the knowledge of the population mean of X are introduced, that is

$$\sum_{i \in A} p_i^X x_i = \mu_X = \sum_{i \in B} p_i^X x_i$$

we obtain that $Kd^{X,I} = 0.12$ and $Kd^{X} = 0.02$.

Figure 1: Population *pdf* and Kernel density estimate of the sample *pdf* of $Y \mid X = 10$.

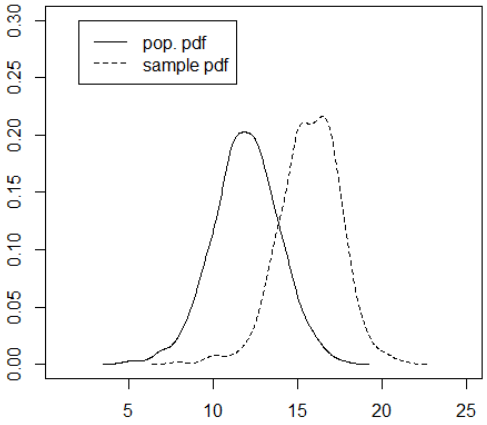


Table 1: *Kolmogorov distances* Kd^{X}, $Kd^{X,I}$ as γ^{A}, γ^{B} vary.

γ^{A}	γ^{B}	Kd^{X}	$Kd^{X,I}$
(0,0,0)	(0,0,0)	0.03	0.03
(0.5,0,0)	(0.5,0,0)	0.04	0.20
(0.5,0.5,0)	(0.5,0,0.5)	0.08	0.29
(0.5,0.5,0.5)	(0.5,0.5,0.5)	0.15	0.34

References

1. Conti, P.L., Marella, D., Scanu, M.: Statistical matching analysis for complex survey data with applications. J. Am. Stat. Assoc., 111, 516, 1715-1725 (2016).
2. D'Orazio, M., Di Zio, M. and Scanu, M.: Statistical Matching: Theory and Practice. Wiley, Chichester, (2006).
3. Hartley, H. O. and Rao, J. N. K.:A new estimation theory for sample surveys. Biometrika, 55, 547-557 (1968).
4. Marella, D. and Pfeffermann, D.:Matching Information from two independent informative samples. J. Stat. Plan. Inference, 203, 70-81.
5. Owen, A. B:. Empirical likelihood ratio confidence intervals for a single functional. Biometrika, 75, 237-249 (1988).
6. Pfeffermann, D., Krieger, A.M., Rinott, Y.: Parametric distribution of complex survey data under informative probability sampling. Stat. Sinica, 8, 1087-1114 (1998).
7. Pfeffermann, D., Sverchkov, M.: Inference under informative sampling. In Handbook of Statistics 29B; Sample Surveys: Inference and Analysis (Eds, D. Pfeffermann and C.R.Rao), Amsterdam: North Holland, 455-487 (2009).
8. Pfeffermann, D.: Modelling of complex survey data: Why model?Why is it a problem?How can we approach it? Survey Methodology,37,2, 115-136 (2011).

Evaluating a Hybrid One-Staged Snowball Sampling through Bootstrap Method on a Simulated Population

Campionamento a valanga ibrido ad uno stadio: una valutazione con metodo bootstrap su una popolazione simulata

Venera Tomaselli[1], Giulio Giacomo Cantone[2]

Abstract: Snowball sampling is a design where a number of individuals is surveyed in a population and then is requested to share the survey tool among their 'social links'. The aim is to recruit other people into the sample. One-stage recruitment means that individuals who are recruited by people already participant are not requested to recruit any others. Although surveys adopting snowball are financially less expensive than alternatives, the reliability of asymptotical estimates through this method is often questioned. A hybrid snowball sampling is designed such that a quota is randomly sampled, and another through snowball. Through a simulation of a demographic setup we found that bootstrap statistics of a one-staged hybrid design asymptotically show no significant difference to random design. This does not extend into situations where the random quota is smaller. We conclude that more complex setups will be needed before to generalise these results.

Abstract: *Il campionamento a valanga è un disegno campionario dove un sondaggio è sottoposto ad certo numero di individui, ed è poi chiesto a questi di condividere il sondaggio tra i loro 'legami sociali' al fine di reclutarne altri nel campione. Il reclutamento ad un livello prevede che chi è reclutato da chi è già partecipante non ne recluti altri. Sebbene i sondaggi a valanga siano meno costosi delle alternative, l'affidabilità delle stime di questo metodo è spesso posta in dubbio. Un campionamento a valanga ibrido è un disegno campionario dove una frazione del campione è estratta a sorte ed un'altra reclutata a valanga. Per mezzo di una simulazione di un setup demografico abbiamo trovato che le statistiche bootstrap di un disegno ibrido ad uno stadio non mostrano differenze significative rispetto all'alternativa casuale. Ciò non si estende ad una situazione con una minor frazione di individui estratti a caso. Riteniamo che setup più complessi siano necessari prima di generalizzare questi risultati.*

Key words: hybrid one-staged snowball sampling, bootstrapping, simulated population.

1 Introduction to snowball sampling

The terminology 'snowball sampling' denotes the methodology of social research where "a small [random] sample of persons [is interviewed], asking who their best friends are, interviewing these friends, then asking them their friends, interviewing these, and so on" (Coleman, 1958, 29). Leo Goodman (1961), who was a colleague of Coleman at University of Chicago, adopted "snowball sampling" to refer to a

[1] Venera Tomaselli, Department of Political and Social Sciences, University of Catania IT, e-mail: venera.tomaselli@unict.it (*corresponding author*).

[2] Giulio Giacomo Cantone, Department of Physics and Astronomy, University of Catania IT, e-mail: prgcan@gmail.com.

mathematical model to "make statistical inferences about various aspects of the relationships present in the population" by "data obtained using an s stage k name snowball sampling procedure" (*idem*, 148).

The aim of Goodman was to formalize conditions such to assume asymptotically unbiased estimates from Coleman's procedure. Goodman's model assumes that *k*, the amount of interviewed people per *s* stages of the chain, is a constant. "Stage 0" or $s = 0$ is the initial sample of *k* cases in the population and is random, instead. Snijders (1992) offered a different formal model with no fixed *k*. Both authors conclude that while a snowball procedure is not a fully randomized method, if 'homophily' among participants is sufficiently low, the sample may be assumed to be representative of a population.

Homophily is a terminology introduced since the '50 in sociological works and later formalized by network scientists as a parametrization of the correlation between the value of the variable in a node and the amount of its edges (Newman, 2010). Nevertheless, the interpretation of homophily is often ambiguous. Crawford, Aronow, Zeng et al. (2017) claim that homophily is often confused with "preferential recruitment" which is when the probability to be drawn in the sample from stage 1 cannot be assumed uniformly distributed.

When Goodman (2011) returned on the topic of snowball sampling, he highlighted an issue: while the model he developed assumed a random primary sample ('stage 0'), many studies performed after lacked this assumption[3]. However, Craig, Hays, Pickard et al. (2013) found that on 7 surveyed panel vendors, 6 had a mean of 20% of their proposed participants involved in at least one other of the 6. This leads to intuitively think that samples can support a certain 'quota' of biased cases within, without a great loss of precision in estimates within a multivariate design.

The research of Etter and Perneger (2000) is very noteworthy in this sense. Authors surveyed a random sample of 1000 residents in Geneva aged 18-70 (primary participants) in 1997. They asked every contacted subject, even those not willing to participate, to mail the questionnaire to all the smoker and ex-smoker residents in Geneva they knew (secondary participants). At the end of the data gathering process, 3,300 residents were mailed with the questionnaire and 1,167 individuals (35%) returned the questionnaire filled. Of these, 578 were primary participants and 566 were secondary participants. According to authors the mean age difference between the samples was only of 1.7 years (*p*-value=0.003). The only other significant difference was in sex ratio (7% difference, *p*-value=0.009) while behavioural traits were reported to be not significantly different in the two groups. As unintended consequence, the authors obtained two samples very similar in size within the one-staged snowball. We call this sampling design 'hybrid one-staged snowball sampling'.

In a study comparing a random statistical approach to a snowball-based approach to estimate morbidity and mortality of the rare disease visceral Leishmaniosis in two districts of the state of Bihar, India in 2011, Siddiqui, Rabidas, Sinha et al. (2016) concluded that snowball approach was found not sensitive enough to be adopted to estimate morbidity of the rare disease. At the same time, authors noticed that comparing costs, snowball approach required 1/6 of man-days and half the financial costs of the random alternative. Since full snowball design could not be suited for many applications, we want to evaluate if adoption of hybrid sampling can overcome losses of accuracy while reducing costs of research.

2 Simulation of a hybrid sampling

Testing procedures on empirical data in population studies raises the following issues:

- usually empirical research is designed for finding information over a phenomenon, not to provide data for evaluative research about methods
- we lack 'true values' on parameters of variables. We can test the hypothesis if two samples are drawn from the same population within confidence level (CL) but this says nothing about which design is better performing on estimation.

We propose instead a combination of two computational methods:

[3] According to the author, this was a result of historical label of 'snowball' mostly for the Coleman's procedure of data collection. Hence, the technique (also referred as 'chain-referral sampling') was associated to qualitative studies on 'rare population' or on 'hidden-populations' (Atkinson and Flint, 2001). Erikson (1979) and Snijders (1992) expressed concern about the actual possibility to randomly sample the stage 0 for hidden populations.

- a simulation model to procedurally generate a virtual dataset of population data from known parameters. Simulated populations can be modeled to fit pre-existing data with a small error (Alfons, Kraft, Templ et al., 2011; Burgard, Kolb, Murkle et al., 2017) or according with a theoretical model i.e. to construct hypotheses
- computational *bootstrap* sampling, falling under the more generic terminology of Monte Carlo methodologies (Gil, Montenegro, González-Rodríguez et al., 2006; Gobet, 2016).

We want to test performance of a sampling design against a standard. Hence, the simulated population does not need to fit empirical data. Still, this virtual population needs to be generated according to non-unrealistic assumptions over expected general outcomes. In particular, assumptions will take form of structural equations made through random variables. Technical sophistications in assumptions of the model may be added once the validity of the simplest models is established.

Bootstrap means that the model is then run many times, and each time estimates are recorded. Hypotheses are tested on the statistics (e.g. the so-called bootstrap-mean) of the distribution of values of the recorded estimates.

The general issues we found for the proposed simulation are two:

Issue 1: How can we simulate non-unrealistic variables in a virtual population?

In order to reflect the complexity of population studies without losing simplicity of linear structural equations, N simulated individual in the population (or 'agent' of the model) will be designed such to show few random variables:

- age: X_0; Gompertz and Weibull are established density functions to model age (Ricklefs and Scheuerlein, 2002). As the Weibull's function has only two parameters, *scale* (λ) and *slope* (k), we adopt it because it is a simpler assumption for our purposes[4].We propose:

$$X_0 \sim Weibull\left(\lambda_{X_0} = 50, k_{X_0} = 3\right). \quad [1]$$

  For ethical issues, data on minors are usually not collected; therefore, any agent with $X_0 < 18$ (~ 4,56% of N) will be removed from the virtual population.
- sex: X_1; a dummy variable with a probability equal to 0.5 per side:

$$p(X_1 = 1) = 0.5 \quad [2]$$

- behaviour: X_2; the research on behavioural metrics is a vast field with many insightful contributions on proper tools to measure behaviour's presence and extension (Kline, 2015). But we aim at keeping unambiguity and simplicity in assumptions before prompting actual 'field knowledge' into it. Therefore, we will just notice the presence/absence of the behaviour into a dummy variable with a probability density. For the proposal of the simulation, we arbitrarily set again 0.5 per side:

$$p(X_2 = 1) = 0.5 \quad [3]$$

- chronic condition: *Y*; the outcome of a structural equation where X_0 of the agent is the baseline, X_1 and X_2 are endogenous factors, β are weights, and ε is a continuous random error variable which sums all exogenous factors of *Y*:

$$Y = \begin{cases} 0, X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon < t_\zeta \\ 1, X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon > t_\zeta \end{cases} \quad [4]$$

t is an integer threshold value such that the probability of *Y*=0 in the whole population asymptotically tends to ζ. To intuitively understand the equation: if we keep out the agents when $X_0 < 18$, set β to be equals to 0 and ε is symmetrical with an expected value equal to 0, then, for ζ =0.5, the value 45 approximates $t_{\zeta=0.5}$. In simple terms, 45 will be the median age of the simulated population, given these parameters.

We can use this information to model our assumptions on β. If we assume:

- normal distribution of ε with μ=0
- age, sex, and behaviour having all equal impact on chronic condition
- $\beta_1 = \beta_2 = 90$

then $t_{\zeta=0.5} = 135$[5].

The threshold *t* can be set equal to 0 by the property of linear transformation of normal distributions (Bryc, 1995). Once $-(t_\zeta)$ is added to μ_ε, we notice that the lower the μ_ε, the higher the ζ. Therefore, we

[4] The Gompertz's assumptions about the parameters may be more suited for fitting synthetic data, instead.

[5] $\boldsymbol{E}(x_0) + \boldsymbol{E}(\beta_1 x_1) + \boldsymbol{E}(\beta_1 x_2) = t_{\zeta=0.5}$.

computationally find a setup pair of values for μ_ε and σ_ε such that the expected value of ζ is approximates a desired value. We want a model where the chronic condition afflicts 1 agent every 20 in the population: for a population of 100,000 agents, decreased to ~ 95445 agents after removal of agents aged < 18, under $\zeta \sim 0.95$, we expect ~ 4775 agents afflicted by the chronic condition. In the next setup, the E(Y) = 0.0505

$$Y = \begin{cases} 0, X_0 + 45(X_1) + 90(X_2) + \varepsilon < 0 \\ 1, X_0 + 45(X_1) + 90(X_2) + \varepsilon > 0 \end{cases} \quad [5]$$

$$\varepsilon \sim Gaussian(\mu_\varepsilon = -295, \sigma_\varepsilon = 90)$$

σ_ε was arbitrarily set such that $\sigma \sim 6(\sigma_{X_0})$.

Issue 2: How can we simulate recruitment in the hybrid sampling?

There are two different processes to simulate a network among individuals of our population. The first is to actually employ a software that connects all the members of the virtual population in a network. The 'primary' sample is randomly drawn, then agents in the primary sample recruits 1 stage of secondary participants among their links. Through this method, both homophily clusters and preferential recruitment are simulated. Unfortunately, this method is unpractical for a high-sized N of population as it requires intense computational resources.

Another approach to simulate hybrid sampling is less computationally intensive:

1. the primary sample ***I*** is drawn randomly and removed from the ***N*** population. Then two new variables are assigned to any *i* agent that is an element of ***I***[6]:
   - Links:
     $$Z_1 \sim Integer\left(ChiSquared\left(df_{Z_1} = 150.5\right)\right) \quad [6]$$
   - Recruitment:
     $$Z_2 \sim Integer\left(ChiSquared\left(df_{Z_2} = 5.5\right)\right)^{7} \quad [7]$$
2. every *i* agent in the first sample randomly draws z_1 other agents j_i from the ***N*** population; j_i agents are not removed from ***N*** – ***I*** but becomes elements of $\boldsymbol{J}_i$
3. for each *i* with a $z_2 > 0$ and all its j_i, a $d_{(i,j)}$ value is assigned. $d_{(i,j)}$ measures statistical distance between *i* and *j* expressed by the value of a structural equation

$$D_{(i,j)}(X_0, X_2): d_{(i,j)} = \left|x_{0_i} - x_{0_j}\right| + \varepsilon_d \quad [8]$$

$$\varepsilon_d \sim \left|\left(Gaussian(\mu_\varepsilon = 0, \sigma_\varepsilon \sim 0)\right) - \left(Gaussian(\mu_\varepsilon = 0, \sigma_\varepsilon \sim 0)\right)\right| \simeq 0^{8}.$$

For each *i*, z_2 elements of $\boldsymbol{J}_i$ are drawn with a probability equal to:

$$1 - \frac{d_{(i,j)}}{\sum d_{j_i}} \quad [9]$$

which is the function of preferential recruitment. In order to reduce endogenous bias introduced by how the structural equation of distance $D_{(i,j)}$ is modelled we perform a re-sample: a sample equal in size of ***I*** is randomly re-sampled into ***I'*** from the ∪(***J***) union of all the *j* agents drawn from each *i*. The union of ***I*** and ***I'*** is the actual 'hybrid sample'.

In this model only age has an impact on statistical distance. Impact of sex and other traits was already a topic in Coleman (1958), but since $\boldsymbol{E}(X_1)$ and $\boldsymbol{E}(X_2)$ are equal per side, any non-complex interaction among endogenous variables in $D_{(i,j)}$ would result in a more 'randomised' estimates in the final sample.

3 Results and conclusions

The simulation and all the procedures were performed with software *R*[9].

[6] This design reflects "layers of friendships" mentioned in Mac Carron, Kaski, and Dunbar (2016). The parameters are then chosen only as standards for the recruitment process.

[7] *Integer* is a function that subtracts to a value its mantissa. This operation was made to force integer values on the random variables. 0.5 was added to the parameters in order to compensate the loss of the mantissa. A different approach is to model the random variables from Poisson's function, or from an exponential one.

[8] While in this model the distance error ε_d representing exogenous factors of $D_{(i,j)}$ is assumed to be equal to 0 by forcing $\sigma_\varepsilon \sim 0$, we kept it in the equation the normal model that in our opinion best fit a random error in distance for future developments.

[9] Packages: base, survival, dplyr, rlist.

The population of N ~ 95445 was simulated only once. The random sample was drawn with a size equal to 1056 agents, which is the minimum size of representative random sample for confidence interval of 3% and a CL equal to 95%. Hybrid sample was drawn from the same population but with a size of ***I*** equal to 528, so the size of the final sample $\boldsymbol{I} \cup \boldsymbol{I'}$ is again 1056.

Both the design were run and recorded 300 times. We noticed that the average amount of ***J*** ~ 2380. Hence, we decided to perform a second hybrid sampling with ***I*** = 264 and ***I'*** = 792. Statistics are reported in table 1 and table 2.

Table 1. Population statistics.

	N	***X₀***		***X₁***		***X₂***		***Y***	
		μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
Simulated Population	95443	46.12	284.86	0.5	0.25	0.5	0.25	0.05	0.047

Table 2. Bootstrap statistics (n=300) of samples in random and hybrid sampling design.

	n	***X₀***		***X₁***		***X₂***		***Y***	
		Mean	***Var***	***Mean***	***Var***	***Mean***	***Var***	***Mean***	***Var***
Random Sampling	300	46.15	0.22	0.49	~0	0.49	~0	0.049	~0
Hybrid Sampling (stage 0 = 528)	300	45.95	0.21	0.5	~0	0.49	~0	0.05	~0
Hybrid Sampling (stage 0 = 264)	300	47.5	0.25	0.5	~0	0.49	~0	0.051	~0

We checked if the difference between the mean of population (46.12) and the bootstrap means of all the variables in the random and hybrid samplings (see Table 2) are significant with *p*-value < 0.05. We found:

- no significant difference in all variables in random design, as expected
- no significant differences in variables of hybrid design (stage 0 = 528)
- significant difference in X_0 and Y in hybrid design (stage 0 = 264).

For this reason, we exclude feasibility of hybrid sampling (stage 0 = 264). On the basis of the results from our experiment, we can assert that exists at least a setup of very simple parameters under which hybrid model can be employed instead of random sampling design.

We think the variances in bootstrap statistics are abysmal because complexity in population data is low and recruitment processes show a low variance (i.e. Chi Square functions are leptokurtic).

4 Future developments

We propose the following topics as future developments of the present study:

1. Complexity in the population's equations: in empirical demographic data, age is not independent from sex. Behaviours are complex: some may be dependent from age and sex while others may not. For future applicative applications, the target behaviour may be modelled as a structural equation. A dummy model may result simplistic, so the output may be rescaled into a multipoint scale, instead. Chronic conditions may be results of non-linear interactions among factors. The final model could be driven from a scientific theory. For internal validity of the model, the relevant parameter is the skewness of the variables in the population: if the variables are not symmetrical, even in a population with zero homophily, where $\boldsymbol{J}_i$ are randomly sampled, and if preferential recruitment is modelled such that agents in the ***I*** subsample 'prefer' to draw other agents that are not statistically distant from them, then we expect an increase in estimates' variance.
2. Complexity in the preferential recruitment: we suppose this point is both the most controversial and the most impactful on biases in asymptotical estimates. We already stated that in order to simulate homophily, we would need to simulate a network. We think a practical compromise could

be to import a network dataset and then test sampling through it, although this incurs in issue mentioned in section 2. We admit we don't know how people recruit other respondents into survey tool, neither we feel like we can generalise too much on this issue. Our general intuitions for future developments are: (i) Z_2 should be platykurtic and with a fat right tail, i.e. an exponential distribution. The fatter the right tail, the more the model is stressed; (ii) the weight in likelihood of recruitment should be positively correlated to a statistical distance.

3. Multi-stage and stage-free models: in order to stress more a reduction in size of 'stage 0' ***I***, more stages of recruitment can be added, so that, i.e. a two-staged hybrid sampling would sample the union of ***I, I', I''***, etc., where the latter stage is recursively not randomly drawn among 'friends' of the precedent stage. If a development of the model simulates a network population with complex variables, a more realistic approach may be free of stages: in other terms, every agent can recruit another, until the union of all sampled agents is equal to target *n*.
4. To evaluate with a regressive model the fitness of bootstrap statistics: in the present study, the validity of the sampling design was tested through significance of difference between bootstrap statistics and population parameters. A different evaluative approach is to estimate fitness of bootstrap statistics in a regressive model for each Y, already known as outcome of structural equation.

References

1. Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Stat Method Appl, 20(3), 383–407.
2. Atkinson, R. & Flint, J. (2001). Accessing Hidden and Hard-to-Reach Populations: Snowball Research Strategies. Social Research Update, 33.
3. Bryc, W. (1995). The Normal Distribution: Characterizations with Applications. Springer-Verlag.
4. Burgard, J. P., Kolb, J.-P., Merkle, H., & Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. AStA Wirtschafts- Und Sozialstatistisches Archiv, 11(3-4), 233–244.
5. Coleman, J. (1958). Relational Analysis: The Study of Social Organizations with Survey Methods. Hum Organ, 17(4), 28–36.
6. Craig, B. M., Hays, R. D., Pickard, S. A., Cella, D., Revicki, D. A., & Reeve, B. B. (2013). Comparison of US Panel Vendors for Online Surveys, J Med Inter Res, 15(11).
7. Crawford, F. W., Aronow, P. M., Zeng, L., & Li, J. (2017). Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling. Am J Epidemiol, 187(1), 153–160.
8. Erickson, B. H. (1979). Some Problems of Inference from Chain Data. Sociol Methodol, 10, 276.
9. Etter, J.F., & Perneger, T. V. (2000). Snowball sampling by mail: application to a survey of smokers in the general population. Int J Epidemiol, 29(1), 43-48.
10. Gil, M. Á., Montenegro, M., González-Rodríguez, G., Colubi, A., & Rosa Casals, M. (2006). Bootstrap approach to the multi-sample test of means with imprecise data. Comput Stat Data An, 51(1), 148–162.
11. Gobet, E. (2016). Monte-Carlo Methods and Stochastic Processes. New York: Chapman and Hall/CRC.
12. Goodman, L.A. (1961). Snowball sampling. Annals of Mathematical Statistics. 32(1), 148–170.
13. Goodman, L.A. (2011). Comment: On Respondent-Driven Sampling and Snowball Sampling in Hard-to-Reach Populations and Snowball Sampling Not in Hard-to-Reach Populations. Sociol Methodol, 41(1), 347–353.
14. Kline, P. (2015). A Handbook of Test Construction (Psychology Revivals). London: Routledge.
15. Mac Carron, P., Kaski, K., & Dunbar, R. (2016). Calling Dunbar's numbers. Soc Netw, 47, 151–155.
16. Newman M (2010). Networks: An Introduction. New York, NY: Oxford University Press.
17. Ricklefs, R. &Scheuerlein, A. (2002). Biological implications of the Weibull and Gompertz models of aging. J Gerontol A-Biol, 57(2), 69-76.
18. Snijders, T. A. B. (1992). Estimation on the Basis of Snowball Samples: How to Weight? BSM, 36(1), 59–70.
19. Siddiqui, N. A., Rabidas, V. N., Sinha S. K., Verma R. B., Pandey, K. P., Singh, V.P., Ranjan, A., Topno, R. K., Lal, C. S., Kumar, V., Sahoo, G.C., Sridhar, S., Pandey, A., Das, P. (2016) Snowball Vs. House-to-House Technique for Measuring Annual Incidence of Kala-azar in the Higher Endemic Blocks of Bihar, India: A Comparison. PLoS Neglected Tropical Diseases, 10(9).

How optimal subsampling depends on guessed parameter values

Sulla dipendenza dai valori nominali dei parametri nel campionamento ottimale

Laura Deldossi and Chiara Tommasi

Abstract For huge amounts of data, subsampling is useful to downside the data volume to get an inferential goal. We suggest to select a sample of observations applying the theory of optimal design. Assuming a relationship between a response variable and some covariates, the idea is to draw a sample from a large dataset so that it contains the majority of the information about the unknown model parameters. For nonlinear models this optimal selection method depends on the unknown parameters and proper values should be guessed. We analyze how the optimal sample depends on these nominal values in the logistic regression model.

Abstract *Quando si dispone di una ingente quantità di dati, ridurne la dimensione - estraendone un campione - può essere utile ai fini di una analisi inferenziale. Se si assume esista un modello che spieghi la relazione tra una variabile di risposta e alcune covariate, la selezione del campione può essere eseguita secondo la teoria del disegno ottimo, per ottenere un campione che preservi la maggior parte dell'informazione sui parametri incogniti del modello. Quando il modello è non lineare, il disegno ottimo dipende dai parametri ignoti ed è quindi necessario fissare dei valori nominali per poterlo utilizzare. In questo lavoro, assumendo un modello di regressione logistica, ci proponiamo di studiare quanto il metodo di selezione basato sul disegno ottimo sia condizionato dalla scelta dei valori nominali.*

Key words: Logistic model, Optimal sample selection, D-optimality, A-optimality

Laura Deldossi
Department of Statistical Science, Università Cattolica del Sacro Cuore, Milano
e-mail: laura.deldossi@unicatt.it

Chiara Tommasi
Department of Economics, Management and Quantitative Methods, University of Milan
e-mail: chiara.tommasi@unimi.it

1 Introduction

Big Dataset are a huge amount of data that are automatically accrued. Since they arise from observational studies, the quality of the Big Data information might not be very good. In addition, in large-sample studies, if the inferential goal is to test the effect of an explanatory variable, then the p-value often leads to the rejection of the null hypothesis. That is, even very small effects can become statistically significant because of the increased power due to the huge amount of data. From here, the idea of selecting a subsample of the Big Dataset to achieve an inferential goal. This topic has been already studied by [5], [3], [8], [9] and [1], among others.

To get a subsample of data, we apply the theory of optimal design instead of considering the most commonly used sampling schemes. Indeed, the connection between the sampling and experimental design had been already explored by [11], [12], [13], [4] and [6], among others.

In [2] we propose an optimal subsample selection strategy – which is called the "Optimal Design Based" (ODB) method - consisting of two steps. First, we identify the "most informative" values of the explanatory variables according to an optimality criterion (these optimal "theoretical" values are not necessarily present in the observed Big Dataset). Then, we select the observations from the full data set that are closer to these "theoretical" optimal values. Hence, this "optimal-sampling" approach enables us to select the most "informative" observations from the Big Dataset.

A selection strategy that is based on D-optimality and linear models is the Information-Based Optimal Subdata Selection (IBOSS) method that was proposed by [8]. The ODB method, unlike IBOSS, can be based on any optimality criterion (herein, we consider the D- and A-criteria) and can be applied also to nonlinear models.

Since in nonlinear models optimal designs depend on the unknown model parameters, herein we study the dependence of our selection procedure on nominal values of the parameters in a logistic regression model.

As a measure of the quality of the optimal sample based on a guessed nominal value, with respect to the ODB sample obtained from the true parameter value, we use the so called design efficiency, which is defined in Section 2.1.

The remainder of this paper is organized as follows. In Section 2 we briefly describe the Optimal Design Based sampling method and the related notation. In Section 3, we develop a simulation study to compare the ODB sample based on a nominal value with the ODB sample based on the true value of the parameter.

2 A sampling rule based on an optimality criterion

We assume that the relationship between the binary response Y and a set of p covariates $x_1, \ldots, x_p$ is a logistic regression model:

$$P(Y_i = 1; \mathbf{x}_i, \theta) = \frac{e^{\mathbf{x}_i^T \theta}}{1 + e^{\mathbf{x}_i^T \theta}}, \quad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ and $\theta = (\theta_1, \ldots, \theta_p)^T$. We are interested in estimating the parameter vector θ as precisely as possible. It is well known that the asymptotic covariance matrix of the maximum likelihood estimator is the inverse of the Fisher Information matrix

$$I(\theta) = \sum_{i=1}^{N} I(\mathbf{x}_i; \theta), \tag{2}$$

where the typical element

$$[I(\mathbf{x}_i; \theta)]_{r,s} = E_Y \left(-\frac{\partial^2 \log P(Y; \mathbf{x}_i, \theta)}{\partial \theta_r \partial \theta_s} \right).$$

When some values of the covariates are repeated then (2) can be written as

$$I(\theta) = N \sum_{i=1}^{k} I(\mathbf{x}_i; \theta) \omega_i,$$

where $\mathbf{x}_i$ are the distinct values of the covariates and ω_i are their frequencies, $i = 1, \ldots, k$.

$$\xi = \begin{Bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_k \\ \omega_1 & \cdots & \omega_k \end{Bmatrix}, \quad 0 \leq \omega_j \leq 1, \ \sum_{j=1}^{k} \omega_j = 1$$

is the so called approximate design and

$$I(\xi; \theta) = \sum_{i=1}^{k} I(\mathbf{x}_i; \theta) \omega_i \propto I(\theta)$$

is the information matrix of the design ξ. The available data are observational, but if it had been possible, the best choice for the values $\mathbf{x}_i$ and ω_i for $i = 1, \ldots, k$ would have been according to an optimality criterion $\Phi[I(\xi; \theta)]$ for precise estimation of θ.

Given a model and a concave optimality criterion, it is always possible to compute an optimum design as

$$\xi_\Phi^*(\theta) = \arg\max_{\xi} \Phi[I(\xi; \theta)] = \begin{Bmatrix} \mathbf{x_1^*} & \cdots & \mathbf{x_j^*} & \cdots & \mathbf{x_k^*} \\ \omega_1^* & \cdots & \omega_j^* & \cdots & \omega_k^* \end{Bmatrix}. \tag{3}$$

A wise "data generator" would have generated $N\omega_j^*$ responses at $\mathbf{x_j^*}$, $j = 1, \ldots, k$ and this would have been the "ideal" combination of covariates values to be applied to obtain a precise parameter estimation.

Given a large dataset, we aim at selecting a sample s of $n < N$ observations emulating the optimal design $\xi_\Phi^*(\theta)$. More specifically, since the Φ-optimum design (3) gives the best combination of covariate values for the precise estimation of θ, we suggest to select the $n\omega_j^*$ observations that are as closest as possible to

$\mathbf{x}_\mathbf{j}^*$ for $j = 1,\ldots,k$. As a measure of closeness, the Euclidean, Mahalanobis or any other distance can be applied. When $n\omega_j^*$ is not an integer number, then a suitable rounding-off rule can be applied (see for instance [7]). This selection method is called optimal design based (ODB) sampling scheme; for more detail see [2]. The ODB sample is denoted by s_θ because it depends on θ through the optimal design $\xi_\Phi^*(\theta)$.

2.1 A- and D-efficiencies

In this paper, we consider the well-known A- and D-optimality criteria defined as $\Phi_A[I(\xi;\theta)] = -\mathrm{Tr}\left[I(\xi;\theta)^{-1}\right]$ and $\Phi_D[I(\xi;\theta)] = |I(\xi;\theta)|^{1/p}$ respectively. The A-optimum design minimizes the total variation of θ as $\xi_A^*(\theta) = \arg\min_\xi \mathrm{Tr}\left[I(\xi;\theta)^{-1}\right] = \arg\max_\xi \Phi_A[I(\xi;\theta)]$. The D-optimum design minimizes the generalized variance of θ as $\xi_D^*(\theta) = \arg\min_\xi |I(\xi;\theta)|^{-1/p} = \arg\max_\xi \Phi_D[I(\xi;\theta)]$.

Let θ_0 denote the true value for the parameter vector θ. The "goodness" of any sample s with respect to the ODB sample s_{θ_0} can be measured through the following ratios

$$0 \le \mathrm{Eff}_A(s;\theta_0) = \frac{\mathrm{Tr}[I(s_{\theta_0};\theta_0)^{-1}]}{\mathrm{Tr}[I(s;\theta_0)^{-1}]} \le 1 \quad \text{and} \quad 0 \le \mathrm{Eff}_D(s;\theta_0) = \frac{|I(s_{\theta_0};\theta_0)|^{-1/p}}{|I(s;\theta_0)|^{-1/p}} \le 1$$

which are called A- and D-efficiencies, respectively.

3 Simulation experiments

This section concerns some results of a simulation study that aims at analyzing how the ODB sampling approach depends on the nominal value of the parameter vector. $N = 100000$ data have been generated from the logistic model (1) with $p = 3$ and $\theta = \theta_0 = (0.5, 0.5, 0.5)^T$. To find the ODB sample s_θ it is necessary to know the parameter value, to be able to compute $\xi_\Phi^*(\theta)$ in (3). In real-life problems, however, we do not know the true value θ_0 and a nominal value θ must be guessed.

To explore how the choice of nominal value influences the ODB selection rule, we considered θ such that $(\theta - \theta_0)$ takes values accordingly to a 3^3 factorial design. In other words, let θ_l be the l-th component of θ, $l = 1,2,3$, we fixed $\theta_l = \theta_{0l} + k$ with $k = -1, 0, 1$. In this way we are able to measure the loss in efficiency of s_θ with respect to s_{θ_0}, for guessed θ-values which differ from θ_0 in all directions.

In Table 1, A- and D-efficiencies of s_θ with respect to s_{θ_0} are given for different values of θ, when the covariates are generated from three independent $U(-1,1)$ random variables. Table 2 reports the same efficiencies when $(X_1, X_2, X_3) \sim N(\mathbf{0}, I_3)$.

Table 1 A- and D-efficiencies of s_θ with respect to s_{θ_0} with $\theta_0 = (0.5, 0.5, 0.5)^T$, when the covariates are generated from three independent $U(-1;1)$.

Guessed value θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$	Guessed value θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$
-0.5, -0.5, -0.5	1.00	1.00	+0.5, +0.5, +1.5	0.791	0.818
-0.5, -0.5, +0.5	0.92	0.919	+0.5, +1.5, -0.5	0.777	0.791
-0.5, -0.5, +1.5	0.752	0.743	+0.5, +1.5, +0.5	0.793	0.815
-0.5, +0.5, -0.5	0.939	0.925	+0.5, +1.5, +1.5	0.618	0.731
-0.5, +0.5, +0.5	0.925	0.921	+1.5, -0.5, -0.5	0.748	0.746
-0.5, +0.5, +1.5	0.772	0.785	+1.5, -0.5, +0.5	0.778	0.784
-0.5, +1.5, -0.5	0.748	0.742	+1.5, -0.5, +1.5	0.615	0.725
-0.5, +1.5, +0.5	0.775	0.784	+1.5, +0.5, -0.5	0.779	0.786
-0.5, +1.5, +1.5	0.622	0.722	+1.5, +0.5, +0.5	0.792	0.816
+0.5, -0.5, -0.5	0.938	0.916	+1.5, +0.5, +1.5	0.612	0.730
+0.5, -0.5, +0.5	0.937	0.924	+1.5, +1.5, -0.5	0.621	0.725
+0.5, -0.5, +1.5	0.775	0.793	+1.5, +1.5, +0.5	0.618	0.730
+0.5, +0.5, -0.5	0.936	0.919	+1.5, +1.5, +1.5	0.603	0.731

Table 2 A- and D-efficiencies of s_θ with respect to s_{θ_0} with $\theta_0 = (0.5, 0.5, 0.5)^T$, when the covariates are generated from three independent $N(0,1)$.

θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$	θ	$\text{Eff}_A(s_\theta;\theta_0)$	$\text{Eff}_D(s_\theta;\theta_0)$
-0.5, -0.5, -0.5	1.00	1.00	+0.5, +0.5, +1.5	0.984	0.952
0.5, -0.5, +0.5	1.00	1.00	+0.5, +1.5, -0.5	0.970	0.949
-0.5, -0.5, +1.5	0.977	0.952	+0.5, +1.5, +0.5	0.942	0.926
-0.5, +0.5, -0.5	1.00	1.00	+0.5, +1.5, +1.5	0.875	0.867
-0.5, +0.5, +0.5	1.00	1.00	+1.5, -0.5, -0.5	0.968	0.946
-0.5, +0.5, +1.5	0.967	0.940	+1.5, -0.5, +0.5	0.996	0.955
-0.5, +1.5, -0.5	0.952	0.932	+1.5, -0.5, +1.5	0.921	0.871
-0.5, +1.5, +0.5	0.936	0.931	+1.5, +0.5, -0.5	0.994	0.946
-0.5, +1.5, +1.5	0.889	0.822	+1.5, +0.5, +0.5	0.951	0.952
+0.5, -0.5, -0.5	1.00	1.00	+1.5, +0.5, +1.5	0.894	0.873
+0.5, -0.5, +0.5	1.00	1.00	+1.5, +1.5, -0.5	0.911	0.876
+0.5, -0.5, +1.5	0.975	0.941	+1.5, +1.5, +0.5	0.868	0.850
+0.5, +0.5, -0.5	1.00	0.996	+1.5, +1.5, +1.5	0.698	0.861

3.1 Comments

Comparing the results in Tables 1 and 2, we can observe that the loss in efficiency due to a wrong nominal value of the parameter is always less than 0.4.

Furthermore, it seems that the ODB selection rule is less influenced by the choice of the nominal value when the covariates follow a Gaussian distribution.

However, the "true" ODB sample s_{θ_0} is less informative when the covariates are generated by a normal distribution than when they come from a uniform random variable. This becomes clear if we compare the efficiency of s_{θ_0} with respect to the optimal design ξ^*_Φ (which does not depend on the distribution of the covariates). In the uniform case,

$$\frac{\mathrm{Tr}[I(\xi_A^*;\theta_0)^{-1}]}{\mathrm{Tr}[I(s_{\theta_0};\theta_0)^{-1}]} = 0.925 \qquad \frac{|I(\xi_D^*;\theta_0)|^{-1/3}}{|I(s_{\theta_0};\theta_0)|^{-1/3}} = 0.923$$

while in the Gaussian case,

$$\frac{\mathrm{Tr}[I(\xi_A^*;\theta_0)^{-1}]}{\mathrm{Tr}[I(s_{\theta_0};\theta_0)^{-1}]} = 0.308 \qquad \frac{|I(\xi_D^*;\theta_0)|^{-1/3}}{|I(s_{\theta_0};\theta_0)|^{-1/3}} = 0.362.$$

This means that the per-unit information contained in s_{θ_0} is larger when data come from the uniform distribution; see [2] for a detailed motivation.

In conclusion, when data follow a Gaussian distribution the ODB selection method seems not depending strongly on the nominal value of the parameters, even if the ODB sub-samples are less informative than in the uniform case.

References

1. Campbell, T. and Broderick, T.: Automated scalable Bayesian inference via Hilbert coresets. J. Mach. Learn. Res., **20**, 1–38 (2019)
2. Deldossi, L. and Tommasi, C.: Big Data and model-based survey sampling. http://arxiv.org/abs/2002.04255
3. Drovandi, C.C., Holmes, C., McGree, J.M., Mengersen, K., Richardson, S., Ryan, E.G.: Principles of Experimental Design for Big Data Analysis. Stat. Sci. **32**(3), 385-404 (2017)
4. Fedorov V.V.: Optimal design with bounded density: optimization algorithms of the exchange type. J. Statist. Plann. Inference, **22**, 1-13 (1989)
5. Ma, P. and Sun, X.: Leveraging for Big Data Regression. Comput. Statist.**7**(1), 70–76 (2015)
6. Pronzato L.: On the sequential constructions of optimum bounded designs. J. Statist. Plann. Inference, **136**, 2783-2804 (2006)
7. Pulkesheim, F. and Rieder, S.: Efficient rounding of approximate designs, Biometrika, **79**(4) 763–770 (1992)
8. Wang H. and Yang M. and Stufken J.: Information-Based Optimal Subdata Selection for Big Data Linear Regression. J. Amer. Statist. Assoc., **114**(525), 393-405 (2019)
9. Wang H. and Zhu R. and Ma P.: Optimal Subsampling for Large Sample Logistic Regression. J. Amer. Statist. Assoc., **113**(522), 829-844 (2018)
10. Wynn H.P.: Results in the theory and construction of D-optimum experimental designs, J. of the Royal Stat. Soc. Ser. B, 34, 133–147 (1972)
11. Wynn H.P.: Minimax Purposive Survey Sampling Design. J. Amer. Statist. Assoc., **72**(359), 655-657 (1977)
12. Wynn H.P.: Optimum designs for finite populations sampling. In: Gupta, S., Moore, D. S. (Eds.), Statistical Decision Theory and Related Topics II, 471-478 (1977)
13. Wynn H.P.: Optimum submeasures with applications to finite population. in Gupta, S., Berger, J. (Eds.), Statistical Decision Theory and Related Topics III. Proc. 3rd Purdue Symp., **2**, 485-495 (1982)

Indicators for risk of selection bias in non-probability samples

Indicatori per valutare il rischio di distorsione in presenza di campioni non-probabilistici

Emilia Rocco and Alessandra Petrucci

Abstract In recent years, in surveys, is increased the relevance of the non-probability sampling with which the units included in the sample are determined by an unknown self selection process. The self-selection of the units in the sample raises concerns about the potential for biased estimates of population characteristics and consequently it originates the need of indicators for assessing the degree of bias and for identifying potential adjustment/weigting methods for reducing it. Here, we are motivated by a real problem: the risk of bias in a non-probability web survey on the customers' satisfaction with the services provided by the municipal library system in Florence. We suggest and discuss the combined use of three measures, each one for a specific aspect of the data that impacts on the risk of biased estimates.

Abstract *Negli ultimi anni è aumentta la rilevanza del campionamento non probabilistico come strumento di raccolta dei dati nonostante esso implichi che le unità incluse nel campione siano determinate da un processo di auto-selezione incognito. Tale mancaza di casualità nel processo di selezione delle unità solleva preoccupazione sulla possibilità di ottenere stime distorte dei parametri della popolazione e conseguentemente rende necessario definire degli indicatori per quantificare il rischio di distorsione. In questo contributo, motivati da uno specifico caso reale, un'indagine web sulla soddisfazione degli utenti del sistema bibliotecario municipale di Firenze, suggeriamo l'uso congiunto di tre indicatori, che indagano la plausibilità di tre diverse condizioni necessarie per escludere il rischio di distorsione.*

Key words: auxiliary variables, missing data, relative sample size, representativeness

Emilia Rocco and Alessandra Petrucci
Department of Statistics, Computer Science, Applications "G. Parenti" (DISIA)- University of Florence - Viale G. Morgagni, 59 - 50134 Firenze
e-mail: emilia.rocco@unifi.it - alessandra.petrucci@unifi.it

1 Introduction

In recent years data from non-probability samples are becoming more widely used. Several factors contributed to this propagation: more and more surveys have moved online; large amounts of data are available from self-reported administrative datasets or fast and easily collected via social-media; the cost for probability surveys is steadily raising and far higher than that for non-probability surveys; response rates in probability surveys continue to decline in all modes of adminstration.
Unlike to the probability sampling the chance to be selected is unknown in non probability sampling as well as the random selection is absent. This poses new challenges for inference since renders design-based methods of inference, commonly used in probability surveys, inapplicable and raises concerns about the potential for biased results. The selection bias, that refers to the systematic differences betweens statistical estimate and the population parameter, brings the need of measures /indicators to assess the risk of bias and to identify, if possible, any adjustment method. In order to identify such indicators it is worth to note that the selection bias, depending on the composition of the sample, that can vary greatly from that of the surveyed population, may occur not only when non-probability sampling is adopted but also when a probability sample is initially selected and non-response occurs. For this reason many indicators suggested in literature as indirect measures of non-response bias and many statistical adjustment procedures typically used to adjust for non-response bias can be extended to non-probability samples. However, in doing this, it is necessary to take into account that unlike the non-response set, what is a non probability sample is not univocal defined. Non-probability sampling is a collection of methods very different from each other and with in common only the fact that the participants are chosen or choose themselves so that the chance of being selected is not known. Moreover while non respondents in probability surveys are usually identifiable in the population; this does not happen for some non-probability sampling methodologies. Therefore the possibility to use the indicators of non-response bias existing in literature to assess self-selection bias due to non probability sampling depends on the type of non probability sampling: sometimes some methods may be usable and others no, sometimes their use may require some adjustments/ad hoc adaptations. All existing indicators of the risk of non-response bias, apart from the response rate, are based on the use of auxiliary data. Some in addition to the auxiliary data also involve the survey data (i.e. the data for respondents). In this study we are motivated by a real problem: the risk of bias in a non-probability web survey on the customers' satisfaction with the services provided by the municipal library system in Florence. For this case of study we consider the joint use of two indicators of non response bias, one that involves only auxiliary data and the other that in addition to the auxiliary data also involve the survey data. We discuss how they need to be modified for assessing the risk of bias for this specific non probability sample, how they work in the specific real situation and the opportunity to consider in addition to them a third index. Next section describe preliminaries and the suggested indicators. Section 3 presents the application, discuss the results and concludes with some final remarks.

2 Indicators definition

We assume that for a target population U of N units $(i = 1, ...N)$, $\delta_i \in \{0,1\}$ indicate the selection of unit i in a non probability sample s of size n, y_i is the value of the target variable Y for unit i and the data available for estimation purposes are the values $\{y_i; i \in s\}$ of the survey variable and the values $\{\mathbf{z}_i = (z_{1,i}, ...z_{K,i}); i \in U\}$ of a vector of auxiliary variables that may influence the selection mechanism and/or the survey variable. Moreover we shall suppose that the sample inclusion indicators are independent random variables and that the parameter of interest is the mean of Y.
The first indicator that we consider is the "*R-Indicator*" first suggested by Schouten et al. (2009) and developed in Shlomo et al. (2012). The basic idea of the R-indicator is that a selected subset is representative with respect to the auxiliary variables $\mathbf{z}$ when selection probabilities are constant for $\mathbf{z}$. Relying on this idea, and taking into account that selection probabilities are unknown, but may be estimated as the conditional expectation (under a model) of δ_i given the auxiliary variables, the *R-Indicator* measures the extent to which the estimated selection probabilities $\hat{\rho}(zi)$ vary as follows:

$$\hat{R}_\rho = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i\in U}(\hat{\rho}_i - \hat{\bar{\rho}})^2} = 1 - 2\hat{S}_{\hat{\rho}} \tag{1}$$

where $\hat{\bar{\rho}} = \sum_{i=1}^{N}\hat{\rho}_i/N$. Since it may be shown that $\hat{S}_{\hat{\rho}} \leq \sqrt{\hat{\bar{\rho}}(1-\hat{\bar{\rho}})} \leq 0.5$, then $\hat{R}_\rho$ takes values on the interval $[0,1]$. The response propensities, $\hat{\rho}_i$, are commonly estimated with explicit or implicit models linking the response occurrences to the auxiliary variables, for instance, by using a logistic or a probit regression model.
Assuming that the implicit or explicit model used for estimating the response propensities is "correct", a value of $\hat{R}_\rho$ close to 1 denotes a weak association between the sample selection mechanism and the auxiliary variable. A weak association implies that the risk of selection bias is low whatever is the target statistic as long as it may be assumed that the selection/non-selection process is "Missing at Random" (MAR) given the covariates. Unfortunately, however, a value of $\hat{R}_\rho$ not close to 1 - due to an high association between the selection mechanism and the auxiliary variables - is not able to detect the risk of bias for a survey estimate since it also depends on the relation between the covariates and the target variable. In order to judge the risk of bias for a single parameter, when the estimated value of the R-Indicator is not close to 1, we can compare the respondents' averages of the single variable across the percentiles of the predicted response propensities through the following indicator suggested in Rocco (2019) and called R-Statistic-level-Indicator:

$$\hat{R}_{\bar{y}_{p\rho}} = 1 - \frac{\sum_{h=1}^{H}(\bar{y}_{s_h} - \bar{y}_s)^2 n_{s_h}}{\sum_{i\in r}(y_i - \bar{y}_s)^2} \tag{2}$$

that is obtained through the following steps. First the units in the sample are ordered with respect to the estimated selection probabilities, $\hat{\rho}_i$, $i \in s$. Then the ordered sam-

ple is partitioned into H classes, $s_h\,(h = 1,...H)$, of size n_{s_h} = roughly n_s/H on the basis of $(H-1)$ percentiles of the estimated selection probabilities. Finally, the sample mean of the target variable for each class, $\bar{y}_{s_h} = \sum_{i \in s_h} y_i / n_{s_h}$, and the sample mean for the whole sample, $\bar{y}_s = \sum_{i \in r} y_i / n_s = \sum_{h=1}^{H} \sum_{i \in s_h} y_i / n_s$, are calculated and compared through the above expression of the index.
$\hat{R}_{\bar{y}_{p\rho}}$ takes values on the interval $[0,1]$. Under the assumption that the selection process is MAR, a value of $\hat{R}_{\bar{y}_{p\rho}}$ near to 1 indicates a risk of non-response bias negligible whatever is the value of $\hat{R}_\rho$. Only a low value of $\hat{R}_{\bar{y}_{p\rho}}$ associated with a low value $\hat{R}_\rho$ indicates a high risk of selection bias since this condition happens when the covariates are associated both with the selection indicator δ and with the specific variable of interest y. Therefore, under the MAR assumption for the selection process, the two indicators, $\hat{R}_\rho$ and $\hat{R}_{\bar{y}_{p\rho}}$, used jointly, allow to evaluate the risk of selection bias. However the reliability of this assumption needs to be evaluated. To this end we suggest to use, in addition to $\hat{R}_\rho$ and $\hat{R}_{\bar{y}_{p\rho}}$, a third indicator that measures the goodness of fit of the model used to estimate the response probabilities. A very simple indicator is the proportion of correct predictions of the participation (or non-participation) to the survey and may be calculated through a cross-validation procedure. Obviously this indicator works only when the estimated response propensities vary that is when the covariates are related to the selection indicator δ and consequently $\hat{R}_\rho$ is not close to 1. In this case a value of I_{gf} not close to 1 indicates the presence of a residual component of the selection probabilities not explained by the auxiliary variables. This residual part may or may not depend on the target variable. Therefore, in this case, even if the $\hat{R}_{\bar{y}_{p\rho}}$ is close to 1 it is not possible to exclude the risk of a high bias. On the other hand the risk may be excluded when $\hat{R}_\rho$ is not close to 1 but both I_{gf} and $\hat{R}_{\bar{y}_{p\rho}}$ are close to 1.
In some situations the three indicators are not straight exploitable. In some non probability samples the sample units may be not identifiable in the population, in other words it is not possible to associate to each population unit the selection indicator δ_i. In these cases the three indicators cannot be calculated. A small relative size of the sample is another issue for the use of the $\hat{R}_\rho$. A very small relative size of the sample should produce a sort of smoothing of the estimated selection probabilities and consequently a value of $\hat{R}_\rho$ close to 1 that is not able to assess the representativeness of the sample. Unfortunately these situations must be managed differently in each case. In the next section we discuss this issues and suggest a possible solution for a case study.

3 Application and discussion

The case study is a non-probability web survey that has been carried out (between February 6 and March 4, 2019) by the statistical office of the Florence municipality in order to evaluate the customers' satisfaction with the services provided by the municipal library system. The satisfaction questionnaire was administered via web

to the individuals at least 18 years-old present in one or both the following lists: the subscribers to the newsletter of the Florentine Municipal Libraries and the subscribers to at least one of the Florentine Municipal Libraries. The available e-mail addresses were about 95,000, but at the end only 7,680 completed questionnaires returned. The survey collected information on: the knowledge and use of the services offered by the municipal librarian system; the level of satisfaction with the services actually used; the socio-demographic characteristics of the participants (in particular age, gender, education level and professional condition) and the residence. We are interested in evaluating the representativeness of the survey with respect to population of all the citizens of Florence at least 18 years-old. For this reason we consider only the survey data relative to the Florentine residents. The representativeness of such sample would allow the statistical office of the Florence municipality to use the same set of e-mail addresses for other surveys regarding municipal services. In order to evaluate the usability of the three suggested indicators and their ability to assess the representativeness of the survey data we have considered a pseudo real application. From Census 2011 are available for the population units the same socio-demographic variable collected in the survey. Therefore we have assumed as auxiliary variables gender, age and education level and as response variable the professional condition. We evaluate the risk of selection bias for the following parameters of interest: the proportion of employees, the proportion of housewives and the proportion of retired. The data present both the issues discussed at the end of the previous section: it is not possible to identify the survey participants in the population and the participation fraction is very very small (only 1.5%). The small sample size compared to that of the population, suggests that the characteristics of the population would not change significantly if we replaced it with a fictitious population made up of the union of the population and the sample. This allows us to associate to the units of this new population the indicator variable δ_i which assumes value 1 for the units belonging to the web sample and 0 for the others. Consequently we can estimate the probability of participation in the web survey and the three indicators. However this solution does not solve the second issue. To solve both issues, moving from the well known consideration that a simple random sample reproduce on average the same means and proportions of the population, we suggest the following replicated sampling procedure: m fiction populations are obtained as the union of the web sample and a simple random sample selected from the original population. Then for each of these populations the three indicators $\hat{R}_{\rho}^{(h)}$, $R_{\bar{y}_{p\rho}}^{(h)}$ and $I_{gf}^{(h)}$ $(h = 1, ..., m)$ are calculated. Finally, the indicators $\hat{R}'_{\rho}$, $\hat{R}'_{\bar{y}_{p\rho}}$ and I'_{gf} are obtained as the median of the correspondent distributions over the m replications. The results, shown in Table 1 are obtained considering a size of the simple random samples equal to the number of the web survey participants (that is the participation fraction is chosen equal to 0.5 but different values of the sample fraction from 0.2 to 0.8 have been investigated and produce the same conclusions). Table 1 shows the values of the three indicators and the values of the relative true bias (RB_{True}). The RB_{True} is calculated as the difference between the respondent mean and the true population parameter scaled by the population standard deviation. The low value of $\hat{R}'_{\rho}$ suggests

Table 1 Summaries of results

Parameters	RB_{True}	$\hat{R}'_{\rho}$	I'_{gf}	$R'_{\bar{y}_{pp\rho}}$
Employed proportion	0.279	0.451	0.651	0.838
Household proportion	−0.154	0.451	0.651	0.979
Rertired proportion	−0.144	0.451	0.651	0.818

an important association between the selection process and the auxiliary variables, while the low value of $I^{(h)}_{gf}$ denotes the presence of a residual component of the selection probabilities not explained by the auxiliary variables and since this may or may not depend on the target variable, the values of $\hat{R}'_{\bar{y}_{pp\rho}}$ quite high do not allow to exclude the risk of bias that effectively result from the true values.
It is worth to note that the rather high values of the true relative bias may be also due to the very small relative size of the sample. In fact, as shown in Meng (2018) the exact error of the sample mean $(\bar{y}_{n_s} - \overline{Y}_N)$ is about $\sqrt{N}\rho_{\delta,Y}$ times the standard error of the sample mean under a simple random sample of equal size with $\rho_{\delta,Y}$ denoting the correlation coefficient between the selection indicator δ and the outcome variable Y.
Finally, the results show that, specific indicators need for each type of non-probability sampling methodology and by using them the risk of bias of unweighed mean can be assessed only if the following conditions are satisfied: all the variables correlated with both inclusion in the sample and the outcome of interest are know and measured; there are no kinds of unit with distinct values of the outcome variable that are systematically missing from the sample. When these conditions happen it is also possible to reduce the eventual bias through weighting methods.

References

1. Meng, X.-L.: Statistical paradises and paradxes in big data (I): Law of large populations, big data paradox, and the 2016 Us presidential election. The Annals of Applied Statistics **12**, 685-726 (2018)
2. Rocco, E.: Indicators for monitoring the survey data quality when non-response or a covenience sample occurs. In: Petrucci, A. et al.(eds.) New Statistical Developments in Data Science, pp. 233-245. Springer Nature Switzerland AG (2019)
3. Schouten, B., Cobben, F., Betleehem, J.: Indicators for the representativeness of survey response. Survey Methodology **35**, 101-113 (2009)
4. Shlomo, N., Skinner, C., Schouten, B.: Estimation of an indicator for the representativeness of survey response. Journal of Statistical planning and Inference. **142**, 201-211 (2012)

On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design

Stima di massima verosimiglianza per modelli esponenziali in presenza di un disegno fisso e di un disegno adattivo a due stadi

Caterina May and Chiara Tommasi

Abstract We consider two different design strategies for collecting "optimal" data with the aim of estimating as precisely as possible the vector parameter in a dose-response model. In particular, an exponential model with Gaussian errors is considered, and the maximum likelihood method is applied. Through a simulation study we compare the performance of the the maximum likelihood estimator (MLE) when: a) a locally D-optimum design is used to get a sample of independent observations (fixed procedure); b) a two-stage adaptive experimental procedure is applied to collect data, which are dependent since the second stage D-optimal design is determined by the responses observed at the first stage. In the latter case, the theoretical properties of the MLE are described; differently from the most of the literature, asymptotic theory is applied only in the second stage since the first stage sample size is assumed to be finite.

Abstract *Consideriamo qui due diverse strategie per raccogliere dati "ottimali" allo scopo di stimare con precisione il vettore dei parametri in un modello di risposta alla dose. Consideriamo, in particolare, un modello esponenziale con erori Gaussiani. Mediante uno studio di simulazione confrontiamo l'efficienza dello stimatore di massima verosimiglianza quando: a) si utilizza un disegno localmente D-ottimo ottenendo un campione di osservazioni indipendenti (procedura fissa); b) si utilizza una procedure adattiva a due stadi da cui si ottengono dati che sono dipendenti, dato che il disegno D-ottimo al secondo stadio è determinato dalle risposte osservate al primo stadio. In quest'ultimo caso descriviamo le proprietà teoriche dello stimatore di massima verosimiglianza; al contrario di quanto viene fatto normalmente in letteratura, la teoria asintotica viene qui applicata solo al secondo stadio mentre la dimensione campionaria el primo stadio è considerata fissa.*

Caterina May
Università degli Studi del Piemonte Orientale, Department DiSEI, via Perrone, 18, 28100 Novara (Italy), e-mail: caterina.may@uniupo.it

Chiara Tommasi
Università degli Studi di Milano, Department DEMM, via Conservatorio, 7, 20122 Milano (Italy) e-mail: chiara.tommasi@unimi.it

Key words: maximum likelihood inference, D-optimum design, two-stage adaptive design, exponential model, relative efficiency, simulations

1 Introduction

The exponential model is applied in many different contexts (medical, enviromental, pharmaceutical) to interpret dose-response relationships. A three-parameters exponential model can be written as

$$Y = \theta_0 + \theta_1 \exp(x/\theta_2) + \varepsilon, \quad \varepsilon \sim N(0;\sigma^2) \tag{1}$$

where $\theta = (\theta_0, \theta_1, \theta_2)^T$ is a vector of unknown parameters and $\eta(x,\theta) = \theta_0 + \theta_1 \exp(x/\theta_2)$ denotes the response mean at the dose $x \in \mathscr{X} = [a,b]$. The inferential goal is to estimate θ and thus efficient experimental designs are very important. An experimental design ξ can be defined as a finite discrete probability distribution over $\mathscr{X}$; the information matrix of ξ is

$$M(\xi;\theta) = \int_{\mathscr{X}} \nabla\eta(x,\theta)\nabla\eta(x,\theta)^T d\xi(x), \tag{2}$$

where $\nabla\eta(x,\theta)$ denotes the gradient of the mean response $\eta(x,\theta)$ with respect to θ. A D-optimal design $\xi^*(\theta)$ minimizes the generalized asymptotic variance of the maximum likelihood estimator (MLE) of θ, i.e.

$$\xi^*(\theta) = \arg\max_{\xi\in\Xi} |M(\xi;\theta)|, \tag{3}$$

where Ξ is the set of all designs (see [4] and [1]).

Since $\eta(x,\theta)$ is a non-linear model, the D-optimal design (3) depends on the unknown parameter θ. A common approach to tackle this problem is to use a locally optimal design, where θ in (2) is replaced by a guessed value $\tilde{\theta} = (\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2)^T$; n independent observations are collected according to this locally D-optimal design $\xi^*(\tilde{\theta})$ and then used to compute the MLE.

Another possibility to obtain the data is to adopt a two-stage procedure. At the first stage a locally D-optimal design $\xi^*(\tilde{\theta})$ is applied to collect n_1 responses (with $n_1 < n$), which are used to estimate the unknown parameter. Let $\hat{\theta}_{n_1}$ be the MLE of θ based on first-stage responses. Then, at the second stage, $n_2 = n - n_1$ additional responses are collected according to another locally D-optimal design, $\xi_2^*(\hat{\theta}_{n_1})$, where $\hat{\theta}_{n_1}$ is used in (2) instead of $\tilde{\theta}$. Finally, the MLE is computed employing the whole sample of $n = n_1 + n_2$ data. Let us note that $\xi_2^*(\hat{\theta}_{n_1})$ is a random probability distribution as it depends on the first-stage observations through $\hat{\theta}_{n_1}$; as a consequence, the second-stage observations are not independent on the first-stage ones. Given $\xi_2^*(\hat{\theta}_{n_1})$, however, the second-stage observations are conditionally independent on the first-stage data, and hence it can be proved that the likelihood function is the same in both the following experimental settings (see Sect. 2.1):

1) n independent observations obtained according to $\xi^*(\tilde{\theta})$ (fixed procedure or one-stage);
2) n_1 independent observations accrued according to $\xi^*(\tilde{\theta})$ and then, given $\hat{\theta}_{n_1}$, other $n_2 = n - n_1$ independent responses coming from $\xi_2^*(\hat{\theta}_{n_1})$ (two-stage procedure).

Let $\hat{\theta}_n^{1S}$ and $\hat{\theta}_n^{2S}$ denote the MLEs when the one-stage and the two-stage procedures, respectively, are adopted to collect the data. In this paper we develop a simulation study to compare the performance of $\hat{\theta}_n^{1S}$ and $\hat{\theta}_n^{2S}$.

2 Theoretical properties of the two-stage design

2.1 Likelihood

The total likelihood is

$$\begin{aligned}\mathscr{L}(\theta;\mathbf{y}_{n_2},\mathbf{x}_{n_2},\mathbf{y}_{n_1},\mathbf{x}_{n_1}) &= \mathscr{L}(\theta;\mathbf{y}_{n_2}|\mathbf{x}_{n_2},\mathbf{y}_{n_1},\mathbf{x}_{n_1})\cdot\mathscr{L}(\mathbf{x}_{n_2}|\mathbf{y}_{n_1},\mathbf{x}_{n_1})\cdot\mathscr{L}(\theta;\mathbf{y}_{n_1}|\mathbf{x}_{n_1})\cdot\mathscr{L}(\mathbf{x}_{n_1}) \\ &= \mathscr{L}(\theta;\mathbf{y}_{n_2}|\mathbf{x}_{n_2})\cdot\mathscr{L}(\mathbf{x}_{n_2}|\mathbf{y}_{n_1},\mathbf{x}_{n_1})\cdot\mathscr{L}(\theta;\mathbf{y}_{n_1}|\mathbf{x}_{n_1})\cdot\mathscr{L}(\mathbf{x}_{n_1}) \\ &\propto \mathscr{L}(\theta;\mathbf{y}_{n_2}|\mathbf{x}_{n_2})\cdot\mathscr{L}(\theta;\mathbf{y}_{n_1}|\mathbf{x}_{n_1}) \qquad (4)\end{aligned}$$

From the (4) we can see that the total likelihood for the dependent data of the two-stage design is the same as the likelihood with independent data of the fixed design.

2.2 Asymptotics

In order to obtain the consistency and the asymptotic distribution of $\hat{\theta}_n^{2S}$, the classical approach in the literature is to assume that both the sample sizes n_1 and n_2 grow to infinity (see [6]).

A different approach is considered in [5], where n_1 is fixed and only n_2 goes to infinity; this assumption is more realistic in many experimental situations and, in addition, fixing n_1 should improve the approximation of the finite distribution with the asymptotic one. Despite the second stage observations depend on the first-stage data through $\hat{\theta}_{n1}$, the MLE $\hat{\theta}_n^{2S}$ maintains good properties, as stated in the following theorems (see [5] for the proofs).

Theorem 1. *As $n_2 \to \infty$, $\hat{\theta}_n^{2S}$ converges in probability to the true value θ of the parameter.*

Theorem 2. *As $n_2 \to \infty$, $\sqrt{n}\left(\hat{\theta}_n^{2S} - \theta\right)$ converges in distribution to*

$$\sigma\, M[\xi_2^*(\hat{\theta}_{n_1}), \theta]^{-1/2}\, \mathbf{Z},$$

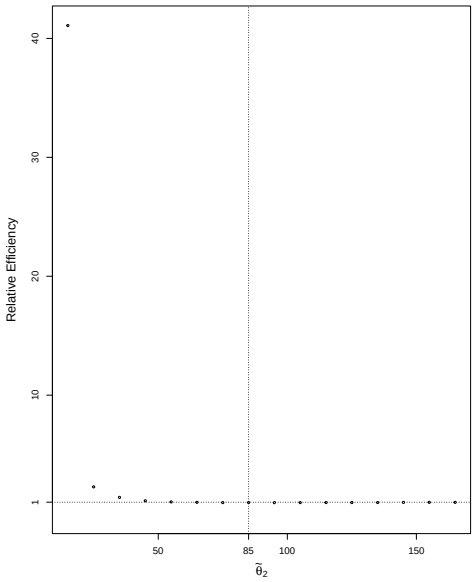


Fig. 1 Relative efficiency $\mathrm{MSE}(\hat{\theta}_n^{1S})/\mathrm{MSE}(\hat{\theta}_n^{2S})$ (on the y-axes) versus different nominal values $\tilde{\theta}_2$ (on the x-axes). The model is exponential with $\theta_0 = -0.08265$, $\theta_1 = 0.08265$, $\theta_2 = 85$ and $\sigma = 0.1$. The vertical line represents the true value of θ_2.

where $\mathbf{Z}$ *is a* 3*-dimensional standard normal random vector independent of the random matrix* $M(\xi_2^*(\hat{\theta}_{n_1}), \theta)$.

Theorem 3. *As* $n_2 \to \infty$*,, the asymptotic variance of* $\sqrt{n}\left(\hat{\theta}_n^{2S} - \theta\right)$ *is*

$$\sigma^2 E\left[\left(\int_{\mathscr{X}} \nabla\eta(x,\theta)\nabla\eta(x,\theta)^{\mathrm{T}}\, d\xi_2^*(\hat{\theta}_{n_1})(x)\right)^{-1}\right] \tag{5}$$

The expression of the asymptotic variance of $\hat{\theta}_n^{2S}$ provided in (5) justifies the use of a D-optimal design to collect the second stage data.

3 Simulations

The goal of the simulation study is to compare the two-stage adaptive procedure with the fixed design in terms of precision of the MLEs, $\hat{\theta}_n^{2S}$ and $\hat{\theta}_n^{1S}$.

The D-optimum design $\xi^*(\theta)$ for the exponential model is provided in [3]. It is a three point design equally supported at the extremes of the experimental domain $\mathscr{X} = [a;b]$ and at

$$x^* = \frac{(b-\theta_2)\exp(b/\theta_2) - (a-\theta_2)\exp(a/\theta_2)}{\exp(b/\theta_2) - \exp(a/\theta_2)}. \tag{6}$$

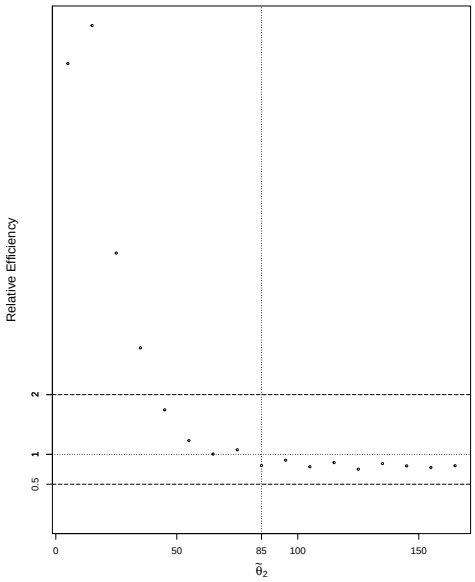


Fig. 2 Relative efficiency $\mathrm{MSE}(\hat{\theta}_n^{1S})/\mathrm{MSE}(\hat{\theta}_n^{2S})$ (on the y-axes) versus different nominal values $\tilde{\theta}_2$ (on the x-axes). The model is exponential with $\theta_0 = -0.08265$, $\theta_1 = 0.08265$, $\theta_2 = 85$ and $\sigma = 0.25$. The vertical line represents the true value of θ_2.

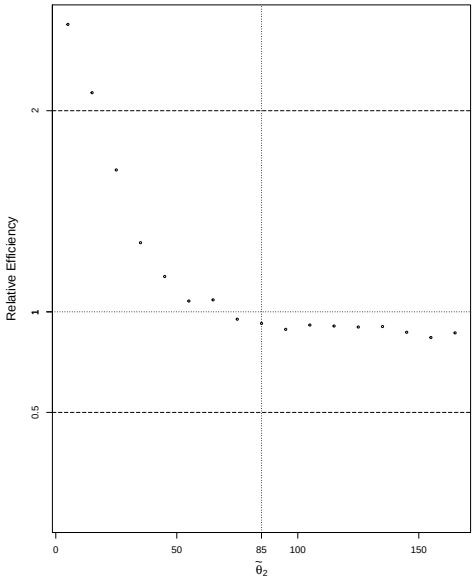


Fig. 3 Relative efficiency $\mathrm{MSE}(\hat{\theta}_n^{1S})/\mathrm{MSE}(\hat{\theta}_n^{2S})$ (on the y-axes) versus different nominal values $\tilde{\theta}_2$ (on the x-axes). The model is exponential with $\theta_0 = -0.08265$, $\theta_1 = 0.08265$, $\theta_2 = 85$ and $\sigma = 0.5$. The vertical line represents the true value of θ_2.

Hence, at each stage, $1/3$ of the observations are taken at a, b and x^*. Note that x^*, and hence $\xi^*(\theta)$, depends only on the non-linear parameter θ_2 of model (1). Herein, we take $a = 0$ and $b = 150$. From model (1) with $\theta_0 = -0.08625$, $\theta_1 = 0.08625$ and $\theta_2 = 85$ and 3 different values for $\sigma = 0.1; 0.25; 0.5$,

1. we generate $n_1 = 30$ independent observations according to $\xi^*(\tilde{\theta}_2)$ to compute the first-stage MLE, $\hat{\theta}_{n_1}$, where $\tilde{\theta}_2 \in (0; 150)$ is a nominal value for θ_2;
2. we generate further $n_2 = 300$ independent observations according to $\xi^*(\hat{\theta}_{n_1})$ to obtain $\hat{\theta}_n^{2S}$;
3. we generate further $n_2 = 300$ independent observations according to $\xi^*(\tilde{\theta}_2)$ to obtain $\hat{\theta}_n^{1S}$.

For each choice of σ and $\tilde{\theta}_2$ we repeat the computation of $\hat{\theta}_n^{1S}$ and $\hat{\theta}_n^{2S}$ 5000 times, to get their Monte Carlo MSEs. Simulations are realized with R package in [2]. Figures 1, 2 and 3 displays the relative efficiency $\mathrm{MSE}(\hat{\theta}_n^{1S})/\mathrm{MSE}(\hat{\theta}_n^{2S})$ for different choices of the nominal value $\tilde{\theta}_2$, and for $\sigma = 0.1; 0.25; 0.5$, respectively.

4 Conclusions

The simulations in Sect.3 show that the two-stage procedure outperforms the one-stage (or fixed) procedure whenever the assumed nominal value $\tilde{\theta}_2$ is much inferior to the true value of θ_2 (in this example $\theta_2 = 85$). For the other values of $\tilde{\theta}_2$, the relative efficiency of the two-stage procedure is around one (never less then 0.5). This behaviour appears to be more pronounced as σ increases. An explanation can be seen in the slope of $x^* = x^*(\theta_2)$ in (6), which is larger for small value of θ_2. Hence, $x^*(\tilde{\theta}_2)$ is far away from the true optimal dose if $\tilde{\theta}_2 < 85$ and replacing $\tilde{\theta}_2$ with the first stage estimate may improve the results.

In conclusion, we suggest to apply the two-stage procedure to collect the data if we do not have enough knowledge about the true value of θ_2, which is often the case in real-life problems.

References

1. A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum experimental designs, with SAS*, volume 34 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 2007.
2. B. Bornkamp, J. Pinheiro, and F. Bretz. *DoseFinding: Planning and Analyzing Dose Finding Experiments*, 2018. R package version 0.9-16.
3. H. Dette, C. Kiss, M. Bevanda, and F. Bretz. Optimal designs for the emax, log-linear and exponential models. *Biometrika*, 97(2):513–518, 2010.
4. V. Fedorov. *Theory of Optimal Experiments*. Acadimic Press, New York, 1972.
5. N. Flournoy, C. May, and C. Tommasi. The effects of adaptation on maximum likelihood inference for non-linear models with normal errors. arXiv:1812.03970, Submitted, 2019.
6. L. Pronzato and A. Pázman. *Design of experiments in nonlinear models*, volume 212 of *Lecture Notes in Statistics*. Springer, New York, 2013. Asymptotic normality, optimality criteria and small-sample properties.

Pseudo-population based resamplings for two-stage design

Ricampionamento basato su pseudo-popolazioni per il disegno a due stadi

Pier Luigi Conti and Daniela Marella and Vincenzina Vitale

Abstract This communication is devoted to resampling based on pseudo-populations for a class of two-stage sampling design. A resampling scheme is defined, and its main theoretical properties are established.
Abstract *La presente comunicazione è essenzialmente dedicata alla definizione e allo studio di uno schema di ricampionamento per un'ampia classe di disegni campionari a due stadi. In primo luogo alcuni risultati asintotici relativi a disegni ad uno stadio vengono estesi al caso di disegni a due stadi. Viene poi definito uno schema di ricampionamento e ne vengono stabilite le principali proprietà teoriche.*

Key words: two-stage sampling, resampling, empirical process, asymptotics.

1 Introduction

The use of resampling methods in finite populations sampling is widespread, and several approaches have been proposed in the literature; cfr. the survey paper by [1], where different approaches are illustrated, and their properties listed. Among the methodologies proposed, the one based on pseudo-populations plays a particularly relevant role, because of its theoretical properties. In fact, as shown in [2], resampling based on pseudo-populations, under appropriate conditions, possesses

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: pierluigi.conti@uniroma1.it

Daniela Marella
Università Roma Tre; Via del Castro Pretorio, 20; 00185 Roma; Italy, e-mail: daniela.marella@uniroma3.it

Vincenzina Vitale
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: vincenzina.vitale@uniroma1.it

a justification similar to that of the "usual" Efron's bootstrap for *i.i.d.* data: as the sample size increases, the large sample distribution of a sampling statistic tends to the same weak limit as the resampling distribution of the statistic itself.

Results in [2], as well as other asymptotic results for finite populations not involved with resampling (cfr. [3], [4], and, for generalizations, [5]) essentially refer to single-stage sampling designs, and do not cover two-stage type designs.

Generalizations of "direct" resampling schemes to multistage sampling designs are studied in the literature (cfr. [1]), although their theoretical properties are not clear, and a transparent justification is still lacking. In particular, in [6] the pseudo-population procedure is generalized to multi-stage designs.

The goal of the present communication is essentially to define a pseudo-population based resampling scheme for a wide class of two-stage designs, and to provide a justification based on asymptotic arguments similar to [2]. For this reason, asymptotic results in the above mentioned paper will be first extended to two-stage sampling design. This involves the consideration of empirical processes based on two-stage sampling. Then, such results will be applied to show that the limit distribution of a sampling statistic, under fairly general conditions, possesses the same limit distribution as the resampled statistic. As already remarked, this is essentially the justification for Efron's bootstrap provided in [7].

2 Basic assumptions

Let $\mathscr{U}_N$ be a finite population of size N, and suppose it is divided into M clusters (primary units) composed by $N_1, \ldots, N_M$ elementary units, respectively, with $N_1 + \cdots + N_M = N$. The symbol $w_g = N_g/N$ denotes the weight of cluster g $(= 1, \ldots, M)$. Let further Y be the character of interest, and denote by $Y_{g,i}$ the value of Y for unit i of cluster g $(i = 1, \ldots, N_g;\ g = 1, \ldots, M)$.

For the sake of simplicity, clusters are assumed to be "random", *i.e.* there exist M bivariate random variables (r.v.s) (X_g, C_g), $g = 1, \ldots, M$, with

$$w_g = \frac{C_g}{\sum_{h=1}^{M} C_h} = \frac{C_g}{M\overline{C}_M}$$

where $\overline{C}_M = \sum_{h=1}^{M} C_h/M$, and $N_g = [Nw_g]$.

First of all, at a cluster level, the bivariate r.v.s (X_g, C_g) are assumed to be independent and identically distributed (*i.i.d*), with $\mu_C = \mathbb{E}[C_g]$, $\mu_X = \mathbb{E}[X_g]$, $g = 1, \ldots, M$.

In the second place, $Y_{g,i}$s are assumed to be identically distributed conditionally on (X_g, C_g), with distribution function (d.f.)

$$F_g(y) = F(y|X_g, C_g) = \mathbb{P}(Y_{g,i} \leq y|X_g, C_g);\ \ g = 1, \ldots, M.$$

Unit i of cluster g is assumed to be independent of units of different clusters, but may be correlated with other units of the same cluster. More precisely, conditionally on X_g, C_g, X_h, C_h the r.v.s $Y_{g,i}$ and $Y_{h,j}$ are assumed to be independent as $h \neq g$. Furthermore, for each unit i of cluster g there exists a set L_i of $|L_i|$ units of cluster g for which

$$Corr(Y_{g,i}, Y_{g,j}|X_g, C_g) = \begin{cases} \rho_{ij} \text{ if } i \in L_i \\ 0 \text{ if } i \notin L_i \end{cases}.$$

As far as the two-stage sampling design is concerned, our assumptions are listed below.

- 1st Stage. A sample $\mathbf{g}$ of m clusters is selected according to an *high entropy* sampling design (cfr. [8], [9], [3], [2]) with first order inclusion probabilities

$$\pi_g = m\frac{X_g}{\sum_{g=1}^{M} X_g} \sim \frac{m}{M}\frac{X_g}{\mu_X}, \ g = 1, \ldots, M.$$

- 2nd Stage. From each cluster $g \in \mathbf{g}$ selected at first stage, a sample $\mathbf{s}_g$ of n_g elementary units is drawn according to simple random sampling without replacement. Samples $\mathbf{s}_g$s are independent conditionally on $\mathbf{g}$. The *total sample* is $\mathbf{s} = (\mathbf{s}_g;\ g \in \mathbf{g})$.

3 Main theoretical results

Consider the indicator function

$$I_{(y_{g,i} \leq y)} = \begin{cases} 1 \text{ if } y_{g,i} \leq y \\ 0 \text{ if } y_{g,i} > y \end{cases}; \ i = 1, \ldots, N_g, \ g = 1, \ldots, M; \ y \in \mathbb{R}.$$

The finite population distribution function (f.p.d.f.) is defined as

$$F_N(y) = \frac{1}{N}\sum_{g=1}^{M}\sum_{i=1}^{N_g} I_{(y_{g,i} \leq y)} = \sum_{g=1}^{M} w_g F_{N_g}(y) \tag{1}$$

where

$$F_{N_g}(y) = \frac{1}{N_g}\sum_{i=1}^{N_g} I_{(y_{g,i} \leq y)}$$

is the d.f. of gth cluster.

To estimate (1), we consider here the (design-based) Hájek estimator

$$\widehat{F}_H(y) = \sum_{g \in \mathbf{g}} \frac{w_g}{\pi_g}\widehat{F}_g(y) \Bigg/ \sum_{g \in \mathbf{g}} \frac{w_g}{\pi_g}$$

where

$$\widehat{F}_g(y) = \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} I_{(y_{g,i} \leq y)}$$

is the empirical distribution function (e.d.f.) of cluster gth.

From now on, we will assume that the probability distribution of X_g is non-degenerate and possesses finite second moment, so that $\mathbb{E}[X_g^2] = \mu_{2x} < \infty$. As far as the population size and the sample size are concerned, our set-up is essentially similar to [2]. In detail, we will assume that

- $M \to \infty$, $m \to \infty$, with $m/M \to f \geq 0$;
- $\min_{1 \leq g \leq M} N_g \to \infty$, $\min_{1 \leq g \leq M} n_g \to \infty$.

Proposition 1. *Under the above conditions, the following results hold.*

- *The relationship*

$$\sqrt{m}(\widehat{F}_H(y) - F_N(y)) = \sqrt{m} \sum_{g \in \mathbf{g}} w_g \frac{1}{\pi_g} (\widehat{F}_g(y) - F_{N_g}(y)) + o_p(1)$$

  holds, where $o_p(1)$ tends to 0 in probability uniformly w.r.t. y.
- *Conditionally on X_gs, C_gs, $Y_{g,i}$, the sequence*

$$\sqrt{m} \sum_{g \in \mathbf{g}} w_g \frac{1}{\pi_g} (\widehat{F}_g(y) - F_{N_g}(y)); \; m \geq 1$$

  converges weakly to a Gaussian process $(W_H(y);\ y \in \mathbb{R})$, with zero mean function ($E[W_H(y)]$ and covariance kernel $K(y,t) = E[W_H(y)W_H(t)]$. Weak convergence takes place in the space $D[-\infty, +\infty]$ of cadlag functions equipped with the Skorokhod distance, and holds for a set of sequences of X_gs, C_gs, $Y_{g,i}$ having probability 1.

As a by-product of Proposition 1, we easily obtain asymptotic normality for a wide class of finite population parameters. Generally speaking, a finite population parameter is a functional $\theta : F_N(y) \to \mathbb{R}$. As an estimator of $\theta(F_N)$, it is intuitive to use a plug-in approach consisting in replacing F_N with $\widehat{F}_H$ into $\theta(\cdot)$. This would lead to the estimator $\widehat{\theta} = \theta(\widehat{F}_H)$ of $\theta = \theta(F_N)$.

The map $\theta(\cdot) : l^\infty[-\infty, +\infty] \to E$ is *Hadamard-differentiable* at F if there exists a continuous linear mapping $\theta'_F : l^\infty[-\infty, +\infty] \to E$ such that

$$\left\| \frac{\theta(F + th_t) - \theta(F)}{t} - \theta'_F(h) \right\|_E \to 0 \;\; as\, t \downarrow 0, \; for\, every\, h_t \to h. \tag{2}$$

The term $\theta'_F(\cdot)$ in (2) is the *Hadamard derivative* of θ at F (cfr. [10]).

Proposition 2. *Under the condition of Proposition* 1*, and if θ is Hadamard - differentiable at $F(y) = \mathbb{E}[F_N(y)]$ with Hadamard derivative $\theta'_F(\cdot)$, then $\sqrt{m}(\theta(\widehat{F}_H) -$*

$\theta(F_N))$ *tends in distribution to a normal random variable with zero expectation and variance* $\sigma_\theta^2 = E[\theta'_F(W_H)^2]$.

4 The resampling procedure

Although results in Propositions 1, 2, are of importance, they cannot be immediately used to construct a confidence interval for a finite population parameter θ, because the asymptotic variance σ_θ^2 generally possesses a complicate structure, depending on the superpopulation distribution of X_gs, $Y_{g,i}$s. For this reason, we resort to a resampling procedure based on a pseudo-population constructed *via* a two-stage mechanism. The construction of such a two-stage pseudo-population is described below.

1. For every sampled cluster $g \in \mathbf{g}$, expand $\mathbf{s}_g$ to a "pseudo-population for cluster g", where each unit $i \in \mathbf{s}_g$ appears $N^*_{g,i}$ times, with $N^*_{g,i} = [N_g/n_g] + \varepsilon_{g,i}$, where $\varepsilon_{g,i}$ is a Bernoulli r.v. with expectation $N_g/n_g - [N_g/n_g]$.
2. Expand each cluster $g \in \mathbf{g}$ to a "clustered pseudo-population", where each cluster $g \in \mathbf{g}$ appears $M^*_g = [1/\pi_g] + r_g$ times, where r_g is a Bernoulli r.v. with expectation $1/\pi_g - [1/\pi_g]$.

In general, the constructed pseudo-population has total size $N^* \neq N$, although it can be shown that N^*/N tends in probability to 1 as M increases. In the sequel, we will denote by $F^*_{N^*}(y)$ the d.f. of the pseudo-population.

From the pseudo-population, L independent pseudo-samples $\mathbf{s}^*_1, \ldots, \mathbf{s}^*_L$ are drawn according to a two-stage design identical to that used to select the sample $\mathbf{s}$ from the actual population. Denote further by $\widehat{F}^*_{H,l}(y)$ the Hájek estimator of $F^*_{N^*}(y)$ based on the pseudo-sample $\mathbf{s}^*_l$, and compute the corresponding estimates of $\theta(\cdot)$:

$$\widehat{\theta}^*_l = \theta(\widehat{F}^*_{H,l});\ \ l = 1, \ldots, L.$$

Next, the L quantities

$$Z^*_l = \sqrt{m}\left(\widehat{\theta}^*_l - \theta(F^*_{N^*})\right) = \sqrt{m}\left(\theta(\widehat{F}^*_{H,l}) - \theta(F^*_{N^*})\right);\ \ l = 1, \ldots, L \tag{3}$$

are computed, as well as their variance

$$\widehat{S}^{2*} = \frac{1}{L-1}\sum_{l=1}^{L}\left(Z^*_l - \overline{Z}^*_L\right)^2, \tag{4}$$

where $\overline{Z}^*_L = \sum_{l=1}^{L} Z^*_l$. Denote further by

$$\widehat{R}^*_L(z) = \frac{1}{L}\sum_{l=1}^{L} I_{(Z^*_l \leq z)},\ \ z \in \mathbb{R} \tag{5}$$

the empirical distribution function of Z_l^*s, and by

$$\widehat{R}_L^{*-1}(p) = \inf\{z : \widehat{R}_L^*(z) \geq p\}, \ 0 < p < 1 \tag{6}$$

the corresponding pth quantile.

Proposition 3. *Under the same conditions as Propositions* 1*,* 2*, the relationship*

$$\sup_z \left| P(\theta(\widehat{F}_H)) - \widehat{R}_L^*(z) \right| \to 0 \tag{7}$$

holds as $L \to \infty$, a.s. w.r.t. X_gs, C_gs, $Y_{g,i}$s, and in probability w.r.t. samples $\mathbf{s} = (\mathbf{s}_g; g \in \mathbf{g})$, and S^{*2} tends in probability to σ_θ^2.

As a consequence, the confidence intervals for $\theta_N = \theta(F_N)$

$$\left[\widehat{\theta}_H - n^{-1/2} R_L^{*-1}(1-\alpha/2),\ \widehat{\theta}_H - n^{-1/2} R_L^{*-1}(\alpha/2)\right] \tag{8}$$

$$\left[\widehat{\theta}_H - n^{-1/2} z_{\alpha/2} \widehat{S}^*,\ \widehat{\theta}_H + n^{-1/2} z_{\alpha/2} \widehat{S}^*\right] \tag{9}$$

both possess asymptotic level $1-\alpha$ as M, m, $\min_g N_g$, $\min_g n_g$, and L increase.

Acknowledgments. Thanks are due to Prof. Paola Vicard, for her suggestions and insightful remarks.

References

1. Mashreghi, Z., Haziza, D., Léger, C.: A survey of bootstrap methods in finite population sampling. Statistics Survey, **10**, 1–52 (2016)
2. Conti, P L., Marella, D., Mecatti, F., Andreis, F.: A unified principled framework for resampling based on pseudo-populations: asymptotic theory. To appear in Bernoulli - 10.3150/19-BEJ1138 (2019)
3. Bertail, P., Chautru, E., Clémençon, S.: Empirical Processes in Survey Sampling with (Conditional) Poisson Designs. Scandinavian Journal of Statistics, **44**, 97–111 (2017)
4. Boistard, H., Lophuhaä, H P., Ruiz-Gazen, A.: Functional central limit theorems for single-stage sampling design. The Annals of Statistics, **45**, 1728–1758 (2017)
5. Han, Q., Wellner, J A.: Complex Sampling Designs: Uniform Limit Theorems and Applications. Technical Report, arXiv:1905.12824 [math.ST] (2019)
6. Chauvet, G.: *Méthodes de bootstrap en population finie*. Ph.D. Thesis, Université de Rennes (2007)
7. Bickel, P J., Freedman, D A.: Some Asymptotic Theory for the Bootstrap. The Annals of Statistics, **9**, 1196–1217 (1981)
8. Hájek J.: Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. The Annals of Mathematical Statistics, **35**, 1491–1523 (1964)
9. Hájek J.: Sampling from a finite population. Marcel Dekker, New York (1981)
10. van der Vaart, A.: Asymptotic Statistics, Cambridge University Press, Cambridge (2000)

Models and methods - Theoretical Issues in Statistical Inference

A new mixture model for three-way data

Una nuovo modello mistura per dati a tre vie

Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato

Abstract In the last years, there has been a growing interest in the analysis of three-way (matrix-variate) data via mixture models. Quite often real data are affected by outliers, and this also occurs in the context of three-way data. A common solution for managing this type of data consists in fitting mixtures of heavy-tailed distributions. Unfortunately, the three-way literature is still limited, with the only matrix variate *t* mixtures recently proposed to cope with this issue. Therefore, in this work we firstly introduce a new heavy-tailed matrix-variate distribution and then we use it within the mixture model setting. The resulting mixture model is finally fitted to a real dataset for illustrative purposes.

Abstract *Negli ultimi anni, vi è stato un crescente interesse verso l'analisi di dati a tre vie (matrix-variati) per mezzo di modelli mistura. Spesso i dati reali sono affetti da osservazioni atipiche, e questo accade anche nel contesto dei dati a tre vie. Una soluzione comune per gestire questo tipo di dati consiste nell'utilizzo di misture di distributioni a code pesanti. Sfortunatamente, la letterature dei dati a tre vie è ancora limitata, con le sole misture di distribuzioni t recentemente proposte per far fronte a questo problema. Quindi, in questo lavoro innanzitutto introduciamo una nuova distribuzione matrix-variata a code pesanti e successivamente la utilizziamo all'interno di un modello mistura. Il conseguente modello mistura viene quindi utilizzato su un dataset reale a scopi illustrativi.*

Key words: three-way data, model-based clustering, heavy-tailed distributions

Salvatore D. Tomarchio
Università di Catania, Dipartimento di Economia e Impresa, e-mail: daniele.tomarchio@unict.it

Antonio Punzo
Università di Catania, Dipartimento di Economia e Impresa e-mail: antonio.punzo@unict.it

Luca Bagnato
Università Cattolica del Sacro Cuore, Dipartimento di Scienze Economiche e Sociali e-mail: luca.bagnato@unicatt.it

1 Introduction

In the last years, there has been a growing interest in the application of model-based clustering techniques to three-way (matrix-variate) data (see, e.g. [3],[4]). Although there are countless examples of clustering for multivariate distributions via mixture models, the corresponding three-way literature is still little developed. This data structure arises when p variables are observed in r different occasions on n statistical units, so that the data can be organized in a three-way array consisting of the following three dimensions: variables (rows), occasions (columns) and units (layers). Therefore, each statistical unit is a $p \times r$ matrix.

Quite often real data are affected by outliers, and this also occurs in the context of three-way data. In this work we focus on mild outliers (see, [1] for further details) that, in the clustering context, are defined as points that deviate from the distribution within a cluster, but they would fit well if the overall within-cluster distribution had heavier tails [2]. A common solution for managing this type of data consists in fitting mixtures of heavy-tailed distributions. In fact, the matrix Gaussian mixture model (MXG-Ms), that can be considered the reference model, is unable to adequately model such data, because of the tails behavior of its mixing components. However, the three-way literature is still limited, and only the matrix t mixture model (MXt-Ms) has been proposed to deal with this issue. For this reason, in this work we firstly introduce a matrix-variate distribution whose tails are heavier than the matrix Gaussian distribution (MXG), and subsequently it used as component distribution in a finite mixture model. This new distribution, called herein matrix shifted exponential normal (MXSEN), generalizes the corresponding multivariate distribution recently introduced in [5]. To obtain the MXSEN, we use the well-known variance mixture model [6], as discussed in Section 2. Its use in model-based clustering via mixture models (MXSEN-Ms) is also illustrated. An application to a real data set is finally shown in Section 3, where the MXSEN-Ms are compared to MXG-Ms and MXt-Ms in terms of goodness of fit and classification performance.

2 Methodology

2.1 The matrix shifted exponential normal

A classical way to make heavier the tails of a Gaussian distribution is by means of the variance mixture model. In a three-way framework, a $p \times r$ random matrix $\mathbf{X}$ is said to follow a variance mixture model if its probability density function (pdf) is

$$f(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\theta}) = \int_{S_h} f_{\text{MXG}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma}/w,\boldsymbol{\Psi}) h(w;\boldsymbol{\theta}) dw, \tag{1}$$

where:

1. f_{MXG} is the pdf of a matrix Gaussian distribution, that is given by

$$\frac{1}{(2\pi)^{\frac{rp}{2}}|\boldsymbol{\Sigma}|^{\frac{r}{2}}|\boldsymbol{\Psi}|^{\frac{p}{2}}}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\right]\right\},$$

   with $p\times r$ mean matrix $\mathbf{M}$, $p\times p$ row covariance matrix $\boldsymbol{\Sigma}$ and $r\times r$ column covariance matrix $\boldsymbol{\Psi}$;
2. $h(w;\boldsymbol{\theta})$ is the mixing pdf, with support $S_h\subseteq\mathbb{R}_{>0}$, depending on the parameter(s) $\boldsymbol{\theta}$.

Therefore, it is a finite/continuous mixture of MXG distributions on $\boldsymbol{\Sigma}$ obtained via a convenient discrete/continuous mixing distribution with positive support. The additional parameters of the mixing distribution govern, in the variance mixture model, the deviation from normality in terms of tail weight.

If in (1), a shifted exponential distribution is chosen as mixing pdf $h(w;\boldsymbol{\theta})$, then the matrix shifted exponential normal pdf can be obtained. In formula,

$$f_{\text{MXSEN}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\theta})=\int_1^{\infty}f_{\text{MXG}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma}/w,\boldsymbol{\Psi})h_{\text{SE}}(w;\theta)dw, \tag{2}$$

where $h_{\text{SE}}(w;\theta)=\theta\exp[-\theta(w-1)]$ denotes the pdf of a shifted exponential on $S_h=(1,\infty)$. After some algebra, the pdf in (2) can be written as

$$f_{\text{MXSEN}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\theta})=(2\pi)^{-\frac{pr}{2}}|\boldsymbol{\Sigma}|^{-\frac{r}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}\theta\exp\{\theta\}\,\varphi_{\frac{pr}{2}}\left(\frac{\delta(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi})}{2}+\theta\right), \tag{3}$$

where $\delta(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi})$ denotes the Mahalanobis distance from $\mathbf{X}$ to the center $\mathbf{M}$ with respect to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, and $\varphi_m(z)$ is the Misra function [7], generalized form of the exponential integral function [8]. Similarly to the corresponding multivariate distribution [5], the closer the values of θ to 0 are, the heavier the tails of MXSEN distribution with respect to the (nested) MXG distribution become.

2.2 Finite mixtures of MXSEN distributions

Clustering and classification aim at finding and analyzing underlying group structures in the data. One common method used for clustering is model-based, and generally makes use of a G-component finite mixture model. A $p\times r$ random matrix $\mathbf{X}$ arises from a parametric finite mixture distribution if its pdf can be written as

$$p(\mathbf{X};\boldsymbol{\Theta})=\sum_{g=1}^{G}\pi_g f(\mathbf{X};\boldsymbol{\Theta}_g) \tag{4}$$

where $\pi_g>0$ are the mixing proportions, such that $\sum_{g=1}^{G}\pi_g=1$, $f(\cdot)$ are the component densities and $\boldsymbol{\Theta}$ contains all the parameters of the model. If in (4) the pdf of the MXSEN distribution in (3) is chosen for all the component densities, then we

obtain MXSEN-Ms that are more flexible with respect to MXG-Ms in modeling data affected by outliers, because of their tail behavior. Parameter estimation is carried on via an ECM algorithm, a variant of the classical expectation-maximization (EM) algorithm, which is a natural approach for maximum likelihood estimation in the context of mixture models.

3 Real data application

The real dataset herein analyzed consists of $n = 75$ study programs taught in the Italian universities. There are $G = 2$ groups in the data, i.e. $n_1 = 33$ bachelor study programs and $n_2 = 42$ master study programs. For each statistical unit the following three indicators ($p = 3$) concerning the academic careers are considered: the percentage of students that continue in the second year of the study program, the percentage of students that finish their studies within the normal duration of the study program and the percentage of dropouts after $t + 1$ years, where t is the normal duration of study program. All the variables are measured for three years ($r = 3$), implying that each study program is a 3×3 matrix.

MXSEN-Ms, and for comparison purposes also MXG-Ms and MXt-Ms, are hence fitted to the data for $G \in \{1, 2, 3\}$, and their results are shown in Table 1. The Bayesian information criterion (BIC) [9], one of the most popular (likelihood-based) model selection criterion, is then used to select the number of groups G. Furthermore, the adjusted Rand index (ARI) [10] and the percentage of missclassified units (ϵ) are adopted to evaluate the classification performance. As shown in Table 1, the

Table 1 BIC values and classification performance of the competing mixture models.

Model	G	BIC	ARI	ϵ
MXG-Ms	3	-2247.253	0.7196	0.0933%
MXt-Ms	2	-2322.750	0.7953	0.0533%
MXSEN-Ms	2	**-2323.933**	**0.8443**	**0.0400%**

BIC selects $G = 3$ groups for MXG-Ms, whereas for MXt-Ms and MXSEN-Ms the correct number of groups is chosen ($G = 2$). This is an indication of the fact that the tails of each component distribution in MXG-Ms are lighter than required, implying that this model overfits the true number of clusters. Comparable issues, but in the multivariate literature, can be found for example in [11]. In any case, according to the BIC, the MXSEN-Ms are the best model in terms of fitting. From a classification point of view, similar conclusion can be drawn. In fact, MXSEN-Ms show the highest ARI and the lowest percentage of missclassified units. Finally, the estimated values of θ for the two groups are $\theta_1 = 0.25$ for the bachelor study programs and $\theta_2 = 0.10$ for the master ones. Being both close to 0, this seems to suggest the presence of mild outliers in the data, particularly for the master programs, outlying the necessity of considering distributions with heavier tails with respect to the MXG distribution.

References

1. Ritter, G.: Robust Cluster Analysis and Variable Selection. In: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, **137**, Chapman & Hall/CRC Press, Boca Raton (2015)
2. Farcomeni, A., Punzo, A.: Robust model-based clustering with mild and gross outliers. TEST, 1–19, Springer (2019)
3. Gallaugher, M.P.B., McNicholas, P.D.: Finite mixtures of skewed matrix variate distributions. Pattern Recognition, 83–93, **80**, Elsevier (2018)
4. Gallaugher, M.P.B., McNicholas, P.D.: Mixtures of skewed matrix variate bilinear factor analyzers. Advances in Data Analysis and Classification, 1–20, Springer (2019)
5. Punzo, A., Bagnato, L.: Allometric analysis using the multivariate shifted exponential normal distribution. Biometrical Journal, In press, Wiley (2020)
6. McLachlan, G. J., Peel, D.: Finite Mixture Models. John Wiley & Sons, New York (2000)
7. Misra, R.D.: On the stability of crystal lattices. ii. Mathematical Proceedings of the Cambridge Philosophical Society, **36**(2), 173–182 (1940)
8. Abramowitz, M., Stegun, I.: Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables. In: Applied Mathematics Series, **55**, Dover Publications, New York (1965)
9. Schwarz, G: Estimating the dimension of a model. The Annals of Statistics, **6**(2), 461–464, (1978)
10. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification, **2**(1), 193–218, (1985)
11. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Statistics and Computing, **10**(4), 339–348, (2000)

A Sequential Test for the C_{pmk} Index

Un test sequenziale per l'indice C_{pmk}

Michele Scagliarini

Abstract We propose a sequential hypothesis test for the process capability index C_{pmk}. We compare the performance of the proposed test with the properties of the most used non-sequential test by performing a simulation study. The results indicate that the sequential test has best results in terms of power function and makes it possible to save a not negligible amount of sample size.

Abstract *Si propone un test sequenziale per la verifica d'ipotesi sull'indice di capacità C_{pmk}. Le proprietà statistiche del test proposto sono confrontate con quelle del test non sequenziale maggiormente usato in letteratura mediante simulazioni. I risultati indicano che il test sequenziale ha risultati migliori in termini di funzione di potenza e consente una non trascurabile riduzione dell'ampiezza campionaria.*

Key words: Brownian motion, Monte Carlo Simulation, non-central *t* distribution, power function, process capability indices, sequential test.

1 Introduction

Process capability indices are well known management tools used for assessing performance of manufacturing processes both at the production level and for business relationships. The index C_{pmk} has been designed by taking into account the process yield (meeting the manufacturing specifications) as well as the process loss (variation from the target). It is defined as $C_{pmk} = \left(d - |\mu - T|\right) \Big/ 3\sqrt{\sigma^2 + \left(\mu - T\right)^2} = \left(d/\sigma - |\xi|\right) \Big/ 3\sqrt{1+\xi^2}$ where μ is the process mean, σ is the process standard deviation, T is the process target, d=($USL-LSL$)/2 (LSL and USL are the specification limits), and ξ=(μ-T)/σ [4]. As a part of

[1] Michele Scagliarini, Department of Statistical Sciences, University of Bologna; michele.scagliarni@unibo.it

contractual agreement it is often necessary to demonstrate that the process capability index C_{pmk} meets or exceeds some particular target value, say C_{pmk0}. Such decision-making problem may be formulated as a hypothesis testing problem: $H_0 : C_{pmk} \le C_{pmk0}$ (the process is not capable) versus $H_1 : C_{pmk} > C_{pmk0}$ (the process is capable).

In this study, starting from some of the results obtained by [2], we propose a sequential test for the index C_{pmk}. Firstly, we review the most used non-sequential test for assessing whether a process is capable or not. Secondly, we analytically obtain the test statistic of the sequential test and describe in detail the testing procedure. Finally, we compare the sequential test properties with those of the non-sequential test by performing an extensive simulation study. The results show that the proposed sequential test has good power behavior and makes it possible to save sample size, which can be translated into reduced costs, time and resources.

2 Hypothesis testing on C_{pmk}

Assuming a normally distributed quality characteristic, $X \sim N\left(\mu, \sigma^2\right)$, [5] proposed a statistical test (PC-test) based on the distribution of the estimator: $\hat{C}_{pmk} = \min\left\{ \dfrac{USL - \bar{X}}{3\sqrt{S_n^2 + \left(\bar{X} - T\right)^2}}, \dfrac{\bar{X} - LSL}{3\sqrt{S_n^2 + \left(\bar{X} - T\right)^2}} \right\} = \dfrac{d - \left|\bar{X} - T\right|}{3\sqrt{S_n^2 + \left(\bar{X} - T\right)^2}}$ where, $\bar{X} = \sum_{i=1}^{n} X_i \Big/ n$ and $S_n^2 = \sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \Big/ n}$ are the maximum likelihood estimators of μ and σ^2 respectively. For the case $T = \left(USL + LSL\right)/2$ the cumulative distribution function of $\hat{C}_{pmk}$ can be expressed as

$$F_{\hat{C}_{pmk}}(x) = 1 - \int_0^{b\sqrt{n}/(1+3x)} G\left(\frac{\left(b\sqrt{n} - t\right)^2}{9x^2} - t^2 \right) \left[\phi\left(t + \xi\sqrt{n}\right) + \phi\left(t - \xi\sqrt{n}\right) \right] dt \qquad (1)$$

for $x>0$, where $b=d/\sigma$, $G(\cdot)$ is the cumulative distribution function of the chi-square distribution χ^2_{n-1}, and $\phi(\cdot)$ is the probability density function on the standard normal distribution. The decision rule is to reject $H_0 : C_{pmk} \le C_{pmk0}$ if $\hat{C}_{pmk} > c_0$ and fail to reject H_0 otherwise. Given the values of the type-I error probability α, the capability requirement C_{pmk0}, the sample size n and the parameter ξ, the critical value c_0 can be obtained by solving the equation

$\int_0^{b\sqrt{n}/(1+3c_0)} G\left(\frac{\left(b\sqrt{n}-t\right)^2}{9c_0^2}-t^2\right)\left[\phi\left(t+\xi\sqrt{n}\right)+\phi\left(t-\xi\sqrt{n}\right)\right]dt=\alpha$. It can be noted that this function depends on the parameter ξ which in real applications is usually unknown. To eliminate the need of estimating ξ [5] studied the behaviour of the critical value c_0 as a function of ξ. The authors found that the condition $|\xi|=0.5$ should provide conservative critical values and for several combinations of α, C_{pmk0}, and *n*. The results on c_0 are reported by the authors [5] in their Tables 1-5.

3 A sequential test for C_{pmk}

Under the assumption that the data came from a multivariate distribution with density function $f(x;\boldsymbol{\theta})$, [2] proposed a general sequential testing procedure for testing $H_0: h(\boldsymbol{\theta})=\mathbf{0}$ versus $H_1: h(\boldsymbol{\theta})\neq\mathbf{0}$, where $h(\boldsymbol{\theta}):\mathbb{R}^d\to\mathbb{R}^q$, with $q\le d$, is a function with first order derivative matrix denoted by $H(\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ unknown. Under the standard regularity conditions for the existence of the multivariate maximum likelihood estimators the author showed that the statistic $W_k=kh\left(\hat{\boldsymbol{\theta}}_k\right)\left[H^t(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta})H(\boldsymbol{\theta})\right]^{-1}h\left(\hat{\boldsymbol{\theta}}_k\right)^t$, where k is the sample size, $\hat{\boldsymbol{\theta}}_k$ is a consistent estimator of $\boldsymbol{\theta}$ and $I(\boldsymbol{\theta})$ is the Fisher information matrix, can be approximated by a functional of Brownian motions. The proposed test statistic is $W_k^*=kh\left(\hat{\boldsymbol{\theta}}_k\right)\left[H^t\left(\hat{\boldsymbol{\theta}}_k\right)I^{-1}\left(\hat{\boldsymbol{\theta}}_k\right)H\left(\hat{\boldsymbol{\theta}}_k\right)\right]^{-1}h\left(\hat{\boldsymbol{\theta}}_k\right)^t$ where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ [2]. The α-level sequential test procedure, truncated at the maximal allowable sample size n_0, is performed as follows:

1. for $k=2,3,...,n_0$ compute of the statistic $W_k^{*(1)}=\sqrt{k/n_0}\sqrt{W_k^*}$;
2. hypothesis H_0 is rejected the first time that $W_k^{*(1)}$ exceeds the critical value w_α;
3. if $W_k^{*(1)}$ does not exceed w_α by n_0 then do not reject H_0.

The maximal sample size n_0 can be decided on the basis of financial, ethical or statistical reasons as, for example, to achieve a desired power level. The critical value w_α, given the type I error probability α, can be obtained from [1]. Let us consider the hypothesis $H_0: C_{pmk}=C_{pmk0}$ versus $H_1: C_{pmk}\neq C_{pmk0}$ and assume that the quality characteristic is normally distributed $X\sim N\left(\mu,\sigma^2\right)$. To obtain a sequential test comparable with the PC-test we further assume the parameter ξ as known. Let us define the function $h(\boldsymbol{\theta})$ as

$h(\boldsymbol{\theta}) = \ln\left(\left(C_{pmk}\right)^2\right) - \ln\left(C_{pmk0}^2\right) = \ln\left[\left(d/\sigma - |\xi|\right)^2 \Big/ 9\left(1+\xi^2\right)C_{pmk0}^2\right]$, where $\boldsymbol{\theta} = \sigma^2$. For $C_{pmk} \geq 0$, H_0 is equivalent to $H_0 : h(\boldsymbol{\theta}) = 0$ and the alternative hypothesis is equivalent to $H_1 : h(\boldsymbol{\theta}) \neq 0$. In the case at hand the statistic W_k^* can be written as $W_k^* = kh^2\left(\hat{\boldsymbol{\theta}}_k\right)\left[H^t\left(\hat{\boldsymbol{\theta}}_k\right)I^{-1}\left(\hat{\boldsymbol{\theta}}_k\right)H\left(\hat{\boldsymbol{\theta}}_k\right)\right]^{-1}$ where $\hat{\boldsymbol{\theta}}_k = S_k^2$ with $S_k^2 = \sum_{i=1}^{k}\left(X_i - \bar{X}_k\right)^2 \Big/ k$. The function $h\left(\hat{\boldsymbol{\theta}}_k\right)$ is thus given by $h\left(\hat{\boldsymbol{\theta}}_k\right) = \ln\left[\left(d/\sqrt{S_k^2} - |\xi|\right)^2 \Big/ 9\left(1+\xi^2\right)C_{pmk0}^2\right]$ and consequently W_k^* can be written as:

$$W_k^* = \frac{k\left(\ln\left[\left(d/\sqrt{S_k^2} - |\xi|\right)^2 \Big/ 9\left(1+\xi^2\right)C_{pmk0}^2\right]\right)^2}{\dfrac{\left[d\left(\sqrt{S_k^2}\,|\xi| - d\right)\right]^2}{\left[S_k^2\left(S_k^2\xi^2 - 2d|\xi|\sqrt{S_k^2} + d^2\right)\right]^2}\,2S_k^4} \tag{2}$$

Therefore, given the value of α and the maximal allowable sample size n_0, the test is performed by computing, for k=2,3,…, n_0, the statistic $W_k^{*(1)} = \sqrt{k/n_0}\,W_k^*$.

Let n_{stop} be the first integer k=2,3,…, n_0 for which $W_k^{*(1)} > w_\alpha$: we reject H_0 if $W_{n_{stop}}^{*(1)} > w_\alpha$; we do not reject H_0 if $W_k^{*(1)}$ does not exceed w_α by n_0. In this framework n_{stop} is the stopping sample size of the test.

4 Comparisons and concluding remarks

In this Section, we investigate the performance of the sequential test and compare it with that of the PC-test. Note that the sequential test is two sided with composite alternative hypothesis ($H_1 : C_{pmk} \neq C_{pmk0}$), while the PC-test is unilateral ($H_1 : C_{pmk} > C_{pmk0}$). In order to correctly compare the statistical properties of the tests, we considered scenarios under H_1 where $C_{pmk} = C_{pmk1}$ with $C_{pmk1} > C_{pmk0}$. In such a way the sequential bilateral test with Type I error probability α can be compared with the non-sequential unilateral test with Type I error probability equal to α_u=α/2. In the research we examined several scenarios, here, for the sake of conciseness, we show the results only for the case where the capability requirement

is $C_{pmk0} = 2$. For $\alpha = 0.02(\alpha_u = 0.01)$, $C_{pmk1} = 2.1(0.05)3.0$ and $n = 50,100,200$ we analytically compute the power function of the PC-test. As far as the sequential test is concerned we considered the values of *n* as the maximal sample size n_0. Therefore, for each value of *n* and C_{pmk1} we generated 50000 replicates from a normally distributed quality characteristic. The aim of these simulations was to determine the empirical power function $\hat{\pi}_s$ of the sequential test and the average n_{avg} of the stopping sample sizes n_{stop} required for the sequential test, with maximal allowable sample size $n_0 = n$, for concluding in favor of H_1. The results are summarized in Figure 1 and Table 1. Figure 1 displays the power functions of the two tests for the values of *n* and C_{pmk1} of interest. Table 1, for some values of C_{pmk1}, contains: π_{PC} the power of the PC-test; $\hat{\pi}_s$ the estimated power of the sequential test; n_{avg} the average of the stopping sample sizes n_{stop}; $SD(n_{avg})$ the standard deviation of the stopping sample sizes n_{stop}.

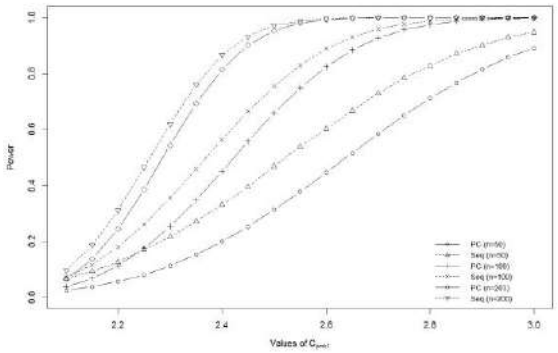


Figure 1: power functions of the PC and sequential tests

Table 1: Summary of the results under H_1 for several values of C_{pmk1} and *n*=50,100,200

C_{pmk1}	π_{PC}	$\hat{\pi}_s$	n_{avg}	$SD(n_{avg})$
		n=50		
2.20	0.0577	0.1255	48.3	5.9
2.30	0.1138	0.2159	47.1	7.3
2.40	0.1993	0.3316	45.5	8.6
		n=100		
2.20	0.1134	0.1800	95.7	11.9
2.30	0.2540	0.3565	91.0	15.9
2.40	0.4512	0.5636	84.2	19.2
		n=200		
2.20	0.2449	0.3111	184.9	29.0
2.30	0.5438	0.6194	165.2	38.4
2.40	0.8159	0.8668	140.3	40.0

From the results reported in Figure 1 and Table 1, it can be observed that the sequential test has better performance than the PC-test. As an example, when the true capability of the process is $C_{pmk} = C_{pmk1} = 2.3$ with n=50 we have that $\pi_{PC} = 0.114$ and $\hat{\pi}_s = 0.216$. When the sample size is n=100 the power functions are $\pi_{PC} = 0.254$ and $\hat{\pi}_s = 0.357$ and for n=200 are $\pi_{PC} = 0.544$ and $\hat{\pi}_s = 0.619$. By examining the averages of the final sample sizes n_{avg} the results show that the sequential test saves a not negligible amount of sample size. As an example, under $H_1 : C_{pmk} = C_{pmk1}$ with $C_{pmk1} = 2.4$ and n=100, the power of the PC-test is $\pi_{PC} = 0.4512$, while with a maximum allowable sample size equal to $n_0 = 100$ the power of the sequential test is $\hat{\pi}_s = 0.5636$, with $n_{avg} = 84.2 (\simeq 85)$. In this case the sequential test saves, on average, 15.8% of the sample size as to the PC-test. We also investigate the behavior of the sequential test under H_0. The results, obtained through a further simulation study, showed that the empirical type I error probability of the sequential test is very close to the nominal α-level. As an example when $H_0 : C_{pmk} = C_{pmk0}$ holds, with n=200 we have that the empirical type-I error probability is $\hat{\alpha} = 0.0021$. In this case $\hat{\alpha}$ is only slightly larger than the nominal level $\alpha = 0.002$.

Before concluding, we would point out some issues that deserve attention and further research. A recent study [3] argued that the PC-test might be not conservative with respect to the type I error probability since the supremum of the α-risk doesn't occur exactly at $|\xi| = 0.5$. We will study the effect of the parameter ξ on both tests in a future piece of research.

Summarizing the results shows that the proposed sequential test allows higher power levels with, on average, smaller stopping sample sizes as compared with the fixed sample size test. We consider these results as valuable, because in a highly competitive context where both cost and quality are relevant, the availability of statistical methods, which make it possible to save sampling size, can be directly translated into saved resources and reduced costs.

References

1. Borodin, A.B., Salminen, P.: Handbook of Brownian Motion-Facts and Formulae. Birkhäuser Verlag, Basel, (1996).
2. Hussein, A., Ahmed, S.E., Bhatti, S.: Sequential testing of process capability indices. Journal of Statistical Computation and Simulation, 82(2), 279--282 (2012).
3. Lepore, A. Palumbo, B., P. Castagliola: A note on decision making method for product acceptance based on process capability indices Cpk and Cpmk. European Journal of Operational Research, 267(1), 393--398 (2018).
4. Montgomery, D.: Introduction to Statistical Quality Control (6th ed). John Wiley & Sons, Hoboken, (2009).
5. Pearn, W.L., Lin, P.C.: Computer program for calculating the p-value in testing process capability index C_{mpk}. Quality and Reliability Engineering International, 18(4), 333-342 (2002).

Probability Interpretations and the Selection of the Most Effective Statistics Method

Sul Significato della Probabilità e la Scelta del Metodo

Paolo Rocchi

Abstract A recent study shows how frequentist and subjective authors often tackle the probability interpretation using philosophical arguments. This paper recalls an attempt to explain the probability meanings by means of a purely mathematical method. Two theorems spell out the testing conditions of long-term events and the single event. Doing so the theorems define the properties of the frequentist and subjective probabilities and prove these probabilities are not irreconcilable as usually credited. Lastly, we derive a precise rule for working statisticians who are called to select the most effective statistics applicable in a project.

Abstract *Questo lavoro affronta il problema dell'interpretazione della probabilità adottando un metodo puramente matematico. Due teoremi illustrano le condizioni per verificare sperimentalmente le probabilità di eventi ripetuti e dell'evento singolo. I teoremi dimostrano che ci sono due diversi modelli di probabilità e non uno soltanto. Le probabilità frequentista e soggettiva non sono inconciliabili come si crede usualmente. Infine viene derivata una regola precisa per selezionare la statistica più efficace in un dato progetto.*

Key words: Frequentism, subjectivism, testability, selection of statistics in a project.

Introduction

Presently statisticians face issues of different weight that could be placed in a scale of severity. It may be said that specialist and minute questions regarding statistical techniques lie at one side; at the other side, there are foundational issues that deal with substantial and general aspects of statistics. This paper deals with two problems of the second group, specifically:

[1] Paolo Rocchi, IBM and Università Luiss; procchi@luiss.it

1) The *selection of the most effective statistics in a project,*
2) The *probability interpretations.*

1) - Experts have the classical and Bayesian statistics at disposal which in principle, should enhance the effectiveness of working statisticians, although the two statistics are divided by various disagreements concerning the scopes to pursue, the data to gather, the techniques to use, etc. [1] There is no shared criterion to select the most effective method in a project, and an investor who pays for a statistical study and requests for the best outcome, does not obtain univocal answers. Experts take different positions: the supporters of the classical [2] and the Bayesian statistics [3] prescribe their own methodology in any situation and do not admit exceptions. Some tend to minimize the divergence and search a common way to estimation [4], to regression [5] or propose hybrid modes [6][7].

2) - The literature usually divides the main views of probability in two. *Objective interpretations* see probability as a feature of reality independent of human cognition; *epistemic interpretations* primarily ascribe probability to human knowledge or belief [8]. The frequentist and subjective (including Bayesian) models are the most popular. The first model supported by Venn, von Mises and Reichenbach, qualifies mass random events whose results are uncertain in detail but the numerical proportion in the long run with respect to a given result can be predicted [9]. The subjective school opened by de Finetti and Ramsey, and consolidated by the Bayesians such as Savage, Schlaifer and Jeffreys, claims that probability derives from an individual's personal judgment or own experience about whether a specific outcome is likely to occur [10]. The two interpretations seem to be irreconcilable from the logical point of view, yet 'dualist' authors – Ramsey, Carnap, Popper and more recently Costantini, Lewis and others – tend to accept both the perspectives. They have devised original ideas and criteria for the comprehensive adoption, however they have not reached a conclusive frame so far.

A recent bibliographical research provides evidence of how probability theorists are strongly influenced by personal opinions and philosophical arguments [11]. In consequence of this discovery, I decided to develop a purely mathematical approach to the interpretation problem. This report recalls the main results – proved and commented in [12] – which suggest a rational criterion even for solving problem 1.

A Mathematical Pathway

Let us recall a few basic assumptions. Probability is the *measure of how likely the random result e will occur*. Given a sample space Ω and an associated sigma algebra Σ, the probability *P* is a function with domain Σ that satisfies:

1. $P(e) \geq 0$ for all $e \chi \Sigma$.
2. $P(\Omega) = 1$.

3. $\left(\bigcup_{i=1}^{\infty} e_{(i)}\right) = \sum_{i=1}^{\infty} P(e_{(i)})$ if pairwise disjoint $e_{(1)}, e_{(2)}, e_{(3)}, \ldots \chi \Sigma$ (1)

The possibility of proving or rejecting a hypothesis by means of experimental trials is inherent to genuine science, thus the probability established in abstract with axioms (1) needs to be tested. It is necessary to discuss how and when the quantity $P(e)$ can be checked. The ensuing theorems illustrate the relations between the experimental frequency F and the probability P of a random event in two distinct situations that are the sequel of repeated trials (e_∞) and the single trial (e_1) in the order.

1.1 Theorems of Large Numbers and a Single Number

Suppose e is an i.i.d variable, the *theorem of large numbers* (TLN) states that as the number of trials grows, the relative frequency almost surely gets closer to $P(e_\infty)$

$$F\left(e_n\right) \xrightarrow{a.s.} P(e_\infty), \qquad as \quad n \to \infty. \tag{2}$$

The *theorem of a single number* (TSN) demonstrates that the relative frequency unfits with the probability of a single random experiment

$$F(e_1) \neq P(e_1), \qquad n = 1. \tag{3}$$

The theorem of large numbers has a conceptual link with the law of large numbers (LLN), but a neat difference divides them.

LLN expresses the convergence of empirical random data toward an expected value. Émile Borel developed a special strong version of LLN where q is the number of empirical occurrences and P is a purely numerical value

$$\frac{q}{n} \xrightarrow{a.s.} P, \qquad as \quad n \to \infty. \tag{4}$$

The statement (4) describes the property of P whereas (2) describes the property of $P(e_\infty)$ because TLN is a theorem about testing and not a general law. In other words, LLN deals with the probability P in abstract, TLN establishes the testing condition for $P(e_\infty)$ that is a particular kind of probability. TLN proves that – at least in principle – the probability of repeated events (called *collective* by von Mises and *series* by Venn) can be verified by means of experiments, hence $P(e_\infty)$ is a parameter that exists in the real world, the *frequency probability* $P(e_\infty)$ is *an authentic physical quantity*.

TSN proves how one cannot test the probability of a single random event. It is not a question of instruments or experimental setting; never ever one can corroborate $P(e_1)$ and therefore it *is not a physical quantity*. This conclusion perfectly matches with the famous aphorism by de Finetti: “Probability does not exist”.

TSN implies that scientists should reject $P(e_1)$ because it is out of control; instead statisticians and even common people are concerned with the probability of a single result. Subjectivists and Bayesians exploit the *semantic value* of $P(e_1)$ and circumvent this obstacle caused by (3). What does 'semantic value' exactly mean?

Semiotics is the science of signs and teaches us that numbers, words, symbols etc. are items of information [13] and as such they can convey meanings. The number $P(e_1)$ does not have any physical significance nonetheless it does not lose the capability of conveying significance. Subjectivists and Bayesians use it to express a *personal credence about the occurrence of* (e_1), namely $P(e_1)$ is a *subjective probability*.

I recall in concise terms the criticism of writers against this model [14]:

1. Subjective probability expresses the personal belief which in principle can be affected by the variety of personal convictions held by an individual. The suspicions of arbitrariness have been raised since the early beginnings.

2. The betting scheme used to ensure the consistent value of $P(e_1)$ sounds somewhat strange in scientific and engineering sectors.

3. Testing is a key criterion for the scientific method but in principle subjective probability is alien to experimental validation. This turns out to be repellent to the science which investigates objective situations and strives for results independent of the observer.

In the light of statement (3), the disapprovals **1, 2** and **3** take completely different significance. In fact, TSN proves that $P(e_1)$ has no physical meaning and remark **3** can but confirm this essential property. TSN leads us to grasp how the subjectivists and Bayesians masters conducted an intelligent plan of action to recycle $P(e_1)$ as epistemic probability. Keynes teaches us how epistemic probabilities are not so easy to assess and to check, hence the manoeuvre of subjectivists and Bayesians necessarily presents the weak sides **1** and **2**. The landscape emerging from TLN and TSN shows how the most critical annotations against the subjective model decrease in importance. It may be said that the above listed remarks and even other negative commentaries give details about the price paid to reuse $P(e_1)$ that otherwise should be rejected from the scientific domain.

The Best Statistics?

1.1 Theorem of Compatibility

The results achieved by means of the theorems of large numbers and a single number can be summarized in the following way:

$$\textit{Frequentist Probability}: P(e_n) \qquad n \to \infty \qquad (7a)$$

$$\textit{Subjective Probability}: P(e_n) \qquad n = 1 \qquad (7b)$$

The *theorem of compatibility* (TC) proves that the frequentist and subjective models do not contradict

$$[P(\infty)] \text{ OR } [P(e_1)]. \quad (8)$$

The hypotheses (7a) and (7b) take two disjoint intervals for n, consequently the frequentist and subjective probabilities apply to distant situations; $P(\infty)$ and $P(e_1)$ do not overlap and thus do not contradict even if they widely differ.

The reader can note how the assumptions (7a) and (7b) play a key role; conversely, conventional constructions who do not specify the hypotheses necessary to apply each model, lead to irreconcilable positions. The founders of the frequentist and subjective schools sustain diverging models but share the same philosophy; they assume that probability has only one meaning. In the place of accurate hypotheses (7.a) and (7.b) they assume there is a sole 'authentic' probability model. This unproven precondition constitutes a sort of hidden axiom raising endless debates and preventing researchers from attacking the core of the interpretation issue. TC disproves this dogmatic assumption and shows how intrinsically probability is multifold.

1.2 How to Select the Method

TLN, TSN and TC regulate the exercise of probability and lead to the criterion for picking the most effective statistics in a given project. Specifically, one reasonably can infer the following rules:

- **a.** *If one means to investigate the long-term event (7a), the one must resort to use the classical statistics.*
- **b.** *If one means to investigate a single event (7b), he must adopt the Bayesian statistics.*

The rules **a** and **b** do have the scope of fixing the superiority of a statistical method over the other; they merely establish a criterion to use the classical and Bayesian statistics in a given project.

Assumption (7a) is consistent with the classical statistical inference that makes propositions about a population, using data drawn from that population with some form of sampling. Bayesian statistics teaches us to update our beliefs in the evidence of new data when we have to forecast a future single event. Note how the Bayesian procedures are not confined to a single observation. When a Bayesian applies to a sequel of repeated events, his conclusions regard each individual case; he addresses the situations one by one.

Comments and Conclusions

A - Gauss, Hilbert and other eminent mathematicians consider simplicity as a virtue of theoretical works, a sign of elegance and not a defect. Other experts claim

that simplicity is a characteristic of mathematics because it is easier to gain enlightenment from a simple proof compared to a complex proof. The theorems presented here sound very simple from the mathematical viewpoint; they offer concise illustrations which should not be deemed as a fault.

B - In terms of methodology, the present research suggests an innovative pathway since it examines the nature of probability by means of the mathematical method and rejects any philosophical approach.

C - Rules **a** and **b** provide a precise guideline to statisticians who are called for selecting the most appropriate methods in a project. Currently writers examine the pros and cons of classical and Bayesian methods using empirical and sometimes personal criteria, instead rules **a** and **b** descend from theorems and not from individuals' ideas.

D - The theorems presented in this paper provide new answers to knotty controversies, specifically:

D.1 TLN and TSN tackle the *problem of testability* that is a typical of science and give precise substance to the probability meanings issue.

D.2 TLN and TSN demonstrate that there are two different models of probability: $P(e_\infty)$ and $P(e_1)$, and not only one.

D.3 TSN enables us to grasp how $P(e_1)$ is reused as subjective probability and makes explicit the intellectual origin of subjectivism.

D.4 TC proves that $P(e_\infty)$ and $P(e_1)$ are not irreconcilable.

References

1. Hand D.J. - Discussion of "Bayesian Models and Methods in Public Policy and Government Settings" by S. E. Fienberg - Statistical Science, 26, no. 2: 227-30 (2011).
2. Bulmer M.G. - Principles of Statistics - Dover Publications (1979).
3. Bolstad W.M.- Introduction to Bayesian Statistics - John Wiley and Sons (2004).
4. Samaniego F.J. - A Comparison of the Bayesian and Frequentist Approaches to Estimation - Springer (2010).
5. Wakefield J. - Bayesian and Frequentist Regression Methods - Springer (2013).
6. Chen L., Yuan A., Liu A., Chen G. - Longitudinal data analysis using Bayesian-frequentist hybrid random effects model - Journal of Applied Statistics, 41(9): 2001-2010 (2014).
7. Raue A., Kreutz C., Theis F.J., Timmer J. (2013) - Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability - Philosophical Transactions of the Royal Society A, 371: 1-18 (1984).
8. Gillies D. - Philosophical Theories of Probability - Routledge (2000).
9. Friedl H., Hörmann S. - Frequentist Probability Theory - In Handbook of Probability: Theory and Applications, Sage Publications: 15-34 (2008).
10. De Finetti B. - Teoria della Probabilità - Einaudi (1970). Translated as Theory of Probability, John Wiley and Sons (1974).
11. Rocchi P. - Four foundational questions in probability theory and statistics - Physics Essays, vol. 30, no. 3: 314-321 (2017) https://arxiv.org/abs/1901.03876
12. Rocchi P. - Janus-faced Probability - Springer (2014).
13. Rocchi P. - Logic of Analog and Digital Machines - 2nd revised ed., Nova Science Publishers, 2012.
14. Kyburg H. - Subjective probability: Criticisms, reflections, and problems - Journal of Philosophical Logic, 7(1): 157-180 (1978).

Robust Composite Inference

Inferenza composita robusta

Mameli Valentina, Musio Monica, Ruli Erlis, Ventura Laura

Abstract Likelihood analyses can be difficult to perform both when the full likelihood is too complex or even impossible to specify and when robustness with respect to data or to model misspecifications is required. To deal both with complex models and robustness, in this paper we propose to resort to a composite robust scoring rule. In particular, we focus on the Tsallis score, which allows us to deal with model misspecifications. Theory and an application are discussed.

Abstract *Le procedure di verosimiglianza possono essere complicate sia quando la verosimiglianza è troppo complessa, o impossibile da specificare, sia quando è richiesta robustezza. Per trattare contemporaneamente sia modelli complessi sia robustezza, si propone di utilizzare una regola di punteggio composita robusta. In particolare, si considera la regola di punteggio di Tsallis, che consente di trattare modelli non correttamente specificati. Teoria e una applicazione sono discusse.*

Key words: *B*-robustness; Complex models; Composite score; Tsallis score.

1 Introduction

Suppose we wish to fit a parametric model $\{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$, based on a random sample $(y_1, \cdots, y_n)$ of size n. The most popular tool for inference on θ is the log-likelihood function $\ell(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$, where $f(\cdot;\theta)$ is the density associated to F_θ. For instance, the maximum likelihood estimator (MLE) is defined as $\hat{\theta} = \arg\max_\theta \ell(\theta)$, and confidence regions with nominal coverage $1-\alpha$ can be con-

Mameli V.
Università degli Studi di Udine, e-mail: valentina.mameli@uniud.it

Musio M.
Università degli Studi di Cagliari, e-mail: mmusio@unica.it

Ruli E., Ventura L.
Università degli Studi di Padova, e-mail: ruli@stat.unipd.it, ventura@stat.unipd.it

structed as $\{\theta : W(\theta) \leq \chi^2_{p;1-\alpha}\}$, where $W(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\}$ is the likelihood ratio statistic and $\chi^2_{p;1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2_p distribution.

However, likelihood inference can be difficult to perform when the full likelihood is too complex or even impossible to specify and, moreover, can behave quite poorly under slight model misspecifications. A possible solution is to resort to suitable pseudo-likelihood functions as a replacement of the full likelihood. Useful examples are given by *composite likelihoods* (see, e.g., [10]), when the full likelihood is computationally cumbersome or when a fully specified model is out of reach, or by *scoring rules* (see, e.g., [2], [3], [4], and references therein), to deal with complex models or model misspecifications.

The aim of this paper is to propose a general approach for dealing with complex models and robustness simultaneously. The proposal is to replace the log-likelihood function with the *scoring rule*

$$S(\theta) = \sum_{i=1}^{n} S(y_i; \theta), \tag{1}$$

where $S(y;\theta)$ is a proper scoring rule (see [4]). The scoring rule (1) yields an unbiased estimating equation for any statistical model and thus it forms a special case of M-estimation [6]. More precisely, we propose to mix the Tsallis scoring rule ([1],[9]), which presents robustness properties, with the composite likelihood, which is still a special case of scoring rule. An example is discussed.

2 Background on proper scoring rules

To deal with complex models or model misspecifications, useful surrogate likelihoods are given by proper scoring rules (see [3] and [4], and references therein). A scoring rule is a loss function which is used to measure the quality of a given probability distribution F_θ for a random variable Y, in view of the result y of Y.

An important example of proper scoring rules is the log-score, which is defined as $S^L(y;\theta) = -\log f(y;\theta)$ [5] and which corresponds to minus the log-likelihood function. In this paper we focus on the Tsallis score ([1], [9]), given by

$$S^T(y;\theta) = (\gamma - 1)\int f(x;\theta)^\gamma d\mu(x) - \gamma f(y;\theta)^{\gamma-1}, \quad \gamma > 1. \tag{2}$$

The Tsallis score gives in general robust procedures [4], and the parameter γ is a trade-off between efficiency and robustness.

Proper scoring rules can also be extended to the case of a random vector in analogy with composite likelihoods [10]. Let $\{Y_k\}$ be a set of marginal or conditional variables with associated proper scoring rule S_k. A proper scoring rule for the random vector Y may be defined as

$$S(\mathbf{y}; F) = \sum_k S_k(\mathbf{y}_k; F_k), \tag{3}$$

where $Y_k \sim F_k$ when $Y \sim F$, and $\mathbf{y}$ and $\mathbf{y}_k$ are the values assumed by Y and Y_k, respectively. Scoring rules (3) are called *composite scoring rules* [3]. Note that when each S_k is the log-score, equation (3) is a negative composite log-likelihood.

The validity of inference about θ using scoring rules can be justified by invoking the general theory of unbiased M-estimating functions, as discussed in [4]. The class of M-estimators is broad and includes a variety of well-known estimators. For example it includes the MLE, the maximum composite likelihood estimator [10], and robust estimators [6] among others.

3 Robust composite scoring rules

Consider independent observations y_i of a random vector $Y_i = (Y_{i1}, \ldots, Y_{iq})$, $i = 1, \ldots, n$, where Y_i has density $f(y_i; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, $y_i \in \mathcal{Y}$. Hereafter, regularity conditions as detailed e.g. in [7] will be assumed. Since the Tsallis score (2) gives in general robust procedures, in this section we focus on this particular scoring rule. The robust composite scoring rule can be defined through K marginal or conditional events $A_k(y_i)$ on $\mathcal{Y}$, giving scoring rule contributions $S_k^T(y_i; \theta) = S^T(A_k(y_i); \theta)$. Then, the robust composite scoring rule is defined as

$$S_c^T(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k S_k^T(y_i; \theta), \tag{4}$$

where w_k, $k = 1, \ldots, K$, are positive weights. If the events are defined in terms of pairs of observations (y_{ir}, y_{is}), from the bivariate marginal density $f_{rs}(y_{ir}, y_{is}; \theta)$, $r, s = 1, \ldots, q$, $r \neq s$, the robust composite scoring rule may be called the robust pairwise scoring rule and is denoted by

$$S_p^T(\theta) = \sum_{i=1}^{n} \sum_{r,s=1, r \neq s}^{q} w_{rs} S^T(y_{ir}, y_{is}; \theta). \tag{5}$$

The validity of inference based on $S_c^T(\theta)$ and $S_p^T(\theta)$ can be justified by using the general theory of scoring rules [4]. Given a robust composite scoring rule $S_*^T(y; \theta)$, where $*$ can be c or p, let us denote by $S_*^T(\theta) = \sum_{i=1}^{n} S_*^T(y_i; \theta)$ the total empirical score (4) or (5). Moreover, let $s_*^T(y; \theta)$ be the gradient vector of $S_*^T(y; \theta)$ with respect to θ, i.e. $s_*^T(y; \theta) = \partial S_*^T(y; \theta) / \partial \theta$. Under broad regularity conditions, the robust composite scoring rule estimator $\tilde{\theta}$ is the solution of the unbiased estimating equation $s_*^T(\theta) = \sum_{i=1}^{n} s_*^T(y_i; \theta) = 0$ and it is asymptotically normal, with mean θ and covariance matrix $V(\theta) = K(\theta)^{-1} J(\theta) K(\theta)^{-\mathsf{T}}$, with $K(\theta) = E_\theta(\partial s_*^T(\theta) / \partial \theta^{\mathsf{T}})$ and $J(\theta) = E_\theta(s_*^T(\theta) s_*^T(\theta)^{\mathsf{T}})$ sensitivity and variability matrices, respectively. The matrix $G(\theta) = V(\theta)^{-1}$ is the Godambe information and its form is due to the failure of the information identity, i.e. $K(\theta) \neq J(\theta)$.

From the general theory of M-estimators, the influence function (IF) of $\tilde{\theta}$ is $IF(y; \tilde{\theta}) = K(\theta)^{-1} s_*^T(y; \theta)$. The estimator $\tilde{\theta}$ is B-robust if and only if $s_*^T(y; \theta)$ is bounded in y [6]. Note that, in general, the MLE has unbounded IF, i.e. it is not

B-robust. Sufficient conditions for the robustness of the Tsallis score are discussed in [1] and [4].

Asymptotic inference on the parameter θ can be based on the Wald-type statistic

$$w_S(\theta) = (\tilde{\theta} - \theta)^{\top} V(\tilde{\theta})^{-1} (\tilde{\theta} - \theta), \tag{6}$$

which has an asymptotic chi-square distribution with d degrees of freedom. In contrast, the asymptotic distribution of the scoring rule ratio statistic $W_S(\theta) = 2\left\{S_*^T(\theta) - S_*^T(\tilde{\theta})\right\}$ is a linear combination of independent chi-square random variables, i.e. $W_S(\theta) \dot{\sim} \sum_{j=1}^{d} \mu_j Z_j^2$, with $\mu_1, \ldots, \mu_d$ eigenvalues of $J(\theta)K(\theta)^{-1}$ and $Z_1, \ldots, Z_d$ independent $N(0,1)$ variables. Adjustments of the scoring rule ratio statistic have received consideration in [4], extending results for composite likelihoods [8]. In particular, using the rescaling factor $A(\theta) = \frac{s_*^T(\theta)^{\top} J(\theta)^{-1} s_*^T(\theta)}{s_*^T(\theta)^{\top} K(\theta)^{-1} s_*^T(\theta)}$, we have [4]

$$W_S^{adj}(\theta) = A(\theta) W_S(\theta) \dot{\sim} \chi_d^2. \tag{7}$$

Analogous limiting results can be shown to hold for inference on the scalar parameter ψ. With the partition $\theta = (\psi, \lambda)$, the scoring rule estimating function is similarly partitioned as $s_*^T(y;\theta) = (s_{*\psi}^T(y;\theta), s_{*\lambda}^T(y;\theta))$, where $s_{*\psi}^T(y;\theta) = (\partial/\partial\psi) S_*^T(y;\theta)$ and $s_{*\lambda}^T(y;\theta) = (\partial/\partial\lambda) S_*^T(y;\theta)$. Moreover, consider the corresponding partitions for K and K^{-1}, and similarly for G and G^{-1}. Finally, let $\tilde{\theta}_\psi$ be the constrained scoring rule estimate of θ for fixed ψ, and let $\tilde{\psi}$ be the ψ component of $\tilde{\theta}_\psi$. A profile scoring rule Wald-type statistic for the ψ component may be defined as $w_{Sp}(\psi) = (\tilde{\psi} - \psi)(G^{\psi\psi})^{-1/2}$, and it has an asymptotic $N(0,1)$ null distribution. Moreover, we have that the asymptotic distribution of the profile scoring rule ratio statistic $W_{Sp}(\psi) = 2\left(S_*^T(\tilde{\theta}_\psi) - S_*^T(\tilde{\theta})\right)$ is $\nu\chi_1^2$, where $\nu = (K^{\psi\psi})^{-1} G^{\psi\psi}$. In view of this, an adjusted profile scoring rule ratio statistic can be computed as [4]

$$W_{Sp}^{adj}(\psi) = \frac{W_{Sp}(\psi)}{\nu} \dot{\sim} \chi_1^2. \tag{8}$$

The adjusted profile scoring rule root can is $r_{Sp}(\psi) = \text{sign}(\tilde{\psi} - \psi) W_{Sp}^{adj}(\psi)^{1/2}$, which has an asymptotic standard normal distribution.

4 An example

We discuss inference on the correlation coefficient of an equi-correlated multivariate normal distribution. This classical example is considered, among others, by [8] and [4]; see also references teherein.

Let $Y = (y_1, \ldots, y_q)$ be a $q-$dimensional random vector with mean μ and covariance matrix Σ, with $\Sigma_{rr} = \sigma^2$ and $\Sigma_{rs} = \rho\sigma^2$ for $r \neq s$, with $r, s = 1, \ldots, q$ and $\rho \in (-1/(q-1), 1)$. The log-likelihood for $\theta = (\mu, \sigma^2, \rho)$ is given by

$$\ell(\theta) = -\frac{nq}{2}\log\sigma^2 - \frac{n(q-1)}{2}\log(1-\rho) - \frac{n}{2}\log 1+\rho(q-1) - \frac{SSW}{2\sigma^2(1-\rho)} - \frac{qSSB+nq(\bar{y}-\mu)^2}{2\sigma^2(1+\rho(q-1))}, \tag{9}$$

where $SSW = \sum_{i=1}^{n}\sum_{r=1}^{q}(y_{ir}-\bar{y}_i)^2$, $SSB = \sum_{i=1}^{n}(\bar{y}_i-\bar{y})^2$, $\bar{y}_i = \sum_{r=1}^{q} y_{ir}/q$ and $\bar{y} = \sum_{i=1}^{n}\sum_{r=1}^{q} y_{ir}/(nq)$. The Tsallis score is

$$S^T(y;\theta) = \frac{(\gamma-1)}{(2\pi)^{\gamma q/2}(\det\Sigma)^{\gamma/2}}\int_{\mathbb{R}^q} e^{-\frac{\gamma}{2}(y-\mu)\Sigma^{-1}(y-\mu)^{\mathsf{T}}}dy + \tag{10}$$

$$- \frac{\gamma}{(2\pi)^{(\gamma-1)q/2}(\det\Sigma)^{(\gamma-1)/2}} e^{-\frac{\gamma-1}{2}(y-\mu)\Sigma^{-1}(y-\mu)^{\mathsf{T}}}. \tag{11}$$

The first order consecutive pairwise Tsallis is obtained by compounding together the bivariate density for pairs of observations (y_{ir}, y_{is}), for $i = 1,\ldots,n$,

$$f(y_{ir},y_{is};\theta) = \frac{e^{-\frac{1}{2\sigma^2(1-\rho^2)}\{(y_{ir}-\mu)^2+(y_{is}-\mu)^2-2\rho(y_{ir}-\mu)(y_{is}-\mu)\}}}{2\pi(\sigma^4(1-\rho^2))^{1/2}}. \tag{12}$$

The pairwise Tsallis scoring rule (with all weights equal to 1) is then equal to $S_p^T(\theta) = \sum_{i=1}^{n}\sum_{r,s=1,r\neq s}^{q} S_p^T(y_{ir},y_{is};\theta)$, where

$$S_p^T(y_{ir},y_{is};\theta) = \frac{\left((\gamma-1)(\gamma)^{-1} - \gamma e^{-\frac{(\gamma-1)}{2\sigma^2(1-\rho^2)}\{(y_{ir}-\mu)^2+(y_{is}-\mu)^2-2\rho(y_{ir}-\mu)(y_{is}-\mu)\}}\right)}{(2\pi)^{(\gamma-1)}(\sigma^4(1-\rho^2))^{(\gamma-1)/2}}.$$

For comparison we also consider the pairwise log-likelihood for $\theta = (\mu,\sigma,\rho)$, given by (see [8])

$$S_p(\theta) = -\frac{nq(q-1)}{2}\log\sigma^2 - \frac{nq(q-1)}{4}\log(1-\rho^2) - \frac{q-1+\rho}{2\sigma^2(1-\rho^2)}SSW + - \frac{q(q-1)SSB+nq(q-1)(\bar{y}-\mu)^2}{2\sigma^2(1+\rho)}.$$

We ran a simulation experiment, with $n = 10$ and $q = 3$, in order to assess the accuracy of the pairwise estimators, when data are generated both from the equicorrelated normal model with parameter $\theta = (0, 0.5, 0.95)$, and also from the equicorrelated normal model with parameter $\theta = (0, 0.5, 0.95)$ contaminated by a uniform distribution on the interval $[1, 10]$. We compare the MLE ($\hat{\theta}$), the pairwise MLE ($\hat{\theta}_p$), the Tsallis estimate based both on the full scoring rule ($\hat{\theta}_T$) and on the pairwise scoring rule ($\hat{\theta}_{pT}$). Table 1 gives the results of the simulation study based on 10000 Monte Carlo trials. Note that, under the central model, the four estimators behave quite similarly. Under the contaminated model, the two estimators based on the Tsallis scoring rule show a reasonable behaviour, while both the estimators

based on the likelihood exhibit poor performances. The pairwise Tsallis estimator shows better performances with respect to the one based on the full scoring rule.

	μ	σ	ρ
$\hat{\theta}$	-0.00459	0.4638	0.9317
	(0.0016)	(0.0011)	(0.0004)
$\hat{\theta}_p$	-0.00459	0.4638	0.9317
	(0.0016)	(0.0011)	(0.0004)
$\hat{\theta}_T$	-0.0046	0.5310	0.9311
	(0.0016)	(0.0012)	(0.0005)
$\hat{\theta}_{pT}$	-0.0046	0.4685	0.9306
	(0.0016)	(0.0011)	(0.0005)

	μ	σ	ρ
$\hat{\theta}$	0.5548	1.718	0.9904
	(0.003)	(0.008)	(0.0001)
$\hat{\theta}_p$	0.5548	1.718	0.9904
	(0.003)	(0.008)	(0.0001)
$\hat{\theta}_T$	0.0037	0.6116	0.9353
	(0.0017)	(0.0022)	(0.0006)
$\hat{\theta}_{pT}$	0.0246	0.5054	0.9314
	(0.0017)	(0.0016)	(0.0006)

Table 1 Point estimates of μ, σ and ρ, with $\gamma = 1.2$: bias (and mean square error) under the central model (left) and the contaminated model (right).

5 Final remarks

In this contribution we illustrate how to derive a composite robust scoring rule, in order to deal both with complex models and with model misspecifications. The behaviour of the proposed composite robust scoring rule is currently under study in other examples.

Acknowledgement. This research work was partially supported by University of Padova (BIRD197903), by PRIN 2015 (grant 2015EASZFS_003) and by the project STAGE of the Fondazione di Sardegna and Regione Autonoma di Sardegna.

References

1. Basu, A., Harris, I.R., Hjort, N. L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**, 549–559 (1998)
2. Dawid, A.P.: Probability forecasting. In: Kotz, S., Johnson, N.L., Read, C.B. (eds.) Encyclopedia of Statistical Sciences, **7**, pp. 210–218. Wiley-Interscience (1986)
3. Dawid, A.P., Musio, M.: Theory and applications of proper scoring rules. Metron **72**, 169–183 (2014)
4. Dawid, A.P., Musio, M., Ventura, L.: Minimum scoring rule inference. Scand. J. Statist. **43**, 123–138 (2016)
5. Good, I.J.: Rational decisions. J. Roy. Statist. Soc. B **14**, 107–114 (1952)
6. Huber, P.J., Ronchetti, E.M.: Robust Statistics. John Wiley and Sons, New York (2009)
7. Mameli, V., Ventura, L.: Higher-order asymptotics for scoring rules. J. Statist. Plann. Inf. **165**, 13–26 (2015)
8. Pace, L., Salvan, A., Sartori, N.: Adjusting composite likelihood ratio statistics. Statist. Sinica **21**, 129–148 (2011)
9. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics.J. Statist. Physics **52**, 479–487 (1988)
10. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Statist. Sinica **21**, 5–42 (2011)

Statistical hypothesis testing within the Generalized Error Distribution: Comparing the behavior of some nonparametric techniques

Verifica d'ipotesi nell'ambito della Distribuzione degli Errori Generalizzata: Confronto dell'efficacia di alcuni test non parametrici

Massimiliano Giacalone and Demetrio Panarello[1]

Abstract This paper's goal is to deal with the issue of hypothesis testing when the errors are assumed to be distributed according to a Generalized Error Distribution. Focus is given to the central tendency parameters, validating the suitability of nonparametric methods in this respect. The present work describes a simulation study aimed at assessing the validity of the Van der Waerden and Wilcoxon tests in the case of data coming from a G.E.D.; in order to compare the statistical power of such tests, we proceed to calculate the usual empirical significance level. The use of test statistics obtained by means of a Van der Waerden test generalized by considering the G.E.D.'s shape parameter provides better results, in terms of statistical power, compared to the Wilcoxon and the classic Van der Waerden tests.

Abstract *Questo lavoro si propone di trattare il problema della verifica di ipotesi nel caso in cui si supponga che gli errori si distribuiscano secondo una Generalized Error Distribution. Si focalizza l'attenzione sui parametri di tendenza centrale, verificando l'opportunità dell'uso di metodi non parametrici. Viene descritto uno studio di simulazione per stabilire la validità dei test di Van der Waerden e Wilcoxon nel caso di dati provenienti da una G.E.D.; per il confronto della potenza di tali test, si procede al calcolo classico del livello di significatività empirico. L'uso di statistiche test ottenute da un test di Van der Waerden generalizzato tenendo conto del parametro di forma della G.E.D. fornisce risultati migliori, in termini di potenza statistica, rispetto al test di Wilcoxon e al Van der Waerden classico.*

Key words: Nonparametric tests, Statistical power, Generalized Error Distribution, Wilcoxon, Van der Waerden

[1] Massimiliano Giacalone, University of Naples "Federico II"; massimiliano.giacalone@unina.it
Demetrio Panarello, University of Udine; demetrio.panarello@uniud.it

1 Introduction

There is a great need for determining how well a model fits the sample data, in various research contexts.

Historically, the first problem of model fitting dates back to the attempt by Galileo Galilei, in the 1600s, of predicting the orbit of a celestial body, with the aim of building a mathematical model and estimate its parameters [1]. In 1806, Carl Friedrich Gauss developed the first method for studying accidental errors, based on the principle of arithmetic mean, known as the Least Squares Method. The axiomatic setting that the Gauss method is based on was slightly modified by the Soviet astronomer Mikhail Subbotin [5], who obtained a more general error distribution law, which best lends itself to describing the majority of phenomena that are observable in reality.

It is very common in Statistics to assume normality of the data, even though this hypothesis is not always fully supportable. Indeed, the normality assumption is definitely restrictive, as it limits the scope of operational situations in which statistical hypothesis testing tools can be used. Nonparametric techniques, on the other hand, are "distribution-free" and can be applied independent of the statistical nature of the data. Indeed, as the actual distribution of any sample is, at best, only partly known, nonparametric tests, despite being less statistically powerful (i.e., they are more likely to lead to Type II errors, thus making the null hypothesis more likely to be accepted), are in fact more appropriate than the parametric ones as they do not require knowledge about the distribution's density function. Nonparametric techniques date back to the introduction of the X^2 test by Karl Pearson, in 1900.

The aim of this paper is to deal with the issue of nonparametric hypothesis testing when assuming that the errors are distributed according to the Subbotin distribution, also known as Generalized Error Distribution (G.E.D.), which constitutes a valid generalization of the Gaussianity assumption [2]. Its density function is:

$$f_p(z) = \frac{1}{2p^{\frac{1}{p}}\sigma^p\Gamma(1+\frac{1}{p})}\exp\left(-\frac{1}{p}\left|\frac{z-M_p}{\sigma_p}\right|^p\right),$$

where $M_p = E(z)$ is the location parameter, $\sigma_p = E\left(|z-M_p|\right)^{\frac{1}{p}}$ is the scale parameter, and p>0 is the shape parameter.

Literature has proposed several nonparametric procedures for testing hypotheses about the centrality parameter. In the following Section, we will introduce three nonparametric tests (Wilcoxon, Van der Waerden, and generalized Van der Waerden test), with the aim of comparing the behavior of such procedures when the sample observations are extracted from a random variable that is distributed according to a law belonging to the Generalized Error Distribution density family.

2 Nonparametric mean hypothesis testing

Let $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$ denote two independent samples extracted from continuous populations. Under the null hypothesis H_0 that the distribution functions coincide, while remaining unspecified, we can count on a single set of random observations, of cardinality $N = n + m$, extracted from the common – though unknown – population. The alternative hypothesis is that the populations are of the same shape but with different measures of central tendency.

Nonparametric tests of this kind are based on ranks, since the ranks of the X's, compared to those of the Y's, contain a piece of information regarding the relative measure of the population's median. The ranks of the X sample observations, in the ordered combination of the two considered samples, would generally be larger than the Y ranks if the population's median of X exceeds the population's median of Y.

Wilcoxon [8] proposed a test in which the unilateral alternative hypothesis $H_1{:}\Theta < 0$ on the centrality parameter is accepted if the rank sum of the X's exceeds the acceptance region, determined based on sample size and significance level; $H_1{:}\Theta > 0$ is accepted if this rank sum is below the acceptance region; and the bilateral hypothesis $H_1{:}\Theta \neq 0$ is accepted when the rank sum is between its smallest and highest values. Therefore, the test can be expressed as:

$$W_N = \sum_{i=1}^{N} iZ_i ,$$

where the Zi's are random variables indicating whether the distributions belong to one sample or the other.

Mean and variance of W_N, under the null hypothesis that refers to the event in which the two distributions are equal, can be calculated as follows:

$$E(W_N) = \frac{m(N+1)}{2} \qquad var(W_N) = \frac{mn(N+1)}{12} .$$

The Van der Waerden [7] test combines the concept of normal scores and the rank sum criterion used in the Wilcoxon test [3]. The classic version of this test allows us to make a generalization of it, suitable when the samples are generated by a Generalized Error Distribution. The X_N test by Van der Waerden is:

$$X_N = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right) Z_i ,$$

where Φ denotes the cumulative distribution function of the standard normal.

The X_N statistics is symmetric with respect to zero and its variance is:

$$var(X_N) = mn \sum_{i=1}^{N} \frac{\left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2}{N(N-1)} .$$

For large samples of N>50, the distribution under the null hypothesis of the standardized X_N can be approximated by the standard normal. Starting from this consideration, it is possible to put, within the computation of the test statistic,

weights reflecting the Generalized Error Distribution percentiles. Such percentiles are known and tabulated, so that this combination feels somehow natural.

That said, it can be interesting to investigate the performance of the test statistic

$$X_{NP} = \sum_{i=1}^{N} \Psi^{-1}\left(\frac{i}{N+1}\right) Z_i$$

,

where Ψ denotes the Generalized Error Distribution's cumulative distribution function, in the event in which the sample observations belong to a G.E.D. population with known shape parameter.

It is possible to notice that the last equation is very similar to the one of the classic Van der Waerden test, except for the rank weighting system, which employs the G.E.D. percentiles rather than the Gaussian ones.

3 Statistical power estimation

As proposed by [6], we estimate the tests' statistical power by means of simulations made on very small samples (N=12, n=m=6) and a dense d_j shifts' set. The simulations are repeated for three different nominal significance levels. We assume p and σ to be known; for the sake of simplicity, we will assume σ to be always equal to 1. All the calculations have been implemented and processed in the statistical environment R, with the aid of the normalp package [4]. From the results (Table 1), it is possible to notice that, especially as the significance level increases, the statistical power of the generalized Van der Waerden test is always equal or higher than the one of the classic Van der Waerden test.

4 Conclusions

The Generalized Error Distribution introduces an extension of the classical theory which is of particular importance in order to analyze the behavior of particular estimators on small samples, and in order to avoid the inaccuracy of reducing any empirical distribution to the normal distribution, regardless of the determination of the p parameter. Until now, the theory relating to the Generalized Error Distribution does not provide adequate and dedicated hypothesis testing tools. With this paper, we aim at paving the way for tackling this central aspect.

It can definitely be concluded that the use of test statistics formulated by considering the value of the data's p parameter represents an improvement in the quality of any statistical study. The results obtained by generalizing the van der Waerden test encourage the deepening of this matter and allow us to glimpse the opportunity of studying test statistics which are suitable for the investigation of other statistical hypotheses.

Table 1: Statistical power of the simulated tests with different significance levels.

		α=0.01			***α=0.025***			***α=0.05***		
d	***p***	***Gen. VdW***	***VdW***	***Wil***	***Gen. VdW***	***VdW***	***Wil***	***Gen. VdW***	***VdW***	***Wil***
0.2	1.0	.009	.009	.009	.027	.026	.015	.059	.055	.052
0.2	1.5	.006	.006	.006	.020	.018	.009	.068	.060	.051
0.2	2.0	.008	.008	.008	.024	.024	.017	.060	.060	.050
0.2	2.5	.016	.015	.015	.022	.022	.016	.060	.060	.054
0.2	3.0	.015	.015	.015	.035	.031	.020	.067	.065	.056
0.6	1.0	.036	.037	.037	.057	.057	.049	.128	.114	.109
0.6	1.5	.033	.031	.031	.089	.088	.059	.146	.138	.125
0.6	2.0	.033	.033	.033	.093	.093	.076	.139	.139	.125
0.6	2.5	.040	.033	.033	.083	.081	.067	.145	.145	.118
0.6	3.0	.054	.043	.043	.083	.075	.057	.180	.180	.162
1.0	1.0	.080	.079	.079	.156	.146	.124	.217	.216	.212
1.0	1.5	.115	.115	.115	.209	.208	.168	.272	.260	.243
1.0	2.0	.120	.120	.120	.203	.203	.160	.301	.301	.284
1.0	2.5	.138	.124	.124	.201	.201	.166	.347	.347	.316
1.0	3.0	.133	.133	.133	.270	.251	.199	.357	.357	.323
1.2	1.0	.156	.149	.149	.215	.212	.185	.372	.308	.284
1.2	1.5	.167	.157	.157	.250	.250	.195	.394	.380	.366
1.2	2.0	.165	.165	.165	.307	.307	.251	.423	.423	.394
1.2	2.5	.219	.191	.191	.319	.318	.260	.460	.460	.430
1.2	3.0	.244	.244	.244	.384	.347	.286	.528	.528	.493
1.6	1.0	.219	.219	.219	.343	.343	.304	.477	.467	.455
1.6	1.5	.273	.273	.273	.464	.464	.405	.604	.588	.566
1.6	2.0	.351	.351	.351	.504	.504	.447	.650	.650	.623
1.6	2.5	.426	.391	.391	.567	.567	.504	.766	.766	.728
1.6	3.0	.468	.468	.468	.648	.611	.537	.779	.779	.749
2.0	1.0	.361	.361	.361	.517	.470	.421	.687	.615	.600
2.0	1.5	.477	.477	.477	.606	.604	.557	.764	.754	.734
2.0	2.0	.539	.539	.539	.715	.715	.656	.843	.843	.820
2.0	2.5	.669	.637	.637	.806	.806	.753	.884	.884	.871
2.0	3.0	.686	.686	.686	.825	.810	.769	.931	.931	.910

Note: Gen. VdW = Generalized Van der Waerden test; VdW = Classic Van der Waerden test; Wil = Wilcoxon test.

References

1. Galilei, G.: Dialogue concerning the two chief world systems–Ptolemaic and Copernican (1632)
2. Giacalone, M., Panarello, D., Mattera, R.: Multicollinearity in regression: an efficiency comparison between Lp-norm and least squares estimators. Quality & Quantity 52(4), 1831–1859 (2018)
3. Luepsen, H.: Comparison of nonparametric analysis of variance methods: A vote for van der Waerden. Communications in Statistics–Simulation and Computation 47(9), 2547–2576 (2018)
4. Mineo, A.M., Ruggieri, M.: A software tool for the Exponential Power Distribution: the normalp package. Journal of Statistical Software 12(4), 1–24 (2005)
5. Subbotin, M.: On the Law of Frequency of Error. Mathematicheskii Sbornik 31(2), 296–301 (1923)
6. Van der Laan, P., Oosterhoff, J.: Experimental determination of the power functions of the two-sample rank tests of Wilcoxon, Van Der Waerden and Terry by Monte Carlo techniques: I. Normal parent distributions. Statistica Neerlandica 21(1), 55–68 (1967)
7. Van der Waerden, B.L.: Order Tests for the Two-Sample Problem (second communication). In: Indagationes Mathematicae (Proceedings) Vol. 56, pp. 303–310. North-Holland (1953)
8. Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6), 80–83 (1945)

Stochastic dependence with discrete copulas

Dipendenza stocastica con copule discrete

Fabrizio Durante and Elisa Perrone

Abstract We here discuss the role of discrete copulas in dependence modeling. We analyze the relationships between copulas and discrete copulas, thereby drawing connections to the popular doubly stochastic matrices and their geometric interpretation. We inquire into the properties of discrete copulas under the assumption of negative dependence constraints. Finally, we discuss how these properties can be exploited to obtain an inferential procedure to detect special dependence properties.
Abstract *Si presenta l'uso di copule discrete nei modelli di dipendenza. Inoltre, si analizza la relazione tra copule e copule discrete, evidenziando legami con matrici doppiamente stocastiche e la loro interpretazione geometrica. Inoltre si studiano le proprietà delle copule discrete aventi un vincolo di dipendenza negativa. Infine, si discute su come tali proprietà possano essere impiegate per ottenere una procedura inferenziale che individui alcuni aspetti di dipendenza stocastica.*

Key words: Copulas, Stochastic Dependence, Discrete copulas, Convex Optimization, Hypothesis Testing.

1 Introduction

Copulas are powerful mathematical tools for constructing marginal-free distributions for multivariate phenomena [2, 4, 10]. The cornerstone of the copula theory is *Sklar's Theorem*, which states that the joint distribution function $F_{\mathbf{X}}$ of a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ in $\mathbb{R}^d$ with univariate margins

Fabrizio Durante
Dipartimento di Scienze dell'Economia, Università del Salento, Centro Ecotekne, 73100 Lecce (Italy), e-mail: fabrizio.durante@unisalento.it

Elisa Perrone
Department of Mathematical Sciences, University of Massachusetts Lowell, 265 Riverside St, Lowell, MA 01854 (USA), e-mail: elisa_perrone@uml.edu

$F_{X_1},\ldots,F_{X_d}$, can be written as

$$F_{\mathbf{X}}(x_1,\ldots,x_d) = C(F_{X_1}(x_1),\ldots,F_{X_d}(x_d)), \tag{1}$$

where the function $C : [0,1]^d \to [0,1]$ is a *d-dimensional copula* [18], which is uniquely identified on the set $\mathrm{Range}(F_{X_1}) \times \cdots \times \mathrm{Range}(F_{X_d}) \subseteq [0,1]^d$. As a result of Sklar's Theorem, when accurate marginal estimates can be obtained, copulas can be used to join them and obtain flexible multivariate statistical models for the data at hand.

In the case of discrete random variables, various copulas can be associated to the same random vector, and Sklar's theorem only identifies the associated *discrete copula*, i.e., the restriction of these copulas on uniform grid domains [5, 8, 9]. Any discrete copula can be extended to a full-domain copula by redistributing the discrete probability mass on each rectangle of the grid [1]; but, obviously, this extension is generally not unique. This well-known identifiability issue suggests caution while bringing copulas into a discrete framework [3]. In spite of this, discrete copulas are useful tools in dependence modeling, and they deserve the attention of the research community.

First, discrete copulas are strictly related to the so-called *empirical copulas*, which are the foundation of rank-based multivariate statistics analysis [14]. This makes them key tools in several applications, such as weather forecasting, where recovering the rank structure of a given multivariate sample is crucial [7, 15, 16]. Besides, the space of discrete copulas admits a representation as a convex polytope, i.e., a bounded convex body which consists of all the points satisfying a list of affine inequalities. As a result, the selection of discrete copulas is amenable to techniques from convex geometry and linear optimization.

Recent progress takes advantage of this polytopal representation to find simple copula models which maximize the entropy and match a prescribed Spearmans ρ [12]. For example, such models have been used in hydrology to generate synthetic data of rainfall totals when missing values occur in the observational data [13]. In [11], the authors introduce the polytopal representations of sub-families of discrete copulas with desirable stochastic properties. These representations are used to adjust the entropy-copula optimization problem to include negative dependence constraints.

In this work, we present recent advances in the study of polytopes of discrete copulas with desirable negative dependence properties, and we discuss their use to detect negative dependence of a multivariate dataset.

2 Discrete copulas and negative dependence constraints

We now assume $p \in \mathbb{N}$, and $I_p := \left\{0, \frac{1}{p}, \ldots, \frac{p-1}{p}, 1\right\}$. A $(p \times p)-$*discrete copula* is a function $C_p : I_p^2 \to [0,1]$ that satisfies the following conditions:

(c1) for all $i \in \{0,\dots,p\}$

$$C_p\left(\tfrac{i}{p},0\right) = C_p\left(0,\tfrac{i}{p}\right) = 0; \;\; C_p\left(\tfrac{i}{p},1\right) = C_p\left(1,\tfrac{i}{p}\right) = \tfrac{i}{p};$$

(c2) for all $i,j \in \{0,\dots,p-1\}$

$$C_p\left(\tfrac{i}{p},\tfrac{j}{p}\right) + C_p\left(\tfrac{i+1}{p},\tfrac{j+1}{p}\right) \geq C_p\left(\tfrac{i+1}{p},\tfrac{j}{p}\right) + C_p\left(\tfrac{i}{p},\tfrac{j+1}{p}\right).$$

As shown in [5], there is a one-to-one correspondence between the discrete copulas and the doubly stochastic matrices, which are matrices with non-negative entries and row and column sums equal to one. Hence, the space of discrete copulas corresponds to the well-known Birkhoff polytope [19]. This correspondence is obtained via a linear transformation: for all $i,j \in \{1,\dots,p\}$, the doubly stochastic matrix $B_p = (b_{i,j})$ associated with a discrete copula C_p is defined as follows:

$$b_{i,j} := p\left(C_p\left(\tfrac{i-1}{p},\tfrac{j-1}{p}\right) + C_p\left(\tfrac{i}{p},\tfrac{j}{p}\right) - C_p\left(\tfrac{i-1}{p},\tfrac{j}{p}\right) - C_p\left(\tfrac{i}{p},\tfrac{j-1}{p}\right)\right). \qquad (2)$$

The entries of the doubly stochastic matrix associated with the discrete copula C_p correspond to the probability mass induced by C_p in each sub-square of the grid. The following example illustrates this property.

Example 1. We now consider the independence copula $C(u,v) = u \cdot v$ for any u,v in the unit interval, which describes the behavior of independent random variables. Assuming $p = 4$, the corresponding discrete copula C_4 is a (5×5)-matrix, whose entries are the values of C in each point of the grid $I_4 \times I_4$. In particular, C_4 and its associated doubly stochastic matrix B_4 are as follows:

$$C_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{16} & \frac{1}{8} & \frac{3}{16} & \frac{1}{4} \\ 0 & \frac{1}{8} & \frac{1}{4} & \frac{3}{8} & \frac{1}{2} \\ 0 & \frac{3}{16} & \frac{3}{8} & \frac{9}{16} & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix}, \quad B_4 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \qquad (3)$$

We note that the independence copula spreads the probability mass uniformly on the unit square. Thus, the entries of the doubly stochastic matrix B_4 are all equal. □

As mentioned in the introduction, there are infinite ways to obtain a full domain copula from a discrete one. A popular approach is through *checkerboard extension techniques*, that is by spreading the probability mass of C_p uniformly in each sub-square of $I_p \times I_p$. The full domain copula obtained in this way is called *checkerboard copula*. The density function of such a copula is a step-wise function whose values are given by the entries of the doubly stochastic matrix associated to C_p. This connection allows for selecting suitable checkerboard copulas by solving simply a convex optimization problem on the Birkhoff polytope [12]. Though, a natural question is how to incorporate desirable dependence properties in the optimization problem by means of suitable constraints.

This question is addressed in [11] through the introduction and study of novel polytopes relevant to copula modeling and discrete geometry. In particular, the authors analyze discrete versions of the *ultramodular copulas*, which represent a form of negative dependence property for bivariate random vectors known as *stochastic decreasingness*.

This negative dependence property is often observed in applications. Indeed, several parametric copulas widely used in finance, hydrology, and environmental sciences, belong to the class of ultramodular copulas. Examples are the Gaussian, Clayton, Ali-Mikhail-Haq, Frank, and Farlie-Gumbel-Morgenstern copula families.

In [11], the authors provide (1) a geometric description of the convex space of discrete ultramodular copulas, and (2) an extension result that shows that ultramodularity is preserved through checkerboard extensions. The combination of (1) and (2) allows for formulating a new class of convex optimization problems for the selection of ultramodular checkerboard copulas in hydrology. In the next session, we discuss how (1) and (2) can be used to develop novel statistical tools for detecting negative dependence of multivariate data.

3 Detecting dependences through discrete copulas

We now consider a bivariate random sample $(X_k, Y_k)_{k=1,\dots,n}$ from the continuous random pair (X,Y). We assume that the underlying distribution of (X,Y) is stochastically decreasing, i.e. $x \mapsto \mathbb{P}(Y > y \,|\, X = x)$ and $y \mapsto \mathbb{P}(X > x \,|\, Y = y)$ are decreasing functions. This condition can be expressed in terms of the copula C_{XY} and is equivalent to the fact that C_{XY} is ultramodular, i.e., it has convex sections in each variable.

In principle, we expect that the stochastic decreasingness of the random vector (X,Y) should be encoded in the discrete copula associated with the sample. In particular, from a geometric perspective, we would expect (X,Y) to be mutually stochastically decreasing if and only if the discrete copula corresponding to the sample is "close enough" to the polytope of ultramodular discrete copulas defined in [11]. Given a sample of size n, we can proceed as follows:

1. Compute the empirical copula EC from the sample;
2. Evaluate the empirical copula on a square grid of fixed partition size $p \times p$ (this results in a discrete copula EC_p) with $p = p(n) < n$;
3. Find the doubly stochastic matrix B_p associated with EC_p by using the linear transformation of Equation (2);
4. Evaluate the distance between B_p and the space of all densities of ultramodular checkerboard copulas of step p (UDC_p), by also matching the value of the empirical Spearman's ρ of the sample:

$$\begin{array}{ll} \text{minimize} & \mathrm{dist}(B_p - U_p) \\ \text{subject to} & U_p \in \mathrm{UDC}_p;\ \rho_{U_p} = \tilde{\rho} \end{array} \tag{4}$$

Intuitively, the smaller the distance, the higher the chance that the random vector (X,Y) is stochastically decreasing. As a consequence, the distance computed in Equation (4) can be used as a test statistic for a significance test whose null hypothesis is *"The given random sample is associated with a ultramodular copula"*. Various tests have been developed in the literature to check stochastic monotonicity [6, 17]. However, to the best of our knowledge, a significance test built on this idea has never been considered. We now present a simulation study to evaluate the performance of the proposed testing procedure.

We consider a distance based on the Frobenius norm, which is the discrete analogous of the L^2-distance often used in Cramér-von Mises-type tests for copulas. The data are simulated from a Gaussian copula, which is ultramodular for negative values of the parameter ρ. We assume a finer partition grid p as the strength of the correlation increases. This adjustment is required since the associated doubly stochastic matrices B_p become sparser. We use a bootstrap approach to compute the p-value of the test by assuming a bootstrap sample of size 1000 and a resampling scheme from a copula in the ultramodular class.

Table 1 shows the percentage of the rejection of the null hypothesis for different significant levels α. The rejection rate is computed over $N = 1000$ independent samples of size $n = 500$, drawn from a Gaussian copula with parameter ρ.

The results are promising and suggest that the presented approach can be used to detect stochastic decreasing properties of a random sample. We are currently working on a follow-up paper where we study the performance of the test under other copula assumptions and settings for the convex optimization problem, alongside further analysis of the power of the test.

p	ρ	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
15	-0.38	10.9 %	6.8 %	0.8 %
25	-0.70	11.5 %	6.0 %	1.6 %

Table 1 The table presents the percentage of the rejection of the null hypothesis over 1000 $N = 1000$ independent samples of size $n = 500$. The samples are drawn from a Gaussian copula with parameter ρ. The test is performed assuming two values of the partition grid p.

Acknowledgements Elisa Perrone has been supported by the Austrian Science Fund (FWF) Project: J3968-N32. Fabrizio Durante has been supported by the project *Stochastic Models for Complex Systems* by Italian MIUR (PRIN 2017, Project no. 2017JFFHSH).

References

1. E. de Amo, M. Díaz Carrillo, F. Durante and J. Fernández-Sánchez. Extensions of subcopulas. *J. Math. Anal. Appl.*, 452 (1), 1–15, 2017.
2. F. Durante and C. Sempi. *Principles of Copula Theory*. CRC/Chapman & Hall, Boca Raton, FL, 2015.
3. C. Genest and J. Nešlehová, J. A Primer on Copulas for Count Data. *ASTIN Bulletin*, 37 (2), 475–515, 2007.
4. H. Joe. *Dependence Modeling with Copulas*. Chapman and Hall/CRC, 2nd edition, 2014.
5. A. Kolesarová, R. Mesiar, J. Mordelová, and C. Sempi. Discrete Copulas. *IEEE T. Fuzzy Syst.*, 14(5):698–705, 2006.
6. S. Lee, O. Linton, and Y.-J. Whang. Testing for Stochastic Monotonicity. *Econometrica*, 77(2):585–602, 2009.
7. X. Li and V. Babovic. A new scheme for multivariate, multisite weather generator with inter-variable, inter-site dependence and inter-annual variability based on empirical copula approach. *Climate Dynamics*, 52, 2018.
8. G. Mayor, J. Suner, and J. Torrens. Copula-like operations on finite settings. *IEEE T. Fuzzy Syst.*, 13(4):468–477, 2005.
9. R. Mesiar. Discrete copulas-what they are. *EUSFLAT-LFA 2005*, 927–930, 2005.
10. R. B. Nelsen. *An Introduction to Copulas*. Springer, 2nd edition, 2006.
11. E. Perrone, L. Solus, and C. Uhler. Geometry of discrete copulas. *Journal of Multivariate Analysis*, 172:162–179, 2019.
12. J. Piantadosi, P. Howlett, and J. Borwein. Copulas with maximum entropy. *Optimization Letters*, 6(1):99–125, 2012.
13. N. F. A. Radi, R. Zakaria, J. Piantadosi, J. Boland, W. Z. W. Zin, and M. A.-z. Azman. Generating Synthetic Rainfall Total Using Multivariate Skew-*t* and Checkerboard Copula of Maximum Entropy. *Water Resources Management*, 31(5):1729–1744, 2017.
14. L. Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009.
15. R. Schefzik, T. L. Thorarinsdottir, and T. Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28:616–640, 2013.
16. R. Schefzik. Multivariate discrete copulas, with applications in probabilistic weather forecasting. *arXiv:1512.05629*, 2015.
17. J. Seo. Tests of stochastic monotonicity with improved power. *Journal of Econometrics*, 207(1):53–70, 2018.
18. A. Sklar. Fonctions de rèpartition à n dimensions et leurs marges. *Publ. de l'Institut de Statistique de Paris*, 8:229–231, 1959.
19. G. M. Ziegler. *Lectures on polytopes*. Springer-Verlag, New York, 1995.

Models and methods - Time Series and Longitudinal Data

Bootstrap test in Poisson–INAR models

Bootstrap test in modelli INAR con innovazioni Poisson

Lucio Palazzo and Riccardo Ievoli

Abstract In this paper we exploit bootstrap–based specification test on the autoregressive coefficient in INteger–AutoRegressive models (INAR) considering the case of homoskedastic Poisson innovations. Performance of four unrestricted bootstrap methods are analyzed through a small–scale Monte carlo exesercise.
Abstract *In questo lavoro si sono indagate le potenzialità di tecniche bootstrap per la verifica di ipotesi sul coefficiente autoregressivo dei modelli INAR, nel caso innovazioni omoschedastiche distribuite come delle v.c. Poisson. Le performance di quattro metodi sono state valutate attraverso uno studio di simulazione.*

Key words: Integer-valued time series, bootstrap methods, specification test

1 Introduction

INteger AutoRegressive models (INAR), introduced in [1], became very popular to model non–negative integer time series. INAR with Poisson innovations (P–INAR) represents one of the most applied version of this model. Moreover, a drawback can be found in the poor performance of asymptotic approximation, requiring large samples to ensure negligible bias of conventional estimators, correct coverage rates and avoid over–rejections in hypothesis testing. Bootstrap methods in INAR have been recently introduced by [3] to obtain more reliable inference in point estimation and confidence bounds.

Lucio Palazzo
Univeristy of Napoli Federico II, Department of Political Science, Via Rodinò 8, e-mail: lucio.palazzo@unina.it

Riccardo Ievoli
Univeristy of Ferrara, Department of Economics and Management, Via Voltapaletto 11 e-mail: riccardo.ievoli@unife.it

In this paper we study, through a simulation study, if and how the bootstrap can improve inference in testing the coefficient of P–INAR(1) model. To the best of our knowledge, this is the first work that analyses the potential of bootstrap in testing P–INAR coefficients especially in small samples which is a common issue in real data.

The paper is organized as follows: Section 2 presents the P–INAR(1) model and its associated specification test. Section 3 introduces the bootstrap schemes while results of a small–scale Monte Carlo simulation are shown in Section 4. Finally, Section 5 concludes with some remarks and possible advances.

2 Testing Coefficient in P–INAR(1)

Consider the following stable P–INAR(1) model, introduced in [1] and defined as

$$X_t = \alpha * X_{t-1} + \varepsilon_t \tag{1}$$

where X is a non–negative (integer valued) random variable, $\alpha \in [0,1)$ and $\{\varepsilon_t\}$ is i.i.d $Pois(\lambda)$ with $\lambda < \infty$. The symbol $*$, i.e. the *binomial thinning* operator is defined as a random sum of i.i.d. random variables $\{Y_i\}$, with $Y_i \sim \text{Ber}(\alpha)$, independent of X_t, such that $\text{E}(Y_i) = \alpha$ and $\text{Var}(Y_i) = \alpha(1-\alpha)$. Aside from the thinning parameter α, the parameter λ in P–INAR(1) λ is the Poisson mean. Conventional MM estimator for α in Equation 1, denoted also as Yule–Walker (YW), is defined as

$$\hat{\alpha}_T = \frac{\sum_{t=2}^{T}(X_t - \overline{X})(X_{t-1} - \overline{X})}{\sum_{t=1}^{T}(X_t - \overline{X})^2} \tag{2}$$

The test statistic, here denoted as τ_{obs}, for the null hypothesis $H_0 : \alpha = \alpha_0$ against the alternative $H_1 : \alpha > \alpha_0$, is

$$\tau_{obs} = \frac{\hat{\alpha}_T - \alpha_0}{\left[\hat{V}(\hat{\alpha}_T)\right]^{1/2}} \tag{3}$$

where $\hat{\alpha}_T$ is defined in expression (2) and $\hat{V}(\hat{\alpha}_T) = T^{-1}[1 - \hat{\alpha}_T^2 + \hat{\alpha}_T \hat{\lambda}_T^{-1}(1-\hat{\alpha}_T)^2]$, with $\hat{\lambda}_T = \hat{\lambda}_T^{MM} = T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t$ is the MM (or YW) estimator for λ and $\hat{\varepsilon}_t = x_t - \hat{\alpha}_T\, x_{t-1}$. In stable P–INAR(1) normalized MM estimators are asymptotically normal. Without loss of generality we consider the case of $\alpha_0 = 0$.

3 Bootstrap Algorithm for testing INAR

Here we present four Bootstrap algorithms in order to test the null hypothesis $H_0 : \alpha = 0$ against the alternative $H_1 : \alpha > 0$. These methods, where we do not impose the null hypothesis (i.e. $\alpha = 0$) in the bootstrap DGP, are suitable to obtain the

distribution of a bootstrap test statistic τ^*, can be classified in two groups. The former is defined as "direct", because sampling (with replacement) from the observed data, while the latter, called "indirect", requires estimation of model parameters and residuals. Two procedures are summarized in the following schemes.

Algorithm 1 (Direct Procedure)**.**
Given a random sample $x_1,\dots,x_T$ of size T, then

Step 1. Sample data randomly with replacement, obtaining the pseudo–sample $x_1^*,\dots,x_T^*$;
Step 2. Compute the bootstrapped test statistic τ^*;
Step 3. Repeat steps 1–2, with $b = 1,\dots,B$ to obtain $\tau_1^*,\dots,\tau_B^*$.

Algorithm 2 (Indirect Procedure)**.**
Given a random sample $x_1,\dots,x_T$ of size T and estimates of the parameter $(\hat{\alpha},\hat{\lambda})$:

Step 1. Sample from bootstrap pseudo–residuals, namely $\varepsilon_1^*,\dots,\varepsilon_T^*$;
Step 2. Create $x_1^*,\dots,x_T^*$ plugging pseudo residuals and estimated parameters in the bootstrap DGP.
Step 3. Repeat B times steps 1–2 producing $\tau_1^*,\dots,\tau_B^*$.

We firstly consider a direct non–parametric approach, denoted as Circular Block Bootstrap (CBB) [5]. Sample is split into overlapping blocks of fixed length, then bootstrapped series are produced drawing resulting blocks at random with replacement. Finally the pseudo–series are circularized including an additional set blocks generated combining initial and final values of the series.

Furthermore, three different methods are illustrated, starting from the general procedure explained in Algorithm 2. A naive approach is to apply an AutoRegressive Bootstrap (ARB) procedure. ARB is very easy to implement because AR and INAR models share the same autocorrelation structure, and ARB usually outperforms asymptotic approximations in finite samples [4]. In order to apply ARB with discrete–valued data, the following rounding scheme is necessary to discretize the bootstrap series: $\tilde{x}_t^{*b} = 0$ if $x_t^* \leq 0.5$ and $\tilde{x}_t^{*b} = \lceil \tilde{x}_t^{*b} \rceil$ otherwise.

Since ARB neglects the discrete nature of data, a possible more efficient approach is given by two methods introduced in [3], which essentially differs in the generation of ε_t^*. The first method is the Parametric Bootstrap (PB) and requires a) the estimation for the thinning parameter and b) the assumption of parametric distribution (i.e. Poisson) depending on a set of parameters used to estimate the true innovations' distribution ε_t from the INAR data. The replicated series are generated as follows:

$$x_t^* = \hat{\alpha} \tilde{*} x_{t-1}^* + \varepsilon_t^*, \quad \forall t = 2,\dots,n \tag{4}$$

Since PB is based on strictly assumptions, a more suitable method is the Semiparametric Bootstrap (SPB). SPB use the estimated P–INAR(1) coefficients, derived by using the same approach introduced in PB, but a nonparametric estimation of the innovations distribution. There are several ways to compute nonparametric estimates ε_t^* of equation (4), we follow [3] adopting the procedure introduced in [2]. Improvements in terms of performances of semi–parametric methods in this framework are

still under debate, although they have some advantages. For instance they can be valid even if some model assumptions, such as normality of residuals, do not hold.

For all considered methods the statistic of interest is computed as follows:

$$\tau^* = \frac{\hat{\alpha}_T^* - \hat{\alpha}_T}{\left[\hat{V}^*(\hat{\alpha}_T)\right]^{1/2}} \tag{5}$$

where $\hat{\alpha}_T^*$ is the bootstrap counterpart of $\hat{\alpha}_T$ obtained through equation (1) and bootstrapped data $x_1^*, \ldots, x_T^*$. Finally, the bootstrap p–value is computed through $\tau_1^*, \ldots, \tau_B^*$ as:

$$p^* = 2 \cdot \min \left\{ B^{-1} \sum_{b=1}^{B} I(\tau_b^* < \tau_{obs}), B^{-1} \sum_{b=1}^{B} I(\tau_b^* > \tau_{obs}) \right\} \tag{6}$$

this is the so–called equitail bootstrap p–value and is more appropriate than the symmetric counterpart, obtaining as $p^* = B^{-1} \sum I\left(|\tau_b^*| > |\tau_{obs}|\right)$, especially when the test statistic of interest is not always nonnegative.

4 Montecarlo Simulation

In order to analyze the behavior of four bootstrap methods, illustrated in Section 3, we generate $M = 2000$ samples following equation (1) and with following DGP: $\varepsilon_t \sim Pois(\lambda)$ with $\lambda = (1,5)$ and increasing sample length. The selected nominal level is equal to 0.05. Empirical size of bootstrapped t statistic is reported fixing $\alpha = 0$, while we choose the following values for the P–INAR(1) coefficient: $\alpha = \{0, 0.05, 0.1, \ldots, 0.8\}$ to evaluate the empirical power. Number of bootstrap replication is set equal to $B = 999$. We compute rejection frequencies for the asymptotic statistic as a benchmark, generating $M = 10000$ dataset under the previously introduced DGP.

Table 1 presents results in terms of empirical size: asymptotic rejection frequencies are often quite far from the nominal level equal, especially when $\lambda = 1$. As expected, ARB and CBB are dramatically ovesized, performing considerably worse than asymptotic approximation, due to their inadequacy in reproduce the real dependency of data in the bootstrap sample. Nevertheless, PB and SPB present rejections which are very closer to the nominal level even in moderately small samples. In particular, SPB seems to outperform PB in most parameter settings. Moreover, when $\lambda = 5$ rejection frequencies of asymptotic approximation are comparable to PB and SPB

In Figure 1 we plot the logit of the rejection frequencies under two different sample sizes ($T = 50, 250$). We found that empirical power associated to PB and SPB always outperform asymptotic approximation under moderately small sample size ($T = 50$). Moreover, when the sample size is moderately large ($T = 250$), empirical power of asymptotic approximation in two considered scenarios is comparable with

the PB and SPB. Surprisingly, only when T is small the ARB presents the same behavior of PB.

Table 1 Rejection frequencies (Empirical size) of t–statistic for the null hypothesis: $H_0 : \alpha = 0$

			$\lambda = 1$		
T	ASY	ARB	CBB	PB	SPB
50	0.032	0.259	0.256	0.062	0.064
75	0.032	0.289	0.215	0.060	0.052
100	0.029	0.319	0.206	0.056	0.065
125	0.033	0.315	0.190	0.051	0.054
150	0.033	0.324	0.203	0.051	0.049
500	0.040	0.347	0.199	0.047	0.056
			$\lambda = 5$		
T	ASY	ARB	CBB	PB	SPB
50	0.045	0.083	0.247	0.067	0.048
75	0.040	0.067	0.218	0.058	0.049
100	0.045	0.067	0.205	0.060	0.046
125	0.040	0.066	0.191	0.055	0.058
150	0.048	0.085	0.196	0.064	0.052
500	0.045	0.318	0.192	0.052	0.049

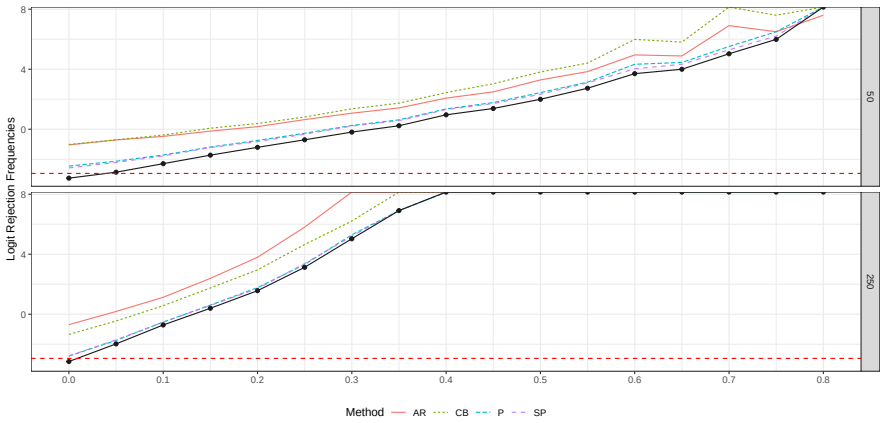


Fig. 1 Logit (empirical) power of t–statistic for different values of α and $\lambda = 1$. Estimation is carried out through Yule–Walker estimator. Red dashed line is $logit(0.05)$, while dotted line represents the empirical asymptotic power.

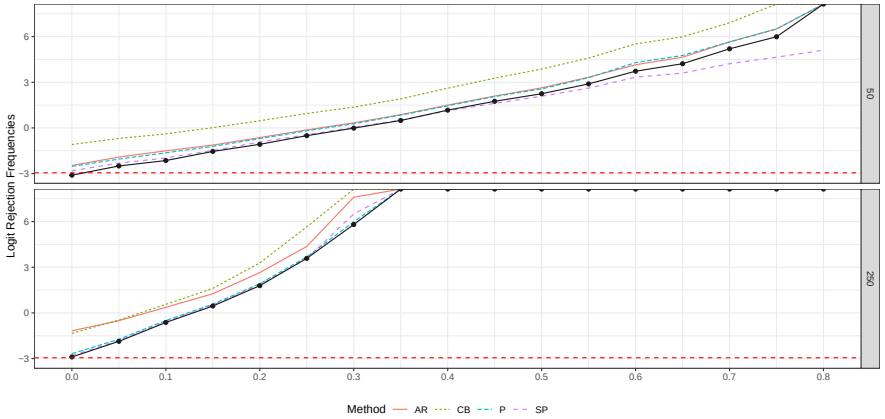


Fig. 2 Logit (empirical) power of t–statistic for different values of α and $\lambda = 5$. Estimation is carried out through Yule–Walker estimator. Red dashed line is $logit(0.05)$, while dotted line represents the empirical asymptotic power.

5 Conclusions

Resampling methods as PB and SBP can outperform asymptotic approximation in testing P–INAR(1) coefficient, even if the sample size is small. PB and SPB can be generalized to the case of more lagged regressors. Proposed bootstrap test may be also exploited considering different INAR marginal distribution, i.e. the Negative Binomial or Geometric, and their associated estimators. On the other hand, conventional methods (CBB and ARB), which not take into account of the discreteness and positiveness of the data, are confirmed unreliable in the context of hypothesis testing. To improve bootstrap inference, some possible extensions can be exploited. First of all, resampling methods involving the null hypothesis in the bootstrap DGP can be exploited. Secondly, other estimators can be implemented in the bootstrap procedure such as conditional least squares and maximum likelihood.

References

1. Al-Osh, M. A., Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR (1)) process. Journal of Time Series Analysis, 8(3), 261-275.
2. Drost, F. C., van den Akker, R., Werker, B. J. M. (2008). Efficient Estimation of Autoregression Parameters and Innovation Distributions forSemiparametric Integer-Valued AR (p) Models (Revision of DP 2007-23) (No. 2008-53).
3. Jentsch, C., Weiß, C. H. (2019). Bootstrapping INAR models. Bernoulli, 25(3), 2359-2408.
4. Kreiss, J. P., Paparoditis, E., Politis, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. The Annals of Statistics, 39(4), 2103-2130.
5. Politis, D. N., Romano, J. P. (1992). A circular block-resampling procedure for stationary data. Exploring the limits of bootstrap, 2635270.

Continuous Time-Interaction Processes for Population Size Estimation

Processi autointerattivi in tempo continuo per la stima della dimensione di una popolazione

Linda Altieri, Alessio Farcomeni, Danilo Alunni Fegatelli and Francesco Palini

Abstract This work presents time-interacting temporal point processes for capture-recapture data with behavioural responses. The proposed model is able to provide unbiased estimates of the population size for both self-exciting and self-correcting behaviours.

Abstract *Questo studio presenta processi di punto temporali autointerattivi. Il modello proposto fornisce una stima non distorta della dimensione di una popolazione sia in caso di processi auto-eccitanti che di processi auto-correggenti.*

Key words: self-exciting process, self-correcting process, time-interaction, behavioural response, population size, capture-recapture data

1 Introduction

A continuous time point process is a stochastic process modelling a list of times of event $\{t_1,\ldots,t_S\}$. The temporal process is associated to a counting process $N(t)$ for the number of events in the interval $(0,t)$, with $N(t) \sim Poisson(\lambda(t))$. The process conditional intensity function is given by

$$\lambda(t) = \lim_{h\to 0} \frac{E[N(t+h) - N(t)|\mathscr{H}(t)]}{h} \tag{1}$$

Linda Altieri
University of Bologna e-mail: linda.altieri@unibo.it

Alessio Farcomeni
University of Tor Vergata, Rome, e-mail: alessio.farcomeni@uniroma2.it

Danilo Alunni Fegatelli
La Sapienza University, Rome e-mail: danilo.alunnifegatelli@uniroma1.it

Francesco Palini
La Sapienza University, Rome e-mail: francesco.palini@uniroma1.it

where $\mathscr{H}(t)$ is the history of event times up to t.

Different processes may be defined by modelling $\lambda(t)$. In this work, we combine the so-called *self-exciting* [1] and *self-correcting* [2] processes, where the occurrence of an event at time t affects the probability of occurrence of another event after t. We apply the resulting general formulation to capture-recapture data [3], based on repeatedly observing subjects over time. The objective is to obtain an estimate of the overall size of the catchable population, including subjects with zero captures.

In continuous time, each subject is at risk of capture in any moment of a fixed-length period, say $(0,T)$. In general, the expected number of captures per subject in the interval $(0,T)$ can be expressed as $E[N(T)] = \int_0^T \lambda(t)\, dt$. It is straightforward to check that $E[n] = N \cdot P(N(T) > 0)$, where n is the number of captured subjects and N is the total amount of subjects in the population, including the unobserved ones. The general form for the Horvitz-Thompson type estimator for N is then $\widehat{N} = \sum_{i=1}^{n} \frac{1}{P(N_i(T)>0)}$ where the probability $P(N_i(T) > 0)$ of observing a subject i in the interval of interest $(0,T)$ depends on the chosen model.

In a generic self-exciting process, the occurrence of an event at time t causes the conditional intensity function to increase. In capture-recapture contexts, a self-exiciting kernel allows events to trigger other events in the short term, the so called *trap-happiness* effect. The intensity is modeled as

$$\lambda_{SE}(t) = \eta + \sum_{k:t_k<t} g(t-t_k), \tag{2}$$

for some smooth self-exciting kernel $g(\cdot)$, where t_k is an event time in $(0,t)$. A standard choice is the exponential kernel $g(t) = \sum_{j=1}^{J} \alpha_j e^{-\beta_j t}$, for some $J \geq 1$. A stationarity condition for the exponential kernel is that $\sum_j \alpha_j/\beta_j < 1$, to avoid explosion of the intensity function. In practice, the exponential kernel with $J = 1$ is often flexible enough to provide good fit in self-exciting patterns.

Assumption (2), though, is not flexible enough to take *trap-shyness* into account, i.e. to consider situations where the conditional intensity function drops after an event. This can be captured by self-correcting kernels, where the intensity increases over time in absence of events, and is decreased by each event. The conditional intensity of a self-correcting process may be expressed as

$$\lambda_{SC}(t) = \eta \exp(-\theta(N(t) - \eta t)),$$

where θ is a parameter controlling the degree of trap-shyness and ηt is a target for the number $N(t)$ of events at time t.

In the following, we combine self-exciting and self-correcting processes. We derive the conditional likelihood for the biased samples obtained with population size experiments, where only units with at least one event are recorded. We then use the resulting parameter estimates within the Horvitz-Thompson estimator to obtain population size estimates.

2 Time-interaction processes

We propose a unified model which extends both self-exciting and self-correcting processes to a general time-interaction process. In the simplest case, we model the conditional intensity of a time-interaction process as

$$\lambda(t) = \alpha \sum_{k:t_k<t} \exp(-\beta(t-t_k)) + \eta \exp\left(-\theta(N(t)-\eta t)\right), \tag{3}$$

where $\alpha, \beta, \eta, \theta > 0$ and $\alpha < \beta$ to avoid explosion. Note that the intensity is not differentiable, but it is continuous in all parameters.

Model estimation relies on the conditional likelihood; the estimated parameters are then used for a Horvitz-Thompson estimator for N, the total population amount. In a generic temporal point process, the log-likelihood for a subject i with capture times $\{t_{i1}, \ldots, t_{ik}, \ldots, t_{iK_i}\}$ in an interval of interest $(0,T)$ is

$$l_i(\lambda) = \sum_{k=1}^{K_i} \log(\lambda(t_{ik})) - \int_0^T \lambda(t)\, dt \tag{4}$$

where $K_i = N_i(T)$ and we assume $t_{i0} = 0$ and $t_{i(K_i+1)} = T$.

For the time-interaction process, the log-likelihood becomes

$$\begin{aligned} l_i(\psi) = \textstyle\sum_{k=1}^{K_i} \log\left[\alpha \sum_{s:t_{is}<t_{ik}} \exp(-\beta(t_{ik}-t_{is})) + \eta \exp\left(-\theta(N(t_{ik})-\eta t_{ik})\right)\right] \\ + \tfrac{\alpha}{\beta} \textstyle\sum_{k=1}^{K_i} \left[\exp(-\beta(T-t_{ik})) - 1\right] \\ - \tfrac{1}{\theta} \textstyle\sum_{k=0}^{K_i} \exp(-\theta k) \left[\exp(\theta\eta t_{ik+1}) - \exp(\theta\eta t_{ik})\right] \end{aligned} \tag{5}$$

where $\psi = (\alpha, \beta, \eta, \theta)$ is a short-hand notation for the parameters of interest.

In absence of selection bias, the log-likelihood for n observed subjects over the interval $(0,T)$, is given by

$$l_n(\psi) = \sum_{i=1}^{n} \left(\sum_{k=1}^{K_i} \log(\lambda(t_{ik})) - \int_0^T \lambda(t)\, dt \right). \tag{6}$$

Naturally, we can only sample subjects with $K_i > 0$. The likelihood above is therefore misspecified, and we resort to the conditional likelihood, which conditions on the event that there are one or more measurements. Because of the Poisson distribution of $N_i(t)$, the probability of observing a subject is

$$\Pr(K_i > 0) = 1 - \exp\left(- \int_0^T \eta \exp(\theta\eta t) dt \right) = 1 - \exp\left[- \frac{\exp(\theta\eta T) - 1}{\theta} \right]. \tag{7}$$

Note that, in this model, $\Pr(K_i > 0) = \Pr(K > 0)$ is the same for all subjects.

Hence, the conditional log-likelihood for n observed subjects is given by

$$l_n^*(\psi) = l_n(\psi) - n\log\left\{1 - \exp\left[-\frac{\exp(\theta\eta T) - 1}{\theta}\right]\right\}. \tag{8}$$

This approach allows not only to estimate the vector of parameters of interest ψ, but also to evaluate the total number of subjects $N = n + n_0$, where n_0 is the number of unobserved individuals belonging to the population of interest.

In order to maximize the conditional log-likelihood, the gradient is needed, which can be derived in closed form; then, the Hessian can be obtained numerically. A Newton-Raphson method proceeds after operating a change of variables so that positive parameters $\psi = (\alpha, \beta, \eta, \theta)$ are replaced by real-valued parameters $\psi' = (\alpha', \beta', \eta', \theta')$, where $\alpha = \exp(\alpha')$, $\beta = \exp(\alpha') + \exp(\beta')$, $\eta = \exp(\eta')$ and $\theta = \exp(\theta')$.

The Horvitz-Thompson type estimator for N in the current model, where the probability of no captures is the same for all subjects, is

$$\widehat{N} = \frac{n}{Pr(K > 0)} = \frac{n}{1 - \exp\left[-\frac{\exp(\theta\eta T) - 1}{\theta}\right]}. \tag{9}$$

2.1 Time-interaction processes with heterogeneity

In addition to behavioural effects, general models for closed population size estimation must consider three further sources of variation in capture rates: time-dependent effects, observed and unobserved heterogeneity ([8, 6]). The most general model is named M_{hotb}, where h stands for the unobserved heterogeneity, o for the observed heterogeneity, t for the time-heterogeneity and b for the behavioral response to capture. Under such model, the intensity function for a generic subject i is:

$$\lambda_i(t) = \alpha_i \sum_{k:t_{ik}<t} \exp(-\beta_i(t - t_{ik})) + \eta t^{\eta-1} \exp\left(X_i'\gamma + \mu_i - \theta_i(N_i(t) - \eta t^{\eta-1})\right), \tag{10}$$

where X_i is a $P \times 1$ vector of subject-specific covariates associated to P regression coefficients γ to capture the observed heterogeneity, while μ_i is a subject-specific random effect measuring the unobserved heterogeneity. The effects α, β, θ and μ may be thought of as following latent class models: they can take values in a discrete set of C latent classes subjects may belong to, with $C < n$. Therefore, $\alpha_i \in \widetilde{\alpha} = (\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_C)$, $\beta_i \in \widetilde{\beta} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_C)$, $\theta_i \in \widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_C)$ and $\mu_i \in \widetilde{\mu} = (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_C)$. For identifiability reasons, we assume the latter vector $\widetilde{\mu}$ is in increasing order and that it sums to zero. The parameter η is under the constraint of being positive as for the simple model (2), called M_b model.

Maximum likelihood estimates for ψ are obtained via an EM algorithm, where starting values are provided for all parameters, respecting the constraints and identifiability. The last step consists in estimating N with a Horvitz-Thompson estimator, that takes random effects into account. Firstly, each subject is assigned to the latent class c with the greatest estimated probability $\widehat{\pi}_c$; such class is denoted as c_i^*.

Other estimated parameters $\widehat{\psi}_i^*$ are selected accordingly. Consequently, the probability $P(K_i > 0)$ is derived for each subject, and the estimator is

$$\widehat{N} = \sum_{i=1}^{n} \frac{1}{P(K_i > 0)} = \sum_{i=1}^{n} \frac{1}{1 - \exp\left[-e^{X_i'\widehat{\gamma}+\widehat{\widetilde{\mu}}_{c_i^*}} \int_0^T t^{\widehat{\eta}+1} e^{\widehat{\widetilde{\theta}}_{c_i^*}} \widehat{\eta} t^{\widehat{\eta}-1} \, dt\right]}. \quad (11)$$

3 Simulation study

We conducted a simulation study to validate the goodness of fit of our model. As competitors, we considered the lower bound nonparametric estimator of Chao (*Ch*) [4], the generalized Zelterman and Chao estimator (*GF*) which also includes time varying covariates and a behavioral effect [5, 7] and model M_t and M_{th} developed in the R package `ctime` [9]. In order to verify the performance of our estimator in different situations, eighty different simulation settings were proposed. For all simulations, we considered two latent classes; the vector parameter $\mu = (\mu_1, \mu_2)$ was fixed to (-7.5,0). For each combination of model parameters, K=1000 datasets were analyzed. Table 3 reports the empirical coverage and the root mean square error (RMSE) for all estimators evaluated for each scenario. As one can see, in almost all cases our estimator (N_{TI}) provides a smaller RMSE. We obtained a higher RMSE only in seven settings in which the impact of the behavioral effect is minor compared to the heterogeneity effect and a parsimonious model was preferred.

4 Conclusions

We proposed a general framework for estimating the population size in closed populations with capture-recapture of subjects in continuous time. Our model, named time-interaction process, applies to situations where a behavioural effect may be assumed, and extends both self-exciting and self-correcting processes. The estimator for the population size is associated to a smaller root mean square error in a variety of situations compared to standard estimators in the literature of capture-recapture data.

Acknowledgements This work has been developed within the framework of the EU project JUST/2010/DPIP/AG/1410: 'New methodological tools for policy and programme evaluation' with the financial support of the Prevention and Information Programme of the European Commission. The contents of this publication are the sole responsibility of the authors and can in no way be taken to reflect the views of the European Commission.
Linda Altieri's work is developed under the PRIN2015 supported project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)' [grant number 20154X8K23] funded by MIUR (Italian Ministry of Education, University and Scientific Research).

Table 1 Simulation results

N=5000														N=500											
Parameters									RMSE					Parameters									RMSE			
α_1	α_2	β_1	β_2	η	θ_1	θ_2	γ_1	γ_2	n/N	N_{TI}	N_{Ch}	N_{GF}	N_{Mt}	N_{Mth}	α_1	α_2	β_1	β_2	η	θ_1	θ_2	γ_1	γ_2	n/N	N_{TI}	N_{Ch}	N_{GF}	N_{Mt}	N_{Mth}
1	0.20	2	0.50	1.25	0.50	1	-3	0	0.55	1200	2419	2304	2486	2436	1	0.20	2	0.50	1.25	0.50	1	-3	0	0.55	124	241	225	248	240
5	0.20	6	0.50	1.25	0.50	1	-3	0	0.55	1207	2444	2260	2514	2371	5	0.20	6	0.50	1.25	0.50	1	-3	0	0.55	137	243	219	251	236
1	0.20	2	0.50	1.25	2	1	-3	0	0.48	463	1866	1842	2060	1898	1	0.20	2	0.50	1.25	2	1	-3	0	0.48	60	185	179	205	187
5	0.20	6	0.50	1.25	2	1	-3	0	0.48	565	1986	1929	2167	1707	5	0.20	6	0.50	1.25	2	1	-3	0	0.48	68	197	189	216	176
1	0.20	2	0.50	1.25	0.50	1	-3	1	0.55	1183	2426	2326	2516	2444	1	0.20	2	0.50	1.25	0.50	1	-3	1	0.55	122	241	226	251	241
5	0.20	6	0.50	1.25	0.50	1	-3	1	0.55	1193	2451	2292	2542	2382	5	0.20	6	0.50	1.25	0.50	1	-3	1	0.55	144	245	224	253	239
1	0.20	2	0.50	1.25	2	1	-3	1	0.48	391	1861	1835	2080	1882	1	0.20	2	0.50	1.25	2	1	-3	1	0.48	54	185	179	207	186
5	0.20	6	0.50	1.25	2	1	-3	1	0.48	500	1981	1928	2184	1695	5	0.20	6	0.50	1.25	2	1	-3	1	0.48	60	198	189	218	174
1	0.20	2	0.50	1.33	0.50	1	-3	0	0.54	1178	2404	2304	2467	2426	1	0.20	2	0.50	1.33	0.50	1	-3	0	0.54	123	240	226	247	239
5	0.20	6	0.50	1.33	0.50	1	-3	0	0.54	1175	2427	2268	2489	2367	5	0.20	6	0.50	1.33	0.50	1	-3	0	0.54	132	242	221	249	236
1	0.20	2	0.50	1.33	2	1	-3	0	0.47	351	1775	1757	1980	1815	1	0.20	2	0.50	1.33	2	1	-3	0	0.47	49	177	172	197	179
5	0.20	6	0.50	1.33	2	1	-3	0	0.47	436	1905	1859	2088	1648	5	0.20	6	0.50	1.33	2	1	-3	0	0.47	60	190	182	208	168
1	0.20	2	0.50	1.33	0.50	1	-3	1	0.55	1147	2405	2323	2490	2430	1	0.20	2	0.50	1.33	0.50	1	-3	1	0.55	122	240	227	249	240
5	0.20	6	0.50	1.33	0.50	1	-3	1	0.55	1167	2433	2296	2518	2378	5	0.20	6	0.50	1.33	0.50	1	-3	1	0.55	143	244	224	252	238
1	0.20	2	0.50	1.33	2	1	-3	1	0.47	284	1764	1746	1998	1795	1	0.20	2	0.50	1.33	2	1	-3	1	0.47	42	176	171	199	179
5	0.20	6	0.50	1.33	2	1	-3	1	0.47	375	1900	1853	2107	1644	5	0.20	6	0.50	1.33	2	1	-3	1	0.47	52	190	183	211	168
1	0.20	2	0.50	1.50	0.50	1	-3	0	0.53	1111	2370	2308	2424	2399	1	0.20	2	0.50	1.50	0.50	1	-3	0	0.53	120	235	226	241	235
5	0.20	6	0.50	1.50	0.50	1	-3	0	0.53	1137	2391	2287	2442	2360	5	0.20	6	0.50	1.50	0.50	1	-3	0	0.53	141	239	224	244	235
1	0.20	2	0.50	1.50	2	1	-3	0	0.43	100	1531	1522	1779	1598	1	0.20	2	0.50	1.50	2	1	-3	0	0.43	33	153	149	177	159
5	0.20	6	0.50	1.50	2	1	-3	0	0.43	200	1701	1675	1895	1490	5	0.20	6	0.50	1.50	2	1	-3	0	0.43	47	170	164	190	151
1	0.20	2	0.50	1.50	0.50	1	-3	1	0.53	1086	2365	2312	2442	2401	1	0.20	2	0.50	1.50	0.50	1	-3	1	0.53	119	236	227	244	237
5	0.20	6	0.50	1.50	0.50	1	-3	1	0.53	2259	2397	2300	2468	2361	5	0.20	6	0.50	1.50	0.50	1	-3	1	0.53	223	238	226	247	235
1	0.20	2	0.50	1.50	2	1	-3	1	0.43	71	1525	1513	1798	1576	1	0.20	2	0.50	1.50	2	1	-3	1	0.43	34	152	148	180	157
5	0.20	6	0.50	1.50	2	1	-3	1	0.43	170	1694	1664	1915	1486	5	0.20	6	0.50	1.50	2	1	-3	1	0.43	44	169	163	191	151
1	0.20	2	0.50	1.75	0.50	1	-3	0	0.51	2361	2317	2292	2360	2349	1	0.20	2	0.50	1.75	0.50	1	-3	0	0.51	193	231	227	236	232
5	0.20	6	0.50	1.75	0.50	1	-3	0	0.51	1940	2342	2293	2376	2330	5	0.20	6	0.50	1.75	0.50	1	-3	0	0.51	208	233	226	237	231
1	0.20	2	0.50	1.75	2	1	-3	0	0.36	942	1068	1062	1396	1171	1	0.20	2	0.50	1.75	2	1	-3	0	0.35	44	105	102	138	114
5	0.20	6	0.50	1.75	2	1	-3	0	0.36	884	1290	1275	1525	1150	5	0.20	6	0.50	1.75	2	1	-3	0	0.36	65	129	125	153	117
1	0.20	2	0.50	1.75	0.50	1	-3	1	0.51	2367	2312	2292	2377	2355	1	0.20	2	0.50	1.75	0.50	1	-3	1	0.51	189	231	226	237	233
5	0.20	6	0.50	1.75	0.50	1	-3	1	0.51	2359	2340	2297	2393	2331	5	0.20	6	0.50	1.75	0.50	1	-3	1	0.51	208	233	226	239	232
1	0.20	2	0.50	1.75	2	1	-3	1	0.36	697	1066	1058	1420	1150	1	0.20	2	0.50	1.75	2	1	-3	1	0.36	45	105	101	141	113
5	0.20	6	0.50	1.75	2	1	-3	1	0.36	818	1296	1278	1549	1146	5	0.20	6	0.50	1.75	2	1	-3	1	0.36	68	128	125	154	117
1	0.20	2	0.50	2	0.50	1	-3	0	0.49	1382	2258	2255	2301	2296	1	0.20	2	0.50	2	0.50	1	-3	0	0.49	219	225	223	230	227
5	0.20	6	0.50	2	0.50	1	-3	0	0.49	1380	2291	2280	2317	2296	5	0.20	6	0.50	2	0.50	1	-3	0	0.49	205	228	226	231	228
1	0.20	2	0.50	2	2	1	-3	0	0.26	639	513	491	914	631	1	0.20	2	0.50	2	2	1	-3	0	0.26	48	51	46	92	63
5	0.20	6	0.50	2	2	1	-3	0	0.26	1072	785	769	1057	701	5	0.20	6	0.50	2	2	1	-3	0	0.26	43	78	75	106	71
1	0.20	2	0.50	2	0.50	1	-3	1	0.49	1610	2249	2245	2312	2301	1	0.20	2	0.50	2	0.50	1	-3	1	0.49	215	225	223	231	228
5	0.20	6	0.50	2	0.50	1	-3	1	0.49	1308	2282	2269	2324	2292	5	0.20	6	0.50	2	0.50	1	-3	1	0.49	214	226	224	231	227
1	0.20	2	0.50	2	2	1	-3	1	0.27	634	545	513	952	634	1	0.20	2	0.50	2	2	1	-3	1	0.27	51	55	49	95	63
5	0.20	6	0.50	2	2	1	-3	1	0.27	1084	805	784	1088	712	5	0.20	6	0.50	2	2	1	-3	1	0.27	39	80	76	109	72

References

1. Hawkes, A. G.: Spectra of Some Self-Exciting and Mutually Exciting Point Processes. Biometrika **58**(1), 83–90 (1971)
2. Isham, V., Westcott, M.: A self-correcting point process. Stochastic Processes and their Applications **8**, 335–347 (1979)
3. Amstrup, S. C., McDonald, T. L., Manly, B. F. J.: Handbook of Capture– Recapture Analysis. Wiley, London eds. (2003).
4. Chao, A.: Non-parametric estimation of the classes in a population. Scandinavian Journal of Statistics **11**, 265–270 (1984)
5. Bohning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., Arnold, M.: A generalization of Chao's estimator for covariate information. Biometrics **69** (2013)
6. Farcomeni, A.: A general class of recapture models based on the conditional capture probabilities, Biometrics, **72**, 116-124 (2016)
7. Farcomeni, A.: Fully general Chao and Zelterman estimators with application to a whale shark population. Journal of the Royal Statistical Society Series C **67**(1), 217–229 (2018)
8. Farcomeni, A. and Scacciatelli, D.: Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy, Annals of Applied Statistics, 7, 2293-2314 (2013)
9. Schofield, M. R., Barker, R. J., Gelling. N.: Continuous-time capture recapture in closed populations. Biometrics **74**(2), 626–635 (2018)

Longitudinal data analysis using PLS-PM approach

Analisi dei dati longitudinali attraverso l'approccio PLS-PM

Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia, Marina Marino

Abstract Longitudinal data over the past 20 years have seen a greater diffusion in the social sciences. Accompanying this growth was an interest in the methods for analyzing such data. Structural Equation Modeling (SEMs) and especially Partial Least Squares Path Modeling (PLS-PM) are a valuable way to analyze longitudinal data because it is both flexible and useful for answering common research questions. The aim of this paper is to demonstrate how PLS-PM can help us to analyze longitudinal data.

Abstract *I dati longitudinali degli ultimi 20 anni hanno visto una maggiore diffusione nelle scienze sociali. Ad accompagnare questa crescita è stato un interesse per i metodi di analisi di tali dati. I modelli ad equazioni strutturali (SEM) ed in particolare il metodo Partial Least Squares- Path Modeling (PLS-PM) sono un prezioso metodo di analizzare i dati longitudinali perché è sia flessibile che utile per rispondere a domande di ricerca comuni. Lo scopo di questa ricerca è di dimostrare come questo approccio può aiutarci ad analizzare i dati longitudinali.*

Key words: Partial Least Squares - Path Modeling, Longitudinal study

Rosanna Cataldo
Department of Social Science, University of Naples Federico II, vico Monte della Pietà, 1, 80138 Napoli, Italy e-mail: rosanna.cataldo2@unina.it
Corrado Crocetta
Department of Economics, University of Foggia, Largo Papa Giovanni Paolo II, 71121 Foggia, Italy e-mail: corrado.crocetta@unifg.it
Maria Gabriella Grassia
Department of Social Science, University of Naples Federico II, vico Monte della Pietà, 1, 80138 Napoli, Italy e-mail: mgrassia@unina.it
Marina Marino
Department of Social Science, University of Naples Federico II, vico Monte della Pietà, 1, 80138 Napoli, Italy e-mail: mari@unina.it

1 Introduction

A longitudinal study refers to an investigation where participant outcomes are collected at multiple times. Longitudinal data are difficult to collect, but longitudinal research is popular. And this popularity seems to be growing. With this comes the subsequent need for good data analysis methods to analyze these special kinds of data. The analysis of change in longitudinal data has attracted considerable attention over past decades in behavioral research [1]; [3]. Researchers have proposed a number of advanced and sophisticated quantitative approaches to address the stability and change of variables over time [5]; [17]. A broad range of statistical methods exists for analyzing data from longitudinal designs. Each of these methods has specific features and the use of a particular method in a specific situation depends on such things as the type of research, the research question, and so on. The central concern of longitudinal research, however, revolves around the description of patterns of stability and change, and the explanation of how and why change does or does not take place [11]. A common design for longitudinal research in the social sciences is the panel or repeated measures design, in which a sample of subjects is observed at more than one point in time. If all individuals provide measurements at the same set of occasions, we have a fixed occasions design. When occasions are varying, we have a set of measures taken at different points in time for different individuals. Such data occur, for instance, in growth studies, where individual measurements are collected for a sample of individuals at different occasions in their development. The model on longitudinal data can be approached from several perspectives, and the model can be constructed as a Structural Equation Model (SEM). According to Baltes and Nesselroade [2], SEM is a valuable way to analyze longitudinal data because it is both flexible and usefulfor answering common research questions. The explicit invocation of latent variables (LVs) afforded by the SEM makes this framework the one most commonly used to implement and analyze longitudinal data [18]; [23]. The SEM framework is often used to study change processes because this framework provides an opportunity to specify multiple latent variables as predictors and outcomes. SEM techniques include two main methods: covariance-based SEM (CB-SEM), represented by LISREL [14] and component-based SEM, with Partial Least Square-Path Modeling (PLS-PM) [9]; [24]. PLS-PM can be used to implement and to analyze the longitudinal data.

2 Longitudinal data analysis with PLS-PM approach

Recently, Roemer [20] has proposed using the component-based approach to SEM - PLS-PM in a longitudinal study [7]; [24]. In accordance with Roemer [20], we posit that PLS path modeling is highly appropriate for an analysis of the development and change in constructs in longitudinal studies, since it offers three favorable methodological characteristics. First, constructs often need to be predicted in evolutionary models [12]; [22]. Secondly, model complexity quickly increases when

development and change need to be analyzed in longitudinal studies. This is due to the larger number of constructs that are measured at different points in time and the respective effects between those constructs [12]. PLS-PM is well suited to dealing with such complex models [8]; [28]. Thirdly, sample sizes can become quite small in longitudinal studies [13]. PLS path modeling is particularly appropriate in such cases [10]; [15]. In Wold's original design of PLS-PM [27] it was expected that each construct would necessarily be connected to a set of observed variables. On this basis, Lohmöller [16] proposed a procedure to treat hierarchical constructs, the so-called hierarchical component model. There are several main reasons for the inclusion of a Higher-Order Construct Model: Higher-Order Construct Models prove valuable if the constructs are highly correlated; the estimations of the structural model relationships may be biased as a result of collinearity issues, and a discriminant validity may not be established. In situations characterized by collinearity among constructs, a Second-Order Construct can reduce such collinearity issues and may solve discriminant validity problems. PLS path modeling allows for the conceptualization of a hierarchical model, through the use of three main approaches existing in the literature: the Repeated Indicators Approach [16], the Two Step Approach [19] and the Mixed Two Step Approach [4]. The Repeated Indicators Approach is the most popular approach when estimating Higher-Order Constructs in PLS-PM [25]. We propose using a Higher Order Model to analyze longitudinal data. An example, with three points in time, is presented in Fig. 1.

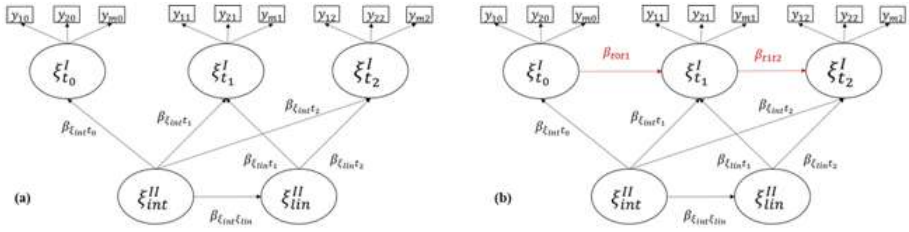


Fig. 1 (a) PLS-PM model - three times with m indicators in each n times; (b) PLS-PM model - three times with m indicators in each n times, with $\xi^I_{t_{i-1}}$ that impact on $\xi^I_{t_i}$

The Higher Order LV ξ^{II}_{int} describes the mean growth, and the LV ξ^{II}_{lin}.the mean slope. ξ^{II}_{int} is reflected in the construct of first order $\xi^I_{t_0}$, $\xi^I_{t_1}....\xi^I_{t_n}$. The construct of second order ξ^{II}_{lin} is reflected in the construct of first order $\xi^I_{t_1}....\xi^I_{t_n}$, The equations of the inner model are:

$$\begin{aligned}
\xi_{lin}^{II} &= \beta_{0lin} + \beta_{\xi_{int}\xi_{lin}}\xi_{int}^{II} + \zeta_{lin} \\
\xi_{t_0}^{I} &= \beta_{0t_0} + \beta_{\xi_{int}t_0}\xi_{int}^{II} + \zeta_{t_0} \\
\xi_{t_1}^{I} &= \beta_{0t_1} + \beta_{\xi_{int}t_1}\xi_{int}^{II} + \beta_{\xi_{lin}t_1}\xi_{lin}^{II} + \zeta_{t_1} \\
&\ldots \\
\xi_{t_i}^{I} &= \beta_{0t_i} + \beta_{\xi_{int}t_i}\xi_{int}^{II} + \beta_{\xi_{lin}t_i}\xi_{lin}^{II} + \zeta_{t_i} \\
\xi_{t_n}^{I} &= \beta_{0t_n} + \beta_{\xi_{int}t_n}\xi_{int}^{II} + \beta_{\xi_{lin}t_1}\xi_{lin}^{II} + \zeta_{t_n}
\end{aligned} \tag{1}$$

where $\beta_{\xi_{int}\xi_{lin}}$ is the strength and sign of the relations between construct ξ_{lin}^{II} and the predictor construct ξ_{int}^{II}; $\beta_{\xi_{int}\xi_{lin}}$ representing the growth mean rate; $\beta_{\xi_{int}t_i}$ is the strength and sign of the relations between construct $\xi_{t_i}^{I}$ and the predictor construct ξ_{int}^{II}; $\beta_{\xi_{lin}t_i}$ is the strength and sign of the relations between construct $\xi_{t_i}^{I}$ and the construct ξ_{lin}^{II}. They indicate how both intercept and slope factors contribute to explaining each time. β_0 is just the intercept term and ζ accounts for the residuals. The intercept term β_0 of each equation should always be non-significant. If we introduce the impact of the LV at i-1 time ($\xi_{t_{i-1}}^{I}$) on the LV at i time ($\xi_{t_i}^{I}$), for its better prediction ($\xi_{t_0}^{I} \rightarrow \xi_{t_1}^{I}$; $\xi_{t_1}^{I} \rightarrow \xi_{t_2}^{I}$;...; $\xi_{t_{n-1}}^{I} \rightarrow \xi_{t_n}^{I}$) as in Fig. 1 (b), the equations of the inner model become:

$$\begin{aligned}
\xi_{lin}^{II} &= \beta_{0lin} + \beta_{\xi_{int}\xi_{lin}}\xi_{int}^{II} + \zeta_{lin} \\
\xi_{t_0}^{I} &= \beta_{0t_0} + \beta_{\xi_{int}t_0}\xi_{int}^{II} + \zeta_{t_0} \\
\xi_{t_1}^{I} &= \beta_{0t_1} + \beta_{\xi_{int}t_1}\xi_{int}^{II} + \beta_{\xi_{lin}t_1}\xi_{lin}^{II} + \beta_{t0t1}\xi_{t_0}^{I} + \zeta_{t_1} \\
&\ldots \\
\xi_{t_i}^{I} &= \beta_{0t_i} + \beta_{\xi_{int}t_i}\xi_{int}^{II} + \beta_{\xi_{lin}t_i}\xi_{lin}^{II} + \beta_{t_{i-1}t_i}\xi_{t_{i-1}}^{I} + \zeta_{t_i} \\
\xi_{t_n}^{I} &= \beta_{0t_n} + \beta_{\xi_{int}t_n}\xi_{int}^{II} + \beta_{\xi_{lin}t_1}\xi_{lin}^{II} + \zeta_{t_n}
\end{aligned} \tag{2}$$

where $\beta_{t_{i-1}t_i}$ that represent the carry-over effects [12]; [6]. A sizeable positive effect means that the individuals' estimations of the construct remain stable over time [6]. In contrast, a small effect means that there has been a substantial reshuffling of the individuals' standings on the construct over time [21]. Finally, a sizeable negative effect means that there has been a reversal of the position of individuals on the structure over time. $\beta_{t_{i-1}t_i}$ contributes to explaining the variability at t time.

As in the CB-SEM framework, the model must be evaluated: first the measurement model and then the structural model. For the measurement model the Dillon-Goldstein's Rho, the mean of communalities and the mean redundancies must be examined. The structural model quality of the inner model must be assessed by examining the following indices: the regression weights, the coefficient of determination (R^2), the redundancy index, and the goodness-of-fit (GoF) statistics [24]. If the structural model quality is well assessed, but one or more *carry-over effects* are negative, this means there are two or more subsamples, with different growth curves. In this case, we suggest splitting the sample into two or more subsamples.

Subsequently, multi-group comparisons could be used to test any differences in the structural path estimates.

3 Conclusions

In this paper we have showed how the PLS-PM approach can be successfully used to analyze longitudinal data. Using PLS-PM we have the best estimation of the measurement model without any problem concerning its identification. The choice of using the PLS-PM is particularly useful for several reasons. First of all, this approach is applicable to small samples and it is capable of estimating rather complex models (with many latent and observable variables) with less restrictive requirements concerning normality and variable and error distributions [10]. Furthermore, PLS-PM approach provides the possibility of working with missing data and in the presence of multi-collinearity. Another advantage of this approach, as compared to other multivariate techniques, is that it examines simultaneous a series of dependence relationship, using a single statistical approach to test the full scope of projected relations [10].

References

1. Ancona, D.G., Okhuysen, G. A and Perlow, L. A: Taking time to integrate temporal research. *Academy of Management Review*, 26 (4), 512-529 (2001)
2. Baltes, P. B. and Nesselroade, J. R: The developmental analysis of individual differences on multiple measures. Academic press (1973)
3. Braun, M. T., Kuljanin, G., and DeShon, R. P.: Spurious results in the analysis of longitudinal data in organizational research. *Organizational Research Methods*, 16(2), 302-330. (2013)
4. Cataldo, R., Grassia, M.G., Lauro, N,C., and Marino, M.:Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators. *Quality & Quantity*, 51 (2), 657–674, Springer (2017)
5. Collins, L. M., and Sayer, A. G.: New methods for the analysis of change. emphAmerican Psychological Association. (2001)
6. Duncan, T.E.,Duncan, S.C. and Strycker, L.A.:An introduction to latent variable growth curve modeling: Concepts, issues, and application, Routledge. 2013)
7. Esposito, Vinzi, V., Chin, W. W., Henseler, J., Wang, H.: *Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications*, Springer, Berlin, Heidelberg, New York (2010)
8. Fornell, C. and Cha, J.:Advanced Methods of Marketing Research, ed. RP Bagozzi, Blackwell, Cambridge. (1994)
9. Henseler, J. and Chin, W. W: A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling, *Structural Equation Modeling*, 17 (1), 82-109 (2010)
10. Henseler, J., Ringle, C.M. and Sinkovics, R.R.:The use of partial least squares path modeling in international marketing. *New challenges to international marketing*, 277–319, Emerald Group Publishing Limited (2009)
11. Kessler, R. C: Greenberg. DF (1981). Linear panel analysis: Models of quantitative change, New York: Academic Press (1986)

12. Johnson, M.D., Herrmann, A. and Huber, F.:The evolution of loyalty intentions, *Journal of marketing*, 70 (2), 122-132, SAGE Publications Sage CA: Los Angeles, CA. (2006)
13. Jones, E., Sundaram, S.and Chin, W.:Factors leading to sales force automation use: A longitudinal analysis. *Journal of personal selling & sales management*, 22 (3), 145-156, Taylor & Francis. (2002)
14. Joreskog, K. G. and Van Thillo, M.: LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. ERIC. (1972)
15. Lauro, N.C., Grassia, M.G. and Cataldo, R.:Model based composite indicators: New developments in partial least squares-path modeling for the building of different types of composite indicators. *Social Indicators Research*, 135 (2), 421–455, Springer (2018)
16. Lohmöller, J.B.:Latent variable path modeling with partial least squares, Springer Science & Business Media (2013)
17. McArdle, J. J. and Nesselroade, J. R: Longitudinal data analysis using structural equation models., *American Psychological Association*. (2014)
18. McArdle, J.J.: Dynamic but structural equation modeling of repeated measures data. *Handbook of multivariate experimental psychology*, 561–614. (1988)
19. Rajala, R. and Westerlund, M.:Antecedents to Consumers' Acceptance of Mobile Advertisements-A Hierarchical Construct PLS Structural Equation Model, 2010 43rd Hawaii International Conference on System Sciences, 1–10, IEEE (2010)
20. Roemer, E.:A tutorial on the use of PLS path modeling in longitudinal studies, *Industrial Management & Data Systems*, 116 (9), 1901-1921,Emerald Group Publishing Limited. (2016)
21. Selig, J.P. and Little, T.D.:Autoregressive and cross-lagged panel analysis for longitudinal data. (2012)
22. Shea, C.M. and Howell, J.M.:Efficacy-performance spirals: An empirical test, *Journal of Management*, 26 (4), 791-812, Sage Publications Sage CA: Thousand Oaks, CA (2000)
23. Stoel, R.D., van den Wittenboer, G. and Hox, J.:Methodological issues in the application of the latent growth curve model, *Recent developments on structural equation models*, 241–261, Springer (2004)
24. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., Lauro, C. N.: PLS Path Modeling. *Computational Statistics and Data Analysis*, 48 (1), 159 - 205 (2005)
25. Wilson, B.:Using PLS to investigate interaction effects between higher order branding constructs. *Handbook of partial least squares*, 621–652, Springer (2010)
26. Wold, H.:Path models with latent variables: The NIPALS approach. *Quantitative sociology*, 307-357, Elsevier (1975)
27. Wold, H.:Soft modeling: the basic design and some extensions. *Systems under indirect observation*, 2, 343 (1982)
28. Wold, H.:Partial least squares. S. Kotz and NL Johnson (Eds.), Encyclopedia of statistical sciences (vol. 6), Wiley, New York (1985)

Long-memory models for count time series
Modelli a memoria lunga per serie di dati di conteggio

Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto

Abstract In this paper we analyze two different approaches for modeling dependent count data with long-memory. The first model we consider explicitly takes into account the integer nature of data and the long-range correlation, while the second model is a count-data long-memory model where the distribution of the current observation is specified conditionally upon past observations. We compare these two different models by looking at their estimation and forecasting performances.
Abstract *In questo lavoro analizziamo due diversi approcci per modellare la dipendenza di lungo periodo in serie storiche di dati di conteggio. Il primo modello considera esplicitamente la natura dei dati e la correlazione di lungo periodo, mentre il secondo è un modello a memoria lunga per dati di conteggio in cui la distribuzione dell'osservazione attuale viene specificata condizionatamente alle osservazioni passate. Il confronto fra questi due approcci è fatto tramite uno stidio Monte Carlo che confronta le performance di stima e previsione.*

Key words: count time series, long-memory, GLM, estimation, forecasting

1 Introduction

Recently, there has been a growing interest in studying nonnegative integer-valued time series and, in particular, time series of counts. In some cases, the discrete values of the time series are large numbers and may be analyzed using continuous-valued models such as ARMA with Gaussian errors. However, when the values are small,

L. Bisaglia
Dept. Statistical Sciences, University of Padova, e-mail: luisa.bisaglia@unipd.it

M. Caporin
Dept. Statistical Sciences, University of Padova, e-mail: massimiliano.caporin@unipd.it

M. Grigoletto
Dept. Statistical Sciences, University of Padova e-mail: matteo.grigoletto@unipd.it

as in the case of counting processes, the usual linear ARMA processes become inappropriate for modeling and forecasting purposes since they would invariably produce non-integer forecast values. One of the most common approaches to build an integer-valued autoregressive (INAR) process is based on a probabilistic operator called binomial thinning, as reported in Al-Osh and Alzaid (1987) and McKenzie (1985) who first introduced INAR processes. A different approach is based on the generalized linear models (GLM) advanced by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). This framework generalizes the traditional ARMA methodology allowing for more flexible dynamics also coherent with count data time series (for details and references, see, for example, Fokianos, 2016).

Long-memory (LM) processes have proved to be useful tools in the analysis of many empirical time series. One of the most popular processes that takes into account this particular behavior of the autocorrelation function is the AutoRegressive Fractionally Integrated Moving Average process (ARFIMA(p,d,q)), independently introduced by Granger and Joyeux (1980) and Hosking (1981). This process generalizes the ARIMA(p,d,q) process by relaxing the assumption that d is an integer. In particular, when $d \in (0,0.5)$ the autocorrelation function of the process decays to zero hyperbolically at a rate $O(k^{2d-1})$, where k denotes the lag. If $p = q = 0$, the process $\{X_t,\, t = 0,\pm 1,\ldots\}$ is called Fractionally Integrated Noise, FI(d). In the following we will concentrate on FI(d) processes with $d \in (0,0.5)$.

Persistent count time series occur for example in finance when modeling stock market daily trading volumes (e.g. Palma and Zevallos, 2011).
In this work, we analyze two different approaches for modeling dependent count data with long-memory. The first model we consider takes explicitly into account the integer nature of data and the long-range correlation, mixing the INteger AutoRegressive (INAR) model proposed by Al-Osh and Alzaid (1987) and McKenzie (1985) with the Fractionally Integrated (FI) model introduced by Granger and Joyeux (1980) and Hosking (1981). The second model, introduced by Palma and Zevallos (2011), builds on a conditional distribution for count data where the parameters' dynamic is characterized by long-memory. We compare estimation and forecasting performances of the two models by Monte Carlo simulations.

2 LM models for count time series data

2.1 GLM approach

Palma and Zevallos (2011) introduce a model for count time series characterized by long-range dependence. They propose a new class of conditional long-memory models (CLMs), where the conditional distribution of the data, given a data-driven parameter, is explicitly specified. The conditional long-memory process, X_t, can be defined as follows:

$$X_t \mid \mathscr{F}_{t-1} \sim G(\lambda_t, g(\lambda_t)), \quad \text{with} \quad \lambda_t = \mu \sum_{j=0}^{\infty} \pi_j - \sum_{j=1}^{\infty} \pi_j X_{t-j} \tag{1}$$

where $\mathscr{F}_t$ is the $\sigma-$field generated by the information up to instant t, $\{X_t, X_{t-1}, \ldots\}$, $G(\alpha, \beta)$ is a distribution corresponding to a continuous or discrete nonnegative random variable with mean α and variance β, both finite, $g(\cdot)$ is a positive function, μ is a constant and $\{\pi_j\}_{j>0}$ is an absolutely summable sequence of real numbers such that $\pi_0 = 1$ and $\pi_j \approx C j^{-d-1}$ for large positive j and some $d < 0.5$. Therefore, given the information $\mathscr{F}_{t-1}$, X_t has distribution G with conditional mean λ_t and conditional variance $g(\lambda_t)$. Obviously, if G is a distribution corresponding to a discrete nonnegative random variable, we obtain a LM model for count time series data. Even if, from a theoretical point of view, this setup is general enough to allow for the use of different integer distributions, in practice only a Poisson has been used by Palma and Zevallos (2011).

It can be shown (Palma and Zevallos, 2011) that model (1) has a non-Gaussian ARFIMA(p,d,q) representation. In particular, for the ARFIMA$(0,d,0)$ case we have $\pi_0 = 1$ and $\pi_j = \Gamma(j-d)/[\Gamma(j+1)\Gamma(-d)]$, for $j \geq 1$.

For further details about CLM-ARFIMA processes see Palma and Zevallos (2011). In particular, it is shown that CLM-ARFIMA and standard ARFIMA processes share the same correlation structure.

The CLM approach allows using all the tools available for GLM models (see, for instance, Liboschik et al., 2017).

2.2 Models based on the thinning operator

Integer-valued autoregressive (INAR) processes, initially proposed by Al-Osh and Alzaid (1987) and McKenzie (1985), in their most basic form follow the recursion:

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t$$

where '$\circ$'is the thinning operator, defined by $\alpha \circ X = \sum_{i=1}^{X} Y_i$ with $X \in \mathbb{N}$, $\alpha \in [0,1]$ and Y_i is a sequence of i.i.d. count random variables, typically Ber(α), independent of X with common mean α. Hence, α plays the role of thinning probability. Moreover, ε_t is a sequence of i.i.d. discrete random variables with mean μ_ε and variance σ_ε^2. While the INAR(1) and INMA(1) models are defined univocally, for the INAR(p) and INMA(q) models there are additional complexities and different types of INAR(p) and INMA(q) processes might be considered (see Weiss, 2018 for a recent review on this topic). Recently, Weiss (2019) developed the INARMA(p,q) process:

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \beta_1 \circ \varepsilon_{t-1} + \ldots + \beta_q \circ \varepsilon_{t-q} + \varepsilon_t$$

where, to obtain feasible stochastic properties, he assumed independence among all thinnings, independence from the innovations, and independence from $(X_s)_{s<t}$

for the thinning at time t. Combining the ideas of the INARMA model with the fractional integration of Granger and Joyeux (1980) and Hosking (1981), Quoreshi (2014) introduce the INARFIMA model based on the following INMA(∞) representation:

$$X_t = \sum_{i=0}^{\infty} \psi_i \circ \varepsilon_{t-i} \tag{2}$$

where $\psi_0 = 1$ and $\psi_i = \Gamma(i+d)/[\Gamma(i+1)\Gamma(d)]$, for $i \geq 1$, and d is the long memory coefficient. Since the ψ_i in (2) are considered thinning probabilities, then $d \in [0,1]$. Quoreshi (2014) proposes different estimation methods, based on conditional least squares, feasible generalized least squares and the generalized method of moments. In his paper Quoreshi (2014) does not consider the problem of forecasting with the INARFIMA model.

In the present paper, differently from Quoreshi (2014), who adopts the MA(∞) representation, to take into account the long memory and integer nature of data, we propose to consider the INAR(∞) recursion:

$$X_t = \sum_{i=1}^{\infty} \pi_i \circ X_{t-i} + \varepsilon_t \tag{3}$$

that is:

$$\sum_{i=0}^{\infty} \pi_i \circ X_{t-i} = (1 - B^{\circ})^d X_t = \varepsilon_t \tag{4}$$

where $\pi_0 = 1$ and $\pi_j = \Gamma(j-d)/[\Gamma(j+1)\Gamma(-d)]$ for $j \geq 1$, with $d \in (0, 0.5)$. As in Du and Li (1991) the ε_t constitutes a sequence of i.i.d. discrete random variables independent of all counting series, and all thinning operations are mutually independent. The conditional mean of process (3) is given by:

$$E[X_t \mid X_{t-1}, \ldots] = \mu_\varepsilon + \sum_{i=1}^{\infty} \pi_i X_{t-i}$$

and thus the autocorrelation function of X_t is the same of an I(d) process. Moreover, the conditional variance is:

$$V[X_t \mid X_{t-1}, \ldots] = \sigma_\varepsilon^2 + \sum_{i=1}^{\infty} \pi_i (1 - \pi_i) X_{t-i} \, .$$

In practice, only $X_1, \ldots, X_n$, are available, but X_t depends on the infinite past of the process. Therefore, we must approximate X_t with an AR(p), taking p large enough so that $\sum_{i=p+1}^{\infty} \pi_i \circ X_{t-i}$ in (3) is negligible. In our applications, all the available past observations are used.

2.3 Forecasting LM models for count time series

For CLM-ARFIMA models the natural one-step predictor of λ_{n+1} conditional on the past information, $\mathscr{F}_n$, is based on (1) and can be written as: $\hat{\lambda}_{n+1} = \hat{\mu}\sum_{j=0}^{n}\hat{\pi}_j - \sum_{j=1}^{n}\hat{\pi}_j X_{n+1-j}$ where each $\hat{\pi}_j$ depends on the long-memory parameter estimate, $\hat{d}$ (and other parameters, if present). Hence, the predicted conditional distribution is

$$\hat{X}_{n+1} \mid \mathscr{F}_n \sim G(\hat{\lambda}_{n+1}, g(\hat{\lambda}_{n+1}))$$

and the construction of conditional prediction intervals for one-step forecasts is a simple task. For k-step forecasts, with $k > 1$, we have to recursively use previous forecast values, for example: $\hat{\lambda}_{n+2} = \hat{\mu}\sum_{j=0}^{n+1}\hat{\pi}_j - \hat{\pi}_{n+1}\hat{X}_{n+1} - \sum_{j=2}^{n+1}\hat{\pi}_j X_{n+2-j}$ and the predicted conditional distribution is

$$\hat{X}_{n+2} \mid \mathscr{F}_n \sim G(\hat{\lambda}_{n+2}, g(\hat{\lambda}_{n+2}))\ .$$

In this case, the construction of conditional prediction intervals is not immediate and we obtain prediction intervals via computational methods. Forecasting with INARFIMA models is very simple too. Using (3), we have:

$$\hat{X}_{n+k} = \sum_{i=1}^{\infty}\hat{\pi}_i \circ \hat{X}_{n+k-i}\ ,$$

with $\hat{X}_j = X_j$ for $j \leq n$. Also in this case, prediction intervals cannot be directly recovered and we must resort to computational methods.

3 Some simulation results

In this section we provide the results of some Monte Carlo experiments we carried out to asses the estimation performance of different long-memory parameter estimators. Count time series of lengths $n = 500$ and $n = 1000$ were generated from models (1) and (3). The Poisson and Negative Binomial distributions were used for the conditional distribution G in (1) (models CP and CNB) and for ε_t in the INAR model (3) (models PI and NBI). The distribution parameters were chosen so that the amount of over-dispersion in the different cases is comparable. The functions we use are written in the `R` language and are available upon request from the authors. Table 1 shows the average estimates (and their standard deviations, in parentheses) over 1000 Monte Carlo replications. The estimation methods considered are maximum likelihood (ML), Geweke and Porter-Hudak (GPH) and Whittle (WH). Simulation results show how the long memory parameter d is, for all models, correctly estimated on average, with the ML and WH methods yielding comparable standard deviations, while the GPH method performs considerably worse.

		NBI				CNB			
d	*n*	OD	ML	GPH	WH	OD	ML	GPH	WH
0.15	500	1.5	0.14 (0.035)	0.16 (0.169)	0.14 (0.036)	1.1	0.14 (0.036)	0.16 (0.165)	0.14 (0.037)
	1000		0.15 (0.026)	0.15 (0.136)	0.15 (0.026)		0.14 (0.026)	0.15 (0.132)	0.15 (0.026)
0.35	500	1.4	0.34 (0.037)	0.36 (0.168)	0.34 (0.039)	1.5	0.34 (0.037)	0.36 (0.168)	0.34 (0.039)
	1000		0.35 (0.025)	0.39 (0.141)	0.35 (0.026)		0.34 (0.026)	0.35 (0.141)	0.35 (0.027)
0.45	500	2.9	0.43 (0.030)	0.47 (0.181)	0.44 (0.034)	2.2	0.43 (0.033)	0.46 (0.169)	0.44 (0.036)
	1000		0.44 (0.024)	0.48 (0.141)	0.45 (0.027)		0.44 (0.024)	0.46 (0.144)	0.45 (0.027)
		PI				CP			
d	*n*	OD	ML	GPH	WH	OD	ML	GPH	WH
0.15	500	5.5	0.14 (0.038)	0.15 (0.165)	0.14 (0.038)	1.1	0.14 (0.037)	0.15 (0.169)	0.14 (0.038)
	1000		0.15 (0.026)	0.16 (0.137)	0.15 (0.026)		0.15 (0.025)	0.15 (0.135)	0.15 (0.025)
0.35	500	2.5	0.34 (0.035)	0.37 (0.166)	0.35 (0.036)	1.4	0.34 (0.036)	0.35 (0.182)	0.34 (0.038)
	1000		0.35 (0.026)	0.38 (0.146)	0.35 (0.027)		0.34 (0.025)	0.35 (0.138)	0.35 (0.026)
0.45	500	3.8	0.43 (0.030)	0.48 (0.166)	0.45 (0.034)	2.7	0.43 (0.031)	0.47 (0.175)	0.44 (0.035)
	1000		0.44 (0.022)	0.48 (0.145)	0.45 (0.025)		0.44 (0.023)	0.47 (0.131)	0.45 (0.025)

Table 1 Estimation of the long memory parameter d for series generated from different models, having comparable over-dispersions (OD). The considered models are the Poisson INAR (PI), Conditional Poisson (CP), Negative Binomial INAR (NBI) and Conditional Negative Binomial (CNB). The estimation methods are maximum likelihood (ML), the Geweke and Porter-Hudak estimator (GPH) and the Whittle estimator (WH). Results show the average estimates and their standard deviations (in parentheses) over 1000 Monte Carlo replications.

References

1. Al-Osh, M. A. and A. A. Alzaid: First order integer-valued autoregressive INAR(1) process. Journal of Time Series Analysis **8**, 261–275 (1987)
2. Du, J.G. and Y. Li: The integer-valued autoregressive INAR(p) model. Journal of Time Series Analysis **12**, 129–142 (1991)
3. Fokianos, K.: Statistical analysis of count time series models: A GLM perspective. In Davis R., Holan S., Lund R., and Ravishanker R. (eds.), Handbook of Discrete-Valued Time Series, pp. 3-27. Chapman & Hall/CRC (2016)
4. Granger, C. and R. Joyeux: An introduction to long-range time series models and fractional differencing. Journal of Time Series Analysis **1**, 15–30 (1980)
5. Hosking, J.: Fractional differencing. Biometrika **68**, 165–176 (1981)
6. Liboschik, T., K. Fokianos, and F. R.: tscount: An R package for analysis of count time series following generalized linear models. Journal of Statistical Software **82**, 1–51 (2017)
7. McCullagh, P. and J. Nelder: Generalized Linear Models. London, U.K. Chapman & Hall (1989)
8. McKenzie, E.: Some simple models for discrete variate time series. Water Resources Bulletin **21**, 645–650 (1985)
9. Nelder, J. and R. Wedderburn: Generalized Linear Models. Journal of the Royal Statistical Society, Series A, **135**, 370–384 (1972)
10. Palma, W. and M. Zevallos: Fitting non-Gaussian persistent data. Applied Stochastic Models in Business and Industry **27**, 23–36 (2011)
11. Quoreshi, A: A long-memory integer-valued time series model, INARFIMA, for financial application. Quantitative Finance **14**, 2225–2235 (2014)

Combining multiple frequencies in Realized GARCH models

Combinazione di frequenze in modelli Realized GARCH

Naimoli Antonio and Storti Giuseppe

Abstract This paper proposes extensions of the Realized GARCH model of Hansen et al. (2012) by incorporating information from multiple realized volatility measures computed at different sampling frequencies to achieve an optimal trade-off between bias and efficiency. Future volatility forecasts are determined by a weighted average of the considered realized measures, where the weights are time-varying and adaptively determined according to the estimated amount of noise and jumps. This specification aims to reduce bias effects related to the different sampling frequencies at which the realized measure are computed.

Sommario *L'obiettivo di questo lavoro è quello di estendere il modello Realized GARCH di Hansen et al. (2012) incorporando informazioni provenienti da molteplici misure di volatilità realizzate calcolate a diverse frequenze di campionamento al fine di raggiungere un trade-off ottimale tra distorsione ed efficienza. La volatilità futura viene determinata come media ponderata delle misure realizzate considerate, usando pesi variabili nel tempo e determinati in modo adattivo in base all'errore di misura stimato. Questa specificazione mira a ridurre gli effetti della distorsione legati alla frequenza di campionamento con cui vengono calcolate le misure realizzate.*

Key words: Realized GARCH, realized variance, measurement error, sampling frequency, volatility forecasting.

Naimoli Antonio
Università degli Studi di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovannni Paolo II, 132, 84084 Fisciano (SA), Italy. e-mail: anaimoli@unisa.it

Storti Giuseppe
Università degli Studi di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovannni Paolo II, 132, 84084 Fisciano (SA), Italy. e-mail: storti@unisa.it

1 Introduction

Accurate estimation of volatility is of crucial importance in investment and risk management activities. The wide dissemination of high-frequency financial data has inspired researchers to develop ex-post volatility estimators to measure the quadratic variation of asset prices. It is well recognized that the use of intra-day data in dynamic models for forecasting volatility allows for more accurate forecasts.

In the absence of microstructure noise and measurement errors, Andersen et al. (2003) show that the Realized Variance (RV), using all available data, produces a consistent estimate of the latent Integrated Variance.[1] This suggest that the RV should be computed by using tick-by-tick data or intra-daily returns sampled at the highest possible frequency. However, due to market microstructure frictions, RV tends to be biased as the intra-daily sampling frequency increases. This implies a trade-off between bias and variance related to the working sampling frequency used.[2] This trade-off is the reason why realized volatility measures are often based on intraday returns sampled at a frequency ranging from 5 to 30 minute. However, although in most settings the 5-minute aggregation is sufficient to control the effects of microstructure noise, it may not be sufficient to filter out the effect of jumps. This leads to the use of jump-robust estimators[3] or to aggregate intra-daily returns at lower frequencies in computing realized measures. The drawback of using these solutions is a substantial loss of efficiency on days when the jumps are negligible or not present at all. This emphasizes the duality between variability, maximized at higher sampling frequencies, and bias, due to microstructure noise and jumps, vanishing at lower frequencies.

This paper proposes dynamic models for forecasting daily volatility based on the combination of realized volatility measures computed at different frequencies aiming to achieve an optimal "compromise" between bias and efficiency. In particular, in our approach, future volatility forecasts are driven by the variation of a weighted average of multiple realized measures, where the weights vary over time depending on the estimated amount of noise and jumps. The proposed modelling approach is developed on the basis of the Realized GARCH (RGARCH) (Hansen et al., 2012). Here, we focus on models that include two realized variances based on different discretization grids, and therefore the proposed specification is called the *Mixed Frequency* Realized GARCH (MF-RGARCH). The model includes separate measurement equations for each of the realized measure considered. We also develop a variant of the MF-RGARCH characterized by a single measurement equation, the SMF-RGARCH model, in which the dependent variable is a time-weighted average of the "raw" realized measures.

---

[1] The Realized Variance is given by $RV_t = \sum_{i=1}^{M} r_{t,i}^2$, where $r_{t,i} = p_{t-1+i\Delta} - p_{t-1+(i-1)\Delta}$ is the i-th Δ-period intraday return, with p_t the logarithmic price process of a financial asset and $M = 1/\Delta$ the sampling frequency.

[2] See, for example, Aït-Sahalia et al. (2005), Zhang et al. (2005) and Bandi and Russell (2008), among others.

[3] See, among others, Barndorff-Nielsen and Shephard (2004b) and Andersen et al. (2012).

2 Mixed Frequency models

In this section, we present extensions of the standard RGARCH aiming to achieve an optimal trade-off between bias and variability by exploiting the information of realized measures based on different sampling frequencies. In our modelling approach, volatility dynamics are given by a time-weighted average of realized measures and, as in Hansen and Huang (2016), models also include different measurement equations for each realized measure considered. Therefore, the resulting model specification is denoted as Mixed Frequency Realized GARCH (MF-RGARCH) model. In particular, the MF-RGARCH is specified as

$$r_t = \sqrt{h_t} z_t \tag{1}$$

$$log(h_t) = \omega + \beta\, log(h_{t-1}) + \gamma log(\tilde{x}_{t-1}) \tag{2}$$

$$log(\tilde{x}_t) = \alpha_t\, log\left(x_t^{(H)}\right) + (1-\alpha_t)\, log\left(x_t^{(L)}\right) \tag{3}$$

$$log\left(x_t^{(j)}\right) = \xi_j + \varphi_j\, log(h_t) + \tau_j(z_t) + u_t^{(j)} \qquad j = H, L \tag{4}$$

where r_t is the daily open-to-close log-return at time t, $x_t^{(H)}$ and $x_t^{(L)}$ denote realized measures computed at a higher and a lower frequency, respectively, h_t is the conditional variance of returns, $z_t \overset{iid}{\sim} (0,1)$ and $u_t^{(j)} \overset{iid}{\sim} (0, \sigma^2_{u,j})$. Consequently, $\log(\tilde{x}_t)$ is a time-weighted average of two log-realized measures based on different sampling frequencies, where the weight α_t is given by:

$$\alpha_t = \alpha_0 + \alpha_1 R_t \text{ and } R_t = \left(RQ_t^{(L)} / RQ_t^{(H)}\right)^{0.5}.$$

Therefore, α_t is a linear function of the ratio between the Realized Quarticity (RQ) measures computed at low and high frequency respectively.[4] The ratio R_t plays the role of a state variable whose impact is related to the level of bias due to noise and jumps. As discussed in Barndorff-Nielsen and Shephard (2004a) and Bandi and Russell (2008), among others, the RQ highlights the sensitivity of fourth moment of returns to outliers and market microstructure noise. They found that these effects tend to be reduced as the returns sampling frequency shrinks. Being the ratio between RQs computed at different frequencies, R_t has the function of correcting for upward and downward bias eventually affecting the involved realized measures.

Furthermore, by collapsing the two measurement equations into one single equation in which the dependent variable is given by the weighted average $log(\tilde{x}_t)$ we obtain the *Single* equation Mixed Frequency Realized GARCH (SMF-RGARCH). Namely, equation (4) is replaced by: $log(\tilde{x}_t) = \xi + \varphi\, log(h_t) + \tau(z_t) + \tilde{u}_t$.

[4] The Realized Quarticity is defined as $RQ_t^{(j)} = \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4, j = H, L$.

3 Empirical Application

In order to assess the relative merits of the proposed modelling approach, we present an empirical application on a set of 4 German stocks: Bayerische Motoren Werke (BMW), Daimler (DAI), Deutsche Telekom (DTE) and Metro Group (MEO). Namely, we consider daily open-to-close log-returns and daily realized variances and quarticities based on 30 second, 1, 5, 10, 15 and 30 minute frequencies. The sample period spans a total of 2791 trading days ranging from 02/01/2002 to 27/12/2012. The raw data have been filtered according to the procedure described in Brownlees and Gallo (2006), where only continuous trading transactions within the regular market hours 9:00 am - 5:30 pm have been used to build the variables of interest. In the following, we refer to $m = minutes$ and $s = seconds$ to denote the sampling frequency of the realized measures employed in our analysis.

The out-of-sample predictive ability of the fitted models has been assessed by means of a rolling window forecasting exercise using a window of 1500 days, where the model parameters are recursively re-estimated every day. Also, we assume a Gaussian specification of z_t and the measurement error of the high and low realized measure, such that $z_t \overset{iid}{\sim} N(0,1)$, $u_t^{(H)} \overset{iid}{\sim} N(0,\sigma^2_{u,H})$ and $u_t^{(L)} \overset{iid}{\sim} N(0,\sigma^2_{u,L})$. The forecasting exercise covers the 2008 credit crisis and the turmoil period related to the instability of the Euro area in the late 2011, ranging from 15 November 2007 to 27 December 2012 for a total of 1298 days.

We consider the standard RGARCH as a benchmark, taking the 5-minute *RV* as volatility measure. Also, we set the 5-minute *RV* as the base frequency in the (S)MF-RGARCH models, then mixed with realized measures computed at higher frequencies of 30 second and 1 minute, and lower frequencies of 10, 15 and 30 minute.[5] The forecast accuracy has been evaluated by the QLIKE loss function, assessing the significance of differences in forecasting performance across different models by the Model Confidence Set (MCS) of Hansen et al. (2011) considering the confidence levels of 75% and 90%. We refer to the QLIKE loss function since it is more powerful in rejecting poorly performing predictors (see e.g. Patton (2011) and Liu et al. (2015) among others), among the class of robust loss functions for volatility forecast evaluation. The QLIKE loss function is given by

$$QLIKE = T^{-1} \sum_{t=1}^{T} \left(log(\hat{h}_t) + RV_t / \hat{h}_t \right). \tag{5}$$

The left panel of Table 1 shows that the QLIKE loss function, using the 5-minute realized variance as volatility proxy, is always minimized by the MF-RGARCH models mixing the 5-minute *RV* with realized measures based on intra-daily returns sampled at lower frequencies. Furthermore, the results of the MCS point out that the MF-RGARCH$^{(5_m,10_m)}$ is the only model that always falls into the MCS at

[5] This choice is due to the fact that it has been provided empirical evidence that, in most settings, the 5-minute frequency ensures better performance in terms of fitting and forecasting than realized measures computed at different frequencies (Liu et al., 2015).

Table 1: Average values of QLIKE loss using 5-min *RV* as volatility proxy (left) and MCS p-values (right). For each stock: in **boldface**: minimum loss; in box models $\in$ 90% MCS and in box models $\in$ 75% MCS.

	Average QLIKE				MCS p-values			
	BMW	DAI	DTE	MEO	BMW	DAI	DTE	MEO
RGARCH	-6.6763	-6.5630	-7.3547	-6.9452	0.0060	0.0903	0.0560	0.0757
SMF$^{(5_m,30_s)}$	-6.6432	-6.5401	-7.2688	-6.9356	0.0000	0.0337	0.0000	0.0250
SMF$^{(5_m,1_m)}$	-6.6618	-6.5390	-7.3099	-6.9391	0.0000	0.0567	0.0013	0.0387
SMF$^{(5_m,10_m)}$	-6.6761	-6.5582	-7.3460	-6.9473	0.0130	0.0817	0.0430	0.0597
SMF$^{(5_m,15_m)}$	-6.6762	-6.5674	-7.3475	-6.9440	0.0050	0.0817	0.0680	0.0543
SMF$^{(5_m,30_m)}$	-6.6747	-6.5631	-7.3523	-6.9484	0.0060	0.0903	0.1307	0.0810
MF$^{(5_m,30_s)}$	-6.6704	-6.5938	-7.3311	-6.9481	0.0005	0.1090	0.0033	0.1577
MF$^{(5_m,1_m)}$	-6.6698	-6.5967	-7.3431	-6.9540	0.0020	0.1090	0.0130	0.3310
MF$^{(5_m,10_m)}$	-6.6840	**-6.6047**	-7.3656	-6.9558	0.5870	1.0000	0.8907	0.3310
MF$^{(5_m,15_m)}$	-6.6845	-6.5844	-7.3604	-6.9571	0.5870	0.0903	0.2483	0.3310
MF$^{(5_m,30_m)}$	**-6.6853**	-6.5991	**-7.3659**	**-6.9597**	1.0000	0.1090	1.0000	1.0000

the confidence level of 75%. The MF-RGARCH models based on the $(5_m, 15_m)$ and $(5_m, 30_m)$ frequency combinations still provide a good performance always entering the MCS at the considered confidence levels, with the exception of the MF-RGARCH$^{(5_m,15_m)}$ for DAI. In contrast, the MF-RGARCH models combining the 5-minute *RV* with realized volatility measures computed at higher frequencies are overall characterized by a lower predictive accuracy since they are only included in the MCS in some isolated cases. On the other hand, the predictive performance of models that include a single measurement equation, the SMF-RGARCH, is not as satisfactory. Finally, even the standard RGARCH never enters in the set of superior models. It is close in entering the 90% MCS for the stock DAI, but it is clearly outperformed by its mixed frequency counterparts.

4 Conclusions

We propose flexible generalizations of the Realized GARCH model of Hansen et al. (2012), where the dynamics of the conditional variance are driven by a weighted average of realized variances computed using intra-daily information at different frequencies. The coefficients of the weighted average are time-varying and adaptively estimated in order to guarantee, in a fully data driven fashion, an optimal bias-variance trade-off. The proposed approach can be used to generate improved conditional variance forecasts and, in an ex-post framework, also to compute an optimized volatility measure. The out-of-sample forecasting exercise shows that the mixed-frequency models can allow for substantial improvements

in terms of forecasting accuracy over the standard Realized GARCH as measured by the QLIKE loss function. The only models entering MCS at the considered confidence levels, for all the analyzed assets, are of the MF-RGARCH type, thus confirming the good predictive power of this class of models.

References

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies 18*(2), 351–416.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*(2), 579–625.

Andersen, T. G., D. Dobrev, and E. Schaumburg (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics 169*(1), 75–93.

Bandi, F. M. and J. R. Russell (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies 75*(2), 339–369.

Barndorff-Nielsen, O. E. and N. Shephard (2004a). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica 72*(3), 885–925.

Barndorff-Nielsen, O. E. and N. Shephard (2004b). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics 2*(1), 1–37.

Brownlees, C. T. and G. M. Gallo (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis 51*(4), 2232–2245.

Hansen, P. R. and Z. Huang (2016). Exponential garch modeling with realized measures of volatility. *Journal of Business & Economic Statistics 34*(2), 269–287.

Hansen, P. R., Z. Huang, and H. H. Shek (2012). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics 27*(6), 877–906.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Liu, L. Y., A. J. Patton, and K. Sheppard (2015). Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics 187*(1), 293–311.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics 160*(1), 246–256.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association 100*(472), 1394–1411.

Models with Time-Varying Parameters for Realized Covariance

Modelli con Parametri Variabili per la Realized Covariance

Luc Bauwens and Edoardo Otranto

Abstract Several recent contributions in econometrics and statistics deal with the dynamic modelling of conditional covariance matrices. To guarantee the positive definiteness of the estimated covariance matrices and to obtain parsimonious models, most of the models proposed use scalar parameterizations that involve a small number of parameters, but have the drawback to impose constraints that may strongly restrict the flexibility of the dynamics of the conditional covariance or correlation process. Using the properties of the Hadamard exponential functions, we develop parsimonious but flexible models, which provide positive definite covariance matrices with different and time varying coefficients for each element of the covariance matrix. Their properties are verified with an empirical exercise, using realized covariance daily data for 29 assets.

Abstract *Vari contributi nell'ambito della letteratura statistica ed econometrica riguardano la modellistica dinamica di matrici di covarianze condizionate. Per ottenere matrici di covarianze stimate definite positive e nello stesso tempo modelli parsimoniosi, la maggior parte delle proposte è basata su parametrizzazioni scalari, che rendono poco flessibile la dinamica dei processi che generano le covarianze o le correlazioni. In questo lavoro vengono sviluppati modelli, basati sulle proprietà della funzione esponenziale di Hadamard, molto flessibili e parsimoniosi, che producono matrici di covarianze definite positive con differenti parametri (variabili nel tempo) per ogni elemento delle matrici. Le proprietà di questi modelli sono verificate con una applicazione empirica su un data set di 29 serie finanziarie.*

Key words: Hadamard Exponential, CAW models, Dynamic Conditional Correlation, Positive definiteness

---

Luc Bauwens
Université catholique de Louvain, e-mail: luc.bauwens@uclouvain.be

Edoardo Otranto
University of Messina, e-mail: eotranto@unime.it

1 Introduction

The dynamic modelling of conditional covariance matrices is the topic of a large number of contributions in financial modelling. The literature started by extending the univariate GARCH model to the multivariate case, developing progressively the family of MGARCH models. Due to the availability of intraday prices and the development of realized volatility measures, attention shifted to the dynamic modelling of realized covariances and correlations. This has resulted in new classes of models for positive definite matrices, such as the Conditional Autoregressive Wishart (CAW) models proposed by Golosnoy et al. (2012). When the number of financial series (n hereafter) is not small (say, more than ten), the estimation of MGARCH and CAW models is difficult due to two problems: 1) the requirement of a parameterization of the dynamic process for the covariance matrix which guarantees its positive definiteness; 2) the "curse of dimensionality" problem (meaning that the number of parameters of such models tends to be large, increasing at least quadratically with n).

A turning point in MGARCH models was signed by the introduction of the DCC model of Engle (2002), who proposed a two-step specification and estimation procedure. The log-likelihood function based on the assumption of multivariate normality can be maximized in two steps. The first step provides the estimation of n univariate GARCH models for the conditional variances of the n series, whereas in the second step, the targeting of the constant matrix term is used before the correlation process parameters are estimated conditionally on the estimates of the first step. The advantage is not only in computational easiness, but also in goodness of fit: the possibility to adopt different parameters to modelize each variance increases the fit of this part (with respect to scalar BEKK), whereas for the correlation part, an assumption of common dynamics (scalar DCC) fixes the number of parameters to two. On the other hand the two-step procedure involves a loss of statistical efficiency because both sets of parameters are estimated using a limited information approach. The previous considerations are valid also for CAW models that use the DCC formulation, as proposed by Bauwens et al. (2016). The desire to modelize large dimensional covariance matrices has favoured the success of the two-step DCC formulation based on variances and correlations, at the expense of the direct joint modeling of the variances and covariances.

Our main contribution is a set of new parameterizations of the CAW model family, extending the existing BEKK-type formulation of Golosnoy et al. (2012) and the DCC-type formulation of Bauwens et al. (2016). The proposed new parameterizations imply a specific impact parameter of the lagged realized covariance (or correlation) on the next conditional covariance (or correlation) for each asset pair; moreover these impact parameters are time-varying. They nevertheless guarantee the positive definiteness of the conditional covariance (or correlation) matrix with simple parametric restrictions, while keeping the number of parameters fixed or at most near to n. In brief, they are more flexible than existing scalar or rank-1 BEKK and DCC versions, while adding a single scalar parameter to these models, hence, they remain parsimonious. This proposed parameterizations use the element–by–

element (Hadamard) exponential function of a matrix to define the impact parameter matrix of the lagged realized covariances (or correlations) of the conditional process. The model proposed is more general than the one proposed in Bauwens and Otranto (2020); it deals with realized volatility, uses different parameterizations involving the Hadamard exponential operator and extends the previous approach to the modelization of covariance matrices in a 1-step estimation framework.

2 A class of new models

Let $\mathbf{C}_t$ denote the $(n \times n)$ realized covariance matrix of day t ($t = 1, \ldots, T$), and $\mathscr{I}_t$ the information set at time t, consisting of the past values of $\mathbf{C}_t$. In the CAW framework, the conditional distribution of $\mathbf{C}_t$ is a $n-$dimensional central Wishart with $\nu (> n)$ degrees of freedom; in symbols:

$$\mathbf{C}_t | \mathscr{I}_{t-1} \sim W_n (\nu, \mathbf{S}_t / \nu) \tag{1}$$

where $\mathbf{S}_t$, of dimension $n \times n$, is the (positive semidefinite) expected value of the conditional distribution of $\mathbf{C}_t$. There is a large set of options to specify $\mathbf{S}_t$, inspired by the MGARCH literature; for example, Golosnoy et al. (2012) adopt a BEKK model. The diagonal BEKK-CAW model is the following process:

$$\mathbf{S}_t = \left(1 - \bar{a} - \bar{b}\right) \bar{\mathbf{C}} + \mathbf{A} \odot \mathbf{C}_{t-1} + \mathbf{B} \odot \mathbf{S}_{t-1} \tag{2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are unknown positive semidefinite symmetric matrices of parameters, $\bar{a}$ and $\bar{b}$ are the averages of the elements of $\mathbf{A}$ and $\mathbf{B}$ respectively, and $\bar{\mathbf{C}}$ is the sample mean of the realized covariances $\mathbf{C}_t$. We call the models belonging to this family the *COV* models. Bauwens et al. (2016) specify $\mathbf{S_t}$ in the CAW framework using the DCC approach of MGARCH. They name the model "Re–cDCC". It consists of n univariate models for the conditional variances, and a scalar DCC model for the realized correlation matrix. Denoting with $\mathbf{D}_t$ a diagonal matrix with elements given by the diagonal elements of $\mathbf{S}_t$, and $\mathbf{R}_t = \mathbf{D}_t^{-1} \mathbf{S}_t \mathbf{D}_t^{-1}$, the Re–cDCC model is defined as the following set of equations:

$$\begin{aligned}
&\mathbf{R}_t = \tilde{\mathbf{Q}}_t^{-1/2} \mathbf{Q}_t \tilde{\mathbf{Q}}_t^{-1/2}, \\
&\mathbf{Q}_t = \left(1 - \bar{a} - \bar{b}\right) \bar{\mathbf{R}} + \mathbf{A} \odot \left(\tilde{\mathbf{Q}}_{t-1}^{1/2} \mathbf{D}_{t-1}^{-1} \mathbf{C}_{t-1} \mathbf{D}_{t-1}^{-1} \tilde{\mathbf{Q}}_{t-1}^{1/2} \right) + \mathbf{B} \odot \mathbf{Q}_{t-1}, \\
&\tilde{\mathbf{Q}}_t = \text{diag}(\mathbf{Q}_t),
\end{aligned} \tag{3}$$

where for any square matrix $\mathbf{X}$, $\text{diag}(\mathbf{X})$ is the diagonal matrix obtained by setting to zero all the off-diagonal elements of $\mathbf{X}$; $\bar{\mathbf{R}}$ is the sample correlation matrix, computed from the realized correlations. We call the models belonging to this family the *COR* models.

Bauwens and Otranto (2020), in the framework of MGARCH conditional correlation models, provide extensions of the scalar DCC model of Engle (2002), where

the elements of $\mathbf{A}$ depend in a nonlinear way on the lagged conditional correlations. In particular, in their model, called NonLinear AutoRegressive Correlation (NLARC) model, the effect of the lagged conditional correlations enters through the element–by–element (Hadamard) exponential function, denoted by $\exp^{\odot}$. The objective of adding flexibility in models (2) and (3), while maintaining a parsimonious parameterization, can be obtained by extending and generalizing the Bauwens and Otranto (2020) NLARC parameterization to the CAW model family. The matrix $\mathbf{A}$ becomes time-varying and is denoted by $\mathbf{A}_t$ in the sequel. Three parameterizations of the time-varying matrix $\mathbf{A}_t$ for (2) and of (3) are introduced below. Similar versions can be adopted for $\mathbf{B}$. The three proposed parameterizations are:

$$\begin{array}{ll} S(Scalar): & \mathbf{A}_t = a\exp^{\odot}(\phi_A\mathbf{M}_t) = a\mathbf{J}_n \odot \exp^{\odot}(\phi_A\mathbf{M}_t) \\ R1(Rank-1): & \mathbf{A}_t = \mathbf{a}\mathbf{a}' \odot \exp^{\odot}(\phi_A\mathbf{M}_t) \\ EO(Equal \quad Off-diagonal): & \mathbf{A}_t = \mathbf{A}_v \odot \exp^{\odot}(\phi_A\mathbf{M}_t), \end{array} \tag{4}$$

where $\phi_A \geq 0$, $\mathbf{M}_t$ is a positive definite symmetric matrix known at date t, $a \in (0,1)$ is a scalar in the first parameterization, and in the second one, $\mathbf{a}$ is an n-dimensional vector in which each element is in $(0,1)$, $\mathbf{J}_n$ is a matrix of ones. In the third parameterization, $\mathbf{A}_v$ is a matrix with diagonal elements $a_1,\ldots,a_n$, and equal off-diagonal (EO) elements, all equal to a_c, with $0 \leq a_c \leq a_i < 1$ for each $i = 1,...n$. This constraint provides a sufficient condition for $\mathbf{A}_v$ to be positive semidefinite. Notice that if $\phi_A = 0$, $\exp^{\odot}(\phi_A\mathbf{M}_t)$ is equal to $\mathbf{J}_n$, so that $\mathbf{A_t}$ is constant, being equal to $a\mathbf{J}_n$ (scalar model), $\mathbf{a}\mathbf{a}'$ (rank-1 model), or $\mathbf{A}_v$ (EO model). Two time-varying versions of $\mathbf{M}_t$ are used in the Hadamard exponential function of $\mathbf{A}_t$ when $\phi_A > 0$:

$$\begin{array}{l} Pt: \mathbf{M}_t = \mathbf{P}_{t-1} - \mathbf{J}_n, \\ Rt: \mathbf{M}_t = \mathbf{R}_{t-1} - \mathbf{J}_n, \end{array} \tag{5}$$

where $\mathbf{P}_{t-1}$ is the realized correlation matrix obtained by transforming the realized covariance matrix $\mathbf{C}_{t-1}$ into a correlation matrix, and $\mathbf{R}_{t-1}$ is the conditional correlation matrix. In the *COV* models, the latter is obtained by transforming $\mathbf{S}_{t-1}$ into a correlation matrix, and in the *COR* models, it is the matrix defined in the first line of (3). Each matrix $\mathbf{A}_t$ obtained by combining (4) and (5) is the Hadamard product of a positive definite (S case) or semidefinite matrix ($R1$ and EO) with strictly positive diagonal entries and a positive definite matrix ($\exp^{\odot}(\phi_A\mathbf{M}_t)$), so that it is a positive definite matrix (see Lemma 3 in Bauwens and Otranto, 2020).

3 Empirical application

High-frequency data for 29 stocks of the Dow Jones Industrial Average (DJIA) index have been used to compute a time-series of daily realized covariance matrices; the 30th stock was dropped since it is not permanently in the index during the sam-

ple period. The data source is the TAQ database. The sample period is 3 January 2001 to 16 April 2018, resulting in 4319 observations.[1]

Seventeen models have been estimated: nine COR and nine *COV* models in the *S*, *R*1 and *EO* parameterizations, each one with $\mathbf{M}_t = \mathbf{0}$ and the two other specifications in (5); the $COV-EO-Pt$ model, after estimation, is equal to the $COV-EO$ model, so we will not consider it longer. For the non-scalar models, we reduce the number of coefficients applying the grouping algorithm described in Bauwens and Otranto (2020). Details about estimates and grouping are available on request. Here we comment only the evaluation of models in terms of loss functions based on the comparison of fitted and realized covariance matrices. The losses adopted are the Quasi-Likelihood loss (QLIK), the Frobenius loss function (FL) and Global Minimum Variance Portfolio (GMVP) (see Bauwens and Otranto, 2020). Figure 1 shows the values of the three losses for all models, calculated for the full covariance matrix, only for the variances, only for the correlations. The points circled in green identify the models belonging to the MCS with non significant differences of the losses at the 5% nominal size of the tests. Some comments:

1. QLIK loss indicates that the Exponential Hadamard extension provides, in general, significant improvements in the models.
2. COR models shows a better performance than COV models, but this depends essentially on the better fitting of the variance part.
3. Applying the MCS only on the correlation part, COV models perform better in terms of GMVP (which is a financial loss) and sometimes also in terms of MSE.

4 Concluding remarks

We have proposed a new class of models for the realized covariance matrix, with the characteristics of a great flexibility with respect to previous models, providing the possibility, with a few set of parameters, to estimate different dynamics for each element of the realized covariance matrix. This is obtained thanks to the particular parameterization of the coefficients, based on the Hadamard exponential function, which possesses the nice property to guarantee the positive definiteness of the matrix of parameters. The models proposed belong to the family of CAW models for realized covariances and show a better performance with respect to the classical CAW models. We have performed also some out–of–sample analysis (not reported here) with similar results. We have proposed several parameterizations to estimate the new models, but they are not exhaustive; each positive definite matrix $\mathbf{M}$ in (4) is a potential solution to parameterize the model. One of the main advantages of this model is the possibility to consider different dynamics for each element of the co-

[1] Thanks to Professors Lyudmila Grigoryeva (University of Konstanz) and Juan-Pablo Ortega (University of St.Gallen) for providing the datasets for the DJIA companies and to Oleksandra Kukharenko (University of Konstanz) for handling the data and computing the realized covariance matrices from the TAQ data.

variance matrix with a few set of parameters. The approach can be easily extended to the Multivariate GARCH case.

References

1. Bauwens, L., Braione, M., Storti, G. (2016). Forecasting comparison of long term component dynamic models for realized covariance matrices. Annals of Economics and Statistics (123/124), 103-134.

2. Bauwens, L., Otranto, E.: Nonlinearities and regimes in conditional correlations with different dynamics. Journal of Econometrics, forthcoming (2020). https://doi.org/10.1016/j.jeconom.2019.12.014

3. Engle, R.F.: Dynamic conditional correlation—a simple class of multivariate GARCH models. Journal of Business and Economic Statistics (20), 339-350 (2002)

4. Golosnoy, V., Gribisch, B., Liesenfeld, R.: The conditional autoregressive Wishart model for multivariate stock market volatility. Journal of Econometrics (167), 211-223 (2012)

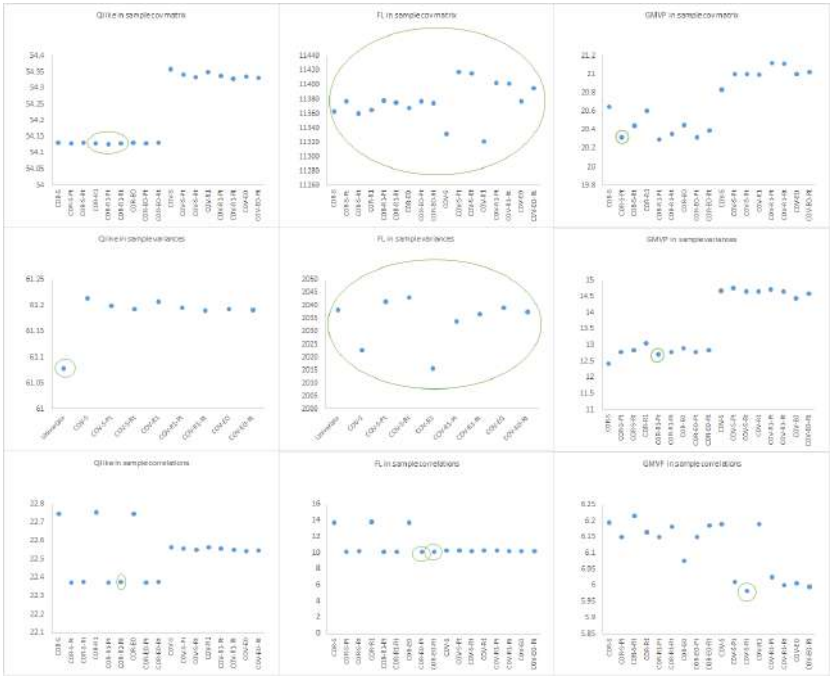


Fig. 1 In–sample evaluation: values of QLIKE, FL and GMVP loss functions and 95% MCS.

Pitman-Yor mixture models for survival data stratification

Modelli mistura basati sul processo di Pitman-Yor per stratificazione di tempi di sopravvivenza

Riccardo Corradin, Luis Enrique Nieto Barajas & Bernardo Nipoti

Abstract Stratification of survival data is a convenient tool to account for heterogeneity across groups. While strata are usually determined through a categorical covariate, the resulting definition of a stratum might not be supported by the data. Here we propose a way of detecting the best possible stratification via a nonparametric mixture model defined by means of a Pitman-Yor process and a logistic kernel. The performance of our proposal is investigated with a simulation study.

Abstract *La stratificazione di dati è uno strumento utile ad introdurre eterogeneità in un modello di analisi della sopravvivenza. Sebbene solitamente gli strati siano definiti tramite una covariata categorica, la stratificazione che ne risulta potrebbe non essere supportata dai dati. Proponiamo qui un modo per definire la miglior stratificazione possibile, tramite un modello mistura nonparametrico definito da un processo di Pitman-Yor e un kernel logistico. Uno studio di simulazione permette di valutare l'efficacia della nostra proposta.*

Key words: Accelerated failure time models, Bayesian nonparametrics, Mixture models, Pitman-Yor process, Stratification, Survival analysis

Riccardo Corradin
Department of Economics, Management and Statistics, University of Milano Bicocca
e-mail: riccardo.corradin@unimib.it

Luis Enrique Nieto Barajas
Departamento de Estadística, ITAM
e-mail: lnieto@itam.mx

Bernardo Nipoti
Department of Economics, Management and Statistics, University of Milano Bicocca
e-mail: bernardo.nipoti@unimib.it

1 Introduction

When modelling the hazard function of time-to-event data, stratification is a convenient tool to account for heterogeneity across groups. While strata are usually determined through a categorical/discrete covariate, it might be possible that the resulting definition of a stratum is not supported by the data. In addition, the best stratification might be the result of a combination of several covariates. Here we propose a way of defining the best possible stratification via a nonparametric mixture model. The nonparametric literature for this type of problems is, to the best of our knowledge, rather scarse with two notable exceptions being [6] and [1].

Modelling properties of nonparametric mixture models have been extensively investigated and their application has proved successful in many fields (see, e.g., [5]). In this work we consider a nonparametric mixture with mixing measure distributed as a Pitman-Yor (PY) process ([8, 9]). Our choice was motivated by the balance the Pitman-Yor process nicely strikes between modelling flexibility and tractability, in both mathematical and computational terms (see [3]).

The rest of the paper is organised as follows. In Section 2 we define the notation and introduce our model. A simulation study is presented in Section 3, while some concluding remarks are made in Section 4.

2 Pitman-Yor mixture for survival data

Let $\boldsymbol{T} = (T_1, \ldots, T_n)$ be a vector of time-to-event observations defined on some space $\mathscr{T} \subseteq \mathbb{R}^+$, and denote by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ the set of corresponding censoring random variables, defined so that $\delta_i = 1$ if the i-th event is observed exactly and $\delta_i = 0$ if it is right-censored. That is, $\delta_i = 1$ indicates that i-th event has occurred at time T_i, while $\delta_i = 0$ means that the i-th event has not occurred before T_i. We further assume that individual information is available and summarised by a d-dimensional set of covariates, denoted by $\boldsymbol{x}_i = (x_1, \ldots, x_d)$ for the i-th individual, and defined on some space $\mathscr{X} \subseteq \mathbb{R}^d$.

We consider an accelerated failure time framework, that is we assume that for each observation the effect of the covariates is multiplicative in the lifetime (see, e.g. [2]). If we denote by h_0 the baseline hazard function, common across different observations, then the accelerated failure time assumption allows us to write the individual hazard, for any $i = 1, \ldots, n$, as

$$h_i(t \mid \boldsymbol{x}_i) = \mathrm{e}^{\boldsymbol{\theta}_i^\intercal \boldsymbol{x}_i} h_0\left(\mathrm{e}^{\boldsymbol{\theta}_i^\intercal \boldsymbol{x}_i} t\right). \tag{1}$$

Moreover, if we consider the positive random variables T_0 and T_i, with $i = 1, \ldots, n$, with distributions respectively characterized by the hazard functions h_0 and h_i, then (1) implies that that $T_i = T_0/\mathrm{e}^{\boldsymbol{\theta}_i^\intercal \boldsymbol{x}_i}$. A logarithmic transformation allows us to write $Y_i = \log T_i = \mu_i + \boldsymbol{\theta}_i^\intercal \boldsymbol{x}_i + \zeta_i Y_0$, where $\mu_i = \mathbb{E}[\log T_i]$ and Y_0 is a random variable with zero-mean and unit variance. In this work we further assume that Y_0 has log-logistic

distribution, which implies that also Y_i is log-logistic, and that T_i is distributed according to a logistic accelerated failure time. Henceforth we denote by $f(y;\mu,\boldsymbol{\theta},\zeta)$ and $S(y;\mu,\boldsymbol{\theta},\zeta)$, respectively, the density and the survival functions of a random variable $Y = \mu + \boldsymbol{\theta}^\intercal \boldsymbol{x} + \zeta Y_0$, with Y_0 being a log-logistic distributed random variable with zero-mean and unit variance.

We consider a nonparametric mixture model with mixing measure distributed as a PY process, characterized by a discount parameter $\sigma \in [0,1)$, a strength parameter $\vartheta > -\sigma$, and a diffuse probability measure P_0 defined on $\Gamma = \mathbb{R}^{d+1} \times \mathbb{R}^+$, space where the individual vectors of latent parameters $\gamma_i = (\mu_i, \boldsymbol{\theta}_i, \zeta_i)$ live. The mixture model can then be expressed as

$$\tilde{f}(y) = \int_\Gamma f(y;\gamma,\boldsymbol{x})^\delta S(y;\gamma,\boldsymbol{x})^{1-\delta} \tilde{p}(\mathrm{d}\gamma) \tag{2}$$

where $\tilde{p} \sim PY(\sigma,\vartheta;P_0)$. Model (2) can be equivalently written in hierarchical form as

$$\begin{aligned} Y_i \mid \gamma_i, \boldsymbol{x}_i, \delta_i &\overset{\text{ind}}{\sim} f(y_i \mid \gamma_i, \boldsymbol{x}_i)^{\delta_i} S(y_i \mid \gamma_i, \boldsymbol{x}_i)^{1-\delta_i}, && i = 1,\dots,n \\ \gamma_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, && i = 1,\dots,n \\ \tilde{p} &\sim PY(\sigma,\vartheta;P_0). \end{aligned} \tag{3}$$

The model specification is completed by defining P_0 as the independent product of a Gaussian distribution for μ, a d-dimensional multivariate Gaussian distribution for $\boldsymbol{\theta}$, and inverse-gamma distribution for ζ.

The almost sure discreteness of the PY process induces ties between the individual latent variables γ_i, with positive probability. We denote by $(\gamma_1^*,\dots,\gamma_k^*)$ the set of $k \leq n$ distinct values appearing in $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_n)$ and by $(n_1,\dots,n_k)$, with $\sum_{j=1}^k n_j = n$, their frequencies. This property can be conveniently exploited to identify homogeneous strata within the data: observations sharing the same latent value γ_j^* can be interpreted as belonging to the same stratum. A simulation scheme for the posterior distribution can be obtained by marginalizing out the distribution of $\tilde{p}$ from model (3), as displayed in the next proposition, which we state after introducing the notation $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$.

Proposition 1. *Let $\boldsymbol{Y} = (Y_1,\dots,Y_n)$ be a set of observable random variables from model* (3)*. Then the joint density of $(\boldsymbol{\gamma},\boldsymbol{Y})$, conditionally on $\boldsymbol{x}_{1:n}$ and $\boldsymbol{\delta}$, is given by*

$$\frac{\prod_{j=1}^{k-1}(\vartheta + j\sigma)}{(\vartheta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \prod_{i \in C_j} \left[f(y_i \mid \gamma_j^*, \boldsymbol{x}_i)\right]^{\delta_i} \left[S(y_i \mid \gamma_j^*, \boldsymbol{x}_i)\right]^{1-\delta_i},$$

where the sets $C_j = \left\{ i \in \{1,\dots,n\} : \gamma_i = \gamma_j^* \right\}$, for $j = 1,\dots,k$.

Proposition 1 represents the building block for devising a Gibbs sampling algorithm analogous to those presented in [4, 7] for non-conjugate mixture models. Inference on the stratification of the survival times can be carried out based on the

posterior sample generated via MCMC, and a point estimate can be obtained by following the variation of information criterion of [10].

3 Simulation study

We carried out an extensive simulation study to evaluate the ability of our model to correctly detect the partition of a simulated dataset, thus providing a useful stratification. To this end we considered three different specifications of model (3), which were used for both generating and analysing the data. Namely:

1. model with all regression coefficients equal to 0 (special case of model (3)),
2. model with common regression coefficients across groups (modification of (3)),
3. model with group-specific regression coefficients, that is model (3).

We considered three different sample sizes, $n \in \{90, 180, 270\}$, and for each n we generated 50 datasets from each one of the three models. For each considered scenario, survival times were generated from a mixture composed by three components, and thus with a known partition of the data. All the generated datasets were analysed by implementing the same three models. Models' performance was then assessed by comparing the true partition with the estimated stratification. Such comparison was made by using the RAND index as a measure of similarity between partitions.

Figure 1 summarises the results we obtained. Each panel refers to one of the nine considered combinations between data generating processes and models implemented to estimate the stratification. In each panel, the sample size is on the x-axis while on the y-axis is the value taken by the RAND index. Some comments are in order. Whenever the data are simulated from model 3 (that is with group-specific regression coefficients), model 1 (null regression coefficients) and model 2 (common regression coefficients) are unable to correctly estimate the partition of the data. On the other hand, model 3 seems to perform well for every considered data generating process. At the same time, it is interesting to observe that when data are simulated from model 1 (or model 2), the increased flexibility of model 3 results in a slightly worse performance if compared to that one of model 1 (or model 2).

4 Discussion

We proposed a Bayesian nonparametric approach for stratifying survival data and ran a simulation study where three different specifications of the proposed model were compared. The results of our study indicate that, at least for the scenarios considered, the more flexible model performs uniformly well. On the contrary, the simpler models, with common regression coefficients, should be considered for applications only in presence of prior evidence supporting this assumption. While we presented a mixture model based on the PY process and a logistic kernel, the same

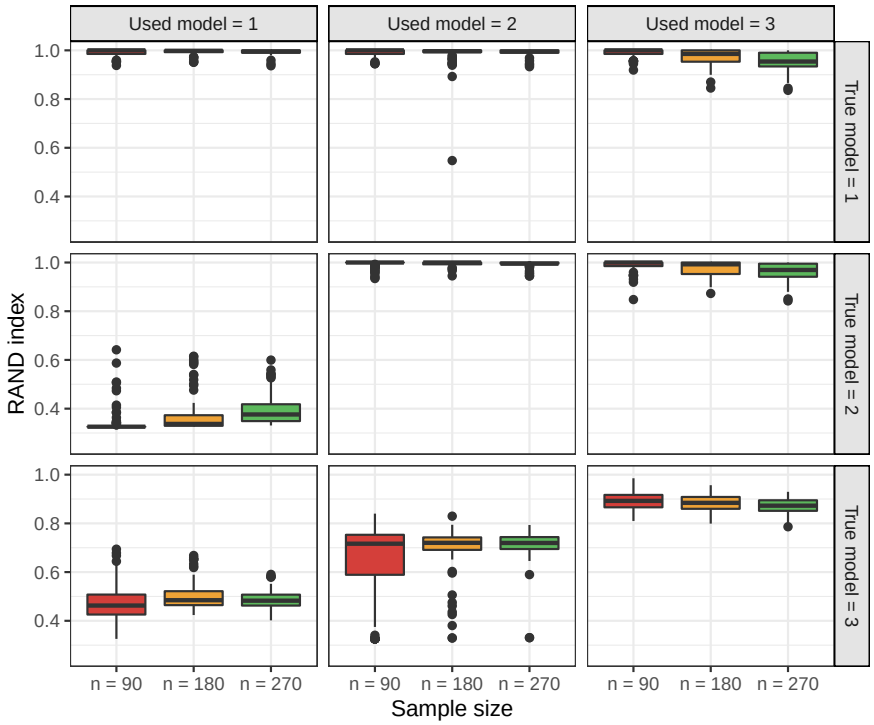


Fig. 1: Boxplots for the values taken by the RAND index, based on 50 replicates per scenario. The study considered three sample sizes $n \in \{30, 60, 90\}$, three data generating processes (one per row) and three fitted models (one per column).

approach can be investigated in greater generality by considering other families of discrete nonparametric priors, such as normalized random measures and Gibbs type priors, and by choosing a different baseline distribution. Moreover, the robustness of the proposed stratification procedure to data censoring is of prominent interest when real-world applications are considered. All these research directions are currently the subject of ongoing work.

References

1. R. Argiento, A. Guglielmi, and A. Pievatolo. Estimation, prediction and interpretation of NGG random effects models: an application to Kevlar fibre failure times. *Statistical Papers*, 55(3):805–826, August 2014.
2. D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.
3. P. De Blasi, S. Favaro, A. Lijoi, R. Mena, I. Prnster, and M. Ruggiero. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37, 10 2015.

4. S. Favaro and Y. W. Teh. Mcmc for normalized random measure mixture models. *Statist. Sci.*, 28(3):335–359, 08 2013.
5. S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert. *Handbook of mixture analysis*. Chapman and Hall/CRC, 2019.
6. M. Kyung, J. Gill, and G. Casella. Estimation in dirichlet random effects models. *Ann. Statist.*, 38(2):979–1009, 04 2010.
7. R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
8. M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, Mar 1992.
9. J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25(2):855–900, 04 1997.
10. S. Wade and Z. Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.*, 13(2):559–626, 06 2018.

Prediction is not everything, but everything is prediction

La previsione non è tutto, ma tutto è previsione

Leonardo Egidi

Abstract Prediction is an unavoidable task for data scientists, and over the last decades statistics and machine learning became the most popular 'prediction weapons' in many fields. However, prediction should always be associated with a measure of uncertainty, because from it only we can reconstruct and falsify the model/algorithm decisions. Machine learning methods offer many point-predictions, but they rarely yield some measure of uncertainty, whereas statistical models usually do a bad job in communicating predictive results. According to the Popper's falsificationism theory, natural and physical sciences can be falsified on the ground of wrong predictions: though, for social sciences this is not always true. We move then to a weak instrumentalist philosophy: predictive accuracy is not always constitutive of scientific success, especially in social sciences.

Abstract *La previsione è un compito inevitabile per gli scienziati dei dati e nel corso degli ultimi decenni statistica e metodi di apprendimento automatici sono diventate le più popolari 'armi di previsione di massa' in molti ambiti. Tuttavia, la previsione dovrebbe sempre essere associata a una misura di incertezza, poiché solo da essa possiamo ricostruire e falsificare le decisioni dell'algoritmo. I metodi di apprendimento automatico offrono molte previsioni puntuali, ma raramente producono qualche misura di incertezza, mentre i modelli statistici di solito comunicano male i risultati predittivi. Secondo la teoria della falsificazione di Popper, le scienze naturali e fisiche possono essere falsificate sulla base di previsioni errate: tuttavia, per le scienze sociali ciò non è sempre vero. Passiamo quindi a una debole filosofia strumentalista: l'accuratezza predittiva non è sempre costitutiva del successo scientifico, specialmente nelle scienze sociali.*

Key words: Prediction, Popper's falsificationist philosophy, Weak instrumentalism, Predictive accuracy, Machine learning

---

Leonardo Egidi
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: legidi@units.it.

1 Introduction

As motivated by the falsificationism approach (Popper, 1934) and many philosophers of science, prediction has a primary role in the progress of science; however, this is often a controversial argument—see Kuhn (1962) and Lakatos (1976) for some criticisms. Popper argues that theories, in order to be scientific, must be falsifiable on the ground of their predictions: wrong predictions should perhaps push the scientists to reject their theories or to re-formulate them, conversely exact predictions should corroborate a scientific theory. Popper's philosophy is instrumentalist in a strong sense (Hitchcock and Sober, 2004) when applied to physical and natural sciences: predictive accuracy is constitutive of scientific success, not only symptomatic of it, and prediction works as a confirmation theory tool for science.

Since the 1940s, with the growing availability of fast computers and the use of simulation routines, science expanded its boundaries and extended the existing frameworks in new directions; think, for instance, at the Manhattan project in Los Alamos, when the problem of neutron diffusion in fissionable material allowed Stanislaw Ulam and Nicholas Metropolis to invent and develop Markov Chain Monte Carlo Methods through the ENIAC computer. Over the last decades, statistics and machine learning became the most popular 'prediction weapons' for both social and natural sciences, including frameworks such as weather's forecasting, presidential elections, planets' motions, global warming, Gross Domestic Product, etc. However, there is often a clear separation between these two fields: statistics is usually seen as a discipline which extracts information from the current data, whereas machine learning is usually designed to predict new events. Though, many times the right weapons are embraced by the wrong people. The predictive power in statistics is an elegant small gun, with good properties and small bullets, whereas in machine learning is a bazooka, with devastating effectiveness and big bullets.

The main novelty of this paper is the introduction of the *weak instrumentalist* position for prediction, under which predictive accuracy is constitutive of scientific success only when the underlying statistical methods are falsifiable and transparently designed to predict out-of-sample events. In other way said, there are many contexts, especially in social sciences, where falsification through the prediction's fallacy should be replaced by a more consistent idea of falsification. We believe this position may be beneficial for the so-called 'hard sciences' as well.

In Section 2 we state the steps required to formulate a scientific theory and review the role of prediction for science. In Section 3 we focus on prediction for statistics and machine learning, and we introduce the weak instrumentalist philosophy.

2 It is prediction part of the scientific design? Prediction as a confirmation theory approach

The main stages required to formulate a scientific law are summarized by Russell (1931) as follows: (1) observation of some relevant facts; (2) formulation of a hypothesis underlying and explaining the facts above; (3) deduction of some consequences from this hypothesis. In his opinion, the modern scientific method is born with Galileo Galilei, father of the law of falling bodies, and with Johannes Kepler, who discovered the three laws of planetary motion. Then, the law of universal gravitation of Isaac Newton embodied the two previous theories, whereas the theory of the general relativity of Albert Einstein generalized the Newton's theory. Thus, in the last 500 years, physics—and, more generally, science—advanced by falsification and generalization of the previous theories, by providing new and more exciting theories to predict new natural facts and highlighting the confirmation nature of prediction. In general, as Hitchcock and Sober (2004) argue, mathematical descriptions of the invariant behaviour of a physical phenomenon are essentially predictive: further experiments and observations can validate these theories.

However, the link of prediction with the scientific laws is in our opinion more ambiguous than what people are usually inclined to think. The following questions arise: is prediction a central step in science? Is prediction a relevant aim of science? A negative answer to the first question could be seen in disagreement with some *instrumentalist* scientists, who would claim that, from an instrumental perspective, predictive success is not merely *symptomatic* of scientific success, but it is also *constitutive* of scientific success (Hitchcock and Sober, 2004). A more sophisticated answer could be: prediction is not explicitly part of the formulation of a scientific hypothesis (1)–(3) *at the time the law is posed*, but it becomes relevant and relevant as science advances.

Steps (1)–(3) are widely used by social scientists and statisticians to build consistent theories about human and social behaviours: recently, it emerged clearly the need to build quantitative population's laws with the aim to mimic the physical nature's laws. However, the role played by prediction in social sciences is more obscure (Popper, 1944) and much controversial than for natural sciences, though data scientists are every day more and more asked to build 'weapons of mass prediction' in many social contexts. Perhaps, the actual outcome may be far away from the predictions: Trump's win in the 2016 US Presidential Elections, Brexit, and Leicester's Premier League's win were very low-probability events, but they all occurred during 2016. Can all of these rare events falsify the finest algorithms and models designed to not predict their occurrence? Our naive and tentative answer is no, they can't.

3 The role of prediction in statistical learning

3.1 From the observed to the observable

Statistics has always been thought as the *science of inference*, or *science of estimates*, and inference is always seen as a separate task from prediction. As statisti-

cians, we are often faced with a double task: first, creating a sound mathematical model to accomodate the data and retrieve useful inferences for our parameters—at the time being, we make no distinction here between classical and Bayesian statistics; second, using this model to make predictions, but this is rarely accounted by the statisticians in a transparent way.

Prediction moves from the observed to the unobserved, being the action designed to forecast future events without requiring a full understanding of the data-generation process. Each person is more or less confident with the weather's predictions or with presidential election predictions, but rarely that person is aware of the underlying statistical model required to produce that forecast, unless he is a statistician/data scientist. In such a view, inference seems hard and obscure, and prediction easy and close to the people. This is often a paradoxical argument, since the inference is often associated with the *explanation* of the problem, and should be relevant and available to the majority of the population. However, parameters do not exist in the real life, they are some fictitious and technical devices to explain and approximate the complexity. Rather, only prediction links the observed with the observable and is accessible to the people: it is not a matter of parameters' interpretation, but a check of the discrepancy between observed and future events. This is the reason why we provocatively claim that *prediction is not everything, but (almost) everything, in the real life, could be the object of a predictive action.*

In our practice, prediction should not be assimilated to 'take a rabbit out of a hat', but looking at its inherent uncertainty. If we are framed in a Bayesian context, we intend the unobserved and future values to come from a posterior predictive distribution, which incorporates the intrinsic uncertainty propagating from the parameters—summarized by the posterior distribution—to the observable future values. In the same spirit, when comparing models we support predictive information criteria which provide a measure of predictive accuracy based on the posterior predictive accuracy, such as the Watanabe Information Criterion (WAIC) and Leave-one-out Information Criterion (LOOIC).

3.2 The two cultures

As brilliantly argued by Breiman et al (2001), there are two cultures in the use of statistical modeling to reach conclusions from data: a stochastic data model consisting of predictors, parameters and random noise to explain the response variable is adopted by the data modeling culture; whereas a function of the predictors to predict the response variable is assumed by the algorithmic modeling culture, also named machine learning (ML) culture. The two approaches strongly differ in their validation: goodness-of-fit tests vs. predictive accuracy on out-of-sample data. It is evident that the data modeling culture—linear regression, generalized linear models, Cox model, etc.—is aimed at extracting some information about how nature is associating the response variable to the dependent variable, whereas the algorithmic

culture—decision and classification trees, neural nets—is more oriented to predict future values of the response variable given the values of the predictors.

Data scientists are used to train their procedures on the *training set*, which is chosen at the beginning in many possible ways. However, a small change in the dataset can cause a large change in the final predictions, and some adjustments are often required to increase the algorithm's robustness. To alleviate this lack of robustness, in the mid-1990s some data scientists argued that by aggregating many algorithms and perturbing the training set, using bagging (Breiman, 1996), boosting (Freund et al, 1996) or random forests (Ho, 1995), dramatically increased the predictive accuracy of the trees, by decreasing the variance.

3.3 ML scientists are strong instrumentalist, statisticians are weak instrumentalist

As emerges from this quick overview, the only rationale to evaluate the goodness of an algorithmic modeling procedure is to look at its predictive accuracy on out-of-sample/future data. 'Shaking the training set' became popular to ensure lower variance and higher accuracy, with the data scientist apparently ready to do 'whatever it takes' to improve over the previous methods. From a philosophical and scientific point of view, algorithmic modelers are *strong instrumentalist*, since for them the predictive accuracy carried out by their algorithms is constitutive—and not only symptomatic—of the broader scientific success.

Evaluating a model/algorithm in light of its ability to predict future data is not shameful at all; conversely, it is beneficial in many areas where a parametric stochastic model failed to be really generative and useful. However, even if predictions of future data are good tools to falsify a posed theory, many times ML techniques lack of a general and valid theoretical framework. For instance, the number of predictors at each split of a random forest is a tuning parameters, but in practice the best values for these parameters will depend on the problem. Predictions should corroborate or reject an underlying theory, but if the method (the theory) is tuned and selected on the ground of its predictive accuracy, the theory to be falsified is bogus, and not posed in a transparent way.

As statisticians and (data) scientists, our efforts should be addressed to produce good, transparent and well posed algorithms/models, and make them falsifiable upon a strong check (Gelman and Shalizi, 2013). Our skepticism regards the role of prediction in falsifying our models, for such a reason we would claim to be *weak instrumentalists*: predictions and predictive accuracy are central tasks of science, but only sometimes they are constitutive of scientific success.

To summarize the above discussion and other arguments not directly treated in this paper, we collect in Table 1 the main points which follow from the weak instrumentalist philosophy.

Table 1: Weak instrumentalism summary

General science
1. Predictive accuracy is not always constitutive of scientific success
2. Scientific falsification on the ground of wrong predictions is sometimes misleading, especially in social sciences (Trump's election, Leicester win, Brexit)
3. Supposedly valid scientific theories should exist before the future data have been revealed
4. Prediction is not explicitly part of the formulation of a scientific hypothesis at the time the law is posed, but it becomes relevant and relevant as science advances
Statistics
5. Take care of variability in the statistical predictions
6. If necessary, go beyond the distinction between inference and prediction, and consider a joint model for data, parameters and future data (falsificationist Bayes)
7. Rather than reasoning in terms of variance and bias, reason more in terms of predictive information criteria and posterior predictive distribution
Machine Learning
8. 'Shaking the training set' to improve predictive accuracy is an obscure step
9. Avoid to tune the algorithm with the only task to improve predictive accuracy
10. To be falsifiable, ML techniques need to be transparently posed

Acknowledgements I want to deeply thank Jonah Gabry (Columbia University) for his rich and precious comments about this topic. He will co-author the next extended version of this paper.

References

Breiman L (1996) Bagging predictors. Machine learning 24(2):123–140

Breiman L, et al (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16(3):199–231

Freund Y, Schapire RE, et al (1996) Experiments with a new boosting algorithm. In: icml, Citeseer, vol 96, pp 148–156

Gelman A, Shalizi CR (2013) Philosophy and the practice of bayesian statistics. British Journal of Mathematical and Statistical Psychology 66(1):8–38

Hitchcock C, Sober E (2004) Prediction versus accommodation and the risk of overfitting. The British journal for the philosophy of science 55(1):1–34

Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, IEEE, vol 1, pp 278–282

Kuhn TS (1962) The structure of scientific revolutions. Chicago and London

Lakatos I (1976) Falsification and the methodology of scientific research programmes. In: Can theories be refuted?, Springer, pp 205–259

Popper K (1934) The logic of scientific discovery. Routledge

Popper K (1944) The poverty of historicism, ii. a criticism of historicist methods. Economica 11(43):119–137

Russell B (1931) The scientific outlook. Routledge

The Generalized Dynamic Mixtures of Factor Analyzers for clustering multivariate longitudinal data

Misture dinamiche di modelli fattoriali generalizzate per la classificazione di dati longitudinali

Francesca Martella, Antonello Maruotti, Francesco Tursini

Abstract This work introduces the Generalized Dynamic Mixtures of Factor Analyzers (GDMFA) approach for clustering high-dimensional longitudinal data. The proposed model can be seen as an extension of the Gaussian mixture model where individuals are allowed to move between components over time and, within each component, local dimensional reduction is performed. Temporal dependence is modelled through a first-order finite-state Markov chain. The model parameters have been estimated through an Alternating Expected Conditional Maximization (AECM) algorithm and the performance of the GDMFA model is discussed on the equitable and sustainable well-being (BES) of Italian territories data set. The results are encouraging and would deserve further discussion.

Abstract *In questo lavoro vengono introdotte le Misture Dinamiche di Modelli Fattoriali Generalizzate (GDMFA) per la classificazione di dati longitudinali. Il modello proposto puó essere visto come un'estensione del modello mistura Gaussiano dove simultaneamente gli individui possono muoversi tra le componenti durante il periodo di osservazione, e all'interno di ogni componente, viene effettuata una riduzione dimensionale locale. Inoltre, per catturare la dipendenza temporale, le componenti sono associate attraverso un processo di Markov. I parametri del modello vengono stimati attraverso un algoritmo AECM mentre l'utilitá e la validitá del modello GDMFA é discussa su dati relativi il benessere equo e sostenibile dei territori Italiani (BES). I risultati sono incoraggianti e meritano ulteriori sviluppi.*

Key words: Factor Analyzers, Dimensionality reduction, Hidden Markov Models

---

Francesca Martella, Francesco Tursini
Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, P.le Aldo Moro, 5 - 00185 Roma
e-mail: francesca.martella@uniroma1.it
e-mail: f.tursini@gmail.com

Antonello Maruotti
Dipartimento Giurisprudenza, Economia, Politica e Lingue moderne, Lumsa Universitá, Via pompeo Magno, 22 - 00192 Roma
e-mail: a.maruotti@lumsa.it

1 Dynamic Mixtures of Factor Analyzers

The key idea of the Dynamic Mixtures of Factor Analyzers (DMFA) introduced by [4] consists of assuming that the joint distribution of a multivariate time series can be approximated by a dynamic mixture of lower-dimensional Gaussian densities in the factor space, and that, the mixture parameters evolve according to a latent homogeneous Markov chain. In details, let $\{\mathbf{Y}_t, t = 1,\ldots,T\}$ be a sequence of multivariate continuous observations, where $\mathbf{Y}_t = \{Y_{t1},\ldots,Y_{tP}\}$ and Y_{tp} represents the p-th response variable at time t ($p = 1,\ldots,P; t = 1,\ldots,T$). Moreover, let $\{S_t, t = 1,\ldots,T\}$ be a unobservable (hidden) process defined on the state space $\{1,\ldots,K\}$ following a first-order Markov chain, that is:

$$\Pr(S_t \mid S_1,\ldots,S_{t-1}) = \Pr(S_t \mid S_{t-1}) \tag{1}$$

characterized by

- the time-homogeneous $(K \times K)$-dimensional transition probability matrices Π containing the time-varying transition probabilities $\{\pi_{k|j}\}$ defined as

$$\pi_{k|j} = \Pr(S_t = k \mid S_{t-1} = j) \qquad t = 2,...,T,; j,k = 1,...,K \tag{2}$$

  where k and j refers to the current and previously visited state, respectively;
- the initial probabilities $\pi = \{\pi_k\}$ defined as:

$$\pi_k = Pr(S_1 = k) \qquad k = 1,...,K. \tag{3}$$

The DMFA model assumes that $\{\mathbf{Y}_t, t = 1,\ldots,T\}$ is a state-dependent process for which the conditional independence property holds, that is:

$$f(\mathbf{Y}_t|\mathbf{Y}_1,\ldots,\mathbf{Y}_T,S_1,\ldots,S_T) = f(\mathbf{Y}_t|S_t) \tag{4}$$

and where

$$f(\mathbf{Y}_t \mid S_t) = \mathrm{N}_P(\mu_k, \Lambda_k\Lambda_k' + \Psi_k), \tag{5}$$

i.e. the conditional distribution of $\mathbf{Y}_t \mid S_t$ is a multivariate Gaussian distribution with state-dependent mean vector μ_k and covariance matrix $\Lambda_k\Lambda_k' + \Psi_k$, where Λ_k is a $P \times q$ nonsparse real matrix of state-specific factor loadings, and $\Psi_k = \mathrm{diag}(\psi_{k1},\ldots,\psi_{kP})$ represents a positive definite matrix containing the error variances ($k = 1,\ldots,K$). The above specification implies that conditional on the state k, $\mathbf{Y}_t$ is specified by a factor analysis model as follows:

$$\mathbf{Y}_t = \mu_k + \Lambda_k\mathbf{f}_{tk} + \mathbf{e}_{tk} \tag{6}$$

where $\mathbf{f}_{tk}$ is a q-dimensional vector of state-specific latent factors drawn from $\mathrm{N}_P(\mathbf{0}, \mathbf{I}_q)$, and $\mathbf{e}_{tk}$ is a p-dimensional vector of state-specific error terms drawn from $\mathrm{N}_P(\mathbf{0}, \Psi_k)$, that is assumed to be independent of $\mathbf{f}_{tk}$. The model parameters are esti-

mated through an ML approach using a three-step Alternating Expectation Conditional Maximization (AECM) algorithm [5].

1.1 Extension to multivariate longitudinal data

To deal with multivariate longitudinal sequences, the DMFA approach introduced in the previous section needs to be extended to accommodate the peculiarities of such kind of data. These come in the form of three-way data: the first dimension identifies individuals, the second dimension identifies variables and the third one identifies time occasions. In this respect, we propose the Generalized Dynamic Mixtures of Factor Analyzers (GDMFA), where the parsimonious HMM is now defined by:

- an observed process $\{\mathbf{Y}_{it}, i = 1,\ldots,n; t = 1,\ldots,T\}$ where $\mathbf{Y}_{it} = \{Y_{it1},\ldots,Y_{itP}\}$ and Y_{itp} represents the p-th response variable given by the i-th individual at time t $(i = 1,\ldots,n; p = 1,\ldots,P; t = 1,\ldots,T)$ such that the following property holds:

$$f(\mathbf{Y}_{it}|\mathbf{Y}_{i1},\ldots,\mathbf{Y}_{iT},S_{i1},\ldots,S_{iT}) = f(\mathbf{Y}_{it}|S_{it}) \tag{7}$$

- a hidden state-dependent process $\{S_{it}, i = 1,\ldots,n; t = 1,\ldots,T\}$ defined on the state space $\{1,\ldots,K\}$ such that the following property holds:

$$\Pr(S_{it} \mid S_{i1},\ldots,S_{it-1}) = \Pr(S_{it} \mid S_{it-1}). \tag{8}$$

Moreover, as before, we defined the initial probabilities $\pi_k = Pr(S_{i1} = k)$ $(i = 1,...,n; k = 1,...,K)$ and the transition probability matrix $\Pi = \{\pi_{k|j}\}$, where $\pi_{k|j} = Pr(S_{it} = k \mid S_{it-1} = j)$ $(i = 1,...,n; t = 1,...,T,; j,k = 1,...,K)$. In line with the DMFA, we assume that conditionally to the k-th state, the random vector $\mathbf{Y}_{it}$ is modelled as

$$\mathbf{Y}_{it} = \mu_k + \Lambda_k \mathbf{f}_{itk} + \mathbf{e}_{itk} \tag{9}$$

where $\mathbf{f}_{itk}$ is a q-dimensional vector of state-specific latent factors drawn from $\mathrm{N}_P(\mathbf{0}, \mathbf{I}_q)$, and $\mathbf{e}_{itk}$ is a p-dimensional vector of state-specific error terms drawn from $\mathrm{N}_P(\mathbf{0}, \Psi_k)$, where $\Psi_k = \mathrm{diag}(\psi_{k1},\ldots,\psi_{kP})$, which is assumed to be independent of $\mathbf{f}_{itk}$. In other words, an observation i in state k follows a multivariate Gaussian density with state-dependent mean vector μ_k and covariance matrix $\Lambda_k\Lambda_k' + \Psi_k$. Note that, following the idea of [1], constraints across clusters on the Λ_k and Ψ_k matrices provide a class of 8 different models which will be presented.

In order to perform maximum likelihood estimation of the model parameters on the basis of the response vector $\mathbf{Y}_{it} = \{Y_{it1},\ldots,Y_{itP}\}$, we propose to use the same AECM algorithm as for DMFA. Specifically, this algorithm is a variant of the classical Expectation-Maximization (EM) algorithm [3] using different definitions of missing data at different stages. It is based on the complete-data log-likelihood, i.e., the log-likelihood of the observations (the incomplete data) plus the states (the missing data). Thus, we introduce two indicator variables $\mathbf{z}_{it} = (z_{it1},\ldots,\ldots,z_{itK})$ e $\mathbf{w}_{it} = (w_{it11},\ldots,\ldots,w_{itKK})$ such that $z_{itk} = 1$ if unit i is in state k at time t

and 0 otherwise, and $w_{itjk} = 1$ if individual i is in to state j at time $t-1$ and in state k at time t, 0 otherwise. Moreover, we partition the set of unknown parameters $\Theta = \{\mu_k, \pi_k, \pi_{k|j}, \Lambda_k, \Psi_k, j,k = 1,\ldots,K\}$ into two disjoint subsets $\Theta_1 = \{\mu_k, \pi_k, \pi_{k|j}, j,k = 1,\ldots,K\}$ and $\Theta_2 = \{\Lambda_k, \Psi_k, k = 1,\ldots,K\}$. One iteration of the AECM algorithm consists of two cycles; one E-step and one CM- step (one for each sub-vector of Θ) for each cycle. At the first cycle, we define the state labels as missing data, and the complete data log-likelihood function has the following form:

$$l_{c_1}(\mathbf{y},\mathbf{z},\mathbf{w}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i1k}\log(\pi_k) + \sum_{i=1}^{n}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=2}^{T} w_{itjk}\log(\pi_{k|j}) + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T} z_{itk}\log(f(\mathbf{y}_{it} \mid S_{it} = k)). \tag{10}$$

where

$$f(\mathbf{y}_{it} \mid S_{it} = k) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_{it}-\mu_k)'(\Lambda_k\Lambda_k' + \Psi_k)^{-1}(\mathbf{y}_{it}-\mu_k)\right]}{(2\pi)^{P/2} \mid \Lambda_k\Lambda_k' + \Psi_k \mid^{1/2}}. \tag{11}$$

The expected value of the complete-data log-likelihood is therefore

$$\begin{aligned} H_1(\Theta_1, \Theta_1^{(m)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{i1k}\log(\pi_k) + \sum_{i=1}^{n}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=2}^{T} \hat{w}_{itjk}\log(\pi_{k|j}) \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T} \hat{z}_{itk}\log(|\Lambda_k\Lambda_k' + \Psi_k|) \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T} \hat{z}_{itk}\mathrm{tr}\left\{(\tilde{\Sigma}_k(\Lambda_k\Lambda_k' + \Psi_k)^{-1}\right\}, \end{aligned} \tag{12}$$

where

$$\tilde{\Sigma}_k = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} \hat{z}_{itk}(\mathbf{y}_{it}-\mu_k)(\mathbf{y}_{it}-\mu_k)'}{\sum_{i=1}^{n}\sum_{t=1}^{T} \hat{z}_{itk}}. \tag{13}$$

is the state-specific sample covariance matrix for all $k = 1,\ldots,K$. Thus, the E-step of the first cycle consists of calculating the conditional expectations $\hat{z}_{itk} = \mathrm{E}(z_{itk} \mid \mathbf{y}_{i1},\ldots,\mathbf{y}_{iT})$ and $\hat{w}_{itjk} = \mathrm{E}(w_{itjk} \mid \mathbf{y}_{i1},\ldots,\mathbf{y}_{iT})$ given the current values of the parameters and the observed data. On the other hand, at the first CM-step, maximizing the expected complete-data log-likelihood with respect to μ_k, π_k and $\pi_{k|j}$, we obtain, respectively:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} \hat{z}_{itk}\mathrm{y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T} \hat{z}_{itk}}, \tag{14}$$

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \hat{z}_{i1k}}{n}, \tag{15}$$

and

$$\hat{\pi}_{k|j} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\hat{w}_{itjk}}{\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=2}^{T}\hat{w}_{itjk}}. \tag{16}$$

At the second cycle of the algorithm, we consider the state labels and the factors as missing data, therefore the complete-data log-likelihood is now as follows:

$$\begin{aligned} l_{c_2}(\mathbf{y},\mathbf{z},\mathbf{f}) = C + \sum_{i=1}^{n}\sum_{k=1}^{K}\Bigg[& -\frac{\sum_{t=1}^{T} z_{itk}}{2}\log(|\Psi_k^{-1}|) - \frac{\sum_{t=1}^{T} z_{itk}}{2}\mathrm{tr}(\Psi_k^{-1}\hat{\Sigma}_k) \\ & + \sum_{t=1}^{T} z_{itk}(\mathbf{y}_{it}-\mu_k)'\Psi_k^{-1}\Lambda_k\mathbf{f}_{itk} - \frac{1}{2}\mathrm{tr}(\Lambda_k'\Psi_k^{-1}\Lambda_k\sum_{t=1}^{T} z_{itk}\mathbf{f}_{itk}\mathbf{f}_{itk}')\Bigg] \end{aligned} \tag{17}$$

where C is a constant. The expected complete-data log-likelihood is then given by

$$\begin{aligned} H_1(\Theta_2,\Theta_2^{(m)}) = C & + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\hat{z}_{itk}}{2}\log(|\Psi_k^{-1}|) \\ & - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\hat{z}_{itk}}{2}\mathrm{tr}(\Psi_k^{-1}\tilde{\Sigma}_k) \\ & + \sum_{i=1}^{n}\sum_{t=1}^{T}\hat{z}_{itk}(\mathbf{y}_{it}-\mu_k)'\Psi_k^{-1}\Lambda_k\mathrm{E}(\mathbf{u}_{itk}\mid\mathbf{y}_{it},\mu_k,\Lambda_k,\Psi_k) \\ & - \frac{1}{2}\mathrm{tr}\left(\Lambda_k'\Psi_k\Lambda_k\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{z}_{itk}\mathrm{E}(\mathbf{u}_{itk}\mathbf{u}_{itk}'\mid\mathbf{y}_{it},\mu_k,\Lambda_k,\Psi_k)\right). \end{aligned} \tag{18}$$

The estimates of Λ_k and Ψ_k matrices can easily be derived from (18). In the most general case, i.e. without any constraints on Λ_k and Ψ_k, we have

$$\hat{\Lambda}_k = \tilde{\Sigma}_k\gamma_k'\Delta_k^{-1}, \qquad \hat{\Psi}_k = \mathrm{diag}(\tilde{\Sigma}_k - \hat{\Lambda}_k\gamma_k\tilde{\Sigma}_k) \tag{19}$$

where

$$\gamma_k = \Lambda_k'(\Lambda_k\Lambda_k'+\Psi_k)^{-1}, \qquad \Delta_k = \mathbf{I}_q - \gamma_k\Lambda_k + \gamma_k\tilde{\Sigma}_k\gamma_k'. \tag{20}$$

In a similar manner, the estimates of Λ_k and Ψ_k can be easily derived under the different imposed constraints (not shown here for sake of brevity). The AECM algorithm iteratively updates the parameters until convergence to maximum likelihood estimates of the parameters. Note that the quantities $\hat{z}_{itk}$ and $\hat{w}_{itjk}$ can be computed recursively ([2]) by defining the forward variable $\alpha_{itk} = f(\mathbf{y}_{i1},\ldots,\mathbf{y}_{it},S_{it}=k)$, which represents the probability of the individual i of seeing the partial sequence ending in state k at time t, and the corresponding backward variable $\beta_{itk} = f(\mathbf{y}_{it+1},\ldots,\mathbf{y}_{iT} \mid S_{it}=k)$. Thus, we can express $\hat{z}_{itk}$ and $\hat{w}_{itjk}$ in terms of forward and backward variables by

$$\hat{z}_{itk} = \frac{\alpha_{itk}\beta_{itk}}{\sum_{k=1}^{K}\alpha_{itk}\beta_{itk}}, \tag{21}$$

and

$$\hat{w}_{it\,jk} = \frac{\pi_{k|j}\alpha_{it-1,k}f(\mathbf{y}_{it} \mid S_{it}=k)\beta_{itk}}{\sum_{j,k=1}^{K}\pi_{k|j}\alpha_{it-1,k}f(\mathbf{y}_{it} \mid S_{it}=k)\beta_{itk}}. \tag{22}$$

As a by-product of the estimation procedure, we classify the i-th individual to the k-th cluster if $\hat{z}_{itk} = \max(\hat{z}_{it1}, \ldots, \hat{z}_{itK})$.

2 Conclusions

A new model-based Gaussian clustering for multivariate longitudinal data is presented. Specifically, the proposed model allows individuals to move between clusters during the period of observation as well as local dimensional reduction within each cluster. In this way, the proposed approach is particularly suitable for high-dimensional data where parsimonious models should be used in order to reduce the general (heteroscedastic) model and, accordingly, the number of parameters to be estimated. We describe an AECM algorithm for estimating model parameters and we discuss the performance of the proposed model on the equitable and sustainable well-being of Italian territories dataset (BES dei territori, ISTAT). It is well known the importance of measuring equitable and sustainable well-being to evaluate the progress of a society not only from an economic, but also from a social and environmental point of view. In this prospective, from 2018 ISTAT provides a system of 55 BES indicators at local level illustrating the 12 domains relevant for the measurement of well-being: Health, Education and training, Work and life balance, Economic well-being, Social relationships, Politics and Institutions, Security, Subjective well-being, Landscape and cultural heritage, Environment, Innovation, research and creativity and Quality of services. Our aim will be to identify how evolve distinctive characteristics of well-being in various territories during a relatively long time period. The results are interesting and like in other studies, not always consistent with the stereotype of the rich North vs poor South.

References

1. Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 3, 803–821 (1993)
2. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurringin the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**, 164–171 (1970)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM-algorithm. J. Royal Stat. Soc. **39**, 1–38 (1977)
4. Maruotti, A., Bulla, J., Lagona, F., Picone, M., Martella, F.: Dynamic Mixture of factor analyzers to characterize multivariat air pollutant exposures. Ann. Appl. Stat. **11**, 3, 1617–1648 (2017)
5. Meng, X.L., van Dyk, D.A.: The EM Algorithm? an Old Folk-song Sung to a Fast New Tune. J. R. Statist. Soc.B **59**, 3, 511–567 (1997)

Trends and long-run relations in cointegrated time series observed with noise

Trends e relazioni di lungo periodo in serie storiche cointegrate osservate con rumore

Angelica Gianfreda, Paolo Maranzano, Lucia Parisio and Matteo Pelagatti

Abstract Classic tests for stationarity of time series and of cointegration may fail when data are observed with dominant noise. We show that results of standard ADF are biased towards more stationarity, while the Johansen's test cointegration test may produce unrealiable results and generate false long-run signals. We show that data filtering improves the performance of standard tests and so it should become a good practice when dealing with very noisy datasets. We prove the effectiveness of different filtering strategies using simulated series.

Abstract *I test classici per stazionarietà e cointegrazione di serie storiche possono portare a conclusioni errate quando i dati sono nascosti da rumore dominante. In questo paper mostriano come i test ADF e Johansen siano distorti in presenza di tali caratteristiche. In particolare il test ADF segnala in modo eccessivo la stazionarietà delle serie, mentre il test di Johansen evidenzia un eccesso di relazioni di lungo periodo rispetto a quelle che realmente costituiscono il processo. Applicare tecniche di filtraggio migliora le performance dei tests e pertanto dovrebbero diventare buona prassi quando si affrontano situazioni simili. I risultati teorici sono validati tramite simulazioni Monte Carlo.*

Key words: Filtering, Cointegration, Stationarity, Averaging, Time series

1 Introduction

Many applied papers study the dynamics of multivariate time series with the aim of finding some long-run relationship governing the behavior of the observed variables. In this stream of research cointegration among time series and mean reversion are very useful and common tools to design

Angelica Gianfreda
University of Modena-Reggio Emilia, Via J. Berengario, 51, Modena, Italy e-mail: angelica.gianfreda@unimore.it

Paolo Maranzano
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, Italy e-mail: p.maranzano@campus.unimib.it

Lucia Parisio
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, Italy e-mail: lucia.parisio@unimib.it

Matteo Pelagatti
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, Italy e-mail: matteo.pelagatti@unimib.it

analysis strategies when dealing with data characterized by significant uncertainty. In this paper we stress that tests for stationarity of time series and of cointegration may fail when data show some unpleasant characteristics like high variability, leptokurtosis and extreme values. In particular, we show that results of standard ADF and Johansen's tests are biased towards more stationarity. In these cases cointegration analysis may produce unrealiable results and generate false lung-run signals. It is therefore extremely important to treat data appropriately when the scope of the analysis is to retrieve long run components governing the behavior of the series. We show that data filtering improves the performance of standard tests and so it should become a good practice when dealing with very noisy datasets. We prove the effectiveness of different filtering strategies using simulated series.

2 Why the ADF and related tests fail when integrated time series are observed with strong noise

Let us consider a simple random walk plus white noise model:

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{WN}(\sigma_\varepsilon^2) \qquad \mu_t = \mu_{t-1} + \eta_t, \quad \eta_t \sim \mathsf{WN}(\sigma_\eta^2), \tag{1}$$

where the notation $\varepsilon_t \sim \mathsf{WN}(\sigma_\varepsilon^2)$ is to be read as "ε_t is a white noise sequence with variance σ_ε^2". Moreover, let $\lambda = \sigma_\eta^2/\sigma_\varepsilon^2 \geq 0$ be the *signal-to-noise ratio*. It is straightforward to prove that the process y_t has the reduced ARIMA$(0,1,1)$ form

$$\Delta y_t = \eta_t + \varepsilon_t - \varepsilon_{t-1} = \zeta_t - \theta\zeta_{t-1}, \qquad \zeta_t \sim \mathsf{WN}(\sigma^2) \tag{2}$$

with $\theta = 1 + \frac{\lambda - \sqrt{\lambda^2+4\lambda}}{2}$ and $\sigma^2 = \frac{\sigma_\varepsilon^2}{\theta}$.

When the signal-to-noise ratio is zero the MA coefficients is $\theta = 1$, the unit root operator cancels out with the MA operator and y_t turns out to be just a white noise sequence.

When λ is close to zero, the MA coefficient approaches 1 from below. In this case, the exact cancellation does not take place, however in small samples the process y_t is almost indistinguishable form a white noise. Such a MA process still has the purely AR representation $\Delta y_t = \zeta_t + \theta\Delta y_{t-1} + \theta^2\Delta y_{t-2} + \theta^3\Delta y_{t-3} + \ldots$, however, this representation cannot be well approximated by an AR(p) process with small p because θ^j approaches zero very slowly. Now, most unit root tests deriving from the Dickey-Fuller test such as ADF [6], ADF-GLS [2], Johansen [4] are based on autoregressive approximations and, if the y_t is generated as above with λ close to zero, then they are severely over-sized [3].

3 Filtering

We propose two types of time series filters that can improve the performance of unit root and cointegration tests:

1. reducing the frequency of the time series by taking averages, for instance by working on weakly means of daily or hourly electricity prices;

2. extracting the level component using the smoother in an unobserved component model (UCM) containing trend, noise and, possibly, seasonal components.

3.1 Frequency reduction by averaging

For $t = 1, 2, \ldots$, let y_t be defined as in (1) and let

$$\bar{y}_t = \frac{1}{m}\sum_{i=0}^{m-1} y_{t-i}, \qquad \bar{\eta}_t = \frac{1}{m}\sum_{i=0}^{m-1} \eta_{t-i}, \qquad \bar{\varepsilon}_t = \frac{1}{m}\sum_{i=0}^{m-1} \varepsilon_{t-i}, \tag{3}$$

where $\bar{y}_t$ is sampled over the set of time points $t \in \mathcal{T} := \{m, 2m, 3m, \ldots\}$. Then, over the time+ set $\mathcal{T}$

$$\bar{y}_t - \bar{y}_{t-m} = \bar{\eta}_t + \bar{\varepsilon}_t - \bar{\varepsilon}_{t-m},$$

with $\bar{\varepsilon}_t$ white noise sequence with variance σ_ε^2/m and $\bar{\eta}_t$ MA(1) process with $\mathsf{VAR}(\bar{\eta}_t) = \sigma_\eta^2\left[\frac{(m-1)(2m-1)}{3m} + 1\right]$, $\mathsf{COV}(\bar{\eta}_t, \bar{\eta}_{t-m}) = \sigma_\eta^2 \frac{(m-1)(m+1)}{6m}$ and $\mathsf{COR}(\bar{\eta}_t, \bar{\eta}_{t-m}) = \frac{(m-1)(m+1)}{2(m-1)(2m-1)+3m}$.

Moreover, the process $\bar{y}_t$ is ARIMA(0, 1, 1) for $t \in \mathcal{T}$ with moving average coefficient given by $\frac{-1+\sqrt{1-4\rho^2}}{2\rho}$ where the first-order autocorrelation is given by $\rho = \frac{\lambda\frac{(m-1)(m+1)}{6m} - \frac{1}{m}}{\lambda\left(\frac{(m-1)(2m-1)}{3m}+1\right)+\frac{2}{m}}$

3.2 Signal extraction in a unobserved component model

An alternative way to reduce the noise in the process y_t defined in Equation 1 is by estimating the random walk component μ_t by projecting it on the linear span of $y_0, y_1, \ldots, y_s$ where s is either equal to t or to $n-1$. This operation can be easily carried out by stating the model in state-space form and running the Kalman filter (for $s = t$) and smoother (for $s = n-1$) on the level component μ_t [1, 5]. The Kalman filter and smoother for μ_t are a linear filters whose weights are different for every t: the former is only backward looking,

$$\mu_{t|t} = \sum_{i=0}^{t} w_{ti} y_{t-i}, \qquad t = 0, 1, \ldots, n-1, \tag{4}$$

while the latter is two-sided,

$$\mu_{t|n-1} = \sum_{i=t}^{t-n+1} \tilde{w}_{ti} y_{t-i}, \qquad t = 0, 1, \ldots, n-1. \tag{5}$$

Nonetheless, when t is not too close to 0, the Kalman filter for the RW plus noise model can be well approximated by its steady state version, say $\tilde{\mu}_t$, given by the recursive filter

$$\tilde{\mu}_t = \gamma y_t + (1-\gamma)\tilde{\mu}_{t-1}, \qquad t = 0, 1, \ldots, n-1, \tag{6}$$

where $\gamma = \frac{\sqrt{\lambda^2+4\lambda}-\lambda}{2}$.

Similarly, when t is not too close to 0 or $n-1$, the smoother can be well approximated by its steady-state version, say $\hat{\mu}_t$, which is given by the backward recursion on $\tilde{\mu}_t$

$$\hat{\mu}_t = \gamma\tilde{\mu}_t + (1-\gamma)\hat{\mu}_{t+1}, \qquad t = n-1, n-2, \ldots, 0. \tag{7}$$

Assume that the variances σ^2_ε and σ^2_η of the process defined in equation (1) are known and $\lambda = \sigma^2_\eta/\sigma^2_\varepsilon$ is the signal-to-noise ratio. Then, $\tilde{\mu}_t$ is a random walk and $\hat{\mu}_t$ is an ARIMA$(1,1,0)$ process with autoregressive coefficient $\phi = 1-\gamma$.

4 Simulation experiments

To verify the empirical effects of the frequency-reduction and UCM filters on the ADF and Johansen tests, we performed a set of Monte Carlo simulations. Each experiment evaluates the performances of ADF and Johansen tests under the null hypothesis, in order to evaluate the size of the tests. The first set of experiments aims at assessing the performance of the ADF test under the random walk plus (leptokurtic) noise model. We simulate time series from a RW buried in leptokurtic noise and each simulation experiment is characterized by a different combination of noise-to-signal ratio and kurtosis value. The data generating process (DGP) for the observation y_t is

$$y_t = \mu_t + \sqrt{c}\,\varepsilon_t, \qquad \varepsilon_t \sim i.i.d.(0,1) \qquad \mu_t = \mu_{t-1} + \eta_t \qquad \eta_t \sim \text{NID}(0,1),$$

where ε_t is the leptokurtic noise generated by a standardized Student's t with ν degrees of freedom (DF) and $c = \lambda^{-1}$ is the fixed parameter identifying the noise-to-signal ratio. The number of DF governs the thickness of the tails of the noise component: the lower the DF, the larger the kurtosis.

For both experiments on ADF and Johansen's test, we simulate 10000 paths of length 1095, corresponding to 3 years of daily observations, for all of the possible pairs of noise-to-signal ratio c in $\{0,1,2,...,10\}$ and degrees of freedom ν in $\{3,6,9,12\}$.

On each of the simulated time series we apply the mean-filter (mean), the Kalman filter (ucm_flt) and the smoother (ucm_smo). The ADF statistic (with drift and number of lags selected by AIC) was computed on every simulated time series and the empirical rejection rates for a nominal size of 5% are represented in Figure 1.

Looking at the four graphs we can conclude that the ADF applied to the filtered time series has size close to the nominal one, the size of the ADF test applied to the smoother (ucm_smo) time series is equal to the nominal size for all the considered noise-to-signal ratios and that the thickness of the tails of the noise distribution has virtually no effect on the size of the ADF.

Similarly to the univariate case, we developed a simulation scheme for multivariate time series in order to evaluate the statistical properties of Johansen's cointegration test in presence of leptokurtik noise. Data are simulated according to a VECM with r cointegrating relations and $k = 4$ underlying times series augmented by a leptokurtic noise term. The noise is randomly generated by a standardized Student's t random variable with ν degrees of freedom and affect directly the VECM through the noise-to-signal ratio. We performed the simulation analysis on the number of cointegrating relations detected by the test considering the case of $r = 1$. For the simulation ex-

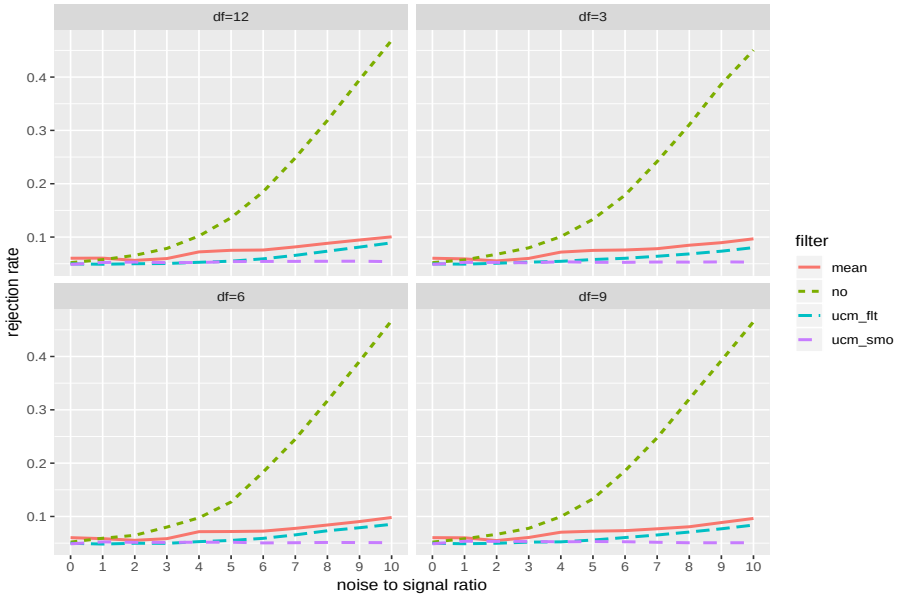


Fig. 1 Actual size of the ADF test for a nominal size of 5%.

periments involving the Johansen test we used the following two DGP: the cointegrating relations $\Delta\boldsymbol{\mu_t} = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\mu}_{t-1} + \boldsymbol{\varepsilon_t}$ with $\boldsymbol{\alpha}' = [0\ 0.1\ -0.1\ 0.1]$ and $\boldsymbol{\beta}' = [1\ -1\ 1\ -1]$ and the noisy series $y_{it} = \mu_{it} + \gamma_i z_{it}$ with $\gamma_i^2 = c\frac{VAR(\Delta y_{it})}{VAR(x_i t)}$ and where z_{it} is a leptokurtic noise and c is a fixed parameter representing the noise-to-signal ratio. The three linear filters are then applied to the series, rising three further series sharing a common underlying process. The number of cointegrating vectors is finally tested on each simulated quartet using the Johansen's trace test. For each pair of degrees of freedom and noise-to-signal ratio and for each sequential value r^*, we computed the test's size (rejection rate) as the proportion of tests rejecting the null hypothesis over the total number of simulations. For values of r^* lower than the real one, the expected rejection rate should be the closest possible to 1; while it should tend towards zero approaching or overcoming the true value. We also estimate the selection rate as the unitary complement with respect to the rejection rate and use it to compare the test performance. High selection proportions should be associated to values of r^* close to the true one and low when r^* is enough lower from it.

Figure 2 shows the empirical selection rates when the real number of cointegrating vector is $r = 1$. Until the noise-to-signal ratio remains close to zero and independently by the degrees of freedom, both filtered and raw data perform similarly and no real advantages in pre-filtering are visible. But as soon as the signal-to-noise ratio gets unbalanced, the selection rates change considerably. Kalman filter and smoother are able to guarantee selection proportions in all situations, in particular the filter remains stable with values above 90%. The figure shows also that for highly unbalanced noise-to-signal ratios, just the filter is able to maintain acceptable performances, while the other filters, and obviously raw data, are not very effective.

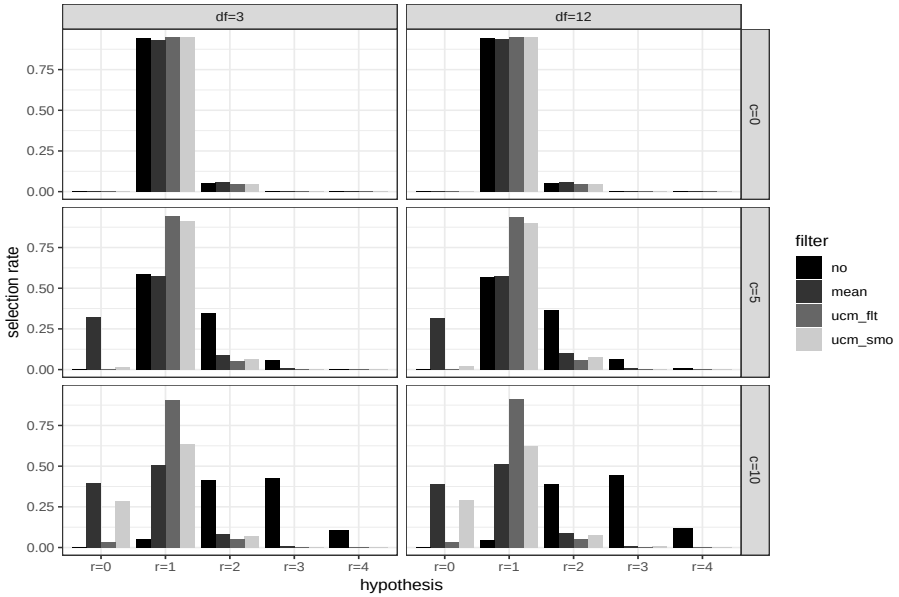


Fig. 2 Selection rate when $r = 1$.

5 Conclusions

The results assess that all the filtering methods improve the size of the ADF and Johansen tests, but the second one is the most effective for selecting the number of cointegrating relations. This kind of filtering should become routinary when testing for integration and cointegration dealing with series affected by high-frequency noise to avoid dominating spourious results.

References

1. J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.
2. G. Elliott, T. J. Rothenberg, and J. H. Stock. Efficient tests for an autoregressive unit root. *Econometrica*, 64(4):813–836, 1996.
3. J. W. Galbraith and V. Zinde-Walsh. On the distributions of Augmented Dickey–Fuller statistics in processes with moving average components. *Journal of Econometrics*, 93(1):25–47, 1999.
4. S. Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580, 1991.
5. M. M. Pelagatti. *Time Series Modelling with Unobserved Components*. Chapman and Hall / CRC, 2015.
6. S. E. Said and D. A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 12 1984.

Population and society

A dimensionality assessment of refugees' vulnerability through an Item Response Theory approach

Un approccio Item Response Theory per l'analisi di dimensionalità della vulnerabilità dei rifugiati

Simone Del Sarto[1] and Michela Gnaldi[2] and Yara Maasri[3] and Edouard Legoupil[3]

Abstract The present work deals with a key challenge for international agencies in charge of dealing with refugees' assistance, that of identifying clusters of refugees characterised by different vulnerability levels – so that the most vulnerable can benefit from support interventions (i.e., cash, livelihood activities etc.) – by at the same time accounting for the multidimensionality of the refugees' vulnerability condition. To this aim, we profile refugees on account of their vulnerability condition relying on a multidimensional and discrete Item Response Theory approach, whose main advantages consist in the possibility of clustering refugees into groups (latent classes) by concurrently taking into account the multidimensionality of the latent trait (i.e., vulnerability). First results of analyses carried out on extensive data provided the United Nation Higher Commission for Refugees highlight that vulnerability is a multidimensional condition, made of several sub-dimensions, some related to basic needs, and others to economic well-being and survival/social activities.

Abstract *Il presente lavoro tratta un topic rilevante per le agenzie internazionali che si occupano di assistenza ai rifugiati: identificare gruppi di rifugiati caratterizzati da diversi livelli di vulnerabilità – cosicché il gruppo più vulnerabile può beneficiare di interventi di sostegno (contanti, attività di sostentamento, ecc.) – e allo stesso tempo considerare la multidimensionalità della condizione di vulnerabilità. Il nostro obiettivo è quindi ottenere profili di rifugiati in base alla loro condizione di vulnerabilità, mediante un approccio IRT multidimensionale e discreto, il cui principale vantaggio consiste nella possibilità di raggruppare rifugiati in gruppi (classi latenti), tenendo conto della multidimensionalità del tratto latente (vulnerabilità). I primi risultati ottenuti su dati forniti dall'Alto Commissariato delle Nazioni Unite per i Rifugiati evidenziano che la vulnerabilità è una condizione composta da più dimensioni, alcune riguardanti il consumo di cibo e altre riferite al benessere economico e ad attività sociali e di sopravvivenza.*

Key words: Refugees' vulnerability, dimensionality, Item Response Theory, UNHCR

[1] Department of Agricultural, Food and Environmental Sciences, University of Perugia
e-mail: simone.delsarto@unipg.it

[2] Department of Political Science, University of Perugia
e-mail: michela.gnaldi@unipg.it

[3] United Nations High Commissioner for Refugees (UNHCR)
e-mail: maasri@unhcr.org
e-mail: legoupil@unhcr.org

1 Background

Due to the increase in humanitarian crises in the last several years, humanitarian agencies have to prioritise their interventions, by directing the most appropriate kind of assistance towards different groups, including the most socio-economically vulnerable. As one of those agencies, the primary purpose of the United Nations High Commissioner for Refugees (UNHCR) is to safeguard the rights and well-being of people who have been forced to flee, and strive to secure lasting solutions for them.

In this context, the present work deals with the multidimensional construct of refugees' vulnerability. Vulnerability as a concept is defined differently in the humanitarian context, depending on the mandates of the humanitarian agencies, the intended impact and the definition of the population to be assisted. For example, the World Food Programme measures food security though the Food Security Index [6], which brings together elements related to food consumption (quality and frequency), food expenditure share and specific coping mechanisms. As per UNHCR's mandate of providing international protection from persecution and refoulement[1], a vulnerability analysis and identification of refugees eligible for assistance will inevitably look at protection impact for any targeting strategies. UNHCR usually identifies risk profiles and protection vulnerabilities through the Specific Needs approach [5]: the agency has developed over 100 codes that assist frontline staff, who meet, interact with and assist refugees, to identify conditions and situations that can put a refugee at risk.

Item Response Theory (IRT) framework is particularly suitable for the context at issue, as it allows for the investigation of a latent phenomenon starting from a manifest observation of it (such as the response to a specific questionnaire). In this regard, by analysing data coming from a socio-economic assessment conducted by UNHCR in a refugee camp, we explore the dimensionality of this dataset in order to define refugees' vulnerability in terms of several dimensions contributing to its measurement. In doing this, we rely on a latent class multidimensional IRT model [2], in particular its version proposed for ordinal responses [1]. Given the latent class assumption, households may be partitioned into a certain number of groups, allowing us not only to identify the most vulnerable group(s), but, also, to characterise all the groups in terms of vulnerability multiple sub-dimensions. This latter point is crucial in order to identify a targeted assistance strategy.

This paper is organised as follows: in Section 2 we describe the data at issue and the statistical model employed for our purpose. Section 3 shows the results obtained from our analyses, while some concluding remarks are given in Section 4.

[1] Under Article 33(1) of the 1951 Refugee Convention relating to the Status of Refugees, refoulement is defined as the forcible return of refugees "to the frontiers of territories where his/her life or freedom would be threatened on account of his/her race, religion, nationality, membership of a particular social group or political opinion." (https://www.unhcr.org/3b66c2aa10)

2 Data and statistical model

We use data from a socio-economic assessment of refugees in Mauritania's Mberra camp, carried out by UNHCR during September-November, 2017. In fact, since 2012, the outbreak of armed unrest in northern Mali has led to major population movements to neighbouring countries, among which Mauritania hosts the largest number of refugees in Mberra camp.

The data at issue refer to 12,747 households and are collected through computer-assisted face-to-face interviewing by 124 enumerators and five supervisors. The dataset initially includes more than 200 variables (categorical and continuous) about the following contents: food consumption, item expenditure, sources of income, coping strategies, credit/debt, housing characteristics, asset and livestock ownership, social participation, household priority needs and preferences. After a careful inspection of the dataset record track and looking at the response rates, some variables have been discarded, while some others have been aggregated. The final dataset contains 23 categorical ordinal variables, summarised in Table 1.

As far as the statistical model is concerned, we use a latent class multidimensional IRT model for polytomous responses. Let Y_j be the response to item j, having l_j ordinal categories; the conditional probability of a generic response $y = 0, \ldots, l_j - 1$ is defined as $\phi_{j,y}(\boldsymbol{\theta}) = P(Y_j = y | \boldsymbol{\Theta} = \boldsymbol{\theta})$, where $\boldsymbol{\Theta}$ is the latent trait vector (D components) and $\boldsymbol{\theta}$ is its realisation. We assume a discrete distribution with k support points for it, which is equivalent to suppose that the population under study is divided into k latent classes, then units (i.e., households in our case) in the same class have the same latent trait level (i.e., a dimension of vulnerability).

We use the following model parametrisation:

$$g_y[\boldsymbol{\phi}_j(\boldsymbol{\theta})] = \gamma_j \Big(\sum_{d=1}^{D} I_{jd} \theta_d - \beta_{jy} \Big), \;\; j = 1, \ldots, J, \;\; y = 0, \ldots, l_j - 1$$

where $\boldsymbol{\phi}_j(\boldsymbol{\theta}) = [\phi_{j,0}(\boldsymbol{\theta}), \phi_{j,1}(\boldsymbol{\theta}), \ldots, \phi_{j,l_j-1}(\boldsymbol{\theta})]^\top$ and γ_j and β_{jy} are the discrimination parameter and difficulty levels of item j. Finally, I_{jd} is an indicator variable, equal to 1 if item j contribute to measure latent trait d, and 0 otherwise.

According to the definition of link function $g_y(\cdot)$ and to the constraints adopted for the item parameters, different model specifications may arise. In this work, we use a global logit link function with no constraints for the item parameters, allowing us to consider a latent class multidimensional graded response model [1].

In order to investigate the dimensionality of refugees' vulnerability, we need to find how many dimensions (or latent traits) this construct is made of and the items composing each of these dimensions. By looking at the model specification, this translates into determining the dimension of latent vector $\boldsymbol{\Theta}$ (i.e., D) and the specification of indicator variables I_{jd} for each item $j = 1, \ldots, J$ and each dimension $d = 1, \ldots, D$. To this aim, we adopt an exploratory hierarchical clustering algorithm [2, 3], which allows us to consider all the possible combinations of items (or groups of items), from the most restrictive model (one dimension for each item) to the most general (the undimensional model).

Table 1 Descriptive statistics of the variables contained in the final dataset

Label	**Description**						
	Last week consumption of	*Never*	*Sometimes*	*Often*	*Always*		
cer_7d	Cereals	0.001	0.006	0.009	0.985		
tub_7d	Tubers	0.682	0.211	0.054	0.052		
leg_7d	Legumes	0.764	0.128	0.042	0.066		
fr_veg_7d	Fruits and vegetables	0.645	0.154	0.072	0.129		
meat_egg_7d	Meat and eggs	0.063	0.233	0.180	0.525		
fat_7d	Fat	0.006	0.009	0.016	0.969		
milk_7d	Milk	0.097	0.158	0.138	0.607		
sugar_7d	Sugar	0.016	0.018	0.018	0.948		
cond_7d	Condiments	0.015	0.017	0.015	0.954		
tea_coff_7d	Tea and coffee	0.039	0.018	0.017	0.925		
honey_7d	Honey	0.972	0.011	0.003	0.013		
other_7d	Other foods	0.932	0.052	0.007	0.010		
	Monthly expenditure (€)	*Q1*	*Median*	*Mean*	*Q3*		
food_exp	Food	437	770	919	1,189		
non_food_exp	Non-food	254	499	693	872		
		1	*2*	*3*	*4+*		
income_sources	Income sources	0.349	0.589	0.059	0.003		
		None	*Stress*	*Crisis*	*Emerg.*		
coping	Coping strategies	0.291	0.562	0.099	0.048		
		No	*Yes*				
debts	Taking out debts	0.080	0.920				
cred_local	Access to local market credit system	0.451	0.549				
jewellery	Possession of jewellery	0.857	0.143				
mobile_phone	Possession of mobile phone	0.553	0.447				
animals	Possession of animals	0.657	0.343				
assoc_group	Member of an associative group	0.950	0.050				
		Never	*Rarely*	*Sometimes*	*Often*	*Very often*	*Always*
tea_somebody	Having a tea with somebody in the camp?	0.220	0.245	0.193	0.173	0.066	0.102

3 Results

In this section, we initially report the results of our analyses as regards the dimensionality assessment of refugees' vulnerability, using the dendrogram shown in Figure 1. We decide to use $k = 5$ latent classes in order to properly differentiate profiles of households, given the objective of a targeted assistance. According to the Bayesian Information Criterion [4], we can cut the dendrogram in correspondence of the dashed line, allowing us to select five groups of items (hence, $D = 5$).

Each of these groups consists in a different dimension characterising refugees' vulnerability. For example, dimensions from 3 to 5 clearly represent vulnerability

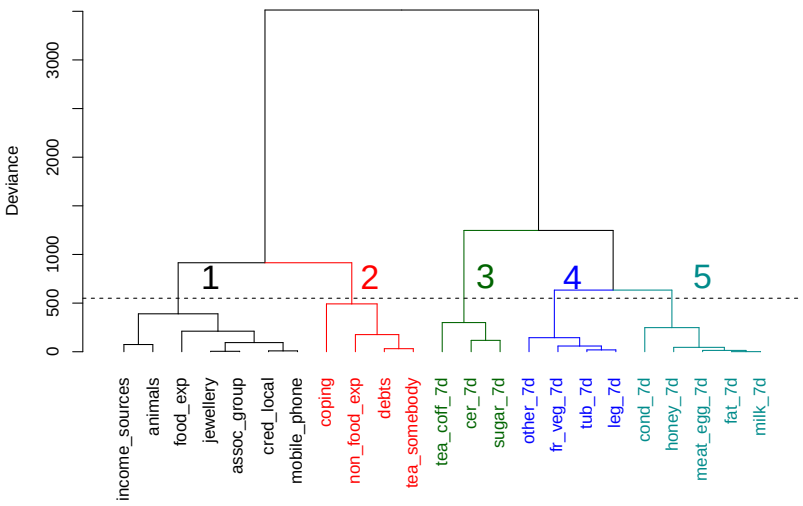


Fig. 1 Dendrogram resulting from the hierarchical clustering algorithm of the variables in the dataset.

sides connected with better food consumption. In particular, dimension 5 clusters animal-based foods while dimension 4 includes plant-based foods. They provide different nutritional components (fats and protein vs. carbs and fiber) and one group is harder to acquire (those in dimension 5) than the other. Moreover, dimension 3 includes foods that do not have high nutritional value but they are still highly consumed.

As regards the non-food dimensions, the first one (dimension 1 in Figure 1) appears as, say, the economic side of vulnerability, since it includes the possession of assets, the amount of income sources and food expenditure, as well as membership in an associative group, which in this context could be a cooperative of farmers, for example. Finally, we can say that dimension 2 mainly represents survival capacity, through specific coping strategies as well as by means of purchase of non-food goods, together with social integration opportunities within the community.

Taking into account these vulnerability facets, in Table 2 we report the latent trait estimates for each of the five latent classes (labelled from A to E), in order to describe refugees' household profiles. As we can see, the most vulnerable profile, identified by the lowest latent trait estimates (latent class A), includes almost 5% of the refugees' households and is characterised by low levels of food consumption (particularly far from the rest of the households), especially as regards dimensions 3 and 5. On the other hand, a quarter of the households belongs to the least vulnerable group (latent class E), having the highest estimates with respect to all the five vulnerability dimensions.

Table 2 Latent trait estimates for each latent class and each dimension, along with the prior probability estimates of each class ($\hat{\pi}$).

	Dimension					
Latent class	1	2	3	4	5	$\hat{\pi}$
A	-1.24	-0.95	-3.89	-0.95	-2.09	0.057
B	-1.60	-1.74	-0.33	-0.72	-0.86	0.161
C	0.04	0.35	0.24	-0.99	-0.63	0.329
D	-0.15	-0.40	0.18	1.01	0.75	0.212
E	1.44	1.26	0.65	1.17	1.27	0.241

4 Conclusions

This work stems from the need to target assistance to refugee households, which includes different types of assistance according to profiles. As such, refugees' vulnerability is a latent and multidimensional construct, since it may be considered as composed by several and related dimensions. In order to investigate such dimensionality, an Item Response Theory approach is carried out, by exploiting a survey administrated directly on a refugees' camp in Mauritania during the year 2017. A hierarchical clustering algorithm applied on the available variables is performed with the aim of exploring how items cluster together and how many dimensions contribute to define the vulnerability construct. Main findings reveal that in this case five dimensions characterise refugees' vulnerability, related to survival and basic needs, coping strategies, food consumption, as well as socio-economic well-being and social activities. Keeping in mind these dimensions and exploiting the latent class approach of the model at issue, we are able to highlight the different vulnerabilities that exist among the different profiles, including the most socio-economically vulnerable, in order to alert humanitarian agencies towards a focused intervention.

Acknowledgements This work is the result of a collaborative research agreement between the University of Perugia, Department of Political Science, and UNHCR. The authors would like to thank Dr Osama Abdelhay for his contribution to the research by having explored the problem, tested alternative statistical options and methodologies and recommended IRT as the methodology for in-depth research; Ms Irina Conovali to have coordinated the research efforts at UNHCR's side and many other UNHCR staff who have provided their time, knowledge and insights about the refugee situations in the Middle East and North Africa Region.

References

1. Bacci, S., Bartolucci, F., Gnaldi, M.: A class of multidimensional latent class IRT models for ordinal polytomous item responses. Commun. Stat.-Theory Methods **43**, 787–800 (2014)
2. Bartolucci, F.: A class of multidimensional IRT models for testing unidimensionality and clustering items. Psychometrika **72**, 141–157 (2007)
3. Gnaldi, M., Del Sarto, S.: Time use habits of Italian Generation Y: dimensions of leisure preferences. Soc. Indic. Res. **138**, 1187–1203 (2018)
4. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. **6**, 461–464 (1978)
5. UNHCR: UNHCR Emergency Handbook, 4th edition (2015). Available at https://emergency.unhcr.org/
6. World Food Program: Consolidated Approach to Reporting Indicators of food security. Technical guidance note, 2nd edition (2015).

Accounting for Interdependent Risks in Vulnerability Assessment of Refugees

Rischi Interdipendenti per la Misura della Vulnerabilità dei Rifugiati

Daria Mendola, Anna Maria Parroco, Paolo Li Donni

Abstract The United Nations' Vulnerability Assessment Framework (VAF) of refugees encompasses a set of indicators of the living conditions in refugee camps in Lebanon and Jordan; it assumes the independence among its components. In this paper we maintain the importance to account for existing interdependencies, and provide a definition of vulnerability for high income countries. The proposed "vulnerability scale", based on the estimated joint risk of social isolation, economic deprivation and bad health, is a useful tool to address interventions toward selected groups of more vulnerable refugees. Analyses are based upon the survey of refugees carried on in Germany in 2016. Germany is the first country in Europe for the number of hosted refugees.

Abstract Per la misura della vulnerabilità dei rifugiati le Nazioni Unite hanno proposto il modello VAF, che adotta un insieme di indicatori delle condizioni di vita nei campi profughi in Libano e Giordania. Il VAF assume l'indipendenza delle sue componenti. In questo lavoro si propone invece di tener conto delle interrelazioni tra le componenti della vulnerabilità e si ragiona sul concetto di vulnerabilità nei paesi ad alto reddito. La scala di valutazione della vulnerabilità qui proposta- basata sulla stima del rischio congiunto di isolamento sociale, povertà e cattiva salute- costituisce un utile strumento per indirizzare le azioni di sostegno verso i gruppi maggiormente a rischio vulnerabilità. Le analisi sono basate sull'indagine sui rifugiati fatta in Germania nel 2016, il primo paese in Europa per numero di rifugiati accolti.

Key words: trivarite logit, poverty, health, social isolation, VAF, Germany, IAB-BAMF-SOEP.

[1] Daria Mendola, Dipartimento SEAS, Università di Palermo; daria.mendola@unipa.it;
Anna Maria Parroco, Dipartimento SPPEFF, Università di Palermo; annamaria.parroco@unipa.it;
Paolo Li Donni, Dipartimento SEAS, Università di Palermo; paolo.lidonni@unipa.it.

1 Introduction and literature review

The UNHCR -United Nations High Commissioner for refugees- data [15] assessed in 2019 25.9 million of refugee worldwide in 2018. Most of them fled in neighbourhood countries. Refugees living in camps, such as those in Jordan, experience hard living conditions [5]. Worst conditions are experienced in Lebanon, where in 2015 the Government stopped the UNHCR's registration of refugees (but not the entrance) in order to make not visible the huge amount of hosted refugees (16.7% of its population according to the last UNHCR registration).

In recent years, a growing number of refugees arrived also to Europe and several countries are struggling in processing the high number of applications for international protection. The 2018 was a peak year in Europe, following the preceding 5 years of steep increase. Eurostat database counts for 1,127,690 new refugees in EU-28 at 31st December 2018 (up to 1,635,289 considering also holders of subsidiary protection), of these 662,954 were hosted in Germany (up to 888,016 including subsidiary protection). Recently emerging conflicts in West Asia and in several African countries make reasonable to think that these flows and stocks will increase rapidly in the next future.

While there are several academic and institutional studies on the quality of life and refugees' vulnerability in medium-low income countries, mostly on Lebanon and Jordan (e.g. [11, 5]), very few studies deal with the vulnerability of refugees and asylum seekers in Europe (see e.g., [2, 3, 8, 9, 10]).

In 2017 the UNHCR [13,14] developed the Vulnerability Assessment Framework (VAF). It is a scoreboard for targeting individuals for intervention. VAF encompasses 10 thematic areas, through 65 indicators, covering a wide spectrum of needs of refugees in camps (health, shelter, food security, documentation, education, economic deprivation, etc.). Noteworthy the UN's VAF hardly fits in high income countries where the dimensions of vulnerability go beyond that of basic needs.

Although this growing literature, still little has been done to study the vulnerability of refugees and asylum seekers in high-income countries. In qualitative study [2] on Iraqi and Iranian refugees in Greece, the authors discuss about adequacy of several indicators for measuring vulnerability at household and individual level; [3] measures vulnerability among refugees and asylum seekers living in informal settlements in Italy; [4] focuses on the assessment of the environmental health conditions and associated vulnerability of migrant residents in the Calais (France) refugee camp, analysing a set of indicators but not providing a synthetic measure of multidimensional vulnerability; [9] focuses on two health outcomes of forced migrants living in informal settlements in Italy and find that they are associated with both personal and settlement characteristics. None of these studies (except [3]) takes into account the correlations between items of vulnerability, which introduce a "double counting/weighting" effect in the measurement of vulnerability.

We propose a measure of vulnerability of refugee communities in high-income countries that deals with this correlation by modelling the joint probability of

experiencing interdependent risks. Our analysis focuses on Germany, since it is the 5-th hosting countries for refugees worldwide ([15]) and first in Europe.

In the following, we first briefly introduce our theoretical and then statistical model along with data on refugees living in Germany, then we discuss some results (primarily our vulnerability scale) and reflections on future research paths.

2 Models and Data

The Theoretical Model. Vulnerability is the "*state of high exposure to certain risks, combined with a reduced ability to protect or defend oneself against those risks and cope with their negative consequences*" ([12], p. 210). Our theoretical framework acknowledges that these risks are often highly interrelated, producing a cumulative disadvantage that goes beyond the sum of parts. In the context of high income countries, particularly Germany (the European country with the highest presence of refugees), where basic needs are usually provided to refugees under the mandate of the Genève Convention, we assume that main risks are those of social isolation, economic deprivation and bad health.

The Statistical Model. In order to accomplish for the interdependence of these risks, we estimate a trivariate logit model to evaluate how individual and household characteristics are associated with the probability of experiencing each risk, and make inference also on the residual association between pairs of risks, conditionally to a set of selected covariates. Hence, we jointly model: a) the univariate marginal distribution by assuming a linear model for each logit in equation (1), where Y_1 is the outcome variable "social isolation" (that is: feeling very/often socially isolated), Y_2 refers to "perceived economic difficulties" (i.e. being very/somewhat concerned about finances) and Y_3 to "bad health" (i.e. reporting poor or bad health); and b) the marginal association between each pair of responses, modelled by a set of odds ratios described in equation (2). We include a core set of variables explaining all the observed outcomes (household composition -size and marital status-, education -three levels-, employment -yes/no/seeking-, years since migration, having any form of international protection- and nationality groups- most likely to remain (Afghans, Syrians, Eritreans, Iraqis) vs. others-) and specific sets pertaining to each single risk Y_k (frequency of social relations, knowledge of German language, health problems limiting social relations for Y_1;allowances and benefits received, disability for Y_2; physical and mental health impairments-four variables- for Y_3) The whole matrix of covariates is here referred as *Z*.

The model for the risk of social isolation particularly includes language skills, social network and health measures; the risk of perceived financial difficulties is modelled including also benefits dependence, health and quality of housing measures, the one for the risk of bad health includes quality of housing and state benefit dependence.

Data are from the IAB-BAMF-SOEP Survey of Refugees in Germany that is a longitudinal survey of people who entered Germany between 2013 and 2016 and applied for asylum, whatever the result of the application. It includes information on individual socio-demographic characteristics and household level information.

The survey is provides yearly interviews of household members aged 18 and over. We exploit the first wave of the survey (2016). Our sample is made of 3,072 adults, with a prevalence of men (62%), a mean age of 33.6 years, with four nationalities (Afghan, Eritrean, Iraqi, and Syrian) accounting for about 82% of the sample. Among them, only 59% were granted any form of international protection - refugee status (73.66%), international protection, status of tolerance- while the remaining 41% is lacking of this status (among these 85.67% are asylum applicants with a pending request).

$$\log \frac{\Pr(Y_k = 1)}{\Pr(Y_k = 0)} = -\gamma_k + z_k^{'}\tau_k \qquad k=1, 2, 3 \qquad (1)$$

$$\lambda_{hk} = \log \frac{\Pr(Y_h = 1, Y_k = 1 \mid z)\Pr(Y_h = 0, Y_k = 0 \mid z)}{\Pr(Y_h = 0, Y_k = 1 \mid z)\Pr(Y_h = 1, Y_k = 0 \mid z)} \qquad \forall\, h, k \in \{1,2,3\} \qquad (2)$$

3 The Vulnerability Scale

Results from the trivariate logit reveals a significant conditional association between the three risks supporting the hypothesis of interdependent risks. Particularly, from equation (2): $\lambda_{y_1y_2}|z = 2.11$, $\lambda_{y_1y_3}|z = 1.31$, $\lambda_{y_2y_3}|z = 1.46$ are all statistically significant at 1%.

A useful by-product of the model in Section 2 are the predicted probabilities of experiencing one, two or three risks jointly (p_{ihk}). Hence, for each individual we can estimate 2^3 different probabilities, corresponding to the combination of presence/absence of three dichotomous risks. Table 1 provides a synthetic tools for grading risks, assumed as measures of different levels of vulnerability. It shows mean estimated probabilities over the selected sample and their standard deviations for each vulnerability profile, defined, in the second column, in terms of presence (1) or absence (0) of each of the three risks.

Mean probabilities in Table 1, column 3, can be assumed also as a measure of the (model) predicted incidence of each of eight vulnerability profiles, while categories introduced in column 1 can be used as a criterion for prioritizing interventions. The condition of severe vulnerability refers to a very small amount of individuals who may be the target for the very first intervention. On the opposite side of our scale, the level of low vulnerability may concern about 15% of individuals and, in this case, less (or no special) aid is presumably needed. Comparing the other different levels of our vulnerability assessment scale, the

highest probability is associated with economic deprivation that, controlling for the other two risks, has a mean value of 0.54. Other conditions have probabilities relatively low, not exceeding 10%.

Table 1: Vulnerability scale

Level of Vulnerability	***Isolated-deprived-bad health***	***Mean p_{ihk}***	***Standard Dev.***	***Min***	***Max***
Severe	111	0.019	0.035	0.000	0.470
High	101	0.018	0.030	0.001	0.336
	110	0.069	0.044	0.001	0.395
	011	0.074	0.111	0.002	0.633
Moderate	100	0.099	0.049	0.003	0.420
	010	0.536	0.172	0.010	0.925
	001	0.031	0.049	0.000	0.339
Low	000	0.154	0.060	0.004	0.434

The upper tail (last decile) of the distribution of p_{111} (the joint probability of experiencing all three risks, i.e. the condition of severe vulnerability) may be assumed as the target population for more urgent policy intervention. In this group about 64% have no international protection; 60.26% have no education and 28.66% have higher education; 78% live in household with 4 or less members; 57.65% has at least one child in the household.

4 Conclusion

Refugees can experience vulnerability in country of origin, while they are on route or in the destination country, which has a responsibility in the infringement of migrants' human rights and on migrant's welfare standards and social inclusion ([1]). In this paper, we propose a new approach for measuring vulnerability of refugees and asylum seekers, by means of the estimated joint effect of three risks they can experience in hosting high-income countries (social isolation, economic deprivation, bad health).

Although international laws recognize all refugees as vulnerable *per se*, an emerging strand of literature argues that once arrived in hosting countries, and assisted by welfare state, not all refugees are still vulnerable (see the Judge Sajo's dissenting opinion [6] against the European Court of Human Rights' qualification of asylum seekers as a particularly underprivileged and vulnerable population *per se*).

Our empirical appraisal to analyse vulnerability allows for partitioning refugees hosted in high income countries, and supported by welfare states, in some selected subgroups according to their vulnerability profiles. They should be the targeted for early interventions. Even if not considered here for the sake of brevity, it is possible to detect what risk/s is/are more likely to make them more exposed, hence more vulnerable, allowing addressing interventions toward specific hampering conditions.

Indeed, factors predicting vulnerability often change depending upon the way in which vulnerability is measured. Scoreboard approaches, ignoring the double counting effect implied by adding highly interdependent factors of risks (such as it is the case for the VAF exercise), may not provide a correct priority ranking of people needing aid.

Our approach can be extended in several directions. First of all, intensity of risk can be modelled by using a set of ordered rather than binary discrete risk indicators. Secondly, residual association can be allowed to depend also on a set of covariates, capturing in this way a sort of "propagation" effect between risk dimensions.

References

1. Atak, I., Nakache, D., Guild, E., Crépeau, F. (2018) Migrants in vulnerable situations and the Global Compact for Safe Orderly and Regular Migration. Queen Mary University of London, School of Law, Legal Studies Research Paper n. 273/2018.
2. Black, R. (1994) Livelihoods under Stress: A Case Study of Refugee Vulnerability in Greece, 7 J. Refugee Studies, 360
3. Busetta, A., Mendola, D., Wilson, B., and Cetorelli, V. (2019). Measuring vulnerability of asylum seekers and refugees in Italy, Journal of Ethnic and Migration Studies, DOI: 10.1080/1369183X.2019.1610368
4. Dhesi, S., Isakjee, A. and Davies, T. (2018). "Public Health in the Calais Refugee Camp: Environment, Health and Exclusion." Critical Public Health, 28(2): 140–152.
5. El-Khatib, Z., Scales, D., Vearey, J., and Forsberg, B. C. (2013). Syrian refugees, between rocky crisis in Syria and hard inaccessibility to healthcare services in Lebanon and Jordan. Conflict and health, 7(1), 18.
6. European Court of Human Rights (2009) Grand Chamber judgment in the case of Paladi v. Moldova (application no. 39806/05), n.189, 10.3.2009.
7. Krafft, C., Sieverding, M., Salemi, C., & Keo, C. (2018, April). Syrian refugees in Jordan: Demographics, livelihoods, education, and health. In Economic Research Forum Working Paper Series (No. 1184).
8. Kohlenberger, J., Buber, I., Rengs, B., and Al Zalak, Z. (2016). A social survey on asylum seekers in and around Vienna in fall 2015: Methodological approach and field observations (No. 6/2016). Vienna Institute of Demography Working Papers.
9. Mendola, D., and Busetta, A. (2018). Health and Living Conditions of Refugees and Asylum-Seekers: A Survey of Informal Settlements in Italy. Refugee Survey Quarterly, 37(4): 477–505.
10. Stewart, E. (2005). Exploring the Vulnerability of Asylum Seekers in the UK. Population, Space and Place, 11 (6): 499–512.
11. Verme, P., Gigliarano, C., Wieser, C., Hedlund, K., Petzoldt, M., and Santacroce, M. (2015). The welfare of Syrian refugees: evidence from Jordan and Lebanon. The World Bank.
12. United Nations (2001) Report on the World Social Situation.
13. UNHCR (United Nations High Commissioner for Refugees). (2017). Vulnerability Assessment Framework Guidance Note. Geneva: UNHCR.
14. UNHCR (United Nations High Commissioner for Refugees). (2018). Jordan – Vulnerability Assessment Framework 2017 – Population Survey Report – Sector Vulnerability Review (July 2018). Geneva: UNHCR.
15. UNHCR (United Nations High Commissioner for Refugees). (2019) Database of refugees. http://popstats.unhcr.org/en/overview

Active ageing in China: What are the domains that most affect life satisfaction in the elderly?

Invecchiamento attivo in Cina: Quali domini influenzano maggiormente la soddisfazione per la vita negli anziani?

Ilaria Rocco

Abstract Given the rapidly increasing life expectancy in China, more attention has recently been focused on the concept of active ageing as a more appropriate goal for policymakers and on the importance of monitoring its changes over time. The aim of the present work is to further explore the domains of the Active Ageing Index taking into consideration the perspectives of older persons.
The analysis was carried out utilizing the data produced by the fourth wave of the "China Health and Retirement Longitudinal Study". Generalized Structural Equation Model was applied to investigate the implications of the dimensions of active ageing on the perceived life satisfaction and to identify the domains more urgently requiring interventions.
Abstract *Dato il rapido incremento della speranza di vita in Cina, una attenzione crescente da parte della letteratura è stata dedicata al processo di invecchiamento attivo come obiettivo prioritario per i responsabili politici e al monitoraggio dei suoi cambiamenti nel tempo. Il presente lavoro è volto ad approfondire i domini dell'Active Ageing Index prendendo in considerazione la prospettiva delle persone anziane. L'analisi è stata condotta utilizzando i dati prodotti dalla quarta rilevazione dell'indagine "China Health and Retirement Longitudinal Study". È stato applicato un modello ad equazioni strutturali generalizzato per studiare l'implicazione delle dimensioni dell'invecchiamento attivo sulla soddisfazione percepita e per identificare i settori che richiedono interventi più urgenti.*

Key words: Active ageing, life satisfaction, elderly

Ilaria Rocco, Dipartimento di Scienze Statistiche, Sapienza Università di Roma; Istituto di Neuroscienze, CNR, Padova; email: ilaria.rocco.90@gmail.com

1 Introduction

Due to the dramatic decline in fertility rates and the rapid increase in life expectancy, the world's population is quickly growing older. This is particularly true in China, where those aged 65 and older make up 12% of the population and will probably make up 26% by 2050 [9].

A rapidly increasing life expectancy can be considered a positive step forward only if the additional years are characterized by good health and a satisfactory quality of life. If that is the case, human beings will be able to look forward to growing old if governments, international organizations, and the civil society promote "active ageing" policies and programmes that enhance the health, participation and security of older citizens [10].

The World Health Organization (WHO) adopted the term "active ageing" in the late 1990s to convey a more inclusive message with respect to "healthy ageing" and to recognize the factors in addition to health care that affect how individuals and populations age [5]. In 2002, the WHO defined it as "the process of optimizing opportunities for health, participation and security in order to enhance quality of life as people age" since it was sure that applying an active ageing framework onto a healthy ageing promotion program could more effectively solve the challenges of an ageing population [10].

Based on the WHO's definition of active ageing, in 2012 the United Nations Economic Commission for Europe experts developed the Active Ageing Index (AAI) as a tool to measure and monitor active ageing [11]. The AAI covers four domains of active ageing: employment, social participation, independent/health/secure living, and capacity and enabling environment. It includes 22 indicators collected from different questionnaires with scores ranging from 0 to 100, with the latter representing the best possible result. The index has been used in 28 European countries [11] and has recently been adopted by other countries, including China [12] where the overall AAI score in 2014 was 37.3, which was higher than the EU average (33.9).

According to its developers [11], the index has the following advantages: it is an appealing list of factors offering several analytical possibilities; it is useful for evaluating "the untapped potential of older people" in each country; it enables disaggregating the overall indices into domain-specific ones; it facilitates numerical interpretation for use by a wide audience.

Although the AAI is one of the main tools for monitoring active ageing policies particularly in the European context, several limitations in its construction and interpretation have been identified. A study argued that the AAI is an incomplete tool for policymaking purposes, since policymakers cannot be sure whether they should promote certain activities because the AAI does not clarify to what extent these activities are valued by older people or the degree of freedom they have in achieving valuable "doings and beings" [8].

The aim of the present work is to further explore the domains of the AAI in an "emic perspective", that is taking into consideration the perspectives and preferences

of older persons [4] so as to investigate which dimensions of active ageing influence more individual's overall life satisfaction.

2 Data and methods

The analysis of active ageing domains was carried out utilizing the data of the sample survey on "China Health and Retirement Longitudinal Study" (CHARLS), a biennial multistage longitudinal survey investigating the middle-aged and older population collected by the National School for Development (China Center for Economic Research) together with the Institute for Social Science Survey at Peking University. After a pilot study examining 2 provinces in 2008 [2], the baseline national wave of CHARLS was fielded in 2011; it included about 10,000 households and 17,500 individuals in 150 counties/districts and 450 villages/resident committees. We focused on the fourth wave which was carried out in 2015 and gathered data about 21,095 individuals. The present work was particularly interested in the respondents aged 65 years and above.

Life satisfaction was evaluated through the question: "How satisfied were you with life as a whole?" answered using a 5-point Likert scale. This single item to measure life satisfaction was proved to be a reliable and valid method and was widely used [1, 3].

The dimensions of the active aging considered were those from the original AAI [11]. However, not all the variables used to operationalize the dimensions in the original index were included in the measurement model: four variables were not available or computable in CHARLS dataset (i.e. physical exercise, relative median income, remaining and healthy life expectancy); the lifelong learning was available but not included due to the low percentage in the sample (<1.0%); the educational attainment, that referred to the past history of the elderly, was not suitable of variation, hence it was used as control variable with age and gender.

The "Employment" (EMP) domain of the AAI framework was represented by the employment status. We defined as employed people who were engaged in agricultural work or worked at least one hour the week before the interview.

The "Participation in society" (PIS) domain included the participation in voluntary activities; the provision of care to children/grandchildren; the provision of care to older adults or disabled relatives; the participation in political activities.

The "Independent, healthy and secure living" (IHSE) domain was measured considering: access to health care, independent living arrangements, poverty risk, material deprivation and physical safety.

Finally, the "Capacity and enabling environment for active ageing" (CEE) domain included: mental wellbeing, use of ICT, and social connectedness. The description and the percentage distribution of the variables used to measure each domain are reported in Table 1.

In order to test the viability of the AAI framework, a confirmatory factor analysis was conducted on each of the four above-described domains. To estimate

the effect that the AAI domains have on the individual quality of life, the Structural Equation Modelling (SEM) approach was adopted.

SEM is a very general statistical modelling technique, widely used in the behavioural sciences, which combine the strengths of factor analysis and multiple regression in a single model that can be tested statistically.

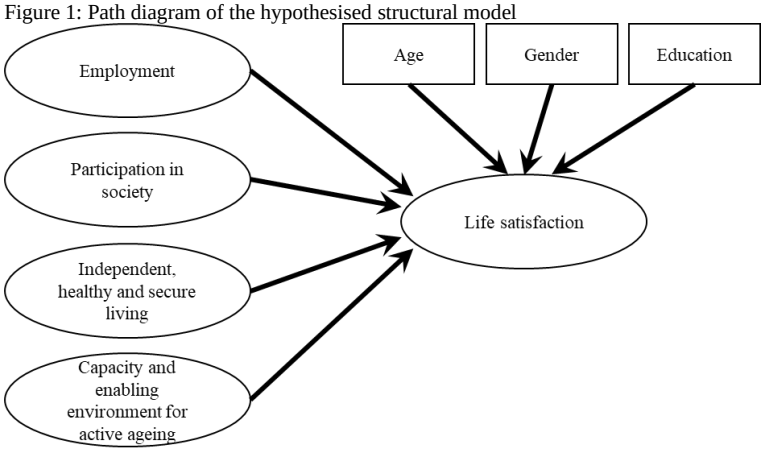
SEM has been further extended to the generalized form (GSEM) for the evaluation of categorical and time-to-event outcomes [7]. In this work, the model was analysed using the ordinal distribution and logit link.

The model consisted of two parts: the measurement model, where the four AAI domains were reflected by their respective variables [6]; and the structural model (in Figure 1) representing the relationships between the AAI domains and the life satisfaction. Age, gender and education were included as control variables. Weighted data have been used to account for sampling design.

The statistical analyses were performed using the SAS 9.4 software (SAS Institute, Cary, NC, USA) and Stata version 13 (StataCorp, USA).

Although 21,095 individuals were surveyed in the 2015 CHARLS study, 5,508 respondents were 65 or older, had complete data on satisfaction variables and valid individual sample weight: 49.4% were men, the mean age was 72.0 years (SD = 5.8) and 10.6% completed at least high school. Only 7.7% of the respondents were not very satisfied or not satisfied at all, while 43.0% were completely or very satisfied.

Figure 1: Path diagram of the hypothesised structural model



3 Results and preliminary conclusion

The results of the GSEMs estimation are reported in Table 1. It emerged that two AAI domains (EMP and PIS) do not have a significant effect on the perceived life satisfaction (p-value>0.05).

Table 1: Description and percentage distribution of the variables used to measure each AAI domain. Estimate (coefficient and p-value) of the measurement and structural model

Measurement model				
Domain	Variables (Description)	%	Coef.	p-value
EMP	Employment status (% Engaged in agricultural work; or Worked at least one hour last week)	38.5	-	-
PIS	Participation in voluntary activities (% Provided help to family/friends/neighbours; or Did voluntary or charity work)	10.5	0.2088	<0.001
	Provide care to children/ grandchildren (% Provided care to children/ grandchildren)	33.4	0.0355	<0.001
	Care to older adults or disabled relatives (% Provided care to older adults or disabled relatives)	2.1	0.0390	<0.001
	Participation in political activities (% Went to a sport, social, or other kind of club; or Took part in a community-related organization)	10.0	0.0719	<0.001
IHSL	Access to health care (% have access health care when ill last month, or who had not accessed health care but this is not due to poor delivery or provision of care)	96.3	0.0432	<0.001
	Independent living arrangements (% who live alone or in a couple)	78.8	0.0993	<0.001
	No poverty risk (% had a household income above the median income (=9000 yuan))	49.9	0.1099	<0.001
	No severe material deprivation (% had at least one of the following equipment: refrigerator, washing machine, TV, air conditioner.)	89.3	0.0889	<0.001
	Physical safety (% did not feel fearful most of the time (5-7 days))	95.1	0.0646	<0.001
CEE	Mental wellbeing (% had good mental wellbeing)	77.3	0.2844	<0.001
	Use of ICT (% Used the Internet)	3.6	0.0146	<0.001
	Social connectedness (% Interacted with friends or Played Ma-jong/chess/cards, or went to community club)	42.3	0.0353	0.002

Structural model			
Path	.	Coef.	p-value
Age → Life Satisfaction		0.0247	<0.001
Gender (ref: Male) → Life Satisfaction		0.1333	0.018
Education (ref: Up to lower secondary) → Life Satisfaction		-0.4532	<0.001
EMP → Life Satisfaction		-0.0285	0.627
PIS → Life Satisfaction		0.0690	0.091
IHSL → Life Satisfaction		0.3458	<0.001
CEE → Life Satisfaction		0.6296	<0.001
Cutpoints	c1	-2.3316	<0.001
	c2	-0.7822	0.049
	c3	2.2135	<0.001
	c4	4.7619	<0.001

The IHSL and CEE domains positively affect how older people perceive their life satisfaction. Therefore, the most important aspects for active ageing are: the mental wellbeing, the independence, both in term of financial resources and living arrangement, the safety, and the accessibility of the health care needed.

The application of SEM was able, taking into account several latent dimensions, to simultaneously estimate the implication of the active ageing dimensions on the perception of the individual life satisfaction. This exploration of the domains in an "emic perspective" has revealed that not all the dimensions included in the AAI affect the individual's perception of life. Therefore, this allowed the identification of the two domains in which there is a real need felt by older people for an urgent intervention that mobilizes them to make their "contributions" to the respective societies.

References

1. Blanchflower, D.G.: International evidence on well-being. In Measuring the subjective well-being of nations: national accounts of timeuse and well-being, pp. 155–226. University of Chicago Press (2009).
2. Zhao, Y., Strauss, J., Park, A., Shen, Y., Sun, Y.: 2008 CHARLS Pilot: User's Guide. Beijing, China: China Center for Economic Research, Peking University. http://charls.pku.edu.cn/pages/doc/2008_pilot/en.html Updated on February 16, 2013. Accessed on March 5, 2020.
3. Diener, E., Inglehart, R., Tay, L.: Theory and validity of life satisfaction scales. Soc. Indic. Res. 112(3), 497–527 (2013). doi: 10.1007/s11205-012-0076-y
4. Hjelm, M., Holst, G., Willman, A., Bohman, D., Kristensson, J.: The work of case managers as experienced by older persons (75+) with multi-morbidity - a focused ethnography. BMC Geriatr. 15:168 (2015) doi:10.1186/s12877-015-0172-3
5. Kalache, A., Kickbuschm, I.: A global strategy for healthy ageing. World Health 50: 2 (1997).
6. Lai, M.M., Lein, S.Y., Lau, S.H., Lai, M.L. Modeling Age-Friendly Environment, Active Aging, and Social Connectedness in an Emerging Asian Economy. *J Aging Res*. (2016). doi:10.1155/2016/2052380
7. Rabe-Hesketh, S., Skrondal, A., Pickles, A.: Generalized multilevel structural equation modeling. Psychometrika 69: 167–190 (2004).
8. São José de, J.M., Timonen, V., Amado, C.A.F., Santos, S.P.: A critique of the Active Ageing Index, J. Aging Stud. 40, pp. 49-56 (2017).
9. United Nations, Population Division. World Population Prospects 2019. https://population.un.org/wpp/. Accessed on March 5, 2020.
10. World Health Organization: Active Aging. A Policy Framework. Geneve (2002).
11. Zaidi, A., Gasior, K., Hofmarcher, M.M., Lelkes, O., Marin, B., Rodrigues, R., Schmidt, A., Vanhuysse, P., Zolyomi, E.: Active ageing index 2012 concept, methodology and final results, Research Memorandum/Methodology Report, European Centre Vienna (2013). Available at: www.euro.centre.org/data/aai/1253897823_70974.pdf
12. Zaidi, A., Um, J., Xiong, Q., Parry, J.: Active Ageing Index for China Comparative Analysis with EU Member States and the Republic of Korea. Available from: https://www.euchinasprp.eu/images/ProjectMemorabilia/2018Reports/Comparati ve_Study_on_Active_Ageing-Experiences_of_EU_Member_States_for_Policy_ Developments_in_China.pdf. Accessed on March 5, 2020.

Analyzing the waiting time of academic publications: a survival model

Un modello di sopravvivenza per i tempi di accettazione delle pubblicazioni accademiche

Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi, Francesco Porro

Abstract In this paper a survival model is used to perform an analysis of the waiting time to publication for academic articles. The model is a multilevel excess hazard model and it allows to include non-linear and non-proportional effects of the covariates. The analysis is performed by considering covariates at two levels: the first one is the article level, the second one is the journal level.

Abstract *In questo articolo viene utilizzato un modello di sopravvivenza per effettuare un'analisi del tempo di attesa per la pubblicazione di articoli accademici. Il modello utilizzato é un modello multilivello con excess hazard che permette di includere effetti non lineari e non proporzionali delle covariate. L'analisi é condotta considerando covariate a due differenti livelli: articolo e rivista.*

Key words: Peer review, Waiting times, Net survival, Excess hazard model

Francesca De Battisti
Dipartimento di Economia, Management e Metodi Quantitativi - Universitá degli Studi di Milano - via Conservatorio 7, 20122 MILANO, e-mail: francesca.debattisti@unimi.it

Giuseppe Gerardi
Dipartimento di Economia, Management e Metodi Quantitativi - Universitá degli Studi di Milano - via Conservatorio 7, 20122 MILANO, e-mail: giuseppe.gerardi@unimi.it

Giancarlo Manzi
Dipartimento di Economia, Management e Metodi Quantitativi - Universitá degli Studi di Milano - via Conservatorio 7, 20122 MILANO, e-mail: giancarlo.manzi@unimi.it

Francesco Porro
Dipartimento di Statistica e Metodi Quantitativi - Universitá degli Studi di Milano-Bicocca - piazza dell'Ateneo Nuovo, 1, 20126 MILANO e-mail: francesco.porro1@unimib.it

1 Introduction

The topic of waiting time in academic publication decisions is very relevant and interesting. The overall process of submission, especially for top-level journals, and any required revisions is such that many months, if not years, must pass between the submission and the acceptation of an article, if ever there will be one. De Battisti and Manzi [4] put forward some considerations on such issue, and suggested to apply multilevel models to find determinants affecting the waiting time until acceptance. The aim of this paper is to extend these considerations by carrying out a multilevel excess hazard model to analyze the waiting time to publication, working on the hierarchical data. We propose to consider the waiting time for academic publication as survival time, with article as units of interest, and model the effects of potential explanatory factors.

2 The methodology

Survival analysis typically considers the time until an event occurs. We usually refer to the time variable as *survival time*, because it measures the time that an individual has survived over a certain follow up period. We also usually refer to the event as a failure, because the event of interest usually is death, disease incidence, or any other negative individual experience. However, survival time may be for example the time to return to work after an elective surgical procedure, in which case failure is a positive event. As argued in [2], in the context of survival analysis, "the main survival indicator when comparing populations is net survival, that is, the hypothetical survival that patients would experience could they die only from their cancer ([3], [6])". In the general case, beyond the medical framework, the net survival refers to the occurrence of the event under study only because of specific causes. Moreover, the net survival is estimated comparing the all-cause hazard of death experienced by the patients to the general population from which the individuals come. One of the approaches proposed in literature to estimate net survival is modelling the excess hazard (see [2], [5], [7]). Starting from these proposals, in [2] a methodology to estimate an excess hazard regression model with non-linear and non-proportional effects due to the covariates and including a random effect is developed. The excess hazard approach is based on the assumption that the total hazard of the event occurrence, denoted by $\lambda(t,\mathbf{x},\mathbf{z})$, can be decomposed into the sum of a cause-specific hazard, denoted by $\lambda_+(t,\mathbf{x})$, and a hazard due to all the other causes, denoted by $\lambda_P(a+t,\mathbf{z})$ (the latter being usually estimated from general population life tables in the case of death). This means that:

$$\lambda(t,\mathbf{x},\mathbf{z}) = \lambda_+(t,\mathbf{x}) + \lambda_P(a+t,\mathbf{z}),$$

where $\mathbf{x}$ is a vector of prognostic variables, $\mathbf{z}$ is a vector of demographic characteristics, and a is the age at the diagnosis, so that $a+t$ denotes the age at death or at

censoring. The excess hazard is associated with the net survival through the classical relationship between hazard and survival function:

$$S(t) = \exp\left(-\int_0^t \lambda(v)dv\right).$$

Moreover, in [2] a multilevel excess hazard model is proposed. For each individual (or unit) j (with $j = 1,...,n_i$) from cluster (or group) i (with $i=1,...,D$), let t_{ij} denote the observed time-to-event and δ_{ij} be an indicator variable taking the value 1 in case of the event occurrence and 0 in case of censoring. Then the total hazard is:

$$\lambda(t, \mathbf{x}_{ij}, \mathbf{z}_{ij}, w_i) = \lambda_+(t, \mathbf{x}_{ij}) \exp(w_i) + \lambda_P(a+t, \mathbf{z}_{ij}), \quad (1)$$

where w_i denotes a random effect at the cluster level. This model allows to introduce multiple covariate effects. For example, by expressing $\mathbf{x}_{ij}$ by its components, i.e. $\mathbf{x}_{ij} = (x_{1,ij}\ x_{2,ij}\ x_{3,ij})$, the following formula:

$$\log(\lambda_+(t, \mathbf{x}_{ij})) = \log(\lambda_0(t)) + \beta \cdot x_{1,ij} + f(x_{2,ij}) + g(t) \cdot x_{3,ij}$$

represents the logarithm excess hazard, with a linear and proportional effect related to $x_{1,ij}$, a non-linear and proportional effect due to a continuous function f of $x_{2,ij}$, and a non-proportional (time-depending) effect due to $x_{3,ij}$.

3 The dataset

The dataset used in this application is formed by 3489 published papers between 2011 and 2016 in the following top-level statistical journals: *Journal of Statistical Software* (J Stat Softw), *Fuzzy Sets and Systems* (Fussy Set Syst), the *Journal of the Royal Statistical Society - Series A - Statistics in Society* (JRSSA), the *Journal of the American Statistical Association* (JASA), the *Annals of Probability* (Ann Prob), the *Journal of Business Economic Statistics* (JBES), *Advanced in Data Analysis and Classification* (ADAC), and *Biostatistics*. The distribution of the articles is described in Table 1. We apply survival analysis on the waiting time of academic

Table 1 Distribution of articles across journals

Journal	*Publisher*	*Country*	*Eds' country*	*Number of Articles*
JRSSA	Royal Stat.Soc./Wiley	UK	UK	173 (4.96%)
ADAC	Springer	GER	ITA, GER, JAP	129 (3.70%)
JBES	Am.Stat.Ass./ Taylor&Francis	USA	USA	258 (7.39%)
Ann Prob	Inst.Math.Stats./ Bernoulli Society	USA	USA	493 (14.13%)
Biostatistics	Oxford Uni. Press	UK	NED, USA	330 (9.46%)
Fuzzy Set Syst	Elsevier	NED	BEL, FRA, GER, SPA	979 (28.06%)
JASA	Am.Stat.Ass./ Taylor&Francis	USA	USA	741 (21.24%)
J Stat Softw	UCLA Dept.Stats	USA	AUT, SWI, GER	386 (11.06%)

publications, namely the time that elapses between the submission and the publication of an article in a journal. More in details, for each article (uniquely identified by the Document Object Identifier - DOI) the waiting time (variable *Age*, in days) is calculated as the difference between the date of the acceptance when this was available (otherwise the date of the online publication or the date of the final revision) and the date of the submission (always available). The covariates considered in this application and their meanings are:

- *Bayes*: dichotomous variable (1=Bayesian article, 0=otherwise);
- *Month Scopus cit*: the average monthly number of Scopus citations per article;
- *Avg h Index*: the average *h* index of the authors;
- *Junior less 5*: dichotomous variable (1=if the Scopus *h* index of one of the authors is lower than 5, 0=otherwise);
- *Senior more 20*: dichotomous variable (1=if the Scopus *h* index of one of the authors is greater than 20, 0=otherwise);
- *Number author*: number of the authors;
- *USA all*: dichotomous variable (1=if all the authors are affiliated to US institutes, 0=otherwise);
- *USA*: dichotomous variable (1=if at least one author is affiliated to an US institute, 0=otherwise);
- *Same country*: dichotomous variable (1=if all the authors are affiliated to institutes in the same country, 0=otherwise);
- *Same nationality eds*: dichotomous variable (1=if the institutions of the most important author and of the journal editor have the same nationality, 0=otherwise);
- *Age journal*: age of the journal since its first issue;
- *IF*: the 2017 Thomson Reuters impact factor;
- *AI*: the AI index (Article Influence index) that measures the average influence of an article over the first five years after publication.

It is worth remarking that the last three variables refer to the journal level, that is they are second-level variables, meaning that they assume the same value for all the articles published in the same journal. The application of the survival analysis to this particular context requires some adjustments. First, our data refer only to published paper, not to all the submitted papers. For this reason, the event of interest (publication) occurs after a reasonable period of time for all our observed units. This is different from the usual case, in which individuals can be dead or alive at the end of the study period. In order to have a situation similar to the classical one in survival analysis models, we have to censor the data: we have to choose a time interval (which corresponds to the follow up period) to evaluate whether the article is published or not. In this way, it is possible to model the article waiting time. All the articles with a waiting time to publication (variable *Age*) greater than 3 years are censored. For these articles the variable δ (the indicator variable mentioned in Section 2) is set equal to 0. We selected the value of 1095 days (3 years) in order to have a restrained percentage (less than 5%) of censored articles. In this way, the censored articles represent the 4.39% of the total number of articles. The average of the waiting time (with censored data) is 451.49 days, the median is equal to 397.

4 Results and discussion

We applied the mixed effect excess hazard model described in (2) to the aforementioned dataset. We consider the 3489 articles as units, clustering them by the corresponding journal. In this work, following [2], we focus on the excess hazard function $\lambda_+(t,\mathbf{x}_{ij})\exp(w_i)$ of formula (2), and we assume a normal distribution of the random effect w, with zero mean and standard deviation σ.
The aim of this analysis is to identify which variables have a relevant impact on the waiting time to publication. We ran more than one hundred models, by setting different kinds of effect for each covariate. The calculations have been performed by using the R package *mexhaz* (Mixed Effect Excess Hazard Models). Such package provides estimates by MLE method, throught the implementation of numerical methods. We considered the following types of effects: linear and proportional, linear and non-proportional (that is, time-depending), non-linear and proportional, and non-linear and non-proportional. As suggested in [2], the logarithm of the baseline excess hazard and the functions (of time) for the time-dependent effects are modelled by cubic B-splines with two knots at 365 and 1094 days, respectively. For explicative purposes, in the following we report three models with a hierarchical level of complexity:

- Model 1: the effects of all the covariates are linear and proportional;
- Model 2: the effects of the covariates *Avg h Index*, *Number authors*, *IF*, *Age journal* and *AI* are linear and non-proportional. The effect of all the other covariates are linear and proportional;
- Model 3: the effects of the covariates *Avg h Index*, *Number authors*, *IF Age journal* and *AI* are non-linear and non-proportional.The effect of all the other covariates are linear and proportional.

The fittings of the three models are compared by using the Akaike Information Criterion (AIC), (see [1] for details). Table 2 reports the parameter estimates (and their standard errors) for the covariates with linear and proportional effects. For each model also the AIC index is reported. The model with the best fitting is Model 2, and, for this model, the highest Hazard Ratio (HR) is the one related to the covariate *Same nationality eds* and it is given by

$$HR_{Same\ nationality\ eds} = e^{0.064} = 1.066.$$

This value shows that if the nationality of the main author of an article is the same of the journal editor, the hazard rate is, *ceteris paribus*, higher than otherwise: this implies that in this case the waiting time to publication is smaller. Similarly, all other conditions being equal, articles with all US authors present a smaller waiting time than the others, since the corresponding HR is equal to

$$HR_{USA\ all} = e^{0.03} = 1.03.$$

Conversely, the model shows that articles with at least one author affiliated to an US institution, experience a higher waiting time than those with no author affilied to an

Table 2 Parameter estimates and the corresponding AIC for the three models

Variable	Model 1 AIC=45237.52	Model 2 AIC=45002.07	Model 3 AIC=45266.02
Bayes	-0.010 (0.060)	0.002 (0.060)	-0.011 (0.060)
Month Scopus cit	0.005 (0.005)	0.007 (0.005)	0.008 (0.005)
Avg h Index	0.004 (0.002)	LIN-NPH	NLIN-NPH
Junior less 5	0.010 (0.041)	-0.008 (0.041)	0.024(0.044)
Senior more 20	0.012 (0.049)	0.019 (0.050)	0.013 (0.058)
Number authors	-0.042 (0.017)	LIN-NPH	NLIN-NPH
USA all	0.028 (0.076)	0.030 (0.076)	-0.005 (0.076)
USA	-0.067 (0.061)	-0.064 (0.061)	-0.037 (0.060)
Same country	-0.044 (0.051)	-0.045 (0.051)	-0.055 (0.051)
Same nationality eds	0.077 (0.046)	0.064 (0.046)	0.039 (0.046)
Age journal	0.091 (0.012)	LIN-NPH	NLIN-NPH
IF	-0.063 (0.022)	LIN-NPH	NLIN-NPH
AI	-0.277 (0.051)	LIN-NPH	NLIN-NPH
Standard deviation σ	1.698	1.723	2.758

US institution, everything else being equal, as the HR for the variable *USA* is

$$HR_{USA} = e^{-0.064} = 0.938,$$

therefore teams with all US members seem to be more successful than mixed ones.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974)
2. Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A.: A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. Statistics in Medicine **35**, 3066–3084 (2016)
3. Danieli, C., Remontet, L., Bossard, N., Roche, I., Belot, A.: Estimating net survival: the importance of allowing for informative censoring. Statistics in Medicine **31**(8), 775–786 (2012)
4. De Battisti, F., Manzi, G.: Smart Tools for Academic Submission Decisions: Waiting Times Modeling. in Smart Statistics for Smart Applications Book of Short Papers SIS 2019, [edited by] Arbia, G., Peluso, S., Pini, A., Rivellini, G., Ed. Pearson, June 2019, 787–792 (2019)
5. Estve, J., Benhamou E, Croasdale M, Raymond L.: Relative survival and estimation of net survival: elements for further discussion. Statistics in Medicine **9**(5), 526–538 (1990)
6. Perme, MP., Stare, J., Estve, J.: On estimation in relative survival. Biometrics **68**(1), 113–120 (2012)
7. Remontet, L., Bossard, N., Belot, A., Estve, J., and the French network of cancer registries FRANCIM: An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Statistics in Medicine **26**, 2214-2228 (2007)

Clustering of food choices in a large sample of students using university canteen

Profili di scelta rispetto alla composizione del pasto in un campione di studenti che utilizzano le mense universitarie

Valentina Lorenzoni, Isotta Triulzi, Irene Martinucci, Letizia Toncelli, Michela Natilli, Roberto Barale, Giuseppe Turchetti[1]

Abstract In view of the crucial role eating habits represent in academic life and in the development of obesity and chronic condition, both forthwith and in the future, the present study aim at detecting eating choices in a large sample of students accessing canteens serving the University of Pisa, Italy. Based on the frequency of food groups' choice evaluated from cashier transaction and a model-based clustering relying on the beta distribution, 5 different mixture components were identified and each assimilated to a specific behaviour towards food choice. The mixing proportion estimated suggested that the mixture resembling "healthy choice" comprises about 11.2% of the study population while component resembling students with limited choice of vegetables and fruits were the most represented.

Abstract *Dato il ruolo delle abitudini alimentari nella vita accademica e nello sviluppo dell'obesità e delle malattie croniche, il presente studio mira a raggruppare le scelte alimentari in un ampio campione di studenti che utilizzano le mense dell'Università di Pisa. Utilizzando la frequenza di scelta dei cibi rilevata automaticamente dalle transazioni e un metodo di clustering basato sulla distribuzione Beta, sono stati identificate cinque misture assimilabili a diversi "profili" di scelta. Tra queste quella che identifica scelte salutari rappresenta appena rappresenta appena l'11,2% della popolazione in studio, mentre maggiormente rappresentati sono i sottogruppi caratterizzati da una scelta poco frequente di verdure e frutta.*

Lorenzoni V, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy, valentina.lorenzoni@santannapisa.it
Triulzi I, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy, isotta.triulzi@santannapisa.it
Martinucci I, Division of Gastroenterology - Versilia Hospital, Lido di Camaiore, Lucca, Italy, martinucci.irene@gmail.com
Toncelli L, Azienda Regionale per il Diritto allo Studio Universitario, Area Ristorazione, Pisa, Italy, mtoncelli@dsu.toscana.it
Natilli M, Department of Computer Science, University of Pisa, Pisa Italy, michela.natilli@gmail.com
Barale R, Department of Biology, University of Pisa, Pisa, Italy, roberto.barale@unipi.it
Turchetti G, Institute of Management, Scuola Superiore Sant'Anna, Pisa, Italy, giuseppe.turchetti@santannapisa.it

Key words: eating habits, food choices, university students, mixture models

Introduction

During the university period students are exposed to radical changes and to high risk of adverse health behaviours, being less influenced by parents while tending to comply with the behaviour of their peers (Kandiah et al., 2006; Vella-Zarb and Elgar, 2010).

In particular, a branch of literature suggested unhealthy eating and lifestyle habits among university students also associated with increasing prevalence of overweight and obesity(Kapinos et al., 2014; Nelson et al., 2008). Unhealthy eating behaviours that developed during this period may persist into adulthood thus carrying out potential health consequences in the short and long-term; forthwith habits acquired may also interfere with academic achievement and student' retention.

In this context, universities (and their facilities) while playing a pivotal role in improving student engagement also represent a suitable context for the promotion of students' health, appropriate lifestyle with the opportunity to prevent disease and future consequence (Guagliardo et al., 2011; Lupi et al., 2015; Nelson et al., 2008).

Accordingly, this study aims at providing empirical evidence about food choices in a large sample of students using university canteens serving the University of Pisa (Italy) on the basis of data automatically collected from cashier transactions.

Study population and data

This work is part of part of the RASUPEA project: "Le mense Universitarie: ricerca sulle abitudini alimentari dei giovani e d'educazione e prevenzione alimentare". The project was supported by a grant of the PRAF 2012-2015 REGIONE TOSCANA program of the Tuscany Region and aimed at evaluating eating habits and their correlates with gastrointestinal disorders among students enrolled at the University of Pisa, Central Italy.

The population under study is composed of students enrolled in courses at Pisa University who access canteens serving the university of interest in the academic years from October 2015 to September 2016.

Analyses were performed using the database of the Azienda Regionale per il Diritto allo Studio Universitario (DSU) that contained records about all of consumed meals.

Analysis of eating choices was allowed by the fact that canteens serving the University are equipped with an automatic system that records detailed data about

each meal transaction: date and hours, number and type of dishes, student ID and price applied. All these data are collected into a dedicated Oracle database owned by the DSU that whose used in the present analysis to extract records about meal consumption.

On the basis of dish details, foods were classified according the main dish components mainly according to the coding system established by the Eurocode (Ireland and Møller, 2000) and adding fried foods, pies and omelettes as representing "fatty foods" and also pizza and sandwich because identifying a habit related to a quick meal. Accordingly, eleven different food groups were identified and used in the analyses: (1) grains, (2) miscellaneous and soups, (3) meat, (4) fish, mollusks, crustaceans and their products, (5) vegetables, (6) potatoes, (7) fruits, (8) pulses, (9) sugar products, chocolate products and confectionery, (10) fried foods, pies and omelettes, (12) sandwiches and pizza.

Statistical analyses

The frequency of different food-groups consumption was described in terms of the number of times the specific food group is selected over the total number of access.
Given the nature of those variables, constrained in the [0,1] interval and highly skewed, a model-based clustering relying on multivariate beta distribution (Dean and Nugent, 2013) was used to detect different food choices.

Model-based clustering (MC) is a particular approach of cluster analysis, MC is based on finite mixture models and allows for the automatic selection of the number of clusters (Fraley and Raftery, 2002; McLachlan and Peel, 2000).
According to the general notation used in the MC approach, a set of observations on j variables $X = \{x_1, x_2, \ldots, x_j\}$ is assumed to be made of independent and identically distributed samples from some unknown population density $f(x)$ where each group g in the population is represented by a mixture component $f_g(x)$. The population density could thus be assumed as a weighted mixture of these components

$$f(x) = \sum_{g=1}^{G} \pi_g \, f_g(x; \theta_g)$$

where π_g represents the probability of belonging to group g, so that $0 \leq \pi_g \leq 1$ and $\sum_{g=1}^{G} \pi_g = 1$.
Parameter of the mixture model were estimated using the expectation-maximization (EM) algorithm (Dean and Nugent, 2013). The number of G mixtures (i.e., clusters) better fitting data could then be chosen on the basis of the Bayesian Information Criterion (BIC) and modified Integrated Completed Likelihood alternative, BIC-ICL (Bertoletti et al., 2015; Dean and Nugent, 2013).
Each observation was assigned to the component (representing a different eating profile) on the basis of the highest posterior probability of membership.
All analyses were performed using R version 3.3.3.

Main results

The study population consisted of 4,643 students for whom more than 200 thousand meals consumed in between October 2015 and September 2016 were analysed.

In the study group mean age was 23.6±3.6 years and more than half of the students were male (n=2,786).

Overall, frequencies of food-group choices suggested that grains were the most frequently selected to compose the meal, meat and fried food were included in 1/3 of meals each, similarly vegetable and fruit that were chosen in less than 40% of accesses (Table 1).

Table 1: Frequency (as percentage) of food-group choice in the study group

	Median [25-75perc]
Grains	67.9 [43.8-87.5]
Miscellaneous and soups	0.8 [0-8.4]
Meat	40.7 [20.8-58.3]
Fish	3.1 [0-8.3]
Vegetables	35.7 [20.2-53.3]
Pulses	3.9 [0-9.1]
Potatoes	17.6 [10-25.4]
Fried foods, pies and omelette	36.4 [22.1-50]
Sandwich and pizza	11.5 [2.6-25]
Fruits	39.8 [16.5-66.7]
Sugar/chocolate products and confectionery	4.7 [0-12.5]

According to BIC and ICL-BIC values five mixture components better fit the data.
The first component (G1) grouped students with healthy eating habits characterized by a varied choice of all the food groups with a frequent selection of vegetables and fruits; the second (G2) and third component (G3) represented students limiting the selection vegetables and fruits in fewer than half of the meals while exceeding in the choice of proteins, fried foods/pies/omelets, sugars/chocolate products and confectionery (G2) or grains (G3); the fourth component (G4) resembles the quick eaters who frequently had meals of sandwiches or pizza. Finally, the fifth component (G5) identified students who usually had a light meal with grains or a main dish (Figure 1).

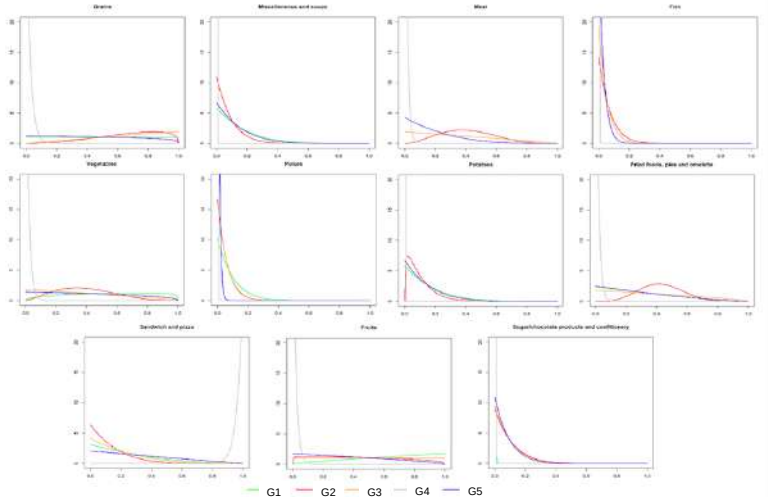


Figure 1: Normalized densities for the five components estimated

The mixing proportion estimated describing the distribution of the diverse eating profile in the study group suggested that G3and G2 were the most represented the study population (42.9% and 32.7% respectively). Healthy eaters were about 11.2% of the study group while he other components, G4 and G5, represented a small proportion of the study population (3.3% and 9.8% respectively).

Conclusions

Findings from the present study outline the relatively low widespread of healthy food choice among university students accessing canteen, while attitudes towards the composition of meal with fatty food is the prevalent behaviour.

Although findings from the present study are limited to students using canteens and could not be immediately extrapolated to the overall group of university students, results are in line with international available evidence on the general population of university students.

Indeed, results corroborate existing evidence and stress the need to implement policies on this “vulnerable” group. Moreover, given the peculiarity of university canteens in Italy that allow students access to meal at relatively limited price, thus guaranteeing access also to healthy food for all category also preventing food insecurity, for the specific context results suggest that allowing students to access healthy foods at affordable price is not sufficient to enable healthy choice rather, inducing changes in individual-making choice require more effort for the design an effective environment enabling consciousness of choices and healthiness.

References

1. Bertoletti, M., Friel, N., Rastelli, R., 2015. Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. Metron. https://doi.org/10.1007/s40300-015-0064-5
2. Dean, N., Nugent, R., 2013. Clustering student skill set profiles in a unit hypercube using mixtures of multivariate betas. Adv. Data Anal. Classif. 7, 339–357. https://doi.org/10.1007/s11634-013-0149-z
3. Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97, 611–631. https://doi.org/10.1198/016214502760047131
4. Guagliardo, V., Lions, C., Darmon, N., Verger, P., 2011. Eating at the university canteen. Associations with socioeconomic status and healthier self-reported eating habits in France. Appetite 56, 90–95. https://doi.org/10.1016/j.appet.2010.11.142
5. Ireland, J.D., Møller, A., 2000. Review of international food classification and description. J. Food Compos. Anal. https://doi.org/10.1006/jfca.2000.0921
6. Kandiah, J., Yake, M., Jones, J., Meyer, M., 2006. Stress influences appetite and comfort food preferences in college women. Nutr. Res. 26, 118–123. https://doi.org/10.1016/j.nutres.2005.11.010
7. Kapinos, K.A., Yakusheva, O., Eisenberg, D., 2014. Obesogenic environmental influences on young adults: Evidence from college dormitory assignments. Econ. Hum. Biol. 12, 98–109. https://doi.org/10.1016/j.ehb.2013.05.003
8. Lupi, S., Bagordo, F., Stefanati, A., Grassi, T., Piccinni, L., Bergamini, M., De Donno, A., 2015. Assessment of lifestyle and eating habits among undergraduate students in Northern Italy. Ann. Ist. Super. Sanita 51, 154–161. https://doi.org/10.4415/ANN-15-02-14
9. McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley Ser. Probab. Stat. New York Wiley. https://doi.org/10.1198/tech.2002.s651
10. Nelson, M.C., Story, M., Larson, N.I., Neumark-Sztainer, D., Lytle, L.A., 2008. Emerging adulthood and college-aged youth: An overlooked age for weight-related behavior change. Obesity. https://doi.org/10.1038/oby.2008.365
11. Vella-Zarb, R.A., Elgar, F.J., 2010. Predicting the freshman 15: Environmental and psychological predictors of weight gain in first-year university students. Health Educ. J. https://doi.org/10.1177/0017896910369416

Cruise passengers' expenditure at destinations: Review of survey techniques and data collection

La spesa dei crocieristi nelle destinazioni: Revisione delle tecniche di indagine e raccolta dati

[1]Caterina Sciortino, [2]Stefano De Cantis, [3]Mauro Ferrante, [4]Szilvia Gyimóthy

Abstract Spending by cruise passengers constitutes an important contribution to the economy of destinations. The issues related to the analysis of the expenditure have been widely debated for many years both in the scientific literature and in the common political debate. Despite 20 years of empirical research on this topic, not consolidated quantification methods exist to inform the debate. This work offers a brief review of the principal research on cruise tourist spending. It analyses the main characteristics of the surveys, reviews the different methods and techniques employed as well as assesses the main results. The article concludes with a discussion of the studies conducted and identifies future research directions.
Abstract *La spesa dei crocieristi costituisce un importante contributo all'economia delle destinazioni. Le questioni relative all'analisi della spesa sono state ampiamente dibattute per molti anni sia nella letteratura scientifica che nel dibattito politico comune. Nonostante 20 anni di ricerca empirica su questo argomento, non esistono metodi di quantificazione consolidati per informare il dibattito. Questo lavoro offre una breve rassegna delle principali ricerche sulla spesa dei crocieristi. Analizzerà le principali caratteristiche delle indagini, esaminerà i diversi metodi e tecniche impiegati e valuterà i principali risultati. L'articolo si conclude con una discussione degli studi condotti e identifica le direzioni di ricerca future.*

Key words: Cruise passengers' expenditure; Data collection methods of expenditure; Survey techniques; Sampling scheme

[1] Caterina Sciortino, Department of Economics Business and Statistics, University of Palermo; caterina.sciortino@unipa.it;
[2] Stefano De Cantis , Department of Economics Business and Statistics, University of Palermo; stefano.decantis@unipa.it
[3] Mauro Ferrante, Department of Culture and Society, University of Palermo, mauro.ferrante@unipa.it
[4] Szilvia Gyimóthy, Department of Marketing, Copenhagen Business School, sgy.marktg@cbs.dk

1 Introduction

The cruise industry experienced important development in recent years, both in terms of expansion and diversification of the sector and of market volume. Over the last twenty-years, cruise tourism represents one of the fastest growing sectors in tourist field [12].

This work contributes to understanding how cruise passengers' expenses were analysed, from a methodological point of view. This is done by mean of a revision of the survey techniques for collecting information on cruise passengers' expenditure at their destination, implemented in various studies that emerged from the literature on this topic. In particular, this paper aims at critically assessing and improving the tools of data collection on cruise passengers spending and addresses one of the most relevant research questions in this area of study: *How is cruise passengers´ expenditure defined, operationalised and measured in tourism research literature?* In order to consider the most relevant articles dealing with an analysis of the spending behaviour of cruise passengers, the following selection criteria have been applied: studies that deal exclusively with cruise tourism and its economic impact; studies which primary aim was to analyse spending behaviour through interviews with cruise passengers; studies that used questionnaire as empirical data collection method. A better knowledge of the way in which cruise passengers' expenditure can be surveyed in practice represents a crucial step in order to determine the economic impact of the cruise sector in terms of costs and benefits [1].

2 Sampling scheme and Questionnaires

A review of relevant literature from the past two decades (1998-2018) were performed, mainly based on the assessment of three macro-themes (Table 1). The first theme addresses questions related with the definition and operationalization of tourist expenditure; the second one analyses issues in data collection procedures. The third theme focuses on the analysis of questionnaires and related survey techniques.

Table 1: Articles included in the review, by authors, year of publication and journal.

#	Authors	**Year**	**Journal**
1	Henthorne T.L.	2000	J. of Travel Research
2	Marušić Z. et al.	2009	Tourism Mar. Environ.
3	Brida J.G. & Risso W.A.	2010	Tourism Analysis
4	Brida J.G. et al.	2012	Tourism Economics
5	Larsen S. et al.	2013	Tour. Manag. Perspect.
6	Parola F. et al.	2014	Res. Transp. Bus. Manag.
7	Brida J.G. et al.	2015	Current Issues in Tourism
8	Marksel M. et al.	2016	Tourism Economics
9	Gargano R. & Grasso F.	2016	J. of International Studies
10	De Cantis S. et al.	2016	Tourism Management
11	Domènech A. et al.	2019	Tour. Plan. Dev.
12	Pino J.F.B. & Tovar B.	2019	Tourism Economics

2.1 Defining cruise tourist expenditure

Cruise passengers differ from tourists. Due to the time spent at the destination (from 2-4 to 8-10 hours) they could be considered as same-day travellers. During the stay period, their expenditure is characterized by a high degree of variability which depends, among other things, from the differences between two main categories of cruise passengers: Guided visitors, i.e. those who buy (almost always on board) a tour package (which usually includes entrance to museums and attractions, tour guide services, transportation costs, etc.) and the independent visitors: i.e. those who decide to visit the destination without buying any package, generally with no pre-defined itinerary. According to the UNWTO “visitor expenditure” is defined “as the total consumption expenditure made by a visitor or on behalf of a visitor for and during his/her trip and stay at destination” [13, p.85]. Similarly, we can define the expenditure of an independent cruise tourist at the destination as the total consumption expenditure made by the visitor or on behalf of the visitor during his/her visit at the destination, during the time spent for the visit. We can identify category-specific expenditure as "the total amount spent for one specific category of consumption good": food and beverage, transportation, shopping and souvenirs, etc. It is clear that the concept of cruise tourist consumption expenditure includes a wide variety of items, from the purchase of consumption goods and services related with their visit at the destination to the purchase of small durable goods for personal use and of souvenirs and gifts for family and friends. However, it is not easy to operationalize this definition, due to several issues. One of these issues is related to the size of the group size to which the total expenditure is referred to. Type and amount of expenditure greatly varies according to group composition, also by considering that some types of expenditure may not comprise all the people of the group, and due to the presence of scale economies in the degree of expenditure, which makes an estimate of “per-capita” expenditure difficult to determine.

2.2 Data collection procedure

Data collection is a crucial stage in empirical research. In the context under analysis temporal and spatial dimension must be clearly defined, especially in the implementation of sampling procedure, as a fundamental stage of the research design. The selection of units to be included in the sample is one of the most delicate procedures under the methodological perspective. In the considered research, the sampling procedures used are manifold, for instance, some have endeavored to use rotation samplings [7], others general random sampling [9], while some others opted for a stratified procedure (Table 2). Another aspect to consider is the interview time: it is possible to divide the research into two macro areas: "Cross-sectional analysis" and "Time series analysis". The first can be considered as a "snapshot" of a certain phenomenon in which the time dimension is not considered. Studies of this type have a clear advantage of greater simplicity and prompt availability of results compared to longitudinal analysis. Other problems in data collection includes: the time span, the choice of the unit under analysis, which represents critical elements.

A second relevant issue is related to the spatial dimension. In the majority of studies considered, the places considered are the ports, the cruise ships and the cities of destination. In the majority of cases the units are interviewed at the end of the visit to the city and just before they return on board in the ship [2,9,12].

Table 2: Sampling scheme, sample size, and target population.

#	*N*	Sampling scheme	Crew Members	Tourists
1	1500	Random pattern	No	No
2	1592	Stratified random	Yes	No
3	1121	Not specified	No	No
4	1361	Convenience random	No	No
5	8371	Not specified	No	Yes
6	127	Accidental sampling	No	No
7	3348	Stratified sampling	Yes	No
8	357	Random sampling	Yes	No
9	5500	Stratified sampling	Yes	No
10	278	Stratified random with selection criteria	No	No
11	161	Random sampling	No	No
12	12578	Two-step stratified	No	No

In table 3, information on study location and on survey period for the papers considered in the review are reported. The most significant problems in data collection are caused by the fact that some cruise passengers remain on board the ship, and this may causes problems in terms of sampling, since the size of the population became unknown. Moreover, some others may decide to come back to the ship for a break, and performing a second visit afterwards, causing issues in the implementation of probabilistic sampling scheme.

Table 3: Study location and survey period.

#	**Study Location**	**Survey period**
1	Ocho Rios, Jamaica	Five years (1993-1997)
2	Croatia	Four months (Jun-Sept 2006)
3	Costa Rica	One-Year (2008)
4	Cartagena de Indias, Colombia	Two Months (Oct-Nov 2009)
5	Bergen, Norway	Summers (2010-2012)
6	West Mediterrean	Spring season (2013)
7	Uruguay	2009-2010 season
8	Port of Koper, Slovenia	One month (Sept 2013)
9	Port of Messina, Italy	Eight-months (March-Oct 2014)
10	Port of Palermo, Italy	April (2014)
11	Tarragona, Catalonia	Three months (Aug-Oct 2017)
12	Canary Islands	Six cruise seasons (2001-2015)

2.3 The questionnaire

The most common survey tool for collecting data on cruise passengers' spending is via questionnaire. The use of questionnaires to evaluate the spending behaviour of

cruise passengers has been a step forward compared to the self-compiled diary and over time, survey techniques and questionnaire building have improved. In the literature examined, the questionnaires are generally divided into sections and each section is dedicated to the collection of information on factors that are likely to affect the spending level at destination, such as psychological, socio-demographic or contextual factors. For example, some surveys include activities carried out and post-consumption items, such as satisfaction [11], others include motivation items at the beginning of the visit [1], while some scholars consider the number of visits to other destination during the same trip, or the number of cruise trips made in the past [2]. Spending is typically measured across standard expenditure categories (food and drinks, tours, souvenirs, transport and others). Only few and more recent publications consider the spatiotemporal characteristics of the visit, such as average time spent in port, the distance travelled or list of places visited [4,5]

As regard the spatio-temporal choice, most of the early studies collected data at one point of time, while others [4,5] divides the data collection into an opening and a closing stage, allowing the interview to take place at separate moments in time (disembarkation vs. embarkation).

3 Main results and conclusion

The empirical results of the various studies conducted offer a certain degree of homogeneity for some factors, uneven for others. Research has shown that as time elapses, money spent in port is increased [7] and there is also a high probability of returning to the destination [1]. Propensity to return is also positively correlated with high satisfaction levels [1], and higher average age of visitors. Repeat visitation affected the overall spending of cruise passengers positively: it was observed that those who had already visited the place, spent significantly more than the so-called “first-time visitors” [6]. An important component that has been deepened by various researchers is the word-of-mouth [10]. The empirical evidence has shown that the opinion of others, the opinion that people have and perceive from the destination, is an important factor. As far as gender is concerned, there is no empirical evidence on significant differences between the two sexes, although some studies have shown that female cruise passengers spend more on average than men [9]. Age differences does not have a great impact in the categories of expenditure [9], instead it presents evidence in terms of total expenditure [5]. Spending by cruise passengers is also associated with the number of visitors: if they are part of a large group, they are likely to spend more money [3]. The size of the group is in fact an important factor, which many consider when it comes to cruise spending, because it has a significant impact on total expenditure. Another socio-demographic aspect considered is nationality, although the results seems to be conflicting. Many authors have shown that the nationality of the cruise passengers does not make significant differences in expenditure [1], others have shown that there are significant differences between expenditure levels among different nationalities [6,9]. Another highly influential factor in cruise passengers' spending is the duration of the visit as well as the distance travelled onshore: those who stay for short periods and/or near the port area

spend less at the destination and spend more money on board [8]. This issue has been deeply investigated, especially through comparison with tourists in general [3], by showing that cruise passengers tend to spend less than tourists.

Research related to the spending of cruise passengers has undergone an extraordinary evolution, accompanied by the development of the cruise sector. Each research, having different populations and samplings, different places and methods, offered a broad investigation of the topic. This work aimed at collating various studies conducted on this topic, placing particular attention to the implementation of survey techniques and of the underlining research hypotheses. The evaluation of the extent to which cruise passengers' spending can be measured and analysed can contribute to the development of policies for cruise tourism management.

References

1. Brida, J. G., & Risso, W. A. (2010). Cruise passengers expenditure analysis and probability of repeat visits to Costa Rica: A cross section data analysis. *Tourism Analysis*, *15*(4), 425-434.
2. Brida, J. G., Bukstein, D., Garrido, N., & Tealde, E. (2012). Cruise passengers' expenditure in the Caribbean port of call of Cartagena de Indias: A cross-section data analysis. *Tourism Economics*, *18*(2), 431-447.
3. Brida, J. G., Bukstein, D., & Tealde, E. (2015). Exploring cruise ship passenger spending patterns in two Uruguayan ports of call. *Current Issues in Tourism*, *18*(7), 684-700.
4. De Cantis, S., Ferrante, M., Kahani, A., & Shoval, N. (2016) Cruise passengers' behavior at the destination: Investigation using GPS technology. *Tourism Management, 52*, 133-150.
5. Domènech, A., Gutiérrez, A., Anton Clavé, S. (2019) Cruise Passengers' Spatial Behaviour and Expenditure Levels at Destination Tourism Planning and Development.
6. Gargano, R., Grasso, F. (2016) Cruise passengers' expenditure in the Messina port: A mixture regression approach Journal of International Studies, 9 (2), pp. 158-169.
7. Henthorne, T.L. (2000) An analysis of expenditures by cruise ship passengers in Jamaica Journal of Travel Research, 38 (3), pp. 246-250.
8. Larsen, S., Wolff, K., Marnburg, E., Øgaard, T. (2013) Belly full, Purse closed. Cruise line passengers' expenditures Tourism Management Perspectives, 6, pp. 142-148.
9. Marksel, M., Tominc, P. and Božičnik, S., (2017). Cruise passengers' expenditures: The case of port of Koper. *Tourism Economics*, *23*(4), pp.890-897.
10. Marušić Z. Horak, S., Tomljenović, R. (2008) The socioeconomic impacts of cruise tourism: A case study of croatian destinations Tourism in Marine Environments, 5 (2-3), pp. 131-144.
11. Parola, F., Satta, G., Penco, L., & Persico, L. (2014) Destination satisfaction and cruiser behaviour: The moderating effect of excursion package *Research in Transportation Business & Management*,*13*, 53-64.
12. Pino, J. F. B., & Tovar, B. (2019). Explaining cruisers' shore expenditure through a latent class tobit model: Evidence from the Canary Islands. *Tourism Economics*, *25*(7), 1105-1133.
13. United Nations and World Tourism Organization. (2014) Recommendations on Tourism Statistics. United Nations publication, Sales No. E.94.XVII.6, New York

Educational integration of foreign citizen children in Italy: a synthetic indicator

Un indicatore sintetico dell'integrazione scolastica degli studenti con cittadinanza straniera

Alessio Buonomo, Stefania Capecchi and Rosaria Simone

Abstract Social inclusion and integration are prominent topics in social sciences, and specifically in the educational field, in order to contrast the experience of inequality conditions at school. Using data from the 2015 national survey on "Integration of the second generation", the paper proposes a synthetic indicator of educational integration of second-generation children in Italy. The chosen methodology is suitable to model subjective assessments on rating scales while accounting for both agreement and heterogeneity in response patterns.

Abstract *L'inclusione sociale e l'integrazione rappresentano temi cruciali nelle scienze sociali ed assumono rilievo specifico nell'ambito dell'istruzione. Utilizzando i dati dell'Indagine Istat sull'integrazione delle seconde generazioni (ISG 2015), lo studio propone un indicatore sintetico di integrazione scolastica che considera sia la percezione del tratto in esame che l'eterogeneitá delle risposte.*

Key words: Model-based indicators, Educational integration, Second-generation, Foreign children, Rating data

1 Introduction

The issue of migrants' integration is a key topic in social sciences, and it is particularly relevant in the educational field to counteract the experience of inequality conditions by children with a migratory background, especially of those born abroad. Focusing on Italy, the number of foreigners has rapidly increased in the last two

Alessio Buonomo
University of Naples Federico II, e-mail: alessio.buonomo@unina.it

Stefania Capecchi
University of Naples Federico II, e-mail: stefania.capecchi@unina.it

Rosaria Simone
University of Naples Federico II, e-mail: rosaria.simone@unina.it

decades from about 350,000 residents in the early nineties to over 5 million, according to the Population Registers. The consequent growing number of their children have attracted the interest of scholars [10].

Due to the fact that the increase of the second generations in Italy is a relatively recent phenomenon, the school environment is a particularly important field of study [3]. Educational and social integration can boost placement into community and labour market [1], especially in the case of immigrants, and thus be beneficial to the whole society. Indeed, studies that focus on immigrant-specific educational inequalities have proven that social and scholastic integration are fundamental in determining educational success [8, 4].

Considering the CI-CUB procedure proposed by Capecchi and Simone [2], the paper aims at introducing an indicator of integration at school for both Italian and second-generation foreign children, distinguishing by nationality. The study provides a synthetic measure able to summarize subjective evaluations expressed by a large number of respondents, relying on CUB models framework [9].

After a brief introduction to the available data, the implemented methodology is illustrated, some results are discussed and further developments of the research are outlined.

2 Data and methods

The "Integration of the second generation" survey (ISG) has been conducted in Italy by the Italian National Statistical Institute [5] in 2015 and offers information on students of secondary schools in Italy. The overall unweighted sample amounted to $68,127$ individuals, of which 52% were students in upper secondary school. In order to deal with a more homogeneous subset, our analyses focus on opinions of students attending higher secondary state schools. Thus, the final target sample consists of $35,427$ unweighted individuals.

We investigate students' assessment - in terms of agreement - to 20 items referred to their perceived conditions/habits gathered in four different batteries: (A) relationships with classmates; (B) homework and study habits; (C) relationships with teachers; (D) family attitudes towards school importance. Variables of interest are comprised in question 14, in section *C* of the questionnaire. Each battery is composed by 5 items, whose responses are expressed on a 5 point Likert scale (1="strongly agree", 5="strongly disagree"). For interpretative purposes, the variables were reversed (with the exception of items A_3, B_3, B_5 and D_4) and organized in order to have the same direction, from the lowest to the highest agreement.

In the target sample, male students are about 50%; 15% of respondents are under the age of 15, and 58% are aged 15-17, whereas 27% are over 17. Italian interviewees represent 53% of the sample. Among foreigners, the three more frequent nationalities are Romanian, Albanian and Moroccan (10%, 7% and 4%, respectively). Foreign children born in Italy (second generation) are 9% of the sample, whereas those born abroad are 38%. Since foreign students represent a heterogeneous group

due to different origins and characteristics, a synthetic measure of their perception of integration may be useful both for research and policy purposes.

In order to illustrate multifaceted and sometimes elusive constructs, as in the case of subjective evaluations and perceptions, the number of composite indicators is growing. In fact, to general readership it may be easier to comprehend a single synthetic measure rather than inspect results from different separate indicators [7].

When dealing with ordinal data, CUB models allow to consider the uncertainty component in response behaviour, usually neglected in standard approaches. In this respect, Capecchi and Simone [2] recently proposed the novel CI-CUB - Composite Indicator CUB - to compute a model-based indicator within the CUB framework [9].

In its simplest version, a CUB model (acronym of Combination of Uniform and shifted Binomial) is implemented on ratings R to single questions to provide a global measure of individual assessment. CUB models are specified via a two component mixture of discrete distributions, one designated to measure the *feeling* towards the item (Binomial), and the other accounting for the inherent *uncertainty* (Uniform).

Then, a CUB model assumes that, for a given number of categories m, the random variable R is specified by the probability mass distribution:

$$Pr(R = r \mid \boldsymbol{\theta}) = \pi \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1} + (1-\pi)\frac{1}{m},\ r = 1,2,\ldots,m.$$

The parameter $(1-\xi)$ of the Binomial and the weight of the uncertainty in the mixture $(1-\pi)$ provide a description of both perception towards the item (which can be indicated as "agreement" with given statements on perceived integration in the present case study), and heterogeneity of the response distribution. Since the two featuring parameters are normalized measures ranging in (0;1], different items and groups can be displayed and compared at a glance in the unit square. Therefore, CUB models establish a different paradigm, providing a unifying tool to interpret and compare the responses given to a number of items, even if measured on scales with different lengths, in a parsimonious way. The possibility of assessing the veracity of the rating evaluations with respect to the analyzed items/dimensions by means of easy-to-interpret graphical outputs is indeed one of the strengths of CUB framework. A comprehensive R library is devoted to CUB model fitting on the official CRAN repository.

In this setting, CI-CUB indicators provide a synthesis of CUB models applied to K items by n respondents: the random variable R_k associated to the *k-th* item follows $R_k \sim \text{CUB}\ (\hat{\pi}_k, \hat{\xi}_k)$, for $k = 1,\ldots,K$. Then, the CI-CUB indicator is a weighted CUB model $\tilde{R} \sim \text{CUB}\ (\tilde{\pi}, \tilde{\xi})$ where

$$\tilde{\pi} = \sum_{k=1}^{K} w_k \hat{\pi}_k, \qquad \tilde{\xi} = \sum_{k=1}^{K} w_k \hat{\xi}_k, \tag{1}$$

with $(\hat{\pi}_k, \hat{\xi}_k)$, for $k = 1,\ldots,K$ being the estimated parameters for the K variables of interest.

A CI-CUB can be seen as a two-dimensional composite indicator for the latent trait under investigation such as, in this case study, the perceived educational integration as measured through the 4 batteries of items (A, B, C, D).

The choice of the weights' system w_k is usually driven by specific circumstances and available information. In the present study, due to the underlying structure of the data along the 4 dimensions, we select the weights expressed by the (normalized) squared factor loadings of the 4 categorical principal components computed on the selected variables.

3 Results

While additive data-driven approaches may lead to neglect an inherent uncertainty component when present, CUB models separately estimated for each of the 20 items of question 14 (here not reported for space constraints) show a medium-high level of agreement, with the uncertainty being substantially not negligible for most of the items. More specifically, item C_1, referred to the perception of an equal treatment received by teachers, shows the highest level of uncertainty with a substantive agreement; on the other hand, item D_4 gets the highest level of agreement with a low uncertainty, meaning that family members do consider studying to be a prominent aspect of life success.

Analysing the whole dataset, the advantages of the implemented CI-CUB indicators are highlighted. In order to provide synthetic probabilistic measures of students' perceived integration, CI-CUB procedure has been implemented according to a twofold trajectory: results are presented graphically in Figure 1 where responses are aggregated by item subgroup and nationality. Notice that the parameter space is narrowed to improve readability, since a large part of the indicators lays in the uncertainty range $0.2 - 0.45$, and the agreement is always higher than 0.5. To save space, details about fitting procedure are not reported.

Left panel of Figure 1 shows the synthetic indicators for each item subgroup, corresponding to the above mentioned batteries (A, B, C, D), as compared with the overall CI-CUB indicator (ALL). For the items altogether (ALL), the level of agreement is quite high with a moderate level of uncertainty.

It seems evident that items referred to family attitudes (D) are those for which the overall level of agreement is the highest, immediately followed by those related to schoolmates (A), both manifesting a low level of uncertainty; estimates for the two subgroups pertaining to study habits (B) and assessments of relationships with teachers (C) get a lower agreement, with a larger uncertainty. Moreover, this representation allows to consider the impact on the overall indicator (ALL) of deleting one subgroup of items at a time. Specifically, removing B and C subgroups, the resulting synthetic indicators ($-$B and $-$C) show higher agreement and lower uncertainty, therefore highlighting the contribution of these two subsets in terms of *feeling*. These results are expected given the mutual positions of the subsets in the parameter space.

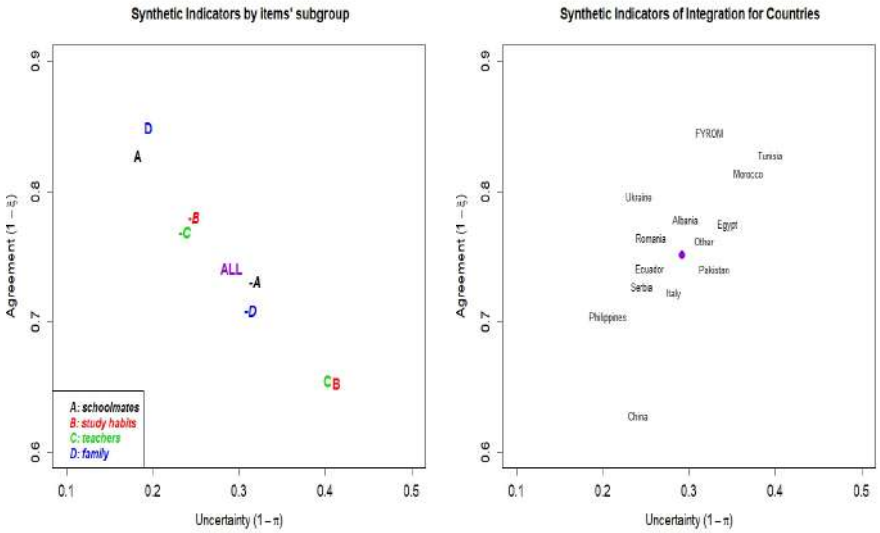


Fig. 1 Synthetic Indicators by item subgroups (left) and nationalities (right)

Right panel of Figure 1 collects CI-CUB indicators by respondents' nationality. In general, the agreement level is higher than 0.6 and the uncertainty spans between 0.10 and 0.43. Specifically, respondents from FYROM and Tunisia manifest the highest levels of agreement. Chinese respondents are quite far from the others in terms of *feeling*. Interestingly, Italian students express one of the lowest level of agreement, followed by respondents from China and Philippines. This finding is consistent with the so-called "Immigrant optimism approach" that assumes that immigrants are a positively selected group: individuals that leave the country of origin are supposed to be more motivated, ambitious, stronger and optimistic as compared to the national counterpart [6].

Table 1 CI-CUB estimated probabilities of the highest agreement

Nationalities	$Pr(\tilde{R}=5)$	Nationalities	$Pr(\tilde{R}=5)$
China	0.165	Egypt	0.308
Philippines	0.237	Romania	0.318
Italy	0.252	Albania	0.320
Serbia	0.260	Morocco	0.356
Pakistan	0.269	Ukraine	0.356
Ecuador	0.272	Tunisia	0.368
Other	0.295	FYROM	0.415

Table 1 presents the estimated probabilities to express the highest rating ($\tilde{R} = 5$), conditional to the nationality, which can be considered as an overall measure of educational integration, as derived by the CI-CUB models. In line with evidence in Figure 1 (right panel), Chinese students have the lowest estimated probability, whereas Macedonians (FYROM) have the highest. It turns out that there is not a clear geographical pattern of the estimated probabilities in terms of agreement, since different and distant countries present very similar estimated probabilities, see: Morocco, Ukraine and Tunisia, or Serbia, Pakistan and Ecuador.

Further developments of the study will follow two directions. From a methodological point of view, the main effort will be devoted to derive the confidence ellipses. For the empirical analyses, more investigation is needed on a geographical basis, as well as on the effects of individual characteristics, such as gender and migratory generation.

Acknowledgements The paper is included in the Research Program on "School inclusion strategies and social cohesion challenges of immigrate immediate descendants in Italy", SCHOOL/GEN2 (corresponding proponent Giuseppe Gabrielli), supported by grants from University of Naples Federico II (2017-2018), D.R. n. 408 07/02/2017 (CUP: E66J17000330001).

References

1. Ballantine, J. H., Hammack, F. M.: The Sociology of Education: A Systematic Analysis (7th ed.). Pearson, New York (2012)
2. Capecchi, S. Simone, R.: A Proposal for a Model-Based Composite Indicator: Experience on Perceived Discrimination in Europe. Soc. Indic. Res. **141**, 95–110 (2019)
3. De Santis, G., Pirani, E., Porcu, M. (eds.): Rapporto sulla popolazione. L'istruzione in Italia. Il Mulino, Bologna (2019)
4. Hadjar, A., Scharf, J.: The value of education among immigrants and non-immigrants and how this translates into educational aspirations: a comparison of four European countries. Journal of Ethnic and Migration Studies **45**, 711–734 (2019)
5. Istat, L'indagine sull'integrazione delle seconde generazioni: obiettivi, metodologia e organizzazione. Roma (2017)
6. Kao, G.,Tienda, M.: Educational aspiration of minority youth. American Journal of Education **106**, 349–384 (1998)
7. OECD, Handbook on Constructing Composite Indicators, Methodology and User Guide. OECD, Paris (2008)
8. Phillips, D.: Minority Ethnic Segregation, Integration and Citizenship: A European Perspective. Journal of Ethnic and Migration Studies **36**, 209–225 (2010)
9. Piccolo, D., Simone, R.: The class of CUB models: statistical foundations, inferential issues and empirical evidence. Stat. Methods Appl.**28**, 389–435 (2019)
10. Strozza, S., De Santis, G., (eds.): Rapporto sulla popolazione. Le molte facce della presenza straniera in Italia. Il Mulino, Bologna (2017)

Estimating the Change in Housework Time of the Italian Woman after the Retirement of the Male Partner: An Approach Based on a Two-Regime Model Estimated by ML

Come Stimare la Variazione del Tempo Dedicato al Lavoro Domestico dalla Donna in Italia, dopo il Pensionamento del Partner. Un Approccio Basato sulla Stima di Massima Verosimiglianza di un Modello a Due Regimi.

Giorgio Calzolari, Maria Gabriella Campolo, Antonino Di Pino and Laura Magazzini

Abstract Using Italian data from the Time Use Survey (Istat), in this study we analyse the influence of the retirement of the male partner on the housework time of the woman in Italy. As a novelty, our ML estimation of the across-regime correlation parameter, obtained applying a Two-Regime Switching model, allows us to better interpret the relationship between retirement of man and the woman's commitment in housework.

Abstract *Utilizzando i dati italiani sull'Uso del Tempo (Istat), abbiamo analizzato l'influenza del pensionamento del partner maschile sul tempo dedicato alle faccende domestiche dalla donna in Italia. Come novità, la nostra stima ML di un parametro di correlazione tra regimi, ottenuta specificando un modello a switching endogeno, consente di interpretare meglio la relazione tra pensionamento dell'uomo e l'impegno della donna nel lavoro domestico.*

Key words: effects of retirement, bargaining process, across-regime correlation

[1] Giorgio Calzolari, Università di Firenze; email: giorgio.calzolari@unifi.it

Maria Gabriella Campolo, Università di Messina; email: mgcampolo@unime.it

Antonino Di Pino, Università di Messina; email: dipino@unime.it

Laura Magazzini, Università di Verona; email: laura.magazzini@univr.it

1 Introduction

The aim of this study is to provide a convincing specification and estimation of the relationship between a man's retirement and the female partner's commitment in housework time.

As several previous studies found, commitment of the woman in housework generally decreases if the male partner is retired [1,3,5,6, for Italy; 9, for France). However, if a woman with higher bargaining power wants a more equitable distribution of housework time with her partner, this can induce the male partner to retire earlier to have more time to devote to work in the home and family. The opposite occurs if the woman has lower bargaining power.

Neglecting these endogenous aspects of the decision to retire may lead to an overstatement of the effect of a man's retirement on the time devoted to housework by a woman with a higher bargaining power and, conversely, an understatement of the effect of a man's retirement on the housework time of a woman with lower bargaining power.

Battistin et al. (2009) and Ciani (2016) faced the problem of influence of the endogenous choice of the agents between one of the two "regimes" (retirement or not) by adopting a "regression-discontinuity" approach. Campolo and Di Pino (2020) provided an identification strategy of the endogenous bargaining performing a Generalized Structural Equation Model (GSEM) as predictor of woman's bargaining power. In this study, we identify the endogenous latent effects on retirement in the across-regime correlation parameter, obtained by applying the Maximum Likelihood estimation procedure to an endogenous switching model, recently introduced by Calzolari and Di Pino (2017). Following this method, we specify two regression equations, each one referred to one of the two regimes. In the regime 1 we estimate the domestic work of the women whose male partner is retired, in the regime 2 we estimate the domestic work of the women whose partner is not retired. The across-regime covariance (or correlation) between the errors of the two outcome equations is here estimated simultaneously with the coefficients and variances of both regression equations. We use this parameter to identify a possible endogenous latent effect of housework on the retirement of the male partner. In particular, we assume that the unexplained components of the outcomes in both regimes may include latent factors related to the choice of the man to retire. A positive sign of the across-regime correlation indicates if the agent (the woman, in our case) gains an "absolute" (dis)advantage in both regimes [7]. Conversely, a negative sign of this parameter means that endogenous self-selection leads the agents to prefer the regime in which they have a comparative advantage in terms of their skills (in our case, we assume that the woman has more bargaining power if the male partner is retired).

However, the across-regime correlation parameter is generally not empirically identifiable as a consequence of the selection rule adopted specifying a Two-Regime switching model; namely the dependent variable referred to an observation cannot be jointly observed in both regimes. Calzolari and Di Pino (2017) solved the identification difficulty by modeling a selection rule based on the difference between

the outcome gained by the agent in the chosen regime (outcome is observed) and the outcome potentially expected in the alternative regime (outcome is latent). More details on the methodology are reported below in Sect. 2.

Following the ML procedure of Calzolari and Di Pino (2017), we use Italian data from the Time Use Survey (Istat) to estimate domestic work of Italian women in both regimes (partner is or is not retired). The estimated across-regime correlation will provide a convincing explanation of the influence of latent factors on male retirement. As post-estimation results, we obtain the probabilities to retire of the male partners by the log-Likelihood function, and we use these probabilities as propensity scores to estimate, by matching, the "treatment parameters" measuring the impact of retirement of man on the housework time of the woman. Estimation results are discussed in Sect. 3.

2 Data and Methods

The sample is drawn from the Italian Time Use Survey 2008-2009 provided by Istat (Italian National Institute of Statistics). In our analysis, the sample was composed of 3126 married or cohabiting women with a male partner whose age is ranged between 50 and 65 years (age compatible with eligibility for retirement). The men were either employed (2096 observations) or retired (1030 observations).

The dependent variable adopted in both equations of the Two-Regime model is the time (minutes) devoted by the woman to domestic work on a weekday. Explanatory variables used in the model provide demographic and socioeconomic information on the subjects: the woman's age (Woman's age), the woman's education expressed in years of school (Woman's edu), the health status of man (Man's health: 1= if man suffers from long-term health problems or chronic disease, 0=otherwise) and of woman (Woman's Health), the woman's working status (Woman retired: 1=yes; 0=otherwise), a dummy variable that assume the value 1 if the family received paid help in domestic work (Help received: 1=yes; 0=otherwise) and the daily time (in minutes) devoted to leisure activities by woman (Woman's leisure). Following Stancanelli and Van Soest (2012), and Ciani (2016), we introduced in the regressors set a second order polynomial indicating the man's eligibility for retirement (in the years 2008-2009, the man achieved the eligibility for retirement if he was aged 58 or more). Moreover, as in the study of Campolo and Di Pino (2020), we use as predictor of woman's bargaining power a latent construct of economic and behavioural factors (Bargaining-MIMIC), previously obtained applying a procedure known as Multiple Indicators and Multiple Causes Model (MIMIC) of a Generalized Structural Equations Model (GSEM) [8].

For the empirical analysis we adopt the ML estimator of Calzolari and Di Pino (2017) of an endogenous switching model with two regression equations whose dependent variables (outcomes) are mutually exclusive in a cross-sectional framework, and where selection is simply based on the choice of the larger outcome.

$$y_{1i} = \mathbf{x'}_{1i}\boldsymbol{\beta}_1 + u_{1i} \quad \text{if observed in regime 1; otherwise latent} \qquad (1)$$

$$y_{2i} = \mathbf{x'}_{2i}\boldsymbol{\beta}_2 + u_{2i} \quad \text{if observed in regime 2; otherwise latent} \qquad (2)$$

A relevant characteristic of this model is that the two dependent variables, y_{1i} and y_{2i}, are explicitly factors in the choice of the regime. For each individual, y_{1i} - y_{2i} represents the net gain (or net loss) from the choice between two options.

The error terms u_{1i} and u_{2i}, given by: $u_{1i} = y_{1i} - \mathbf{x'}_{1i}\boldsymbol{\beta}_1$ and $u_{2i} = y_{2i} - \mathbf{x'}_{2i}\boldsymbol{\beta}_2$ are assumed to be normally distributed with zero mean and variances σ_1^2 and σ_2^2. The identification of the across-regime covariance, σ_{12}, is based on the probability of a subject to gain the outcome of the two regimes, given by the probability density of the observed outcome, multiplied by the conditional probability that the other outcome (latent) is smaller than the observed. The two set of explanatory variables**,** $\mathbf{x'}_{1i}$ and $\mathbf{x'}_{2i}$, can be partially different in each regime. A partially different specification of the regressors in each equation may be helpful to identify the choice of the regime. In many cases, the regressors appear significant only in one among the regimes. Therefore, to improve the performance of the estimation procedure, may be convenient to include these variables only in the equation in which their contribution is significant.

We obtain the following contribution of the i-th observation to the log-likelihood:

$$\ln L_i = -\frac{(y_{1i} - \mathbf{x'}_{1i}\boldsymbol{\beta}_1)^2}{2\sigma_1^2} - \frac{1}{2}\ln\sigma_1^2 + \ln\Phi\left(\frac{(y_{1i} - \mathbf{x'}_{2i}\boldsymbol{\beta}_2) - \frac{\sigma_{12}}{\sigma_1^2}(y_{1i} - \mathbf{x'}_{1i}\boldsymbol{\beta}_1)}{\sqrt{\sigma_2^2 - \sigma_{12}^2/\sigma_1^2}}\right)$$

$$-\frac{(y_{2i} - \mathbf{x'}_{2i}\boldsymbol{\beta}_2)^2}{2\sigma_2^2} - \frac{1}{2}ln\sigma_2^2 + ln\Phi\left(\frac{(y_{2i} - \mathbf{x'}_{1i}\boldsymbol{\beta}_1) - \frac{\sigma_{12}}{\sigma_2^2}(y_{2i} - \mathbf{x'}_{2i}\boldsymbol{\beta}_2)}{\sqrt{\sigma_1^2 - \sigma_{12}^2/\sigma_2^2}}\right) \qquad (3)$$

where σ_{12} is the covariance between the error terms of both regimes, known as the across-regime covariance. Φ() is a normal cumulative distribution function (cdf) used to specify the contribution to the likelihood of censoring y_{1i} or y_{2i}.

Applying this *Two-Equation ML* procedure, the across-regime correlation parameter ρ_{12} (and the across regime covariance σ_{12}) can be directly estimated under the assumption of endogenous selection. After estimating the model's parameters, the following probabilities could be calculated:

-probability of being in regime 1 (male partner is retired):

$$\Phi\left(\hat{u}_{1i}\right)=\Phi\left(\frac{y_{1i}-\mathbf{x'}_{1i}\hat{\boldsymbol{\beta}}_1}{\hat{\sigma}_1}\right) \tag{4}$$

-probability of being in regime 1 (male partner is retired) for observations in regime 2 (male partner is not retired):

$$1-\Phi\left[\frac{\left(y_{1i}-\mathbf{x'}_{2i}\hat{\boldsymbol{\beta}}_2\right)-\frac{\hat{\sigma}_{12}}{\hat{\sigma}_1^2}\left(y_{1i}-\mathbf{x'}_{1i}\hat{\boldsymbol{\beta}}_1\right)}{\sqrt{\hat{\sigma}_2^2-\hat{\sigma}_{12}^2/\hat{\sigma}_1^2}}\right] \tag{5}$$

3 Results

In Table 1 we report the estimation results of the *Two-Equation ML* procedure. In order to better explain the choice of the regime, we adopt a different specification of each regime. In particular, we excluded the variables *Woman's health* (dummy) and *Bargaining-MIMIC* from the explanatory variables of the Regime 1, being these regressors significant only in the equation of Regime 2. Analogously, we dropped out the variable *Woman's leisure time* from the regressors of Regime 2.

Table 1: Woman's domestic work: ML estimation results

	Regime 1 = Man retired			Regime 2 = Man not retired		
	coef.	Std.Err	P value	coef.	Std. Err.	P value
Woman retired: 1= yes	39.70	10.37	***	-32.63	11.90	**
Education of woman	-8.08	0.95	***	-6.84	0.90	***
Woman's age	1.23	0.91		-2.06	0.81	*
Man's health: 1 = Sick	17.55	8.71	*	-8.83	8.52	
Retirement Eligibility of man (*Eligibility*) = 1 if he is 58 years old, at least)	23.27	22.75		-26.28	18.47	
1-Eligibility *(Age-58)	51.43	14.37	***	-14.57	8.68	
1-Eligibility *$(\text{Age-58})^2$	3.24	1.60	*	-1.67	0.93	
Woman's leisure time	-0.19	0.03	***			
Paid help received: 1 = yes	-88.62	17.57	***	-31.78	13.33	*
Woman's Health: 1 = Sick				-13.54	4.70	**
Bargaining-*MIMIC*				-165.0	24.72	***
Constant	336.69	55.41	***	576.91	49.85	***
r_{12} (Across-regime correlation)	**0.612**					

The table shows the estimated coefficients of the equation under the regime 1 (man retired) and the estimated coefficients of the equation under the regime 2 (man not retired). The estimated value of the across-regime correlation ρ_{12} is equal to 0.612. As post-estimation results we calculate: the probability of being in regime 1

(male partner retired) for observations in regime 1 (Eq. 4), and the probability of being in regime 1 for observations in regime 2 (Eq. 5). We use these estimated probabilities to perform a (Propensity Score based) matching procedure in order to calculate the Average Treatment Effects for treated (ATT). The estimated ATT parameter measuring the effect on the woman's domestic work is equal to -26.04 minutes in a day. Namely, housework time of the woman is expected to decrease in mean of 26.04 minutes in a day as a consequence of the retirement of the male partner. This result essentially confirms what previous studies found in terms of reduction of the woman's commitment in domestic work [3,5,6].

However, the novelty of our analysis is given by the interpretation of the effect of retirement in terms of "absolute" or "comparative" advantage for the woman. In particular, the positive sign of the across-regime correlation coefficient indicates that the women gain an "absolute" advantage in both regimes. This means that women, especially those who live in families with higher socioeconomic status, are not constrained by the need to devote a large part of their time to domestic work, regardless of the partner's working condition. However, if they aim to increase their weight in terms of the responsibility of managing the home and family affairs [2], they could encourage the male partner to retire in order of improving his commitment in household affairs.

References

1. Battistin, E., Brugiavini A., Rettore, E., Weber, G.: The Retirement Consumption Puzzle: evidence from a regression discontinuity approach. Am. Econ. Rev. (2009) doi: 10.1257/aer.99.5.2209
2. Bertocchi, G., Brunetti, M., Torricelli, C.: Who holds the purse strings within the household? The determinants of intra-family decision making. J. Econ. Behav. Organ. (2014) doi: 10.1016/j.jebo.2014.02.012
3. Caltabiano, M., Campolo, M. G., Di Pino, A.: Retirement and intra-household labour division of Italian couples: A new simultaneous equation approach. Soc. Indic. Res. (2016) doi: 10.1007/s11205-015-1076-5
4. Calzolari, G., Di Pino, A.: Self-selection and direct estimation of across-regime correlation parameter. J. Appl. Stat. (2017) doi: 10.1080/02664763.2016.1247789
5. Campolo, M. G., Di Pino, A.: Selectivity of Bargaining and the Effect of Retirement on Labour Division in Italian Couples. J. Fam. Econ. Issues (2020) doi: 10.1007/s10834-020-09672-1 (*published online).*
6. Ciani, E.: Retirement, pension eligibility and home production. Labour Econ. (2016) doi:10.1016/j.labeco.2016.01.004
7. Heckman, J.J., Honoré, B.E.: The empirical content of the Roy model. Econometrica (1990) doi: 10.2307/2938303
8. Jöreskog, K.G., Goldberger, A.S.: Estimation of a model with multiple indicators and multiple causes of a single latent variable. J. Am. Stat. Assoc. (1975) doi: 10.1080/01621459.1975.10482485
9. Stancanelli, E., Van Soest, A.: Retirement and home production: A regression discontinuity approach. Am. Econ. Rev. (2012) doi: 10.1257/aer.102.3.600

First and Second Year Careers of STEM Students in Italy: A Geographical Perspective

Primo e Secondo Anno di Carriera degli Studenti Italiani di Area STEM: Una Prospettiva Geografica

Antonella D'Agostino, Giulio Ghellini, and Gabriele Lombardi[1]

Abstract The mobility behaviour of Italian university has feed an increasing interest on the public debate. The very particular geographical characteristics of the country, jointly with the recognized persistence of the economic gap between the Southern and Northern regions, push more and more students in moving from the first macro-region toward the latter. This article will focus on the differences in the performance of those students who decide to move for their studies, against those who remain close to their hometowns. In order to analyse this issue, we conduct multilevel modelling techniques, using administrative microdata from the Italian Ministry of Education, University and Research (MIUR) referring to the first two years of the career of students from the cohort 2014/2015, enrolled in STEM fields.

Abstract *La mobilità degli studenti universitari italiani sta permeando sempre di più il dibattito pubblico. Le peculiari caratteristiche geografiche della nazione, unitamente al persistente gap tra il Nord e il Sud della penisola, spingono una frazione sempre più grande di studenti meridionali a spostarsi verso le regioni settentrionali. Questo lavoro si focalizza sulle differenze di performance tra gli studenti che decidono di emigrare per i loro studi e quelli che decidono di restare a studiare vicino casa. A livello metodologico verrà impiegata una tecnica di modellizzazione multilivello, utilizzando microdati amministrativi provenienti dal Ministero dell'Istruzione, Università e Ricerca (MIUR) riferiti ai primi due anni di carriera degli studenti appartenenti alla coorte di immatricolati 2014/2015 a dei corsi di studio STEM.*

Key words: student mobility, STEM, performance, multilevel model, transfer shock

[1] Antonella D'Agostino, University of Naples "Parthenope", antonella.dagostino@uniparthenope.it;

Giulio Ghellini, University of Siena, ghellini@unisi.it;

Gabriele Lombardi, University of Siena, gabriele.lombardi@student.unisi.it.

1 Introduction

The issue of internal mobility within countries is acquiring a growing interest among scholars, as one of the main events suitable for the analysis of transition through adulthood (see among the others, D'Agostino et al. 2019a; Contini et al. 2015; Dotti et al 2013). In particular, this paper explores the differences in the first academic year's performance among those Italian freshmen newly-graduated at the high school, who have decided to move away for their hometowns for their higher education studies in STEM (science, technology, engineering, and math), or remain in the nearby of their residences. This investigation is particularly important in light of the strong gap between the Northern and Southern regions of the country and it is related to the crucial role played by STEM fields for the local development (D'Agostino et al., 2019b). Indeed, few students who migrate return home to take advantage of their investment and many students stay in the host region to escape low returns on education in their region of origin (Attanasio and Enea, 2019).

In the international literature, several contributions are available concerning about earning a STEM degree and the migration of students pursuing STEM degrees. As an example, Crisp et al. (2009) evidence how, despite to the fact that these fields are acquiring constantly a greater importance in those public programs stimulating higher education attendance, it seems still to be present a difficulty in pursuing successfully a STEM degree by some categories of people, such as women or ethnic minorities. Public support seems to not play any role in mitigating this evidence, while the presence of a strong cultural capital seems to be much more effective. There is an evident stratification in the kind of people who decide to enrol a STEM course, and the growing interest in these fields is globally causing an increasing adoption of strategies in order to stimulate enrolments. This has to come with a bigger attention to the speed of students in adapting themselves to the new context (Lopez and Jones, 2017), with the awareness that the presence of better economic opportunities is the strongest pull factor for STEM students (Gesing and Glass,2019).

Although the importance of studying STEM field in educational analysis from different perspective, the number of empirical analysis on this topic are still poor in Italy. Some exceptions are, for instance, Chise et al. (2019) and Granato (2018), who both highlight the importance of environmental and cultural factors in explaining the main differences in STEM attendance.

Our goal is to contribute towards filling this gap. Precisely, focusing on the specific issue of the so-called *transfer shock* (Hills, 1965). Indeed, movers could suffer a decrease in their own performance after the transition to higher education, namely during their first year of enrolment, then gradually recovering the original achievement. Indeed, all transfer students generally present some degree of transfer shock in the transition from an educational context to another, and a low performance in the first year of studies can be just a sign of the emergence of this phenomenon. The youngest freshmen are usually the most affected by transfer shock, as the lowest performers in previous institutions. More specifically, Glass and Harrington (2002) find that movers in the long run seem to perform better than stayers, but they

experience a drop in their performance during the first semester. Regarding specifically to STEM students, Cejda (1997) finds that freshmen in this area experience a stronger transfer shock with respect to their colleagues in other fields.

2 Data Structure

The analysis is based on micro-data provided by the Italian Ministry of Education, University and Research (MIUR) and collected into the Italian University Student Register (ANS)[1], referring to the Italian university system as a whole. In particular, the database is restricted to 38,773 freshmen enrolled in STEM fields in the academic year 2014/2015 and that have information on their academic credits earned for the first and second year of their university career[2].
The working sample that resulted from this procedure is composed of 77,546 repeated measures clustered in 658 university courses.
The response variable that we use in this paper is the students' academic performance measured using the academic credits earned in each a.y. Indeed, as regulated by the Ministerial Decree 509/1999 by the MIUR, academic credits are a measure used by Italian Universities to estimate the workload required in order to graduate. In particular, we normalized it, as suggested by Leckie (2013)[3]. Firstly, the N observations are ranked basing on their original scores. Then, the standard normal score for the j^{th} ranked observation in the data is calculated as:

$$\boldsymbol{\Phi^{-1}\left[\frac{j-0.5}{N}\right]},$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function. The advantage behind this simple transformation is that it is order preserving and students with the same number of credits will also receive the same standard normal score. As independent variables in the econometric model we used the set of variable available in dataset. The main object of this study concerns the effect of the mobility indicator from South to North. Thus, in order to exploit this effect, we fix stayers and movers from North/Centre as the baseline, comparing them with both movers and stayers from South, separately. Additionally, in our control strategy, the following variables are included: gender (male – baseline), grade earned at high school coded as a dummy variable (grade lower than the 75th percentile – baseline), type of high school classified into three categories: scientific lyceum (baseline), classic lyceum,

[1] Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MIUR – Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze and Napoli Federico II, scientific reference Prof. Massimo Attanasio (UNIPA), Data Source ANS-MIUR/CINECA.
[2] All students enrolled in Italian universities' STEM fields in a.y. 2014/2015 are 51,821. About 13% of these freshmen drops-out the system and about the 3% of them changes course during the first a.y. Another 11% of them has missing information about their performance.
[3] It is worth note that we use this data transformation for using a simple econometric model in the empirical analysis. Nevertheless, more complex approach for such kind of data exist in literature (see for instance, Grilli et. 2016)

technical and or other diploma in secondary education. Then, another binary indicator informs if student, enrolled university later than one year after the end of the high school (less than one year – baseline). Finally, a dummy variable for time effect (a. y. 2014/2015 – baseline) has been considered, aiming at investigating the pattern of student's performance over time and to test the hypothesis of a transfer shock effect through its interaction effect with the mobility indicator.

3 Methodology

In a nutshell, Multilevel Analysis is a typology of hierarchical linear model, which allows to compute regression analysis for data with several nested levels. The underlying idea is that each level should be a potential source of unexplained variability (see among the others, Hox et al. , 2017).
The importance of relying in such a technique in educational studies is widely addressed by Grilli and Rampichini (2009). Consequently, this kind of methodology allows to characterized our three-level longitudinal design that includes repeated measures on students nested in STEM courses (Hair and Fàvero, 2019).
Our preferred specification considers the following three-level variance-components model:

$$\boldsymbol{CFU_{tjk} = \beta_0 + v_k + u_{jk} + \beta_{Time} year_{jk} + \beta_{SS} x_{jkSS} + \beta_{MS} x_{jkMS}}$$
$$\boldsymbol{+\beta_{tSS} year_{jk} \cdot x_{jkSS} + \beta_{tMS} year_{jk} \cdot x_{jkMS} + \sum_{h=1}^{H} \beta_h x_{hjk} + \varepsilon_{tjk}.} \qquad (\mathbf{1})$$

Equation (1) states that CFU_{tjk} (the normalized credit earned) in year t for the student j ($j=1,\dots,J$) in STEM course k ($k=1,\dots,K$) is a linear function of student level explanatory variables **x** and a time indicator variable (year). Whereas, $\beta_{Time}, \beta_{SS}, \beta_{MS}, \beta_{tSS}, \beta_{tMS}$ and β_h $(h = 1, \dots, H)$ are the unknown parameters to be estimated. The errors components $\boldsymbol{\varepsilon_{tjk}}$., v_k and u_{jk} are assumed to be mutually uncorrelated and i.i.d. normally with mean 0 and variance σ_ε^2 , σ_v^2 and σ_u^2, respectively. The main parameter of interest is $\boldsymbol{\beta_{tMS}}$ that indicates whether movers from South actually experiment a transfer shock in the transition toward the new environment. A positive sign of such parameter suggests in which measure they are able to overcome the transfer shock effect in the second year of their STEM studies.

4 Results and Discussion

Table 1 present the results of the model specified in equation 1 (see Model 4). The four models presented differ for the number of variables included in the estimation

process. Findings show that credits earned in the second year tend to decrease. In STEM studies women seems to have lower performances than men. The top-performing students, who come from a Scientific Lyceums, who apply on time for a STEM degree course are more likely to outperform during their university career. Finally, both stayers and movers from South appear to perform worse than their northern colleagues. Nevertheless, the positive estimate of the coefficient of the interaction effect ($\boldsymbol{\beta_{tMS}}$) clearly shows that movers from the South recover part of the performance lost during the transition experimented as freshmen. These results, to be confirmed by some in-depth analysis on previous cohorts of STEM freshmen, might suggest that South of Italy is experimenting a strong loss in its best human capital, emigrating far away for attending higher education studies in STEM.

Table 1: *Estimate results – Dependent Variable is the normalized credit earned.*

	Model 1	Model 2	Model 3	Model 4
Year (baseline 2014/2015)		-0.568***	-0.568***	-0.565***
Gender (baseline male)			-0.0324***	-0.0324***
HS Grade (baseline less than 75 percentile)			0.566***	0.566***
Classic Lyceum			-0.0992***	-0.0992***
Other Diploma			-0.229***	-0.229***
Late (baseline enrolled HE less than one year after HS)			-0.141***	-0.141***
Stayers From South			-0.306***	-0.274***
Movers From South			-0.154***	-0.239***
Year#StaySouth				-0.0649***
Year#MovSouth				0.170***
Constant	-0.0561***	0.228***	0.295***	0.294***
sd(course)	0.348***	0.348***	0.318***	0.318***
sd(student)	0.492***	0.568***	0.508***	0.509***
sd(Residual)	0.785***	0.675***	0.675***	0.673***
K	658	658	658	658
J	38,773	38,773	38,773	38,773
N	77,546	77,546	77,546	77,546

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

References

1. Attanasio, M. and Enea, M., La mobilitá degli studenti universitari nell'ultimo decennio. in Rapporto sulla popolazione. L'istruzione in Italia., De santis G, Pirani (2019).
2. Cejda, B.D., An examination of transfer shock in academic disciplines. Community College Journal of Research and Practice, 21(3):279–288 (1997).
3. Chise, Diana, Margherita Fort, and Chiara Monfardini. "Scientifico! like Dad: On the Intergenerational Transmission of STEM Education in Italy." (2019).
4. Crisp, G., Nora, A., Taggart, A. Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a stem degree: An analysis of students attending a hispanic serving institution. American Educational Research Journal, Vol. 46, No. 4, pp. 924-942 (Dec., 2009).
5. Contini, D., Cugnata, F. and Scagni, A., "From South to North: student internal migration in Italy. Should it be an issue?" paper presented at Population Days 2015, Palermo, 4-6 February (2015).

6. D'Agostino, A., Ghellini, G., Longobardi, S., Out-migration of university enrolment: the mobility behaviour of italian students. International Journal of Manpower, 40(1):56–72 (2019a).
7. D'Agostino A., Ghellini G., Longobardi S., Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy, Electronic Journal of Applied Statistical Analysis (EJASA) 12(4):826-845 (2019b)
8. Dotti, N. F., Fratesi, U., Lenzi, C., & Percoco, M., Local labour markets and the interregional mobility of Italian university students. Spatial Economic Analysis, 8(4), 443-468 (2013).
9. Gesing, Peggy, and Chris Glass. "STEM student mobility intentions post-graduation and the role of reverse push-pull factors." International Journalof Educational Development 65: 227-236 (2019).
10. Glass, J. J. C., Harrington, A. R., Academic performance of community college transfer students and" native" students at a large state university. Community College Journal of Research and Practice, 26(5):415–430 (2002)
11. .Hox, J.J., Moerbeek, M., van de Schoot, R.. Multilevel Analysis: Techniques and Applications, Third Edition. Routledge (2017)
12. Granato, S., Early Influences and the Gender Gap in STEM. mimeo, Queen Mary University of London (2018).
13. Grilli, L., Rampichini, C., Multilevel models for the evaluation of educational institutions: a review. In Statistical methods for the evaluation of educational services and quality of products, pages 61–80. Springer (2009).
14. Grilli , L., Rampichini, C., M , Varriale R., Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts , Statistical Modelling 2016; 16(1): 47–66, (2016)
15. Hair, Jr., J.F., Fàvero, L.P., Multilevel modeling for longitudinal data: concepts and applications. RAUSP Manag. J., Vol. 54 No. 4, pp. 459-489 (2019).
16. Hills, J. R., Transfer shock: The academic performance of the junior college transfer. The Journal of Experimental Education, 33(3):201–215 (1965).
17. Leckie, G., Module 11: Three-level multilevel models - Concepts. LEMMA VLE Module 11, 1-47.http://www.bristol.ac.uk/cmm/learning/course.html(2013).
18. Lopez, C. and Jones, S. J., Examination of factors that predict academic adjustment and success of community college transfer students in stem at 4-year institutions. Community College Journal of Research and Practice, 41(3):168–182 (2017).

Future Scenarios and Support Interventions for the Family: Involving Experts' Participation through a Mixed-Method Research Study

Scenari Futuri e Interventi a Supporto della Famiglia: un Approccio basato sugli esperti

Mario Bolzan, Simone Di Zio, Manuela Scioni and Morena Tartari

Abstract A recent Delphi survey on possible scenarios that will involve the family in the next ten years saw the emergence of some situations judged by experts as particularly relevant. Specifically, these emerging situations are the growing conditioning of women when organizing their family life because of professional commitments, the increasing tendency of young adults to live with their parents and the intensity of solidarity networks between generations. On the basis of some focus groups, eight possible intervention proposals have been identified to deal with any difficulties related to the emerging scenario and, in particular, to support families. In order to rank the proposals according to their efficacy and feasibility, the multi-criteria decision-making technique Analytic Hierarchy Process has been applied.

Abstract *Una recente indagine Delphi sui possibili scenari che interesseranno la famiglia nei prossimi dieci anni ha permesso di evidenziare alcune situazioni, che secondo gli esperti, acquisiranno particolare rilevanza che riguardano il crescente condizionamento per la donna nell'organizzazione della vita familiare a causa degli impegni professionali, la crescente tendenza dei giovani a permanere in famiglia anche quando trovano lavoro e l'intensità delle reti di solidarietà tra generazioni. Sulla base di alcuni focus group, sono state individuate otto proposte di intervento con l'obiettivo di fronteggiare eventuali difficoltà correlate allo scenario ipotizzato e in particolare per sostenere la donna e la famiglia. Infine, con l'obiettivo di creare una classifica di queste proposte è stata applicata la tecnica decisionale multi-criterio detta Analytic Hierarchy Process.*

Key words: family studies, future studies, Delphi, scenario research methodology, AHP

[1] Mario Bolzan, Department of Statistical Sciences – University of Padua; email:mario.bolzan@unipd.it

Simone Di Zio, University "G. d'Annunzio", Chieti - Pescara; email: s.dizio@unich.it

Manuela Scioni Department of Statistical Sciences – University of Padua; email:manuela.scioni@unipd.it

Morena Tartari, Department of Sociology, University of Antwerp, Belgium; email: Morena.Tartari@uantwerpen.be

1 Introduction

The political, social, economic and cultural events affecting our present situation, both nationally and internationally at a macro level will affect, at a micro level, family dynamics and the day-to-day relationships between family members, but how these and other events will influence the intra-family evolution has still to be deciphered. Through the construction of plausible future scenarios, one of the aims of this work is to provide answers to the question: "What will happen to the family in the next 10 years?" The participative process for the definition of the scenarios does not involve predictions for the future, because this is practically impossible with this type of time horizon. Rather, on the one hand, the reasoning concerning the future helps us to understand how we live/think today and, on the other hand, it offers the opportunity to discuss possible policy actions that would be necessary if negative situations are to be prevented and/or to facilitate widely desirable outcomes (FAMILYPLATFORM, 2011; OECD, 2011). Therefore, we are speaking of explorative scenarios that contrast with normative scenarios (Kosow and Gaßner, 2008). In the first case, the goal is to explore the future consequences of possible decisions and actions that could be taken today. We must ask ourselves: “What will happen in the future if today we do or do not do something? What could tomorrow's family be like if we take certain actions today?” Contrastingly, in the case of normative scenarios, a desirable future is depicted as if it were a goal to be achieved, and questions such as “How do we want our future to be? How can we get there?" arise (Gordon and Glenn, 2018). The usefulness of the scenarios lies in the possibility of understanding how the future depends on our actions today and in forcing the debate on the key factors and dynamics of a system. In this way, decision-makers can abandon an action already initiated, correct it, or take new actions.

2 Methods

Over the past three years, a study called "Tomorrow in the Family" has been conducted in the Veneto region—one of the richest and best organized regions in Europe. Several focus groups were conducted with the aim of identifying some areas of interest for the present and the future of the family. A Delphi survey was conducted with a panel of 32 experts, asked to answer a questionnaire containing 41 items divided into seven sections: Parents; Spouses; Extended Family; Children; Housing; Family Models; Policy and Services; Communication; and Solidarity. Each item is a brief statement aimed at describing a specific phenomenon, relating to one of the seven themes under consideration (Di Giulio, Fent, Philipov, Vobecká and Winkler-Dworak, 2013; Bolzan, 2018).
In order to investigate the future development of each item, the experts were asked to provide two assessments using an ordinal scale of 0-100, the first concerning Evolution, that is the spread of the phenomenon indicated in the item, the second regarding Relevance, or its importance. Both evaluations were provided considering the situation hypothesised in 10 years compared to the current time. Scores lower than, close to, or above 50 indicate reduction, substantial immutability, and expansion respectively. The two dimensions, if considered jointly, are intended to represent the future visibility of the phenomenon itself. The survey was carried out in four successive stages, the first through a face-to-face interview and the subsequent ones with a CAWI (Computer Assisted Web Interview) survey method based on the self-administered computerized questionnaire. The 32 experts selected to participate in the Delphi survey are stakeholders, professionals in the sectors of services or production of

family goods, officials, public managers, and scholars who, by culture and professional position, are knowledgeable about family dynamics[1].

The Delphi method serves to differentiate the proposed items and, in some way, to classify them according to their different degree of consensus achieved. On the one hand, there will be items with a full convergence of opinion, on which the experts agree on their future relevance and/or evolution. On the other hand, there will be items with little consensus, which will therefore require further study and analysis, given that the experts do not agree when estimating their relevance and future evolution.

Furthermore, the Delphi survey has made it possible to identify those phenomena that, according to the experts, will be characterized in the future by an important increase in relevance and a medium-high increase in terms of evolution. In practice, by selecting those items on which the experts have had a convergence of opinion regarding their high future evolution, we have the basis of an explorative scenario, because using a future perspective, the key factors and the main dynamics of the "family of tomorrow" emerge. To complete the exploratory scenario, it is necessary to define the possible policy actions to be taken today. Indeed, a mere description of plausible futures does not in itself constitute a scenario, because it must also contain the key decisions that may or may not lead to a possible change in the future (Gordon and Glenn, 2018). For this purpose, a further panel of experts (15) was asked to discuss a scenario characterized by different levels of relevance and evolution that emerged identifying, through concrete examples from their experience, the factors supporting the well-being of family members, in particular as will be indicated below, of the woman, the factors that put her at risk and the factors adverse to her well-being.

The scenario-based research methodology is frequently used in the context of interdisciplinary studies, also known as future studies. A definition shared by some scholars, and applied here, considers the use of scenarios as a 'research tool' that is able to support stakeholders in discussing the relationships between variables, the data that emerges and any indicators and that connects the present to the future, keeping in mind current policies and those that would be necessary for the future (Morgan, 1983; Ramirez et al., 2015). This method has the advantage of reducing any bias inherent in stakeholders from different disciplines, placing the emphasis on social problems and the need to suggest solutions in the public interest (Öborn et al., 2013).

Finally, with the aim of creating a ranking of the factors supporting the family members' well-being and building a priority list of interventions, the identified proposals were subjected to expert evaluation using the Analytic Hierarchy Process (AHP) method.

The Analytic Hierarchy Process is an effective tool for dealing with complex decision-making, and it has been applied in a wide variety of settings. Usually, decision-makers involved in this process are characterized by a personal background and experience pertinent to the problem to be solved. In the AHP, a set of evaluation criteria and a set of alternative options, among which is the best decision to be made, are arranged in a hierarchical structure. First of all, the AHP generates a weight for each evaluation criteria, according to the experts' pairwise comparisons of the criteria. Next, for a fixed criterion, each pair of alternatives is compared and a score is assigned to each of them. Finally, the AHP combines the criteria

[1] More in detail: 10 representatives of the institutions (a Major, some municipality leaders, an alderman, a manager of the Social Department of the Veneto Region, two ULSS managers, a Judge of the Family, a President of the Family Section of a Court, a President of the Juvenile Court); 3 exponents of the school: an headmaster, an education superintendent, a teacher; 9 Representatives of associations/unions: CGIL, CISL, Confcommercio, Confesercenti, National Consumer Union, Coldiretti, Adiconsum; 3 representatives of the volunteering; 7 Professors (sociologists, statisticians, educators, political scientists, demographers).

weights and the alternatives' scores, thus determining a global score for each alternative. Criteria and alternatives are compared in a pairwise way using a scale of absolute numbers known as the Saaty scale, which has been shown to capture individual preferences with respect to quantitative and qualitative attributes (Saaty, 1980). Overall, more than 60 experts will be involved, a relatively large number for this kind of decision tool.

3 Results

The Delphi survey has made it possible to identify items that have the experts' consensus on high values of evolution and relevance and that merit, also with respect to their content, to be the subject of analysis for a reflection regarding possible family support solutions. The three items selected are: (A.4.) "For the mother, the organization of family life will be conditioned by professional commitments"; (G.2.) "The networks of solidarity between generations (elderly, adults, young people) will be intense"; and (B.5.) "Young people will stay with their parents even after they have found a job." The Delphi survey, with respect to the three items considered, provided stability, convergence speed, and consensus level parameters, which all show satisfactory performance. This premise appears important to support the following applications and hypotheses. The conditioning in the organization of family life as a result of a mother's professional life resulted in a significant increase in relevance (Median = 85, Interquartile Range [IR] = 10) and evolution, albeit minor (Median = 70, IR = 15). The phenomenon of the permanence of young people in the family home is also characterized by an important increase in terms of relevance (Median = 80, IR = 20) and a slight increase in terms of evolution (Median = 60, IR = 20). It must be considered that in Italy, the dimension that the phenomenon already assumes today is peculiar: the proportion of "young adults" between 18 and 34 years old, who still live at home with their parents in 2018 (66.1%), is more than 15 percentage points above the average for all 28 EU countries (48.2%) (Eurostat, 2018). Finally, the phenomenon of solidarity networks between generations is characterized by an increase in relevance (Median = 70, IR = 10) and a slight increase in evolution (Median = 60, IR = 5). In summary, the permanence of young people in the family home and the existence of solidarity networks are phenomena that, according to the panel of experts interviewed, will assume greater importance even if their diffusion could remain of the same order of magnitude as today.

Starting from these three key factors, a scenario was hypothesised that factored in the substantial increase in the mother's conditioning when organizing family life because of her professional commitments, a presence similar to today of the solidarity networks between generations, and the permanence of young people in the family home more conspicuous than today. The panel of experts, therefore, identified the following eight intervention proposals to support family members, particularly women, in the context of the hypothesised scenario:

1) Facilitate greater autonomy for young children;
2) Improve public welfare (e.g. availability of services to the individuals, such as the elderly or children);
3) Provide direct financial subsidies to families with children, the elderly, etc.;
4) Strengthen the support networks within the family (e.g. grandparents, relatives, etc.);
5) Strengthen the support networks outside the family (e.g. peers, friends, etc.);
6) Promote a cultural change in family members (father, mother, children) through training to promote awareness of shared responsibilities;
7) Improve corporate welfare to support workers with dependent children and elderly family members;

8) Increase the accessibility and availability of family counselling services for family difficulties of different types (not only psycho-emotional, but also social, economic, etc.).
The application of the AHP to experts at the level of the North East of the country is currently being undertaken, and it is believed to involve at least sixty experts in the three regions. To date, nine experts have participated in the survey. The alternatives were compared according to two different criteria: "Feasibility" and "Efficacy." Regarding the evaluation criteria, experts recognized efficacy as more important (0.626) than feasibility (0.347). Provisional global weights of the intervention proposals to support the considered scenario are shown in Table 1.

Table 1: *Provisional global weights from AHP (n=9)*

Rank	***Intervention Proposals***	w_{global}
1	Promote a cultural change in family members (father, mother, children) through training to promote awareness of shared responsibilities	0.226
2	Strengthen the support networks outside the family (e.g. peers, friends, etc.)	0.158
3	Improve corporate welfare to support workers with dependent children and elderly family members	0.138
4	Increase the accessibility and availability of family counselling services for family difficulties of different types (not only psycho-emotional, but also social, economic, etc.)	0.117
5	Strengthen the support networks within the family (e.g. grandparents, relatives, etc.)	0.106
6	Provide direct financial subsidies to families with children, the elderly, etc.	0.100
7	Improve public welfare (e.g. availability of services to the individuals, such as the elderly or children)	0.084
8	Facilitate greater autonomy for young children	0.071

4 Discussion

The scenario that emerged from expert evaluations through the first Delphi survey is confirmed by the data. In Italy, female employment (15-64 years) has increased steadily over the past 15 years from 45.5% in 2004 to 49.5% in 2018 (ISTAT, 2018). In Italy, unlike in Northern Europe, the increase in female employment did not lead to a corresponding redistribution of personal responsibilities within couples (ISTAT, 2012), so that much of the family and care work remained in the hands of women. The experts' picture hypothesised for the future corresponds consistently with the trend recorded in recent years in terms of a significant increase in the work commitment of women, including those who have families. Family relationships are already the first source of informal support on which working mothers with young children, the elderly, disabled people and, more generally, those who need support rely. To give a figure, the proportion of children usually entrusted to grandparents is approximately 65% (ISTAT, 2011); grandparents are fundamental within the family support network. Husbands/partners do not seem ready to share household chores. As concerns older children, staying with the family of origin represents an increasingly important

and visible phenomenon. By the AHP's first application, we note some interesting results: the first factor with a weight of 0.226—quite distinct from the others—is the promotion of a cultural change in family members. In contrast, in final position with a weight of 0.071 is the autonomy of young children.

References

1. Bolzan, M.: Domani in famiglia. Franco Angeli, Milano (2018)
2. Di Giulio, P., Fent, T., Philipov, D., Vobecká, J., Winkler-Dworak, M.: A Family-Related Foresight Approach. Families and Societies. Working Paper Series. Changing families and sustainable societies: Policy contexts and diversity over the life course and across generations (2013)
3. EUROSTAT: EU-SILC survey. Retrieved from: http://appsso.eurostat.ec.europa.eu/nui/show.do?lang=en&dataset=ilc_lvps08 (2018)
4. FAMILY PLATFORM: Foresight Report: Facets and Preconditions of Wellbeing of Families. Report produced within the framework of the EU's 7th Framework Programme project SSH-2009-3.2.2 "Social platform on research for families and family policies" (2011)
5. Gordon, T. J., Glenn, J.: Interactive Scenarios. In: Moutinho, L., Sokele, M. (eds) Innovative Research Methodologies in Management, pp. 31-61. Palgrave Macmillan, Cham (2018)
6. ISTAT: Aspects of daily life survey. (2011)
7. ISTAT: Uso del tempo e ruoli di genere, Collana Argomenti n.43, Roma. (2012)
8. ISTAT: Labour Force Survey. Retrieved from: http://dati.istat.it/Index.aspx?DataSetCode=DCCV_TAXOCCU1 (2018)
9. Kosow, H., Gaßner, R.: Methods of future and scenario analysis: overview, assessment, and selection criteria. Deutsches Institut für Entwicklungspolitik gGmbH, Bonn (2008)
10. Morgan, G.: Beyond method: strategies for social research. Sage Publications, London (1983)
11. Öborn, J., Bengtsson, F., Hedenus, L., Stenström, K., Vrede, C., Magnusson, U. Scenario development as a basis for formulating a research program on future agriculture: a methodological approach. AMBIO, 42(7), 823–839 (2013)
12. OECD: The future of family to 2030. OECD Publishing, Paris (2011)
13. Ramirez, R., Mukherjee, M., Vezzoli, S., Kramer, A. M.: Scenarios as a scholarly methodology to produce "interesting research". Futures. 71, 70-87 (2015)
14. Saaty, T. L.: The Analytic Hierarchy Process. McGraw-Hill Book Co., New York (1980)
15. van der Heijden, K.: Scenarios: the art of strategic conversation. Wiley, Chichester (2005).

Gender and Monetary Policy Preferences: a Diff-in-Diff Approach

Genere e preferenze di politica monetaria: un'analisi differenze nelle differenze

Donata Favaro, Anna Giraldo and Ina Golikja

Abstract The aim of this article is to assess whether monetary policy preferences of central banks' female chairs differ from male chairs' monetary policy preferences. We study the relationship between the gender of central banks chairs and central banks' attitudes to pursue a more (or less) conservative monetary policy. The analysis is based on a database of 179 central banks observed over the period 1980-2018. We specify a Phillips-curve equation where the inflation rate is regressed on the output gap, trade openness, a dummy that detects central bankers' gender and time. Using a diff-in-diff approach, where the treated countries are the ones who experienced a period of female presidency, we found that women pursue a less conservative monetary policy.

Abstract *In questo lavoro studiamo il rapporto tra il genere dei presidenti delle banche centrali e le "preferenze" delle banche centrali nel perseguire una politica monetaria più (o meno) conservatrice. L'analisi si basa su un database di 179 banche centrali osservate nel periodo 1980-2018 per le quali stimiamo un'equazione della curva di Phillips in cui il tasso di inflazione è spiegato dal gap di produzione, dall'apertura commerciale, da una variabile dicotomica che rileva il genere dei banchieri centrali e dal tempo. Utilizzando un approccio diff-in-diff, in cui i paesi "trattati" sono quelli che hanno vissuto un periodo in cui il presidente della banca centrale era una donna, il nostro studio mostra che le donne a capo delle banche centrali tendono ad implementare politiche monetarie meno conservatrici.*

Keywords: monetary policy, gender, diff-in-diff.

[1] Donata Favaro, Department of Economics and Management "Marco Fanno", University of Padova, donata.favaro@unipd.it

Anna Giraldo, Department of Statistical Sciences, University of Padova, anna.giraldo @unipd.it

Ina Golikja, University of Padova

Introduction

The aim of this article is to assess whether monetary policy preferences of central banks' female chairs differ from male chairs' monetary policy preferences. The literature divides monetary policy makers into two categories: doves and hawks. A "monetary hawk" is a central banker who advocates keeping inflation low as the top priority in monetary policy. In contrast, a "monetary dove" is a central banker who emphasizes other issues, especially low unemployment, over low inflation. According to the existing literature, central banks' chairwomen appear to be more conservative (hawks) than chairmen. Women seem to be more "hawkish" than man in their willingness to use monetary policy tools to fight inflation volatility and to keep inflation at the target level.

There is an increasing interest in the study of how female composition of central banks' committees influences monetary policy. Farvaque et al. (2011) use a sample of nine OECD central banks – the ECB, the Reserve Bank of Australia, the Bank of Canada , the Bank of Japan, the Reserve Bank of New Zealand, the Swedish Riksbank, the Swiss National Bank – to analyse how the composition of central bank committees affect central banks' final outcomes. They estimate a kind of Phillips curve equation of the inflation level (where the regressors are some macroeconomic indicators such as the lag of inflation variation and the output gap, plus country-specific fixed effects) by including additional explanatory variables related to different characteristics of the committee. The gender factor enters the estimation as the share of women in monetary policy committee. The results show a negative coefficient associated with a hawkish attitude: a higher share of females in the boards of central banks contributes in making the committee more prone in lowering inflation levels. Farvaque et al. (2014) use a Data Envelopment Analysis (DEA) to evaluate the performance of central banks in terms of reduction of volatility of both inflation and output. They show that a higher share of women in central banks' boards increases inflation aversion.

More recently, Masciandaro et al. (2016) studied a sample of 112 countries and constructed an index of Gender diversity in Monetary Policy (GMP index) to evaluate the "dovish" and "hawkish" attitude of central banks' committees in relation to the proportion of females in their boards. They ran a regression for the 5-year average inflation rate in function of the GMP index (during 2010 and 2015), the lagged 5-year average inflation rate, the average output gap in the previous 5 years and a vector of country specific factors (level of central bank independence, trade openness, OECD membership). Their results show that a higher presence of females in central banks' committees implies a significant and negative effect on the inflation rate. Thus, they confirm the results of previous research that associates women with a higher degree of risk aversion towards inflation. Masciandaro et al. (2018), by using an augmented forward-looking Taylor rule to include the share of women in monetary policy committees among the regressors, confirm more conservatism in monetary policy as the share of women in central banks' boards increases. Last but not less important, Diouf et al. (2017) confirm that central banks' chairwomen put more emphasis on price stability then output stability, compared to

male chairs, by using a forward-looking Phillips curve equation where the output gap affects the inflation rate with one-year lag. More specifically, women are 73% less attached to the stabilization of output than men.

Taking inspiration from the discussed literature, in this article we propose a study of the relationship between the gender of central banks chairs and central banks' attitudes to pursue a more (or less) conservative monetary policy. The analysis is based on a database of 179 central banks observed over the period 1980-2018. Similarly to other previous studies, we specify a Phillips-curve equation where the inflation rate is regressed on the output gap, trade openness, a dummy that detects central bankers' gender and continental dummies. In addition to these variables, a time trend is included.

The innovative contribution of the study is twofold. First, differently from previous studies we focus on the gender of the president of central banks and not on the female composition of central banks' committees. Second, we employ a Differences-in-Differences methodology (DD). The DD estimation method is a quasi-experimental design (Meyer, 1995), that allows to evaluate the effects of a specific intervention. The method compares the differences or changes in outcome over time between two groups: the group enrolled in a program and a reference group that is "not treated". In our specific case, the "treated" group is represented by countries that had a central bank's chairwoman at some point in the considered period. The nontreated group is given by the rest of the countries. The advantage of DD over other quasi-experimental methods is that it controls for observed and unobserved fixed effects, making possible to compare countries with very different economic situations.

Data and model

Central banks (and relative countries) were selected using the list of central banks provided by the Central Bank Hub on the website of the Bank of Institutional Settlements (BIS). The database includes 179 central banks and related countries. A first delicate decision was how to treat European countries that are part of the European Monetary Union (EMU) nowadays. A possible choice was to exclude these countries from the database. A second choice was to include in the dataset all EMU countries up to the year of the adoption of the Euro and to drop them from the database from that year onwards, when information on the whole EMU area was added to the database. To avoid losing important observations from the database (and to bias the analysis), we chose the second alternative and we kept EMU countries in the database.

Original data on inflation rates and all explicative variables were recovered from the World Bank. Regarding the inflation rate, we also used a second measure recovered by the International Monetary Fund, which describes the inflation rate as the average of consumer prices. The output gap was built using the World Bank GDP expressed in nominal US dollars. After computing the real GDP by means of the GDP deflator (base year 2010), we applied the Hodrick-Prescott (HP) filter to

estimate each country potential GDP and took the difference between real GDP and its potential. Each country's output gap was finally relativized by its real GDP. Regarding trade openness, the index measures the ratio of each country's imports plus exports on GDP.

A critical point was to decide the minimum number of years of a female appointment as central bank president for a country to be considered treated. We decided for a period of at least three years, since the transmission mechanism of monetary policy can lasts for some years

To estimate the effect of the treatment – a central bank having a female president, on the outcome variable – the inflation rate - we use a DD approach that allows comparing differences in the inflation rates of treated and nontreated countries. The assumption that allows the identification of the effect of the treatment is the so-called common trend assumption; this means that, in the absence of the treatment, the inflation rate of treated and nontreated countries would behave similarly.

The DD method can be specified as a regression model (Meyer, 1995) where the observations are countries, treated and not treated, observed in two time points: before the treatment (t=0) and after the treatment (t=1). This implies making some choices. We assume t=0 the year before the appointment of a female chair and t=1 three years after the start of the appointment. Then, treated countries enter the estimation procedure only once in the covered period (1980-2018) whereas nontreated countries enter the estimation more than once. Their "repeated" participation in the estimation procedure is for single time spams of 4 years (as for the treated countries) that do not overlap.

The regression formulation of the DD estimates allows easily obtaining the estimates – the coefficients of interest are estimated by OLS – and the Standard Errors (Angrist and Pischke, 2009).

The general specification is then:

$$Inflation_{ict} = cons + \beta_1 D_t + \beta_2 I_c + \beta_3 D_t * I_c + X'_{ict}\,\boldsymbol{\beta_4} + \varepsilon_{ict}$$

Where i refers to the single country, t refers to time (with $t=0$ and $t=1$) and c indicates whether the country has been treated. D_t is a dummy variable that takes value 1 when $t=1$ whereas I_c takes value 1 when the country is treated. The parameter of interest is the coefficient of the interaction between these two dummies (D_t.*I_c). This parameter captures the situation in which a country i) has a central bank chairwoman and ii) it is observed at $t=1$, that is the time period in which we assume to be able to observe some effect of her monetary policy on the inflation rate. The explicative variables introduced in the model (output gap, trade openness, continental dummies all measured at time $t=0$, together with a trend variable) are summarised in X. The inclusion of the explicative variables is a simple way to adjust for observable differences between treated and nontreated countries (Meyer, 1995).

To take into account the fact that a nontreated country can appear more times in the estimation procedure, Standard Errors are adjusted for clustering at the country level.

Results and discussion

First results are reported in Table 1. As previously explained, control variables (output gap, trade openness, control dummies) are assessed in the year preceding the appointment of a female central banker (t=0). Then, the estimated coefficients are not to be interpreted as in a static model. To be clear, the coefficient of the output gap, for example, is almost significant with a negative sign. This means that when the output gap is positive – and the economy produces more than its potential – the inflation rate is lower than when the output gap is negative. This appears to contradict the predictions of the Phillips curve. Yet, the coefficient represents the relationship between the output gap at t=0 and the inflation rate three years later. Then, the negative coefficient need not worry. Indeed, in these types of model the only variable whose coefficient is relevant is the variable that captures the effect of the treatment, in our case variable D_t*I_c.
Regarding this variable, first estimates detect an upward effect on the inflation rate when a female is appointed as a central banker. Thus, our analysis suggests that women are more "dovish" than men in using monetary policy tools; female central bankers appear to emphasize issues other than low inflation, such as low unemployment and higher growth. This is an interesting and original result compared to the existing literature.

Table 1: Regression model results

Variables	***Coefficients***	***Robust SE***	***pvalues***
constant	1186.361	214.877	0.000
D_t	- 2.372	1.004	0.019
I_c	- 4.547	2.585	0.080
D_t*I_c	4.171	1.912	0.031
Output gap	- 13.820	7.354	0.062
Trade Openness	- 0.218	0.029	0.448
Year	- 0.590	0.107	0.000
Africa	8.131	2.044	0.000
Asia	6.592	2.410	0.007
Europe	5.021	2.503	0.047
Oceania	1.503	1.384	0.279
Center and South America	18.732	4.412	0.000

These first results are robust to different model specifications (with single year dummies instead of a continuous variable "year") and to the choice of different time periods, but still need some sensitivity analysis with respect to different treatment/time specifications. Some sensitivity analysis for the common trend assumption is also necessary. Moreover, to better control for observed heterogeneity, we are planning to apply a matching procedure for pairing treated

countries to similar nontreated countries. Even if a control for time is already in the specification, we also plan to deal with a possible structural change by dividing the 18 year time spam in smaller periods and estimate separate models. More sophisticated models such as a growth model or a latent variable model, including heterogeneity and time dependency, could be interesting future improvements of this work.

References

1. Angrist J.D., Pischke J.S.: Mostly Harmless Econometrics: An Empiricist's Companion, Economics Books, Princeton University Press, (2009)
2. Diouf I., Pépin D.: Gender and Central Banking, ECON MODEL, 61: 193–206 (2017)
3. Farvaque E. Stanek P.: Selecting Your Inflation Targeters: Background and Performance of Monetary Policy Committee Members, GER ECON REV, 2(2): 223-238 (2011)
4. Farvaque E., Stanek P., and Vigeant S.: On the Performance of Monetary Policy Committees. KYKLOS, 67(2): 177–203 (2014)
5. Masciandaro D., Profeta P., Romelli D.: Gender and Monetary Policymaking: Trends and Drivers Gender and Effects, BAFFI CAREFIN Working Papers1512 (2016)
6. Masciandaro D., Profeta P., and Romelli D.: Do women matter in monetary policymaking? Bocconi Working Paper Series n. 88, September 2018. Available at SSRN: http://ssrn.com/abstract=3248310 (2018)
7. Meyer B.D.: Natural and Quasi-Experiments in Economics, J BUS ECON STAT, 13(2): 151-161 (1995)

Headcount based indicators and functions to evaluate the effectiveness of Italian university education

Valutazione dell'efficacia della formazione universitaria italiana attraverso indicatori e funzioni basate sul conteggio del numero di presenze/assenze

Silvia Terzi[1], Francesca Petrarca[2]

Abstract In the present paper we suggest the use of Alkire-Foster (2007, 2011) dual cut-off method to compute a multidimensional performance indicator and of a simple headcount based function (an original suggestion) to detect association among the different dimensions of a performance indicator. In our illustrative application we apply the suggested dual cut-off method and the local concordance curve to synthesize relevant features of student opinions on their study programmes.
Abstract *Nel presente lavoro suggeriamo l'uso del metodo del doppio taglio di Alkire-Foster (2007, 2011) per calcolare un indicatore di performance multidimensionale e di semplice funzione basata sul conteggio delle unità presenti (proposta originale) per misurare l'associazione tra le diverse dimensioni di un indicatore di performance. Nel nostro esempio applicativo utilizziamo l'indicatore suggerito e la curva di concordanza locale per sintetizzare le caratteristiche rilevanti delle opinioni degli studenti sui loro programmi di studio.*

Keywords: Multidimensional Poverty Index, local concordance.

1. Introduction

Alkire-Foster (AF), [1],[2], [3], dual cut-off method was first introduced to measure the Multidimensional Poverty Index (MPI), but it is equally well suited to measure performance and/or quality or customer satisfaction both in business environment and in public organizations [5]. In particular, we suggest its use to evaluate the effectiveness of the Italian university education. Theoretical concepts like satisfaction or performance are generally recognized as multidimensional latent

[1] *Economics Department Roma Tre University, silvia.terzi@uniroma3.it*
[2] *MEMOTEF Department, La Sapienza University Rome, francesca.petrarca@uniroma1.it*

constructs exactly as non-income poverty. To define a synthetic indicator of performance (or satisfaction) we need to aggregate the sub-indicators related to the different dimensions the multivariate indicator is based on. Aggregating different dimensions, it would be extremely useful to embed in the performance indicator some information concerning the association between the single dimensions. This is the main reason why we suggest AF methodology. The second reason is due to the fact that sub-indicators concerning performance are often measured on ordinal scale, when not binary (success/failure).

To complement such information we introduce the notion of local concordance and suggest a *local concordance coefficient* and a *local concordance curve* [6]. Our intent is not to have an overall measure of concordance, i.e. a global indicator, but local concordance coefficients instead, in order to detect different degrees of concordance in the head, tail, or centre of the multivariate distribution of the components of a multidimensional indicator.

In our case study we have a set of several indicators measuring different types of good performance for undergraduate and graduate courses of a certain university, say U. Let us assume that we can set a reference value for each indicator, a benchmark indicating a good-quality standard. Of course, we could count for how many indicators each under-graduate or graduate degree programme reaches the reference value and then define as effective a degree programme that reaches reference values for at least k% of the indicators. This is the basic idea. The overall performance indicator is obtained counting how many degree programmes are well-performing (for each Department or each University). We then proceed with the plot of a local concordance function to see whether the local concordance is high and whether it varies within different windows of the distribution.

The outline of the paper is the following: in section 2 we discuss AF methodology and the head-count based multivariate performance indicator, in section 3 we introduce local concordance and the local concordance curve, in section 4 we illustrate our illustrative application.

2. Alkire-Foster counting approach

Let us assume we have d different performance dimensions (or poverty as in the original context) and no natural definition of an aggregate variable. The different dimensions could be - and indeed will often be - measured on an ordinal scale, such as customer satisfaction/appreciation in the evaluation of the quality of services. For each dimension, one could define a specific threshold/reference value or *cut-off* (as defined by Alkire and Foster, [1]) and identify who is either above or below each of such one-dimensional thresholds.

The second step consists in establishing a second reference value (or second level cut-off) usually indicated with k, to define as multidimensionally effective (poor in the original context) the unit that exceeds the first threshold in at least k dimensions or key-indicators. In other words, the second cut-off value defines how many

successes a unit must record in order to be defined as effective *tout court.* If we set $k=d$ this would lead us to the intersection-based approach, i.e. to consider multidimensionally effective the units that reach or exceed the reference thresholds in all key indicators. Vice versa, if we set the second cut-off value equal to 1 (i.e. $k=1$) this would lead us to the union based setting: an effective unit is successful in any of the key indicators. For $1<k<d$ we have intermediate solutions; and this is one of the advantages of the method.

Let w_j ($j=1, ..., d$) be the weight applied to the j-th dimension, and let $\Sigma w_j = d$, so that the weights w_j of the different dimensions add to the total number of areas d. Let c_i ($i=1, ..., n$) be the weighted number of achievements reached by the i-th unit; choose a performance cut-off k such that $0 < k \leq d$, and define multidimensionally effective the unit whose achievement count c_i is $\geq k$. Let q be the number of effective units and let $c_i(k)$ be the count of the (weighted) achievements only for the effective units. P_0, the performance indicator can be defined as: $P_0 = \Sigma\, c_i(k)/nd$, i.e. the weighted average of the number of achievements in the population. P_0 can also be expressed as a product between two measures: the incidence of effective units (H) and the intensity of achievements (A); more precisely: $P_0 = HxA$, where $H=q/n$ and $A=\Sigma c_i(k)/dq$.

It is logical to expect that, as the cut-off k varies, both the degree of incidence and the intensity will change. More specifically, as the second cut-off k increases, H is reduced because fewer and fewer units will be able to obtain a sufficient number of achievements; but at the same time, the positive variation of k increases A, producing an opposite effect on the final indicator P_0. On the contrary, by choosing a lower value for k, the increase in incidence contrasts with the reduction in intensity, with an uncertain effect on P_0, an effect that depends on the individual univariate distributions.

It is important to underline two other important properties of the methodology proposed by Alkire and Foster: multidimensional monotonicity and decomposability by subgroups. The first property implies that if an additional achievement is recorded for a statistical unit, the overall index increases. The decomposability, on the other hand, is based on the fact that if there were two distinct populations x and y (for example two different Universities, or two different departments or degree programmes of the same University), of n_x and n_y units, the index P_0 referred to the union of the two populations will be the average of $P_0(x)$ and $P_0(y)$ weighted with their respective numerosity.

3. Local concordance

Loosely speaking d variables are concordant if for a unit i, (i =1, … , n) large values on some variables are associated with large values on all the others and conversely for another unit i' small values on some variables are associated with small values on all the others. Perhaps the most widely used coefficient of concordance between 3 or more distributions is Kendall's W [4], designed to assess the agreement

between d raters[1]. This is where we take moves from to introduce the notion of *local concordance* and suggest a *local concordance coefficient* and a *local concordance curve.* Let $\boldsymbol{X}$ be the data matrix of n units and d variables: $\mathbf{X} = (x_{ih})$, $i=1, ..., n$, $h=1, ..., d$. We rank all the observations within each of the d dimensions. Then we partition our ranked observations in contiguous subsets (*slices*) of fixed size s[2]. For each ranking we have a first, a second, …, a last slice. We call *window* the union of the d corresponding slices. To assess local concordance we count how many units are ranked in each window. We have maximum agreement within the r-th window when the s units belonging to the r-th slice of the first distribution also belong to the r-th slices of all the other distributions. So, the smallest the number of units, the greatest the local agreement.

Vice-versa maximum disagreement is when the units belonging to the r-th slice of the h-th distribution do not belong to the r-th slice of any of the other d-1 distributions, so that the total number of units ranked in the window is maximum. Thus if we count the number of units belonging to the r-th window we reach a minimum (s) in case of maximum concordance and a maximum (sd) in case of maximum disagreement. This is the simple and intuitive idea our multivariate association coefficient K_r stems from. It is in fact a concordance coefficient since it assumes co-monotonicity for all variables.

Let C_r be the number of units ranked in the r-th window on at least one of the d rankings. To measure local concordance we can define a *relative indicator:* $K_r = \frac{sd - C_r}{sd - s}$. By computing the local concordance coefficient K_r for all the distinct slices of size s of the multivariate distribution, we can derive and plot a *local concordance curve*.

As for the choice of s, we suggest choosing a significant quantile, for example, some kind of benchmark. Unfortunately, however, the choice of s affects not only the maximum and minimum values of our head count criterion but also what we are defining as agreement. For example setting $s = 3$ means that the first 3 ranks are equivalent: if one unit i is assigned - respectively - ranks 1, 2, 3 on $d = 3$ distributions, we are stating that these rankings agree, coincide; when $s = 10$ we are stating equivalence among the first 10 ranks. In fact, what we call *agreement in rankings* means *agreement among classes of ranks of size s*.

4. An illustrative application

Assume that we want to evaluate student's appreciation of the study programmes held by a certain Department D, based on the results of the questionnaires on student opinions. Suppose that this particular Department offers one undergraduate study course and three masters. The students that attend lectures are asked 13 questions, concerning three distinct macro-environments, relating to the organisation of the course, teaching skills and openness of the lecturer, overall satisfaction.

For each question the student is asked to indicate a degree of satisfaction: totally unsatisfied, rather unsatisfied, rather satisfied, totally satisfied.

---

[1] W = S/maxS, where S is the sum of the squared deviations among the sum of the ranks achieved by each unit, and max S is the value S in case of maximum concordance.

[2] For the sake of simplicity we assume $s \leq n/d$.

We choose to give the same weight to each question; and to set as first level threshold a % of "totally satisfied" at least equal to 50%, i.e. a level of excellence. As far as the second level threshold *k* is concerned, we set *k*=8.

Recalling that $P_0=\Sigma c_i(k)/nd$, is the weighted average of the number of achievements in the population; $H = q/n$ is the incidence of effective units; $A=\Sigma c_i(k)/dq$ is the intensity of achievements; and furthermore P_0= HxA, we computed P_0, H and A for all study courses but also, separately, for the undergraduate course and for the master courses. Then we computed the performance indicator separately for the four different disciplinary areas (that we have called W, X, Y, Z). The results are collected in table 1:

Table 1- Performance indicators for different course levels (Overall, Undergraduate and Master), for four different disciplinary areas (W, X, Y, Z), k=8,

Area	***Study programme***	**P_0**	**H**	**A**
All disciplinary areas	*Overall*	*0,21*	*0,27*	*0,79*
	Undergraduate	*0,22*	*0,29*	*0,75*
	Master	*0,21*	*0,25*	*0,84*
W	*Overall*	*0,29*	*0,38*	*0,77*
	Undergraduate	*0,35*	*0,48*	*0,72*
	Master	*0,24*	*0,29*	*0,83*
X	*Overall*	*0,16*	*0,19*	*0,87*
	Undergraduate	*0,22*	*0,27*	*0,82*
	Master	*0,1*	*0,1*	*1*
Y	*Overall*	*0,19*	*0,22*	*0,85*
	Undergraduate	*0,06*	*0,08*	*0,77*
	Master	*0,52*	*0,6*	*0,87*
Z	*Overall*	*0,04*	*0,06*	*0,62*
	Undergraduate	*0*	*0*	*not a number*
	Master	*0,06*	*0,09*	*0,62*

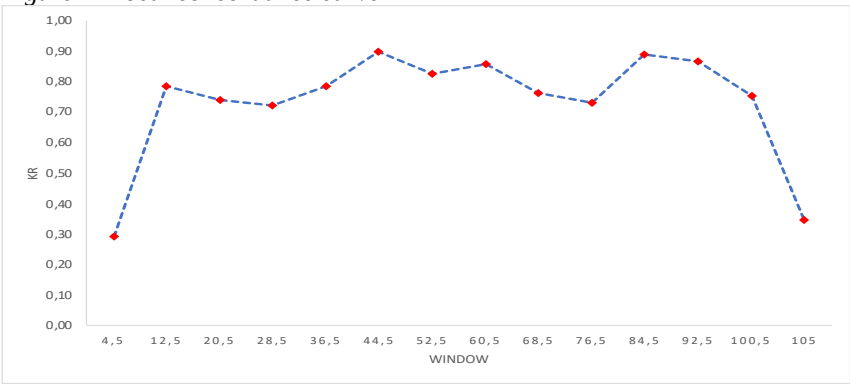
We next tried to see whether the not-so-good performance of disciplinary area Z is critical (i.e. a severely poor performance) or not. So, we defined a different first-order threshold below which each performance can be considered poor: a first order cut-off corresponding to a % of rather satisfied **and** totally satisfied of at least 50% and computed the frequency distribution of the number of thresholds achieved. Most lecturers (81%) reach and exceed all first order cut-off values, all of them achieve at least 8 thresholds. This means that if we maintain the second order cut-off at *k=8* and set this new and more attainable first order cut-off value, all courses in disciplinary area Z are above the standard of adequacy we have defined.

Table 2-Frequency distribution of the number of "new" thresholds for area Z

N. thresholds	8	9	10	11	12	13	Total
Frequency	1	2	1	5	12	91	112

To compute the concordance curve let us set $s = n/d = 8$. It is interesting to note that the local concordance is fairly low in the first and last window and much higher and fairly stable in the central part of the multivariate distribution. In particular, this means that the worst (best) performing units are worst (best) performing only on some dimensions but not on all (as can be seen for disciplinary area *Z* in table 2), whereas for non-extreme evaluations there is fairly high degree of agreement. The overall Kendall's concordance coefficient W = 0,54.

Figure 1- Local concordance curve



References

1. Alkire, S., Foster, J. (2009), Counting and Multidimensional Poverty Measurement, *Journal of Public Economics*, 95 (7-8), 476-487.
2. Alkire, S., Foster, J. (2011), Understandings and Misunderstandings of Multidimensional Poverty Measurement, *Journal of Economic Inequality*,Volume 9, Number 2, 289-314.
3. Foster, J., Greer, J.,Thorbecke, E. (1984), A class of decomposable poverty measures. *Econometrica, 52*(3),761-766.
4. Kendall M.G.,Babington Smith, B.(1939) The problem of m rankings *The Annals of Mathematical Statistics* 10, 275-287.
5. Menghini M., Terzi S, (2016), L'indice di performance ICM per l'analisi della competitività delle imprese.*Rivista economica del mezzogiorno*, vol.1/2016.
6. Terzi S., Moroni L. (2020), Local concordance and some applications, *Social Indicators Research*, https://doi.org/10.1007/s11205-020-02312-z

Identify the speech code through statistics: a data-driven approach

Identificare il codice linguistico attraverso la statistica: un approccio empirico

Andrea Briglia, Massimo Mucciardi, Jérémi Sauvage

Abstract Language is what makes humans a unique species of «symbolic animals» by providing them a way to convey meaning through sounds, and it is undoubtedly one of the pillars of our lives, yet we learn it so spontaneously and effortlessly that it is impossible to remember how we came up in its mastery or to give any account on any stage of its acquisition. Thanks to recent advances in data storage, information visualization and automated processing (*e.g.* data mining), there is a growing interest in cutting-edges researches between statistics and linguistics aimed at unfolding the "linguistic genius" of babies by testing hypotheses mining large spoken longitudinal datasets in order to understand - by means of an inductive procedure - the way each of us learnt his/her language without being aware of it.

Abstract *Il linguaggio è ciò che rende gli esseri umani una specie unica nel suo essere degli "animali simbolici" perché ci fornisce un modo di trasmettere significati attraverso suoni. Esso è indubbiamente di fondamentale importanza nella vita di ognuno, ma lo impariamo in un modo così spontaneo che è impossibile ricordare come ne abbiamo acquisito la padronanza così come è impossibile spiegare una qualsiasi delle tappe del suo apprendimento. Grazie ai recenti sviluppi tecnologici nella memorizzazione, visualizzazione e trattamento automatico di grandi quantità di dati è nato un crescente interesse verso studi che combinano statistica e linguistica per spiegare il c.d "genio linguistico" dei bambini verificando tale ipotesi su corpus longitudinali tramite una procedura induttiva.*

Key words: Phonetic Variation Rate; CHAID Model; First Language Acquisition

[1] Andrea Briglia, Univ. of Montpellier "Paul Valery"; email: abriglia@unime.it;
Massimo Mucciardi, Dep. of Cognitive Science, Univ. of Messina; email: mucciard@unime.it;
Jeremi Sauvage, Univ. of Montpellier "Paul Valery"; email: jeremi.sauvage@univ-montp3.fr

1. Identify the speech code through statistics

The so-called “linguistic genius” of babies (Kuhl, 2010) is a matter of long-standing debate in the scientific community: being adults, we realise how much easier is for a toddler to spontaneously learn every kind of language compared to adults’ struggle to maintain a sufficient mastery of a foreign language required – for instance - to attend international conferences. A number of evidences (Saffran, 2003) show that infants are geniuses because it is hypothetically plausible that they can rely on “statistically biased learning mechanisms” (Saffran, 1996) and on an “automatic” pattern recognition neuronal “device”. So, a fundamental question arises: “What is it about the human mind that allows a young child, merely one year old, to understand the words that induce meaning in our collective minds, and to begin to use those words to convey their innermost thoughts and desires?” (Kuhl, 2010). Every infant is facing a huge challenge: learning a sound system made up of many units that - combined together in an almost infinite set of combinations – gives rise to an arbitrary relationship between sounds’ sequences and meaning. According to Saffran’s metaphor, the task is the following: « You must discover the underlying structure of an immense system that contains tens of thousands of pieces, all generated by combining a small set of elements in various ways. These pieces, in turn, can be combined in an infinite number of ways, although only a subset of those combinations is actually correct. However, the subset that is correct is itself infinite. Somehow you must rapidly figure out the structure of this system so that you can use it appropriately early in your childhood» (Saffran, 2003). In fact, the balance between speed and accuracy in learning a language should be of primary importance in the survival of an individual: for this reason we supposed that human brains have been evolutionarily selected to their specific way of detecting regularities and patterns from the external world in order to retroactively syntonize their cognitive potentialities to the environment (Friston, 2010). The literature on « perceptual attunement» (Fort et al., 2017) demonstrates how early children become familiar with their mother language by focalising their speed and accuracy of the recognition task on what they have been experienced to and – symmetrically – by losing the capacity to readily detect and decode unfamiliar cues. So - as language is acquired through cognitive mechanisms that we could consider to be analogous to statistical engines that store probability distributions and formulate predictions based on means and expectancies on what has been previously stored - our attempt is to try to uncover what we have called “statistical learning” by mining a set of longitudinal *corpora* in french language.

2. Data structure and model

Colaje-Ortolang (2020) is an open access french database, part of the broader CHILDES project (2020): seven children have been recorded in a natural setting one hour every month, from their first months of life approximatively until six years old.

Data are available in three different formats: IPA, orthographic norm and CHAT (acronym for Code for the Human Analysis of Transcription), each of them is aligned to the correspondent video recording, allowing researchers to see the original source and to eventually reinterpret every utterance on their own. The main coding structure of the database consists in the fundamental division between "*pho*" (what the infant really says) and "*mod*" (what the infant should have said according to the adult's standard phonetic/phonological norm): we define "variation" every occurrence in which "*pho*" differs from "*mod*". How much the density of the sampling can influence the range of deductions and generalizations that we could draw from data is a debated question: is there a threshold beyond which the sampling is sufficiently representative and, *a fortiori*, any logical implication from it will be empirically valid? The answer depends on the level of analysis, in other words: the scale at which we want to focus on (Tomasello, 2004). In linguistics there are many scales: from the most basic units such as vowels and consonants to complex syntactical constructions. In a *corpus* such as the MIT Media Lab's pioneering "Human Speech Home project" (Roy, 2006) that consists in 400'000 hours of audio and video recordings, every level of analysis will be granted by a strong empirical support, as nearly everything the infant said has been recorded. The sampling of the "Paris Corpus" (Morgenstern et al., 2012) is obviously many times less statistically representative (one hour every month, that is roughly 0.5 – 1% of what the infant listen and say during the sample period, assuming he is awake about ten hours per day) so it probably could not provide sufficient empirical support to highly specific research on particular lexical phenomena or the emergence of specific syntactic structures, but on the other hand we think it suffices to provide a fair statistical support in order to account to the more general phonetic units' level, as well as the emergence of word categories such as pronouns, articles and determinants, being the probability of finding at least one target from any given sample higher for more basical units (Tomasello, 2006). Further, the age span is wider and – having been recorded 7 infants by using the same research protocol – it gives to researchers an easy way to compare development's intercourses between infants. Goal is to verify whether and how "any variation does not randomly vary into any other, but it rather should follow an underlying pattern, as every variation has an order in itself" (Sauvage, 2015). We first import 4 *corpora* of a single child named "Adrien" at 3 years and 1 month of age (time 22), 3 years and three months (time 24) and then time 27 and time 34. To turn raw data in a computationally and statistically tractable format we unbundle them into a data structure in which every sentence appears on the row side and every word on the column side. In table 1 are summarized the main statistics for 4 *corpora*: we can see how a quantitative increase in the number of words and length of sentences in which these words are combined causes an increase in S.D. that is due to a parallel increase in the lexical variability (type/token ratio) that – in turn - expands the range of possible variations a child can utter.

Table 1: Corpus statistics

Time	Mean	Length	S.D.
22	2.64	343	1.80
24	2.80	324	1.76
27	3.34	580	2.39
34	5.89	641	4.28
Total	3.98	1888	3.32

Mean = average number of words within a corpus; Length = length of the corpus; S.D.= standard deviation of the number of words within a corpus

Consequently, considering a single phrase of a *corpus*, we define "phonetic variation rate" (PVR) the ratio between the number of phonetic variations (NPV), that is the number of differences detected between "pho" and "mod", on the total numbers of words (TNW). In formula, for the phrase "i" and the total numbers of words "j": $PVR_{ij}=NPV_{ij}/TNW_{ij}$. In this way, by appropriately setting the subscript "j", we obtain for each corpus the PVR_j which represents the phonetic variation rate considering a definite number of words "j". Table 2 summarizes the results of the PVR considering j = 1, 2, 3,4,5 and 20 (max number of words in a single sentence.) From table 2 we can see how nonlinearity affects language acquisition: globally, PVR decreases over time but counterintuitive phenomena such as regressions (Morgenstern et. al, 2012) are frequent: it could happen that a child mispronounces something that he had previously correctly pronounced. The same holds for PVR over sentence's lenght: we expect (and observe) that rate increase as the lenght increases, but there are some exceptions to the norm that could require a specific account.

Table 2: Main statistics for PVR by time and number of words

Time	Statistics	PVR_1	PVR_2	PVR_3	PVR_4	PVR_5	PVR_20
22	Mean	0.477	0.556	0.655	0.513	0.667	0.577
	Length	*132*	*62*	*56*	*40*	*18*	*343*
	S.D.	*0.501*	*0.416*	*0.311*	*0.299*	*0.322*	*0.415*
24	Mean	0.494	0.528	0.525	0.538	0.608	0.553
	Length	*79*	*90*	*68*	*39*	*26*	*324*
	S.D.	*0.503*	*0.362*	*0.322*	*0.247*	*0.268*	*0.371*
27	Mean	0.558	0.532	0.563	0.471	0.440	0.483
	Length	*154*	*108*	*87*	*86*	*50*	*580*
	S.D.	*0.498*	*0.388*	*0.284*	*0.247*	*0.239*	*0.359*
34	Mean	0.305	0.281	0.244	0.278	0.260	0.246
	Length	*82*	*57*	*71*	*89*	*63*	*641*
	S.D.	*0.463*	*0.341*	*0.270*	*0.208*	*0.232*	*0.266*

In a second step, we used CHAID (Kass G., 1980) to get a general insight on how PVR changes over time and which kind of phonetic units are correctly articulated and which are not. From the results obtained[2], we can clearly see how time is the main regressor because it splits most part of the *corpus*, then the lenght of sentences

[2] All statistical analyses were performed using R, Excel and SPSS. In the CHAID model, cases are weighted by TNW.

plays a role as well, as we can observe in the *corpus* "time 34", where the fourth word causes the formation of an additional branch to the tree. The main pattern CHAID has detected in a "blind" way is the morphological difference between phonemes: as we can see from the tree table of the CHAID model (table 3), in the node 15 (PVR_20 mean 0.971, variation rate very high) words are longer and contains many "r" and couples of consonants, sounds typically learnt later in development.

Table 3: Tree table for CHAID model (main results - first and last three PVR_20 values)

Node	PVR_20 (Mean)	N	Primary Independent Variable	p-value	Split values
15	**0.971**	68	w_mod_1r	0.000	ɑ̃kɔʁ; sɛlsi; spidœʁma; isi; vjɛ̃; pʁefɛʁ; boku; bɔ̃ʒuʁ; vwatyʁ; vɛʁt; kaʁgo; osito; ɛskaʁgo; pjɛʁo; by; tʁwa; katʁ; sɛ̃k; sis; sɛt; ɔ̃z; duz; tʁɛz; katɔʁz; kɛ̃z; sɛz; dissɛt; dizɥit; diznœf; vɛ̃; vɛ̃teœ; vɛ̃tdø; vɛt; vɛ̃tkat; tɛ; tetɛ; kwɛ̃kwɛ; kwɛ̃; ʁjɛ̃; kɔʁnəmy; flœʁ; vɛʁ
20	**0.918**	255	time	0.000	22
4	**0.880**	490	w_mod_1r	0.000	etɛ̃; ty; sɔʁ; muje; lə; ɑ̃kɔʁ; lwi; sɛlsi; spidœʁma; akʁoʃe; otuʁ; sali; tɔ̃be; uvʁ; dɛʁjɛʁ; pɔʁt; isi; sɥisi; alɔʁ; ɑ̃; adʁijɛ̃; aj; tɛ̃kjɛt; naomi; puʁ; lotʁ; metɛ; zafiʁa; sypɛʁ; desine; mɔ̃tʁ; nunuʁs; dɔʁmevu; ʒak,
30	**0.079**	165	w_mod_1r	0.000	wi; la; ø; œ̃; bɛ̃; komɑ̃; dø; ba; duz; tʁɛz; katɔʁz; dɑ̃; noemi; tɛl; twa; kwa; ə; tjɛ̃; konɛ; ɛm; ka; pe; y; ve; igʁɛk; zɛd; ɛn; potiʁɔ̃; kɑ̃guʁ; s; sɥila; paʁl; tʁo; tabul; taʁɛt
24	**0.033**	152	w_mod_2r	0.000	la; vø; papa; apɛl; bum; dudu; mamɑ̃; sa; lə; akemi; dɔn,
27	**0.025**	119	w_mod_2r	0.000	nɔ̃; le; papa; lə; isi; bys; ʒoli,

while in the node 11 (PVR_20 mean 0.267 – not shown) words are shorter and contain more vowels and bilabials (e.g. "ma","ba") and - more generally - sounds pronounced by using the external part of mouth (easier to learn because infants can spot them by seeing them and thus providing cues for imitation, unlike sounds such as "r" or "l" who are articulated at the bottom of the throat and thus they have to be deducted by the child). We wrote "blind" because CHAID cannot distinguish morphological differences between phonemes, yet it performs a remarkable result simply by calculating interactions between occurrences. In conclusion, in this paper we have shown how the use of the CHAID model could provide us a way to analyse and evaluate child language development in a quantitative manner. Our results are sufficiently coherent to the state of the art of phonetic units acquisition (McLeod et

al. 2018). The main limit is that this technique doesn't take into account morphological differences, as the PVR is calculated on the difference between "pho" and "mod", regardless of what they represent linguistically: in order to overcome this limit we start to use Python to analyse *corpora* according to a predetermined list of phonetic units to track and quantify every variation, then we turn them into a "Multistream graph" (Cuenca et.al, 2018). These are the future directions of our research, once again we are trying to combine statistics and linguistics to try to test whether and how "any variation does not randomly vary into any other, but it rather should follow an underlying pattern, as every variation has an order in itself" (Sauvage, 2015).

References

1. Cuenca E., Sallaberry A., Wang Y., Poncelet P. « *MultiStream : A Multiresolution Streamgraph Approach to explore Hierarchical Time Series* ». IEEE Transactions on visualization and computer graphics, vol.24, no. 12. (2018)
2. Fort M.; Brusini P. ; Carbajal J. ; Sun Y. ; Peperkamp S. "*A novel form of perceptual attunement : Context-dependent perception of a native contrast in 14-month-old infants* ». Developmental cognitive neuroscience 26 , 45-51. (2017)
3. Friston. K. "*The free energy principle: a unified brain theory?".* Nature reviews. Neuroscience. Vol 11. February 2010. 127. Ref to the "Bayesian brain hypothesis"
4. http://colaje.scicog.fr/index.php/corpus cited 20 Feb. (2020)
5. https://childes.talkbank.org/ cited 20 Feb. (2020)
6. Kass, G.V.:. An Exploratory Technique for Investigating Large Quantities of Categorical Data". App. Statist 29(2):119-127 (1980)
7. Kuhl P. K. « *Brain Mechanisms in Early Language Acquisition* ». Neuron 67, (5). 713-727. (2010)
8. McLeod S.; Crowe K.. "*Children's Consonant Acquisition in 27 Languages: A Cross-linguistic Review*". American Journal of Speech-Language Pathology. 1-26. (2018)
9. Morgenstern A.; Parisse C. (2012), « *The Paris Corpus* ». French language studies *22*. 7-12. Cambridge University press. Special Issue. P11
10. Roy. D et al. « *The human speech home project* ». International Workshop on Emergence and Evolution of Linguistic Communication. Springer. Heidelberg. (2006)
11. Saffran J. "*Statistical language learning: Mechanisms and Costraints*". Current directions in Psychological science. 2003 Vol.12 No 4. P 110-114. (2003)
12. Saffran J. R ; Aslin R. N ; Newport E. L ; « *Statistical learning by 8-Month-Old infants* », Science, vol. 274, december. 1926-1928 (1996)
13. Sauvage J. « *L'acquisition du langage : un système complexe* ». L'Harmattan, Louvain la neuve. P103. (2015).
14. Tomasello, M. and Stahl, D. « *Sampling children's spontaneous speech: How much is enough?* » . Journal of Child Language, 31:101–121. (2004).

Inspecting cause-specific mortality curves by simplicial functional data analysis

Ispezione di curve di mortalità specifiche per causa attraverso l'analisi di dati funzionali simpliciale

Marco Stefanucci and Stefano Mazzuco

Abstract Cause-specific mortality rates describe the structure of mortality for a particular country and year. In this work we propose to analyze the dynamics of the rates through years and among countries using Functional Data Analysis (FDA). The particular nature of the data – compositions of a whole – requires nontrivial modifications of standard FDA procedures. Specifically, we applied a simplicial FPCA to data in order to show the main modes of variation. Preliminary results about a clustering are also discussed.

Abstract *I tassi di mortalità specifici per causa descrivono la struttura della mortalità per un dato paese e anno. In questo lavoro proponiamo di analizzare la dinamica dei tassi attrverso gli anni e tra paesi usando l'Analisi di Dati Funzionali (ADF). La particolare natura dei dati – composizioni di un totale – richiede modifiche non banali delle procedure standard dell'ADF. Nello specifico, abbiamo applicato la FPCA simpliciale ai dati per rivelare le principali forme di variabilità. Vengono inoltre discussi anche risultati preliminari riguardo al raggruppamento.*

Key words: Compositional Data; Functional Data Analysis; Cause-specific Mortality Curves.

Marco Stefanucci
Università di Padova, Via Cesare Battisti, 241, 35121 Padova e-mail: stefanucci@stat.unipd.it

Stefano Mazzuco
Università di Padova, Via Cesare Battisti, 241, 35121 Padova e-mail: mazzuco@stat.unipd.it

1 Introduction

The study of the mortality is one of the main branch of demography. A recently published work [6] tries to furnish some insights regarding the decreasing trend of US life expectancy by inspecting mortality by cause. The authors show that midlife mortality caused by drug overdoses, alcohol abuses, suicides, and a list of organ system diseases have particularly increased in the last years. However, the study of cause–specific mortality rates is not straightforward. The key aspect of such quantities is the fact that they represent *compositions*, being part of a whole – the total mortality rate. Recent articles [2, 3] consider this feature of the data and apply models from Compositional Data Analysis (CoDA) [1] literature. However, in this work we are also interested in the evolution over time of cause–specific mortality rates and our approach is to think of these quantities as realizations of smooth *functions*, and apply tools from Functional Data Analysis [5]. The direct application of standard FDA methodology in this context would lead to imprecise analyses and the models have to be modified to take into account the compositional nature of the data. In section 3 we will briefly present a simplicial version of Functional PCA. To be more technical, a simplicial functional datum is a function $\mathscr{C}(t) : I \in \mathbb{R} \to \Delta^n$, from a compact subset of the real numbers to the n-dimensional simplex. Cause–specific mortality rates fall into this category. As an illustrative example, a sample of simplicial functional data taking values in the 2-dimensional simplex is presented in fig 1. This figure could be a simplicial representation of cause–specific mortality rates with only three causes, however a sensible grouping of causes of death would lead to a larger number of categories.

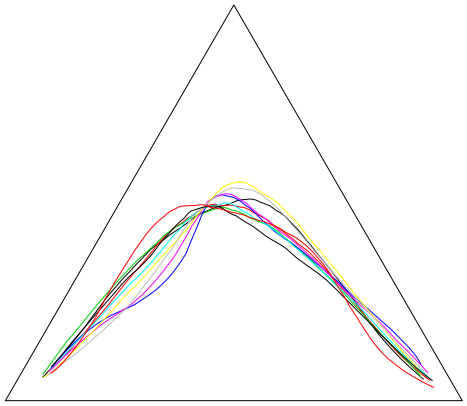


Fig. 1 Synthetic simplicial functional data.

2 Data

The data come from WHO mortality database, a public repository where mortality rates for several European countries and some non-European countries are collected. However, most of these countries does not have a time series long enough to be part of our analysis and we remove them. The final number of countries that entered the study is 22. For these countries we consider age–adjusted rates and age class–specific (i.e. 25–44, 45–64, 65+) rates, distinguishing by sex. The first and last years where data are available are 1959 and 2015, thus the functions are observed on a grid of 57 points. For the sake of simplicity we only consider 8 classes for the mortality: infectious and parasitic diseases, neoplasms (all cancers with the exception of lung cancer), lung cancer, endocrines diseases, circulatory diseases, respiratory diseases, digestive diseases and external causes. A plot of data for some selected countries is produced in fig 2.

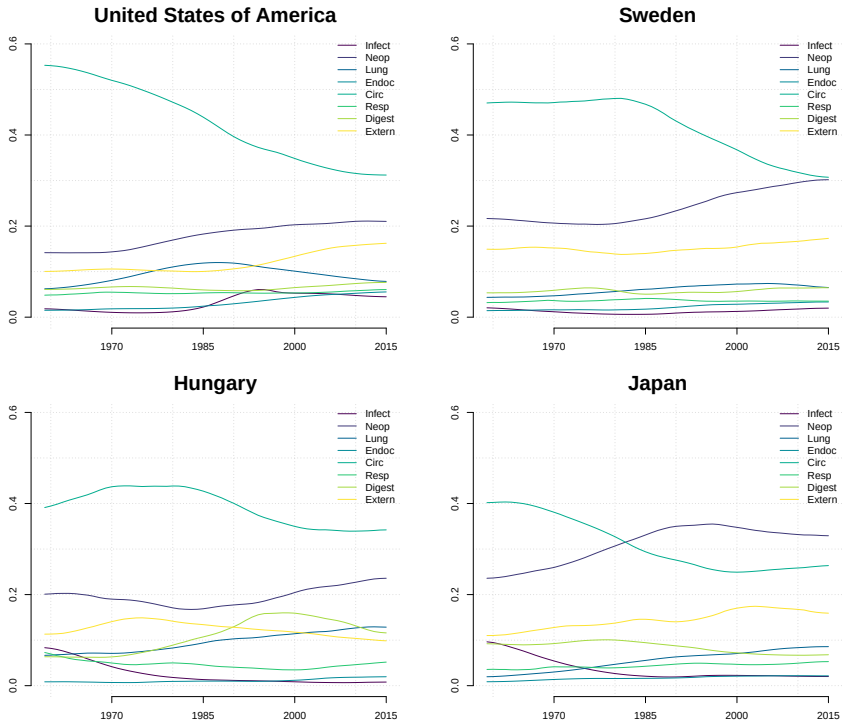


Fig. 2 Cause–specific mortality rates divided by the overall mortality rate for some selected countries in the analysis. Men, age 40–64.

3 Methodology

We first transformed the data by dividing the j-th cause–specific rate of the t-th year $\mu_x^{j,t}$ by the total mortality rate of that year, μ_x^t, obtaining $\tilde{\mu}_x^{j,t}$ (see [4]). These values are proportions that necessary sum to one:

$$\sum_{j=1}^{8} \tilde{\mu}_x^{j,t} = 1$$

This operation is called *closure* and can be further formalized as

$$\mathscr{C}(t) = \mathscr{C}\left\{\mu_x^{1,t}, \mu_x^{2,t}, \ldots, \mu_x^{8,t}\right\} = \left\{\frac{\mu_x^{1,t}}{\mu_x^t}, \frac{\mu_x^{2,t}}{\mu_x^t}, \ldots, \frac{\mu_x^{8,t}}{\mu_x^t}\right\}$$

The object $\mathscr{C}(t)$ is a multivariate functional object that takes values in the 7-dimensional simplex. Thus, as introduced in section 1, we need to take into account this aspect when building a model (i.e. principal components) for this objects. Specifically, thanks to the logratio transform [1] we are able to reduce the analysis to a multivariate functional PCA as illustrated in [5].

4 Results

Here we present preliminary results for males. The overall trends can be spotted in figure 2: neoplasms and circulatory diseases are the main causes of death, but while the former exhibit an increase in the last 50 years, the latter show a constant decrease. Infectious and parasitic diseases had a non-negligible level until the 80s. Lung cancer and endocrines diseases are slowly increasing in the male population. About the variability, the principal components that we computed reveal further interesting aspects of the mortality: most of the variability is due to neoplasms and circulatory diseases with a little contribution of infectious diseases, mainly in the first part of the considered period. This component accounts for about 1/3 of the total variability. Neoplasms and circulatory diseases are dominant also in the second component – that accounts for about 1/5 of the total variability , the only difference being the time period they are linked to: in this case is from mid 80s to now. The third and last component we considered explains about 15% of the total variability and it is mainly related to the overall level of lung cancer and external deaths, with a little contribution of respiratory diseases in the first years.
The projections of the data on the first three components are depicted in fig 3 and reveal a meaningful portrait of the mortality by cause over the years. Preliminary results regarding clustering show coherence with our geographic and demographic knowledge of the countries we considered. Further aspects are still under investigation such, for example, the main modes of variation of the mortality in the female populations.

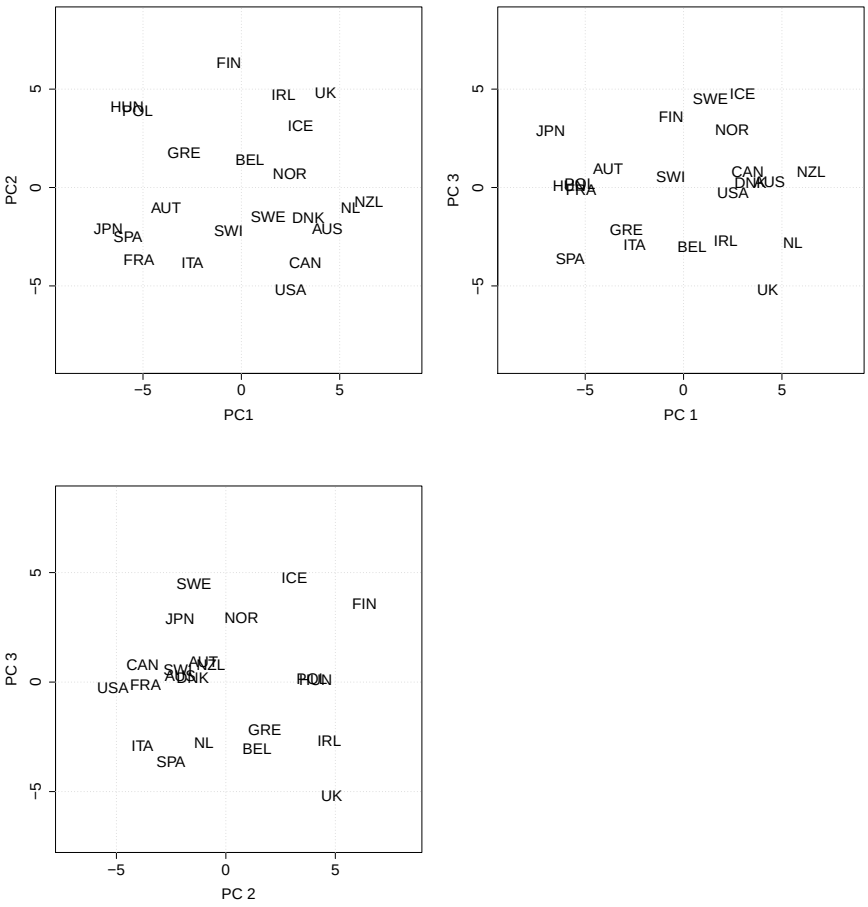


Fig. 3 Scores of the 22 countries in the first three principal components.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society. Series B **44**, 139–177 (1982)
2. Bergeron-Boucher, M.P. et al.: Coherent forecasts of mortality with compositional data analysis. Demographic Research **37**, 527–566 (2017)
3. Kjaergaard, S. et al.: Forecasting causes of death by using compositional data analysis: the case of cancer deaths. Journal of the Royal Statistical Society. Series C **68**, 1351–1370 (2019)
4. Preston, S.H., Heuveline, P., Guillot, M.: Demography. Measuring and Modeling Population Processes, Blackwell Publishing (2001)
5. Ramsay, J.O. and Silverman B.W.: Functional Data Analysis, 2nd edition. Springer, New York (2005)
6. Woolf, S.H. and Schoomaker, H.: Life Expectancy and Mortality Rates in the United States, 1959-2017. Journal of the American Medical Association **322**, 1996–2016 (2019)

Intertemporal decision making and childless couples.

Decisioni intertemporali e coppie senza figli.

Daniela Bellani, Bruno Arpino and Daniele Vignoli

Abstract Economic research has addressed the role of intertemporal decision making in occupational, financial and educational decisions. Very few studies have empirically analysed the association between this preference and the reproductive behaviour. In this study, we use data from the Survey of Household Income and Wealth (SHIW) carried out by the Bank of Italy. In particular, we make use of a question included in the 2004, 2008, 2010 and 2012 waves to examine whether patience has an impact in influencing childlessness in Italy. Results of logistic regressions indicate a non-linear relationship. Individuals with high or low levels of patience are more likely to remaining childless.

Abstract *Studi economici hanno evidenziato l'importanza delle scelte intertemporali nelle decisioni riguardanti la sfera occupazionale, finanziaria ed educativa. Pochi lavori hanno analizzato, da un punto di vista empirico, l'associazione tra le preferenze intertemporali e il comportamento riproduttivo. In questo studio utilizziamo i dati dell'Indagine sui Redditi e la Ricchezza delle Famiglie Italiane della Banca d'Italia per gli anni 2004, 2008, 2010 e 2012 che forniscono informazioni sulle preferenze temporali degli intervistati. I risultati ottenuti dalle regressioni logistiche mostrano che la relazione tra pazienza e non avere figli risulta essere non-lineare. Coloro che mostrano alti o bassi livelli di pazienza hanno una maggiore propensione a non avere figli.*

Keywords: childlessness, time preferences, SHIW.

[1] Daniela Bellani, DISIA, Department of Statistics, Computer Science, Applications, University of Florence; email: daniela.bellani@unifi.it

Bruno Arpino, DISIA, Department of Statistics, Computer Science, Applications, University of Florence; email: bruno.arpino@unifi.it

Daniele Vignoli, DISIA, Department of Statistics, Computer Science, Applications, University of Florence; email: daniele.vignoli@unifi.it

1 Introduction

In the last decades, the literature on fertility has extensively analysed several micro-, meso- and macro-level determinants influencing the decision-making process underlying the fertility timing and quantum choices (see Balbo et al., 2013 for a review).
Values and preferences, socio-economic conditions, demographic characteristics, health, social networks are among the micro-level factors that have been found to impact fertility behaviours. Although to a more limited extent, personality traits have also been analysed as determinants of fertility behaviours (Jokela, 2012).
In this paper we consider the specific trait represented by time preferences. They are strictly related to the trade-off between costs and benefits occurring in different time periods. Sociological and demographic studies of fertility have largely overlooked the role of impatience. One exception is De Paola and Gioia (2017) that studied the association between impatience and marital dissolution. In their study they found that impatient individuals are more likely to experience divorce. This is because they invest less time to find the best match that, in turn, increases the likelihood of separation. They add also that another mechanism that links impatience and marital dissolution is associated with the ability to manage for unanticipated shocks. While patient individuals are more prone to wait for the resolution of marital conflicts, this is not the case for those that prefer immediate (but lower) gains today.
This article gives us important insights for the study of the relationship between time preferences and other demographic behaviours, such as fertility decisions.
Italy represents an ideal context for this analysis given that couples carefully plan fertility behavior (Dalla Zuanna, De Rose, and Racioppi, 2005).

1.1 Hypotheses

One could expect that impatient individuals prefer to avoid or limit the material and immaterial costs of having a child thus reducing their fertility. Since they prefer short-term benefits and they tend to avoid sacrifices for long-term benefits, in a certain sense, they are more aware to immediate costs of fertility (in economic and non-economic terms) than to future gains (children as a security in old age). This implies that the relationship between impatience and the likelihood of remaining childless should take a positive sign (Hyp 1a).
However, it is likely that extremely patient individuals wait for a too long period of time to have a child because they want to achieve the best conditions for the new born. For example, they search for a high level of marital quality and job stability which might affect the likelihood of planning a child. This could increase the likelihood of remaining childless (Hyp 1b).

2 Data and empirical strategy

We use data from the Survey of Household Income and Wealth (SHIW) carried out by the Bank of Italy every two years since mid-sixties. SHIW collects information on consumption, income and wealth in addition to several household characteristics for a representative sample of Italian households drawn in two stages from population registers. The sample used in the most recent waves comprises about 8,000 households (20,000 individuals). From the 1989 wave a rotating panel component has been introduced. The share of panel households on the total has been around 45-50% of the total since 1993. The SHIW provides very detailed information on the demographic and socio-economic characteristics of individuals that belong to the same household unit. Moreover, the questionnaire of the survey includes questions that allow us to precisely capture time preferences of the head of the household. In this sense, SHIW is one of the very few surveys (together with GSOEP, NLSY and PSID) that provide this type of information. The question on time discounting preferences provided by SHIW data has been previously used, for example, to study trust (Albanese et al 2017) or divorce (De Paola and Gioia 2017) and is widely used to elicit impatience from a survey (Frederick et al 2002). It was included in the 2004, 2008, 2010 and 2012 questionnaires. It examines the choice a head of the household would take in a hypothetical situation. He/she has to decide how much money he/she renounces in order to receive a certain amount of money in the present instead after one year. After the description of the hypothetical situation, the respondent has to answer a series of questions about the amount s/he prefers to renounce. More precisely, heads of the households are asked the following question:
Q. Imaging receiving an unexpected inheritance (or lottery) equal to the amount of income that your family earns in a year. Now, imagine that the inheritance is only available after one year.
Would you be willing to sacrifice 10 % of that amount to have immediate access to the remaining 90 %? *yes: go to question (Qa) *no : go to question (Qb)";
Qa. "Would you sacrifice 20%? *yes *no
Qb. "Would you sacrifice 5%? *yes *no

The corresponding indicator operationalizes time discounting preferences as the rate of which the head of the household discounts future utility. We transform these values in rates, thus subtracting to 1 the percentage chosen in order to obtain the money immediately. We then compute the discount rate, thus generating our explanatory variable (as in De Paola and Gioia 2017). It is a continuous variable that can assume values between 0 to 0.25. In the case the respondent prefers a bigger economic reward in the future but smaller in the present, the variable will take low values. In the case he/she prefers higher rewards in the present but lower in the future the variable will take higher values.

We restrict our sample to the panel component of households where the respondent answers at least one time to the question of interest. In case more than one answer is given, we compute the measure of impatience taking the average value of the answers given across the waves. The study of the association between

time discounting preferences and reproductive behaviour covers the period 1995-2016. We consider married and cohabiting couples in a heterosexual partnership. We restrict our sample to female partner aged between 16 and 45 and male partner aged between 16 and 55 (N=760).

2.1 Method

Our dependent variable is the likelihood of remaining childless, i.e. not progressing from zero parity to one parity. Facing the classical trade-off between tractability and flexibility of specification, we model the probability of remaining childless versus progressing to first parity using logistic regression models.
Our main explanatory variable is the time discounting preferences. The majority of individuals who answered the time preferences question declared a value higher than zero of their hypothetical winnings they are prepared to give up as to gain immediate access to the money– only 21.72% declared the value of zero. Among those that declared a value higher than zero, 8,5% declared that they would sacrifice a 20% of the hypothetical winning to receive the sum immediately. We observe a considerable heterogeneity in those that declared a renounce higher than 0% and lower than 20%.

3 Empirical results

We model the probability to remain childless versus having the first child during the observation period. We include in our model time preference and its squared value – since it produces a better fit of the model. We control for the woman's age (and its squared), man's age, year, region of birth, gender of the respondent, female and male partner's educational level (low, medium and high education corresponding to ISCED 1-2, 3-4 and 5-6), credit rejection, liquidity constraints and equivalized income.

In Table 1 we present our main findings. For all specifications (M1, M2 and M3) the coefficient of the squared term of the TDP remains positive and statistically significant, and the size of the estimates does not vary substantially. All in all, our findings show a u-shaped association between impatience and not progressing to the first child. Results show that for both very low (patience) and very high (impatience) respondents, the probability to progressing to first parity is lower.

4 Conclusions

In order to examine whether time discounting preferences influence childlessness we use a nationally representative survey from Italy, the Survey of Household Income and Wealth (SHIW). By estimating logistic regression models, we find that there is a u-shape relationship between the degree of time discounting preferences and remaining childless. In particular, very impatient and very patient individuals have a higher likelihood of not having any child, while the opposite is valid for those with an average value of patience.

Table 1: *Estimates for Variables (N=760)*

	M1	***M2***	***M3***
TDP	-.050**	-.079***	-.076***
	(.023)	(.026)	(.026)
TDP squared	.006***	.009***	.009***
	(.002)	(.002)	(.002)
Age woman		-1.557***	-1.573***
		(.294)	(.292)
Age woman squared		.022***	.022***
		(.004)	(.004)
Woman education. Ref primary			
Secondary education		-.148	-.312
		(.235)	(.254)
Tertiary education		-.118	-.342
		(.293)	(.326)
Sex of the head of the household. Ref: male			
Female		.312	.347
		(.229)	(.238)
Income			.000*
			(.000)
N	760	760	760

*Note. Controls in M2 and M3 are male partner's age, male partner's education, year fixed effects, regional fixed effects. In M3 controls are also credit rejection and liquidity constraints. TDP=time discounting preferences. OR=Odds Ratio. Robust standard error in parenthesis. * p<.1, ** p<.05, *** p<.01*

References

1. Albanese, G., De Blasio, G., Sestito, P.: Trust, risk and time preferences: evidence from survey data. International Review of Economics (2017) doi: 10.1007/s12232-017-02827
2. Balbo, N., Billari, F. C., Mills, M.: Fertility in advanced societies: A review of research. European Journal of Population (2013) doi: 10.1007/s10680-012-9277-y
3. Dalla Zuanna, G., De Rose, A., Racioppi, F.: Low fertility and limited diffusion of modern contraception in Italy during the second half of the twentieth century. Journal of Population Research (2005) doi: 10.1007/bf03031802
4. De Paola, M., Gioia, F.: Does patience matter in marriage stability? Some evidence from Italy. Review of Economics of the Household (2017) doi: 10.1007/s11150-014-9275-4
5. Frederick, S., Loewenstein, G., O'donoghue, T.: Time discounting and time preference: A critical review. Journal of Economic Literature (2002) doi: 10.1257/002205102320161311
6. Jokela, G., De Blasio, G., Sestito, P.: Trust, risk and time preferences: evidence from survey data. International Review of Economics (2017) doi: 10.1007/s12232-017-02827
7. Jokela, G.: Birth-cohorts effects in the association between personality and fertility. Psychological Science (2012) doi: 10.1177/0956797612439067

Italian Households' Material Deprivation: Multi-Objective Genetic Algorithm approach for categorical variables

La deprivazione materiale delle famiglie italiane: un algoritmo genetico multi-obiettivo per dati categoriali

Laura Bocci[1] and Isabella Mingo[2]

Abstract Material deprivation is a complex concept referring to the inability of families to meet certain needs. Some indicators of material deprivation, included in the EU portfolio, are collected by EU-SILC survey on households and individuals through categorical variables. In this study, the large dataset provided by EU-SILC survey collected in Italy in 2017, on a sample of 22,226 households is analysed. The main goal is to identify clusters of Italian households to take into account the multiple aspects of material deprivation conditions, including environmental ones. To that end, a multi-objective genetic algorithm as a clustering technique for categorical data is proposed. The results are compared with those obtained by applying a *K*-means algorithm to latent variables scores.

Abstract *La deprivazione materiale è un concetto complesso che si riferisce all'incapacità delle famiglie di soddisfare determinati bisogni. Alcuni indicatori di deprivazione materiale, inclusi nel portfolio UE, vengono rilevati con l'indagine EU-SILC su famiglie e individui, attraverso variabili categoriali. In questo studio, viene analizzato l'ampio set di dati fornito dall'indagine EU-SILC rilevato in Italia nel 2017, su un campione di 22,226 famiglie. L'obiettivo principale è quello di identificare clusters di famiglie italiane per tenere conto dei molteplici aspetti delle condizioni di deprivazione materiale, inclusi quelli ambientali. A tal fine, viene proposta un'applicazione di algoritmi genetici multi-obiettivo come tecnica di raggruppamento per dati categoriali. I risultati ottenuti vengono confrontati con quelli di un algoritmo K-means applicato a punteggi di variabili latenti.*

Key words: material deprivation, multi-objective genetic clustering algorithm, categorical data.

[1] Laura Bocci, Department of Communication and Social Research, Sapienza University of Rome; email: laura.bocci@uniroma1.it

[2] Isabella Mingo, Department of Communication and Social Research, Sapienza University of Rome; email: isabella.mingo@uniroma1.it

1 On Material Deprivation in the EU: Indicators and Data

Material deprivation (MD) is the second criterion of the social exclusion target group, included in the EU portfolio of commonly agreed indicators in 2009, to better reflect the different living standards in the EU [3]. It is a complex concept which involves several variables based on the enforced lack of some items and referring to the inability of families to meet certain needs. The selected indicators are: 1. arrears in mortgage or rent or utility bills; 2. ability to keep home adequately warm; 3. to face unexpected expenses; 4. to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day; 5. to afford paying a one week annual holiday; 6. to have a colour television set; 7. to have a washing machine; 8. to have a car; 9. to have a telephone. These indicators are provided by European Union Statistics on Income and Living Conditions survey (EU-SILC) through categorical variables: some of them (variables 2, 3, 4, 5) are binary (response modes: Yes / No), while those referred to durable goods (variables 6, 7, 8, 9) have three response modes (Yes / No, cannot afford / No, other reasons) in order to differentiate between individuals who cannot afford a certain good and those who do not have it for other reasons; variable 1 have also three response modes (Yes / Yes, once / Yes, twice or more).

The standard material deprivation indicator adopted in 2009 uses a threshold of at least three deprivations to identify people considered as deprived, while the severe concept adopted in the context of the Europe 2020 strategy sets the threshold at four. The main limitations of this approach are the small number of items and above all the methodological choice based only on the number of deprivations rather than on the combination of the different type of inabilities to meet family needs. Moreover, other items could be added to take into account physical and social environmental indicators, deemed relevant in the analysis of material deprivation. Then, the number of possible combinations of all these multiple aspects becomes increasingly complex.

In this study, we analysed the large dataset provided by EU-SILC survey collected in Italy in 2017 on a sample of N = 22,226 households. The choice of MD variables was driven by the conviction that the analysis of material deprivation should also rely on physical and social environment in which each family lives daily and that affects the quality of life of its members. Therefore, in addition to the 9 variables described above, the following were considered: 10. noise from neighbours or from the street; 11. pollution, grime or other environment problems; 12. crime, violence or vandalism in the area (response modes: Yes/No)[1]. Moreover, some other variables were introduced in the analysis to describe the family clusters. They concern demographic, economic and contextual characteristics of the households: their size, poverty indicator, urbanization level and region.

The main goal is to identify substantially meaningful clusters of Italian households to take into account the multiple aspects of material deprivation conditions. Dealing with categorical variables, suitable algorithms should be adopted. To that end, it is

[1] The consistency of the selected MD variables was tested through Non Linear Principal Component (Promax rotation) which highlighted four different and correlated deprivation dimensions (α=0.934; VAF= 57.86%; λ=6.944): 1) maintenance incapacity (λ_1=2.524); 2) environmental deprivation (λ_2=1.819); 3) deprivation of durable goods (λ_3=1.394); 4) deprivation of household goods (λ_4=1.379).

proposed an application of a Multi-Objective Genetic Algorithm (MOGA) as a clustering technique for categorical data. Finally, the results are compared with those obtained by applying a *K*-means algorithm to latent variables scores.

2 Methods

Categorical data are pervasive in practice. Many real-life datasets are categorical in nature, where no natural ordering within a categorical attribute domain can be found. Most of the clustering algorithms are designed for datasets where the dissimilarity between any two units can be computed using standard distance measures such as Euclidean distance. However, Euclidean distance fails to capture the similarity of data units when attributes are categorical or mixed. In such situations, the clustering algorithms, such as *K*-means, cannot be applied. Some extension of the *K*-means algorithm have been proposed for categorical datasets. Huang [4] proposed the so-called *K*-modes algorithm, in which modes instead means are used to represent the cluster centers. Another extension of *K*-means is the Partitioning Around Medoids (PAM) algorithm [5], or *K*-medoids, where the cluster medoids (i.e., the most centrally located point in a cluster) are determined. A major disadvantage of *K*-means, *K*-medoids or *K*-modes clustering algorithms is that these algorithms often tend to converge to local optimum solutions. Moreover, all these algorithms rely on optimizing a single objective to obtain the partition which may not work equally well for different kinds of categorical data. Hence, it could be profitable to consider multiple objectives that need to be optimized simultaneously. The use of Multi-Objective Genetic Clustering Algorithms (MOGCAs) has emerged as an attractive and robust alternative in such situations. MOGCAs combine the need to optimize different criteria with the capacity of genetic algorithms to perform well in clustering problems, especially when the number of groups is unknown and the dataset to be analysed is very large.

In this paper a MOGCA, integrating the *K*-modes algorithm and simultaneously optimizing two objective functions for automatically partition the large EU-SILC dataset, is applied. A modified version of the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) proposed by Bandyopadhyay et al. [1] as clustering procedure was developed to take into account the specific characteristics of the data. A crucial point in the definition of a MOGCA is the choice of both the encoding of chromosome to represent a solution (partition) and a suitable set of objective functions that are to be optimized simultaneously. In this work a cluster prototype-based encoding of chromosome is adopted for two reasons: a) the large number of units to be clustered does not make possible the use of a different encoding strategy, such as a point-based approach; b) the search for different number of clusters simultaneously. The adopted encoding allows each chromosome to encode a different number of clusters K which may vary in a given pre-specified interval $[K_{min}, K_{max}]$, where K_{min} has to be 2 at least. The chromosome has fixed length K_{max} ($K_{max} << N$) and is coded as a sequence of K_{max} symbols. Each symbol representing a gene may be of two kinds. Type 1: a vector of J coordinates (integer numbers) representing the mode of a cluster.

If units are described in terms of J categorical variables $\{A_1, A_2, \ldots, A_J\}$, the cluster mode is the vector of the attributes each of which corresponds to the most frequent category occurring under the respective variable over all the units belonging to the cluster. Type 2: the character # to mean "don't care". As K must be 2 at least, the symbols # cannot be more than ($K_{max} - 2$). However, chromosomes may possibly not include any "don't care" symbol at all. In general, the chromosome will contain, arranged in any order, K cluster modes and ($K_{max} - K$) symbols #.

Two different sets of fitness functions are used: in the first setting the two functions are the Total Within-Cluster Dissimilarity (*TWCD*) and the Davies-Bouldin (DB) index, in the second one the *TWCD* and the number of clusters K. The Total Within-Cluster Dissimilarity (*TWCD*) is defined as

$$TWCD = \sum_{k=1}^{K_i} \sum_{i}^{N} u_{ik} d(\mathbf{x}_i, \mathbf{c}_k)$$

where $u_{ik} = 1$ if unit $\mathbf{x}_i$ belongs to cluster k, 0 otherwise, $\mathbf{c}_k$ denotes the mode of the k-th cluster and $d(\mathbf{x}_i, \mathbf{c}_k)$ is the Hamming distance between unit $\mathbf{x}_i$ and the centroid $\mathbf{c}_k$ of cluster k, such that $d(\mathbf{x}_i, \mathbf{c}_k) = \sum_{j=1}^{J} \delta(x_{ij}, c_{kj})$ where $\delta(x_{ij}, c_{kj}) = 0$ if $x_{ij} = c_{kj}$, 1 otherwise.

In both settings, the objectives, which are contradictory in nature, must be minimized simultaneously.

NSGAII (Deb et al., 2002) is used as the underlying multi-objective optimization algorithm. An initial random population of chromosomes of size $M = 200$ is created. Then, in the initial population for $i = 1,\ldots,M$, an integer K_i is generated uniformly random in $[K_{min}, K_{max}]$. Therefore, the i-th chromosome encodes K_i cluster modes and each mode is assigned to a gene selected at random within the chromosome. The gene encoding a cluster mode consists of J coordinates each of which is one out of the categories (randomly selected) of the J categorical variables. The genes that are left unassigned are set to the symbol #. The population is sorted based on the non-domination relation. Each chromosome of the population is assigned a fitness which is equal to its non-domination level. A crowding distance-based binary tournament selection strategy [2] is used. The crowding distance operator manages to maintain diversity in the Pareto front. Thereafter, crossover and mutation are used to create a new population of size M, and the process continues. A special crossover and mutation operations are applied to handle the presence of "don't care" symbols in the chromosomes. A two cutting points crossover is adapted. A pair of chromosomes is selected with crossover probability p_c=0.8. Let **R** and **Q** be these two chromosomes. Two integer numbers p and q are chosen at random in the interval $[1, K_{max}]$. If $p < q$ the genes from p through q of the chromosome **R** are exchanged with the genes that in the chromosome **Q** occupy the same positions. However, there is a chance of yielding invalid chromosomes after crossover since the same cluster mode may occur in a child chromosome more than once. Therefore, a penalty function approach is used to handle this situation: an invalid chromosome is given bad fitness values for all the objectives so that it automatically goes out of the competition in subsequent generations. The mutation operator is implemented with a mutation probability p_m=0.01 by generating a new mode or erasing an existing one with equal probability. Whenever mutation takes place, as usually, units are re-allocated to the cluster whose

mode is the nearest one in terms of Hamming distance. Finally, the elitist strategy is implemented. The near-Pareto-optimal chromosomes of the last generation provide the different solutions to the clustering problem.

The multi-objective genetic algorithm generates a set of nondominated trade-off Pareto optimal solutions. None of these solutions can be improved further in any objective value without degrading some other objective value. Among these nondominated set, the final solution is selected using the two cluster validity indices of Dunn and PBM [7], not used as an objective function of the algorithm. Therefore, the cluster modes encoded in this optimal solution are extracted and the partition is obtained assigning each unit to the cluster with the nearest mode in terms of Hamming distance.

3 Main results

The Pareto optimal solutions resulting by applying the MOGCA using both *TWCD* and DB or *TWCD* and *K* as fitness functions were compared. In the first setting, the Dunn and PBM indices suggest two partitions in eight and nine clusters, respectively. Instead, in the second setting the two indices agree in suggesting the same partition in five clusters. These results highlight that the choice of the functions to be optimized represents a key aspect and depends on the data, as stressed in the literature [6].

The partition in five clusters suggested by the above validity indices was analysed to identify substantially meaningful clusters of Italian households. To that end, characteristic categories of MD indicators, demographic, economic and contextual variables are used (Table 1). Taking into account the multiple aspects of material deprivation conditions, the following clusters were detected: 1: Families with maintenance difficulties (23.50%); 2: Not deprived families (42.18%); 3: Moderately deprived families (12.55%); 4: Families with environmental disadvantages (7.57%); 5: Deprived families who own durable goods (14.20%).

Finally, this partition was compared with that obtained by applying a *K*-means algorithm to the latent variables scores (see note 1 p. 2), searching for $K = 5$ clusters and starting from several random solutions. The selected *K*-means solution is the best according to Calinsky and Harabasz index. The two partitions (the one from MOGCA on categorical variables and that from *K*-means on latent variables scores) seem partially consistent (Contingency coefficient = 0.748) and identify some stable clusters which deserve further investigation.

References

1. Bandyopadhyay, S., Maulik, U., Mukhopadhyay, A.: Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, 45 (5), 1506 – 1511 (2007)
2. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Trans. Evol. Comp., 6 (2), 182 – 197 (2002)
3. Guio, A.C.: What can be Learned from Deprivation Indicators in Europe? Eurostat Methodologies

and Working Papers. Luxembourg: Office for Official Publications of the European Communities (OPOCE) (2009)
4. Huang, Z. X.: Extensions to the K-Means Algorithm for Clustering Large Data Sets With Categorical Values. Data Mining and Knowledge Discovery, 2, 283 – 304 (1997)
5. Kaufman, L., Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons (1990)
6. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: A Survey of Multiobjective Evolutionary Clustering. ACM Computing Survey, 47 (4), 1 – 46 (2015)
7. Pakhira, M. K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern Recognition, 37, 487 – 501 (2004)

Table 1: Cluster description - Characteristic categories[1] * (NC means "not characteristic")

	Cluster 1 (23.50%)	**Cluster 2 (42.18%)**	**Cluster 3 (12.55%)**	**Cluster 4 (7.57%)**	**Cluster 5 (14.20%)**
Arrears in mortgage or rent, or utility bills	Yes, once	No	No	No	Yes, once
Ability to keep home adequately warm	Yes	Yes	Yes	Yes	No
Capacity to face unexpected financial expenses	No	Yes	Yes	Yes	No
Capacity to afford a meal with meat, chicken, fish (…) every second day	Yes	Yes	Yes	Yes	No
Capacity to afford paying for one week annual holiday	No	Yes	No	Yes	No
To have a colour TV	NC	Yes	Yes	Yes	Yes
To have a washing machine	Yes	Yes	Yes	Yes	Yes
To have a car	No, cannot afford	Yes	No, other reason	Yes	No, cannot afford No, other reason
To have a telephone (including mobile phone)	No, other reason	Yes	No, other reason	Yes	No, other reason
Noise from neighbours or from the street	No	No	No	Yes	No
Pollution, grime or other environmental problems	No	No	No	Yes	No
Crime violence or vandalism in the area	No	No	No	Yes	No
Household size	1; 5	2; 3; 4	1;5	3	1; 5
Poverty indicator (At risk of poverty)	Yes	No	Yes	No	Yes
Region	Insular South	Northeast Northwest Central	Insular South Central	Northwest Central South	South Insular
Urbanisation Level (area)	Thinly-populated	Thinly-populated Intermediate	Thinly-populated	Densely-populated	Intermediate Thinly-populated

* Test value p<0.000.

[1] Test value is calculated comparing the proportion of a category in the *i*-th cluster and the proportion of the same category in the whole sample. A high test value indicates that the proportion in the cluster differs (positively or negatively) significantly (p<0.05) from that in the sample characterizing the cluster itself. For brevity, Table 1 shows only those categories which differ positively.

LI-CoD Model. From Lifespan Inequality to Causes of Death

Modello LI-CoD. Dalla Lifespan Inequality alle Cause di Morte

Andrea Nigri and Susanna Levantesi

Abstract The evolution of lifespan disparity is gaining a central role in mortality literature. From the beginning of the new millennium, its evolution has led scholars to give more emphasis to longevity and its relationship with causes-of-death evolution. Following this line of research, we propose a novel model aiming to provide the causes-of-death mortality surface, exploiting the relationship between mortality rates by cause-of-death and lifespan variability. Taking advantage of this relationship, and using data from Human Causes of Death Mortality Data Base, our model is able to reconstruct the mortality surface by causes, using a single value of lifespan inequality as input.

Abstract L'evoluzione della lifespan disparity sta acquisendo un ruolo centrale nella letteratura sulla mortalità. Dall'inizio del nuovo millennio, la sua evoluzione ha portato gli studiosi a dare maggiore enfasi alla longevità e al suo rapporto con l'evoluzione delle cause di morte. Seguendo questa linea di ricerca, proponiamo un nuovo modello che mira a fornire la superficie della mortalità per cause di morte, sfruttando la relazione tra i tassi di mortalità per causa di morte e la variabilità della durata della vita. Sfruttando questa relazione e utilizzando i dati dello Human Causes of Death Mortality Data Base, il nostro modello è in grado di ricostruire la superficie della mortalità per cause, usando come input un singolo valore della lifespan disparity.

Key words: Life expectancy, Forecasting, CoD.

Andrea Nigri
Sapienza, Viale Regina Elena 295-G, 00161 Rome, e-mail: andrea.nigri@uniroma1.it

Susanna Levantesi
Sapienza, Viale Regina Elena 295-G, 00161 Rome, e-mail: susanna.levantesi@uniroma1.it

1 Introduction

Over the last two centuries, longevity evolution has had a prominent impact on population dynamics, showing a rapid decline in mortality levels in developed countries. These improvements are driven by medical progress, better living conditions, leading to the infant mortality reduction. On the other hand, scientific achievements against chronic diseases ([8], [12]) lowered the adult age mortality after the second-world war, likewise the recent delaying of mortality at older ages ([14]). As a result, life expectancy at birth has increased over time, without any sign of an impending limit in human life boundaries ([5]). This latter behavior seems to be in lockstep with a significant decrease in lifespan variability, which could be due to different causes-of-death (CoD) composition. In light of that, the understanding of evolution of CoD composition is crucial in predicting the aging process and the population health. Concerning forecasting, cause-specific mortality is often based on predicting cause-specific death rates independently. Only a few methods incorporating dependence among causes have been suggested. An attractive alternative is to model and forecast cause-specific death distributions, rather than mortality rates, as dependence among the causes can be directly incorporated ([3]). Unfortunately, these processes are often not straight and show a lack of parsimony and interpretability. Furthermore, they are based on a rigid structure that does not taken into account the revolution of CoD pattern over time. Our paper contributes to the literature proposing a novel statistical tool able to provide the CoD mortality surface, using the relationship between CoD mortality rates and lifespan variability. In particular, given an age, our model is able to reconstruct the mortality surface by causes, using a single value of lifespan inequality as input. Thus, using data for several countries, we try to explain and forecast the remarkable changes in the transition phases that developed countries have exhibited in the period after the second world war, providing age-cause specific action.

2 Longevity indicators

The constant improvement of BPLE suggests that mortality reductions should not be viewed as a disconnected sequence of unrepeatable revolutions, but rather as a regular flow of continuous progress ([6]). Clearly, mortality is linked to social progress in terms of health, nutrition, education, hygiene, and medicine ([7]). Moreover, mortality improvements have been combined with an increase of lifespan equality in which the increase in the age-at-death corresponds to the increase of the compression of the distribution around its modal value ([4]). While life expectancy has been proven to hide heterogeneity in individual mortality courses, lifespan disparity measures

both uncertainty in the age-at-death distribution and heterogeneity[1] ([10]). In the following equations we provide the formal notation and definitions for both life expectancy and lifespan disparity.

- **Life expectancy**
  Let $S(x,t)$ and $\mu(x,t)$ be two continuous functions with respect to age x and time t, respectively representing the survival probability and the force of mortality of an individual aged x at time t in a given population. We denote $e_{x,t}$ the life expectancy at age x and time t, that is defined as follows:

$$e_{x,t} = \frac{\int_x^\infty S(y,t)dy}{S(x,t)} \tag{1}$$

  where $S(x,t) = \exp\left(-\int_0^\infty \mu(x+\xi,t+\xi)d\xi\right)$ are the survival probabilities.

- **Lifespan disparity**
  We denote $e^\dagger_{x,t}$ the lifespan disparity that is an indicator of the lifespan variation representing the life expectancy lost due to death by an individual aged x at time t ([13]). Formally, its functional form is defined as follow:

$$e^\dagger_{x,t} = -\int_x^\infty S(y,t)\ln S(y,t)dy \tag{2}$$

  The lifespan disparity at birth is: $e^\dagger_{0,t} = -\int_0^\infty S(y,t)\ln S(y,t)dy$. Lifespan disparity varies among populations and over time. It measures the dispersion in the age-at-death: when mortality is highly variable, some individuals will die at a much lower age than the expected age-at-death, contributing many lost years to life disparities; conversely, when mortality is highly concentrated around older ages or the modal age, life disparity decreases.

These two variables show latent behaviors that should be represented by incorporating in forecasting the relationship between lifespan inequality and CoD composition. Our approach aims to introduce a new perspective in the forecasting of CoD mortality surface, thus providing more accurate predictions.

3 Data and model

We use data from Human Causes of Death Mortality Data Base that provides death rates specific for ages and causes of death. This database is coded by using the international classification of diseases (ICD), providing different aggregation levels: full list, intermediate list, and shortlist. Each classification has been developed using the same criteria for all countries, ensuring homogeneity and comparability. For our aims,

[1] In addition to life disparity, other inequality measures have been proposed in literature, e.g. the Gini coefficient and the Keyfitz's entropy ([9], [11]) that appear to be linearly related and negatively correlated to life expectancy at birth ([1]).

we use the shortlist that provides the highest level of ICD aggregation, underlining 10 different macro CoD. Afterward, we perform an additional aggregation as follows:

- **Circolatory:** Heart diseases (I00-I52); Cerebrovascular diseases (G45, I60-I69); Other and unspecified disorders of the circulatory system (I70-I99)
- **Neoplasy:** Neoplasms (C00-D48)
- **Diab:** Endocrine, nutritional and metabolic diseases (E00-E90)
- **External:** External causes (V01-Y98)
- **Perinatal:** Diseases of the genitourinary system and complications of pregnancy, childbirth and puerperium (N00-O99) Certain conditions originating in the perinatal period and congenital malformations/anomalies (P00-Q99, R95)
- **Respiratory:** Acute respiratory diseases (J00-J22, U04); Other respiratory diseases (J30-J98)
- **Infectius:** Certain infectious diseases (A00-B99)
- **Digestive:** Diseases of the digestive system (K00-K93)
- **Other:** Diseases of the skin and subcutaneous tissue, musculoskeletal system and connective tissue (L00-M99); Diseases of the blood and blood-forming organs (D50-D89); Mental and behavioral disorders (F00-F99); Diseases of the nervous system and the sense organs (G00-G44, G47-H95)

Let x be the age, t the year and c a specific CoD. Given a certain age x, the aim of the model is to convert a value of lifespan inequality $e^{\dagger}_{0,t}$ into a list of CoD specific rates $m_{c,t}$ as follows:

$$log(\hat{m}_{c,t}) = \beta_c log(e^{\dagger}_{0,t}) + \delta_c \gamma + \varepsilon_c \tag{1}$$

Where: β_c is CoD-specific pattern of human mortality, δ_c the correction of mortality improvement over CoD, γ the parameter to be optimize and ε_c are the errors such that $\varepsilon_c \sim \mathcal{N}(\mu, \sigma^2)$. In the numerical experiment we use the following input data: $x \in \{0, ..., 100\}$, causes: $c \in \{Circ, Diab, ..., Resp\}$, year: $t \in \{1999, ..., 2013\}$.

The projected values of lifespan inequality and total mortality rate(m_t), which constitute the model's input, can be given by a certain extrapolation method or be the target values resulted by official forecasting (e.g. WHO).

In order to obtain the desired CoD mortality rates we define the following steps, starting from $m_{c,t}$ for a given age x, $e^{\dagger}_{0,t}$ and m_t as input:

- We estimate the slope, β_c, of the linear relation between the logarithmic transformation of lifespan inequality and the CoD specific rates over the observation time t. This is done by using the method of the least-squares, by minimizing the sum of squared residuals: $min \sum_c [log(m_{c,t}) - \beta_c log(e^{\dagger}_{0,t})]^2$;
- We estimate the parameter δ_c by computing the singular value decomposition (SVD) of the matrix of regression residuals, R, obtained in the previous step: $SVD[R] = PDQ^T$. Where D and Q are matrices of left and right singular vectors, and P is a diagonal matrix with singular values along the diagonal. The first term of the SVD, is used for obtaining the estimates of δ_c that can be interpreted as the adjustment of mortality improvement over CoD;
- We compute the mortality rates: $\hat{m}_{c,t} = exp\{\beta_c log(e^{\dagger}_{0,t}) + \delta_c \gamma\}$, where $\gamma = 0$

- Give the total level of mortality, m_t, we optimize the CoD mortality rates finding the optimal value of γ where $\sum_c |\hat{m}_{c,t} - m_t| = 0$.

4 Results

The model is applied to the USA data from 1999 to 2013. Figure 1 shows the model fitting with 95% confidence interval in the year 1999. In order to verify the model's

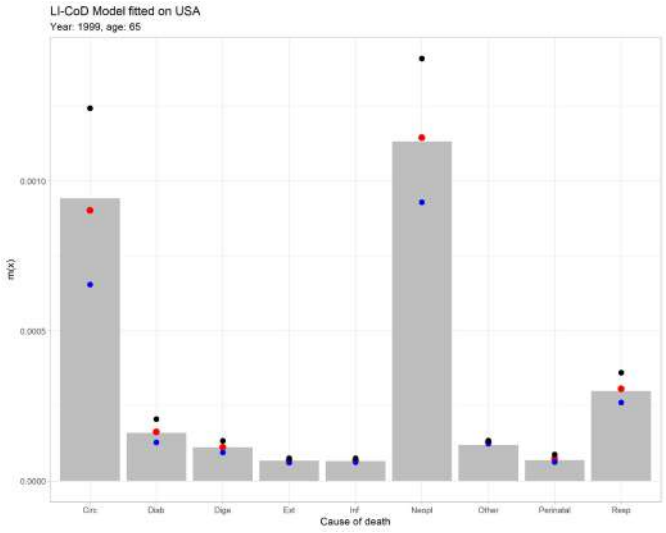


Fig. 1: Model fitting. Predicted values (black dots) and 95% confidence intervals (red and blue dots). Country USA. Year 1999.

accuracy, we perform a backtesting setting the fitting period to 1999-2010 and the forecasting period to 2011-2013. The forecasting results given by the backtesting are compared by the root mean squared error (RMSE) and the mean average error (MAE). Both the values of RMSE and MAE are very low, highlighting the accuracy of the model in predicting CoD mortality rates.

5 Conclusion

In this paper, we provide an innovative model to forecast CoD mortality, able to catch the hidden pattern and the relationship between two summary demographic measures, lifespan inequality and age-specific CoD mortality rates. Our model can have an important application in the context of incomplete data, when the official

Table 1: RMSE and MAE by CoD. Country USA. Years 2011-2013.

CoD	RMSE	MAE
Circolatory	0.000038511	0.000036751
Neoplasy	0.000026127	0.000024954
Diab	0.000007973	0.000007960
External	0.000001244	0.000001218
Perinatal	0.000006181	0.000006170
Respiratory	0.000005397	0.000004857
Infectius	0.000004339	0.000004266
Digestive	0.000002565	0.000002045
Other	0.000007631	0.000006738

registries just provide summary measures and incomplete data. We also suggest further development, modeling the CoD evolution using a Markov chain framework.

References

1. Aburto, J. M., Villavicencio, F., Basellini, U., Kjærgaard, S., and Vaupel J. W. (2020). Dynamics of life expectancy and life span equality. PNAS: 117(10): 5250–5259
2. Human Mortality Database 2018. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 01/01/2019. `https://www.mortality.org`.
3. Kjeargaard, S., Ergemen, Y., Kallestrup Lamb, M., Oeppen, J., Lindahl Jacobsen, R.: Forecasting causes of death by using compositional data analysis: the case of cancer deaths. Journal of the Royal Statistical Society, Series C **68**,5, 1351-1370 (2019)
4. Kannisto, V.: Measuring the Compression of Mortality. Demographic Research **3**, article 6 (2000)
5. Oeppen, J., Vaupel, J.W.: Broken Limits to Life Expectancy. Science **296** (5570): 1029–1031 (2002)
6. Oeppen, J., Vaupel, J. W.: The Linear Rise in the Number of Our Days. Social Insurance Studies, 3. The Linear Rise in Life Expectancy: History and Prospects (2006)
7. Riley J. Rising Life Expectancy: A Global History. Cambridge. Cambridge University Press (2001)
8. Rau, R., Soroko, E., Jasilionis, D., Vaupel, J. W.: Continued Reductions in Mortality at Advanced Ages. Population and Development Review, **34**: 747-768 (2008)
9. Shkolnikov, V.M., Andreev, E.M., Begun, A.Z.: Gini coefficient as a life table function: Computation from discrete data, decomposition of differences and empirical examples. Demographic Research **8** (article 11): 305–358 (2003)
10. van Raalte, A., Sasson, I., Martikainen, P.: The case for monitoring life-span inequality. Science, 1002-1004 (2018)
11. van Raalte, A.A., Caswell, H.: Perturbation analysis of indices of lifespan variability. Demography, **50**: 1615–1640 (2013)
12. Vaupel J. W.: The remarkable improvements in survival at older ages. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences **352**: 1799–1804 (1997)
13. Vaupel J. W.: How Change in Age-Specific Mortality Affects Life Expectancy. Population Studies **40**: 147–157 (1986)
14. Zuo, W., Jiang, S., Guo, Z., Feldman, M.W., and Tuljapurkar, S.: Advancing front of old-age human survival. PNAS **115** (44) 11209-11214 (2018)

Modeling Well-Being through *PLS-SEM and K-M*
Misurare il benessere con il modello PLS-SEM e K-Means

Venera Tomaselli[1], Mario Fordellone[2], Maurizio Vichi[3]

Abstract The concept of "fair and sustainable" well-being (BES) is employed to estimate the effects of 2008-2016 economic crisis on voting behaviour. Many studies suggest overcoming Gross Domestic Product as the only indicator to measure economic growth and development. So, in addition to strictly economic indicators, social welfare and performance of institutions, and public services indicators are used. With the aim to build a composite indicator of well-being, based on BES elementary indicators and domains, *Partial Least Squares-Structural Equation Modeling* and *K-means* clustering simultaneous method is employed to model the well-being classifying territorial micro-areas.

Abstract *Il benessere "equo e sostenibile" (BES) è impiegato per stimare gli effetti della crisi economica 2008-2016 sul comportamento di voto. Numerosi studi suggeriscono il superamento del prodotto interno lordo come unico indicatore per misurare la crescita e lo sviluppo economico. Pertanto, oltre ad indicatori strettamente economici, l'analisi tiene conto degli indicatori BES relativi al benessere sociale ed alle prestazioni delle istituzioni e dei servizi pubblici. Con l'obiettivo di costruire un indicatore composito di benessere, basato su indicatori elementari e domini del BES, il metodo simultaneo Partial Least Squares-Structural Equation Modeling e K-means clustering è utilizzato per l'analisi del benessere e la classificazione di micro aree territoriali.*

Key words: composite indicators, *Partial Least Squares-Structural Equation Models and K-means* method, well-being, territorial micro-areas

[1] Venera Tomaselli, Department of Political and Social Sciences, University of Catania IT; e-mail: venera.tomaselli@unict.it (*corresponding author*).
[2] Mario Fordellone, Department of Statistical Sciences, University "La Sapienza", Rome IT: e-mail: mario.fordellone@uniroma1.it.
[3] Maurizio Vichi, Department of Statistical Sciences, University "La Sapienza", Rome IT: e-mail: maurizio.vichi@uniroma1.it.

1 Introduction

The concept of well-being is very complex and takes into account economic resources, public policies, quality of the environment, and many other topics. Well-being indicators, indeed, are proxies of unobservable, latent dimensions.

By questioning the use of Gross Domestic Product (GDP), the study aims to measure well-being through a multidimensional approach on the basis of the relationships among elementary indicators and domains proposed in the analysis of well-being through BES in order to define the territorial differences of well-being more reliably.

The indicators have been selected from 12 BES domains, grouped in turn into 6 topics: economy and labour market, education, environment, institutional performance, safety and quality of services. The crucial issues are:

- is the economic well-being level the only useful and meaningful measurement to explain the variability of territorial micro-areas?
- are indicators closely related to economic well-being and indicators of well-being related, for instance, to the quality of services and opportunities available in the territory, do not necessarily are correlated?
- is the well-being a composite concept based on a scaling of specific priorities?

In the following section 2, the well-being measurement is discussed; in section 3 the PLS-SEM and K-Means method is presented and in sections 4 the results obtained simultaneously through well-being composite indicator and classification method for the territorial micro-areas are delivered.

2 Well-being measurement

The measures of well-being, economic progress, and social welfare are adopted as drivers for designing public policies by decision makers (Jayawickreme et al., 2012; Layar, 2011; Sachs, 2012). The measures more accurately depict changes not only in individual living standards (Helliwell et al., 2012) but simultaneously also in comprehensive national economic growth (Diener et al., 2009).

Since around 2000 the Organisation for Economic Cooperation and Development (OECD) embarked on a global project to measure the well-being and progress of societies not just and not only through the economic performances. The project was involved in setting up and supporting the Commission on the Measurement of Economic Performance and Social Progress (CMEPSP), established by the President of France, Nicolas Sarkozy in 2008 and led by Stiglitz, Sen and Fitoussi.

The limits of GDP are reviewed in the report of the Commission because the GDP is not believed as a standard of the well-being of societies. The GDP, indeed, does not address economic inequality, happiness, quality of life, wellness, and other crucial societal parameters, and does not integrate environmental services into economic decisions (Stiglitz et al., 2010).

More recently, other scholars (Ven, 2015; Fleurbaey, 2015) have called for a new generation of multifaceted and more comprehensive well-being measures, better able to describe actual living standards and useful for a more accurate design of policies improving efficiency in resources assignment.

A series of measures of well-being, inspired at the Nussbaum-Sen approach to human capabilities and subjective well-being (Nussbaum and Sen, 1993), have been proposed in attempt to go beyond GDP with the aim to broaden the scope of effects in the assessment of policies. For instance, the Human Development Index by UNDP or the Better Life Initiative launched by OECD (2015) and many other approaches are based on the income, health, and education measurements of the countries' performance (for a review, see Fleurbaey and Blanchet, 2013).

From a technical view, many methods are employed to measure the well-being level through composite indicators. Nevertheless, no method is universally valid to select indicators based on theory-driven criteria, measure properly the concept, aggregate and normalise a set of input variables and define a weighting system (OECD, 2008).

The aim is to simplify the analysis of the multidimensional concept according with a formative or reflective measurement model, where elementary indicators are causes or effects of latent variables, respectively (Michalos, 2014; Simonetto, 2012).

Only some studies have focused on the construction of well-being composite indexes to evaluate and compare well-being specifically across the Italian provinces. Mazziotta and Pareto (2019) have obtained a global well-being index by aggregating 11 composite indices with AMPI (Mazziotta and Pareto, 2016) and have ranked Italian provinces for each dimension of well-being and have given an overall ranking.

Also Calcagnini and Perugini (2019) have proposed a composite indicator of well-being for the Italian provinces (NUTS-3) based on the methodology of the regional Index of Regional Quality of Development (QUARS) to analyse the extent to which the socio-economic heterogeneity in individual and contextual features within region affects the well-being among adjacent provinces.

In Italy, indicators of well-being have being used more and more for policy-making reasons at national, but also regional or local levels involving public institutions. From a theoretical point of view, the relationship of well-being assessment with policy-making process in healthcare, education, and training, or local services is the rationale to analyse the well-being measures at local level.

Since the policies of local governmental authorities have a direct and huge impact on the socio-economic context, the assessment of living standards at provincial level allows to evaluate economic, environmental, and social needs of the citizens at any level of government in order to implement and design decentralised policies. Following this, for territorial micro-areas the economic dimension could not be so crucial.

3 PLS–SEM-KM for composite indicator building

Partial Least Squares (PLS) methodologies are algorithmic tools with analytic proprieties aiming at solving problems about the stringent assumptions on data, e.g., distributional assumptions that are hard to observe in real life.

Tenenhaus et al. (2005) try to better clarify the terminology used in the PLS field through an interesting review of the literature, focusing the attention on the Structural Equation Models standpoint.

Given the $n \times J$ data matrix $\mathbf{X}$, the $n \times K$ membership matrix $\mathbf{U}$, the $K \times J$ centroids matrix $\mathbf{C}$, the $J \times P$ loadings matrix $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{L}}]$, and the errors matrices $\mathbf{Z}$, $\mathbf{E}$ and $\mathbf{D}$, the PLS-SEM-*K*M model simultaneously identify a SEM model for variables and a partitioning *K*M model for units according to the following three equations:

$$\begin{aligned} \mathbf{H} &= \mathbf{H}\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Xi}\boldsymbol{\Gamma}^{\mathrm{T}} + \mathbf{Z}, \\ \mathbf{X} &= \mathbf{Y}\boldsymbol{\Lambda}^{\mathrm{T}} + \mathbf{E} = \boldsymbol{\Xi}\boldsymbol{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{E}, \\ \mathbf{X} &= \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} = \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}_{\mathrm{H}}\boldsymbol{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}_{\mathrm{L}}\boldsymbol{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{D} \end{aligned} \tag{1}$$

subject to constraints: *(i)* $\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda} = \mathbf{I}$; and *(ii)* $\mathbf{U} \in \{0,1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Thus, the PLS-SEM-KM model includes the PLS-PM and the clustering *K*M equations (i.e., $\mathbf{X} = \mathbf{U}\mathbf{C}$ and then, $\mathbf{Y} = \mathbf{X}\boldsymbol{\Lambda}$ becomes $\mathbf{Y} = \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}$).

The proposed methodology shows some important advantages with respect to the other proposed approaches for both cluster analysis and composite indicator construction: it is a new simultaneous approach.

In fact, identifies the best homogenous partition of the objects, represented by the best statistical relationships among latent and observed variables (Fordellone and Vichi, 2018; 2020).

4 Results: well-being composite indicator

The conceptual structure (Giovannini et al., 2012) of the BES considers 9 domains related to aspects that directly influence well-being, plus 3 instrumental or context domains.

In each well-being domain (ISTAT, 2018), elementary indicators and a synthesis through composite indicators related to each domain are integrated.

Our dataset consists in 109 units (Italian provinces) and 16 indicators organized in 9 different domains. Table 1 shows the relationships among MVs and LVs and the loadings estimated through PLS-SEM-KM.

Through the application of PLS-SEM-KM model, we have identified 3 homogeneous groups of Italian provinces. The optimal number of clusters identified corresponds to the maximum value of the pseudo-*F* function (around 1.1).

From this analysis, we can see that the theoretical polarity associated to each observed variable (see Table 1) is well described by the measurement PLS approach.

Whereas only the *Overcrowding of prisons* variable shows a non-significant loading, in fact also the sign is not correct.

Table 1: Measurement models matrix estimated through PLS-SEM-KM.

MVs	LV1	LV2	LV3	LV4	LV5	LV6	LV7	LV8	LV9	WB
MV1	0.795	0	0	0	0	0	0	0	0	0.721
MV2	0	0.750	0	0	0	0	0	0	0	0.619
MV3	0	-0.678	0	0	0	0	0	0	0	-0.917
MV4	0	0	-0.876	0	0	0	0	0	0	-0.938
MV5	0	0	-0.887	0	0	0	0	0	0	-0.881
MV6	0	0	0.920	0	0	0	0	0	0	0.870
MV7	0	0	-0.917	0	0	0	0	0	0	-0.914
MV8	0	0	0	-0.669	0	0	0	0	0	-0.569
MV9	0	0	0	0	0.050	0	0	0	0	0.202
MV10	0	0	0	0	0	-0.373	0	0	0	-0.417
MV11	0	0	0	0	0	-0.216	0	0	0	-0.140
MV12	0	0	0	0	0	-0.434	0	0	0	-0.412
MV13	0	0	0	0	0	0	0.105	0	0	0.226
MV14	0	0	0	0	0	0	0	0.277	0	-0.088
MV15	0	0	0	0	0	0	0	0.691	0	0.704
MV16	0	0	0	0	0	0	0	0	0.748	0.822

In Table 2 the path coefficients' matrix of the estimated structural-PLS model is shown.

Table 2: Structural model matrix estimated by PLS-SEM-KM.

	LV1	LV2	LV3	LV4	LV5	LV6	LV7	LV8	LV9	WB
LV1-Health	0	0	0	0	0	0	0	0	0	0.324
LV2-Education and training	0	0	0	0	0	0	0	0	0	0.053
LV3-Work-life balance	0	0	0	0	0	0	0	0	0	0.666
LV4-Economic well-being	0	0	0	0	0	0	0	0	0	0.247
LV5-Policy	0	0	0	0	0	0	0	0	0	0.117
LV6-Safety	0	0	0	0	0	0	0	0	0	0.163
LV7-Cultural heritage	0	0	0	0	0	0	0	0	0	0.067
LV8-Environment	0	0	0	0	0	0	0	0	0	0.192
LV9-Innnovation and research	0	0	0	0	0	0	0	0	0	0.198
Well-being composite indicator	0	0	0	0	0	0	0	0	0	0

From the structural model we can see that the Italian provincial well-being is highly affected by the LV3-*Work-life balance* construct (0.666), followed by LV1-*Health* (0.324) and LV4-*Economic well-being* (0.247) constructs. Whereas, the LV6-*Cultural heritage* (0.067) and LV2-*Education* (0.053) constructs are

dimensions with non-significant coefficients. The overall fit of the structural model is good with a $R^2 = 0.74$. In terms of clustering results, the three clusters identified by the PLS-SEM-KM algorithm describe three groups of well-being, i.e., *high*, *medium*, and *low* levels. In this way is very easy to classify the Italian provinces through a "BES ranking".

5 Conclusions

The BES 2018 report by ISTAT has confirmed that in 2015 and in 2016 an improvement in many areas of well-being has been observed, even if territorial differences remain stable both in levels and dynamics.

The differences appear in some cases as real structural differences between North and South of Italy. Then, the different well-being levels measured through a composite well-being indicator in the territories allow to explain the variability in territorial micro areas.

References

1. Diener, E., Lucas, R., Schimmack, U., & Helliwell, J. (2009). Well-Being for public policy. Oxford, UK: Oxford University Press. doi.org/10.1093/acprof:oso/9780195334074.001.0001.
2. Fleurbaey, M. (2015). Beyond income and wealth. Rev. Income Wealth, 61(2), 199-219.
3. Fleurbaey, M., & Blanchet, D. (2013). Beyond GDP: Measuring welfare and assessing sustainability. Oxford University Press.
4. Fordellone, M., & Vichi, M. (2018). Structural Equation Modeling and simultaneous clustering through the Partial Least Squares algorithm. arXiv preprint: arXiv:1810.07677v1.
5. Fordellone, M., & Vichi, M. (2020). Finding groups in structural equation modeling through the partial least squares algorithm. Comput Stat Data An doi: https://doi.org/10.1016/j.csda.2020.106957
6. Giovannini, E., Morrone, A., Rondinella, T., & Sabbadini, L. L. (2012). L'iniziativa CNEL-ISTAT per la misurazione del Benessere Equo e Sostenibile in Italia. Autonomie locali e servizi sociali, 1, 125-136.
7. Helliwell, J. F., Layard, R., & Sachs, J. (2012). Some policy implications. In J. F. Helliwell, R. Layard & J. Sachs (Eds.) World Happiness Report (pp. 90-107), New York, NY: The Earth Institute, Columbia University
8. ISTAT (2018). BES 2018: Il benessere equo e sostenibile in Italia. Istituto Nazionale di Statistica.
9. Jayawickreme, E., Forgeard, M. J., & Seligman, M. E. (2012). The engine of well-being. Rev Gen Psychol, 16(4), 327-342. doi.org/10.1037/a0027990.
10. Mazziotta, M., & Pareto, A. (2016). On a generalized non-compensatory composite index for measuring socio-economic phenomena. Soc Indic Res, 127(3), 983-1003.
11. Mazziotta, M., & Pareto, A. (2019). Use and misuse of PCA for measuring well-being. Soc Indic Res, 142(2), 451-476.
12. Nussbaum, M. & Sen, A. (1993). The Quality of Life. Oxford: Oxford University Press.
13. OECD (2015). How's life? 2015 Measuring well-being. Paris, FR: OECD Publishing. (Available online: http://dx.doi.org/10.1787/9789264121164-en).
14. Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2010). Mismeasuring our lives: why GDP doesn't add up. New York, NJ: The New Press.
15. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., & Lauro, N. C. (2005). PLS path modelling. Comput Stat Data An, 48(1), 159-205.
16. Ven, P. (2015). Introduction to the symposium on new measures of well-being: Perspectives from statistical offices. Rev Income Wealth, 61(1), 1-3.

News life-cycle: a multiblock approach to the study of information

Il ciclo di vita delle notizie: un approccio multiblock allo studio dell'informazione

Rosanna Cataldo, Marco Del Mastro, Maria Gabriella Grassia, Marina Marino and Rocco Mazza

Abstract Media are instruments with a high capacity to penetrate inside the socio-cultural fabric. In this paper we examine the informative function of media, focusing in particular to its products the news. The aim of this paper is to study the life cycle of a collection of news, inside a defined medial ecosystem. In particular, we analyze two levels of the medial market, the product (the news), and the channel (the sources). The methodology we chosen is the multiple factor analysis (AFM), which allows us to scan multiple data tables that measure sets of variables collected on the same observations.

Abstract *I media sono strumenti con un'elevata capacitá di penetrazione all'interno del tessuto socio-culturale. In questo contributo prendiamo in esame la funzione informativa dei media, soffermandoci in particolare sui prodotti: le news. Lobiettivo del paper é quello di studiare il ciclo di vita di un collettivo di notizie allinterno di un ecosistema mediale delimitato. In particolare esaminiamo il ciclo di vita delle notizie selezionate applicando un metodo che ci permette di analizzare due livelli del mercato mediale, il prodotto (le notizie) e il canale (le fonti). Il metodo utilizzato é lanalisi fattoriale multipla (AFM), la quale permette di analizzare datasets composti da piú blocchi di variabili raccolte sulle medesime unitá statistiche.*

Key words: Multiple Factor Analysis, News, Media

Maria Gabriella Grassia
University of Naples Federico II e-mail: mgrassia@unina.it

Marco Del Mastro
AGCOM e-mail: m.delmastro@agcom.it

Marina Marino
University of Naples Federico II e-mail: marina.marino@unina.it

Rosanna Cataldo
University of Naples Federico II e-mail: rosanna.cataldo2@unina.it

Rocco Mazza
University of Naples Federico II e-mail: rocco.mazza@unina.it

1 Introduction

News are a product of media in the information market. The aim of this work is to study the life cycle of news inside a defined medial ecosystem of a collection of news. Media are instruments with a high capacity to penetrate inside the socio-cultural fabric. The versatility and adaptability which characterizes them allow to easily adjust to the reference markets orientation. They actually process any genre and depending on their peculiarities and the degree of interaction they allow to realize, they are used to carry out a variety of functions: entertainment, information, promoting sales of both products and services, job training, sharing and social relations too. Its our aim to analyze the informative function in particular: this is fundamental due to the fact that media offer to its public products aimed at the satisfaction of their consumption need, and subsequently create new orientations and ideas (Strmberg, 2004; Anderson, McLaren 2012; Larcinese, 2012; Della Vigna, Kaplan, 2007; Bernhardt et Al., 2008; Chiang, Knight, 2011). We distinguish between hard and soft news: with hard news we refer to news used to describe topics that are usually urgent, important and consequential, like politics, international affairs, and economic news (Reuters Istitute, 2016); Soft news concern entertainment, celebrities and lifestyle (Mills-Brown, 2014). This distinction is essential because the object of our research is hard news. Our contribution to the state of the art is the description of the cycle of life of the news selected on the basis of a method that will allow us to analyze two levels of the medial market, news and sources. The following pages show the work organized as such: In the section (2) we introduce the different modes of gathering and selecting data; in section (3) we propose our methodological approach and the design of our research; and in section (4) we feature the main results and conclusions.

2 Methodological approach

The data-gathering used for the analysis is divided in two phases: 1. firstly, we selected the most popular news to include in the collection; 2. Secondly, we selected the sources and we extracted the frequency of transmission of each singular news from these. In the first phase we used the tool Google Trends in order to identify the news to analyze and gather for the collection: this choice is made in order to align our paper to and give a contribution to the Web Search Studies (Zimmer, 2010; Roberg, 2013; Trevisan, 2014), and also for counting on the great potentialities of this big data tool. Using this tool, we identified, in a time span that goes from January 2018 to July 2018, the 40 news with the highest search frequency each month. We divided these in equal parts between 8 categories of topics: 1. News report; 2. Culture; 3. Economy; 4. Foreign news; 5. Politics; 6. Sciences; 7. Entertainment; 8. Sport. This phase produced a list of 280 news. In the second phase a we used a tool to extract the frequency transmission of news through sources during the referred time span. This tool is Volocom, a web platform that allowed us to use a suite of services

for tracking media. Thanks to this new platform it is possible to search a specific news in query and receive as response its frequency inside one or more designated sources. In order to be effective, research needs two parameters: the selection of sources and the period of reference. As far as it concerns sources, the selection takes into account the main channels that produce news and both off and on-line environments: News agency, Facebook Radio, Facebook News Agency, Twitter Radio, Twitter NewsAagency, TV, Newspapers, Facebook TV, Facebook Newspapers, Twitter tv, Twitter Newspapers, Native digital newspapers, Radio, Facebook Native digital newspapers, Radio site, Twitter Native digital newspapers. In order to describe the news life cycle, it is essential to define the time span that describes transmission. This time frame has been operated with a division in three phases: 1. incubation of the news the 15 days in which the latter starts building up; 2. the moment in which the news is launched, named to; 3. Spreading of the news, the 30 days in which the news travels across different media. Overall, of each news we extracted its frequency during 46 days; t_0, being the fundamental moment of the recording, deserves a series of clarifications. It can be absolute and coincide exactly with the event transmitted by the first subject who spread the news registered by Volocom; or it may be the case that the zero-time its not absolute, but relative to a news that has been covered previously. In this case we refer to the very first moment the subject launched the news as it has been subsequently spread. We queried Volocoms database using API REST interface, programming a cycle of queries in Python. Once we have finished the gathering procedure, we built the data matrix assembling each news on one single string and putting in a column informative sources with the related 46 days. Therefore, the dataset is composed by 16 blocks of columns, one for each source. The methodology chosen is multiple factorial analysis (MFA), that allows to analyze multiple data tables that measures sets of variables collected on the same observation. These methods allow studying the relationship between the observations, the variables, and tables (Escofier and Pags, 1984). This is a generalization of Principal Component Analysis (PCA) and a part of the multitable (or multiblock) PCA family (Smilde et al, 2003). The technique allows: 1. To analyse some variables data tables measured on the same observations; 2. to compute a set of common factor scores; 3. to plot each of the original data sets into a new space to analyse communalities and discrepancies (Abdi et al, 2013). The method breaks up into two phases:

1. In the first phase we have N tables and for each P_n variables collected. First we compute generalized PCA on each of the N tables and extract the first singular value for each table (which is the square root of an eigenvalue). Successively we divide each table by its first singular value, with this procedure we normalize the tables (Abdi et al, 2007).
2. Starting from the concatenated normalized tables, the procedure compute a generalized PCA.

Table 1 Sources' absolute contributions

Contributions (%):			Contributions (%):		
	F1	F2		F1	F2
Q	**6,8798**	1,1822	FB-R	1,6876	**6,8627**
AgI	**8,8171**	2,8645	FB-TNatD	**8,6566**	3,0211
R	**5,1794**	0,4479	TW-Q	8,6365	**9,1357**
TNatD	**8,1954**	2,9396	TW-R	0,0185	**4,0089**
TV	**5,4857**	0,5213	TW-TNatD	6,2208	**9,6907**
FB-TV	5,2070	**5,5016**	TW-TV	4,8987	**13,4934**
FB-AgI	5,4744	**15,8786**	TW-AgI	6,2783	**16,4564**
FB-Q	**9,1370**	7,1038	TW-Infl	**9,2272**	0,8915

3 Results

The reading of the correlations between the variables representing the 46 days examined, and the factors extracted actually help us to create a description of the life cycle of these news. The results that follows are partial and related to a selection of the analyzed sources that will help us to create a division based on types useful for the reading of other channels too. Newspapers show a life cycle in which the various moments are positioned around the positive semi-axis of the first factor, with a distribution spread between the second and fourth sector; Time-zero presents a positive correlation increasing with the first dimension compared to the second; the relations we can read closer to the point help to characterize a type of information that tends to extend in the time span. Communication agencies are another source worthy of study: in this case the variables form two distinct blocks one flattened on the first factor with time zero and the following moments, and another one on the second factor formed by the variables preceding the moment in which the news are launched. The variables furthest from date t_0 are located near the origin, meaning a type of communication that is fast, wearing out rapidly around the point zero. About the social networks sources, in the figure there are the correlations between the moments registered regarding the Twitter official accounts of the information agencies. The correlations of the first axis refer to the days following the emission, while on the second we find the previous days. The most significant points with the highest correlation consist of the moments positioned the closest to t_0. Its evident that in this case the profile that emerges for the previous source its even more defined, the life cycle is sharply divided between the two moments preceding and following the breaking of the news. At the same time the latter tends to wear out in the days closest to its emission, without a period of incubation and insight Now lets take a look at the relations established between the sources. The analysis has shown that the absolute contributions in table 2. On the basis of such values it results that the sources that are involved the most in the defining of the first axis are daily newspaper, communication agencies, radio, digital native publications, TV,

Fig. 1 Correlation plot: a. Newspapers; b. News Agency; c. Twitter News Agency

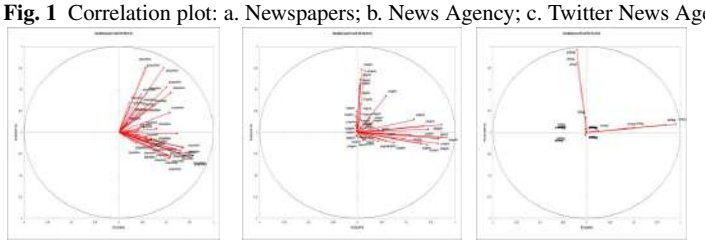
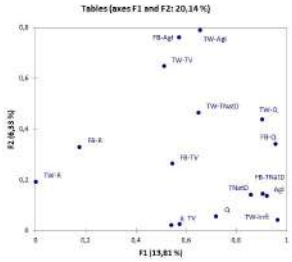


Fig. 2 Sources' coordinates:



Facebook newspaper, FB digital native publications, TW influencer. While for the second axis the most fruitful sources are FB TV, FB communication agencies, FB Radio, TW newspapers, TW Radio, TW digital native publications, TW TV, TW communication agencies. On the first dimension the information more programmed and structured corresponding to the traditional sources is followed by the digital native publications with the related FB pages and TW influencers. Whereas the second dimension is more coherent, characterized exclusively by social sources. The chart displayed represents the projection of sources on new factorial dimensions, the points that contribute the most to the first dimension are distributed more compactly, while the social sources on the second dimension are more distant between them. This refers a weaker relation between the latter and greater variety between distributions.

4 Conclusion

Going back to the aims of this research, the purpose was to describe the life cycle of news through their journey across the chosen sources. Applying MFA it has been possible to recognize three different kind of cycles, basing on the relations established between each moment, and pair them with different channels:

- Incubator sources, these are characterized by life cycles that tend to extend toward the extremes of the registered time period. The news lives a period of in-

cubation and subsequently is launched on the market where it lasts for the great part of the selected time frame. Sources that can be gathered under this category are the traditional ones (newspaper, TV, radio), with a rich media service offer suitable to manage the different life phases of the news.

- Trigger sources, that launch the news on the market but exert over it a limited control in a brief time span. These sources lack a period of incubation and the tail of the moments following the emission in the market its slightly extended. The present picture is similar to the news agency.
- Flash sources, these cycles are very fast and focused on the moment of the broadcasting of the news. In this case the channel launches the news that is immediately absorbed, and it shortly dissolves inside the network of users. The social sources manage cycle of this kind.

References

1. Escoer B, Pags J.: Multiple factor analysis. Comput Stat Data Anal. **18**:121140 (1990)
2. Smilde AK, Westerhuis JA, de Jong S.: A framework for sequential multiblock component methods. J Chemometr. **17**:323337 (2003)
3. Abdi H., Valentin D.: Multiple factor analysis. In: Salkind NJ, ed. Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage; 657663 (2007)
4. Abdi H., Williams L. J., Valentin D.: WIREs Comput Stat, 5:149179. doi: 10.1002/wics.1246 (2013)
5. Strmberg, D.: Mass Media Competition, Political Competition, and Public Policy, Review of Economic Studies, **71**, pp. 265-284. (2004)
6. Gentzkow, M.A., Shapiro, J.M.: Media Bias and Reputation, Journal of Political Economy, **114**: pp. 280-316 (2006)
7. Anderson, S.P., McLaren J.: Media Mergers and Media Bias with Rational Consumers, Journal of the European Economic Association, **10**: pp. 831-859 (2012)
8. Larcinese, V.: The Instrumental Voter Goes to the News-Agent: Information Acquisition, Marginality, and the Media, Journal of Theoretical Politics, **19 (3)**: pp. 249-276 (2007)
9. Della Vigna, S., Kaplan, E.: The Fox News Effect: Media Bias and Voting, Quarterly Journal of Economics, **122**: pp. 1187-1234 (2007)
10. Bernhardt, D., Krasa, S. and Polborn, M. K.: Political Polarization and the Electoral Effects of Media Bias, Journal of Public Economics, **92**: pp. 1092-1104 (2008)
11. Chiang, C.F., Knight, B.: Media Bias and Influence: Evidence from Newspapers Endorsements, Review of Economic Studies, **78 (3)**: pp. 795-820 (2011).
12. Reuters Istitute, Distinct between Hard and Soft News, http://www.digitalnewsreport.org/survey/2016/hard-soft-news-2016/ (2016).
13. Mills-Brown L.: Soft News. Enciclopedia Britannica, https://www.britannica.com/topic/soft-news-ref1198559.
14. 42Zimmer M: Web Search Studies: Multidisciplinary Perspectives on Web Search Engines. In: Hunsinger J., Kalstrup L., Allen M., International Handbook od Internet Research, pp. 507-521 (2010)
15. 42Roberg R.: Digital Methods, MIT Press, Cambridge (MA) (2013).
16. 42Trevisan F.: Search Engines: From social science objects to academic inquiry tools, in First Monday, **19 (11)** (2014).
17. 27Broersma M., Graham T.: Twitter ad a news source. How Dutch and British newspapers use tweets in their news coverage, in Journalism Practice, **7 (4)**: pp. 446-464
18. WIREs Comput Stat. 5:149179. (2013) doi: 10.1002/wics.1246

Short-term rentals in a tourist town

Affitti brevi in una città turistica

Silvia Bacci, Bruno Bertaccini, Gianni Dugheri, Paolo Galli, Antonio Giusti, Veronica Sula

Abstract Short-term rents of rooms or apartments in the most important tourist cities have significantly increased in the last decade, also thanks to spread of dedicated smartphone apps and web platforms. These types of accommodation now constitute a significant part of the tourist accommodation facilities, however this under-regulated development has introduced some advantages but also some drawbacks, making it necessary to get an exact measure of this phenomenon. Official statistics are quite lacking in this sector, so we used, as alternative sources of information, two of the main short-term rentals brokerage platforms (AirBnB and Booking) to estimate the extent of the phenomenon in Florence.

Abstract *Gli affitti a breve termine di camere o appartamenti nelle più importanti città turistiche si sono sviluppati considerevolmente nell'ultimo decennio, anche grazie a Internet. Questi tipi di alloggi costituiscono ora una parte significativa dell'offerta ricettiva, ma questo sviluppo sotto-regolamentato presenta oltre a vantaggi anche alcuni inconvenienti. È importante avere una misurazione precisa dei numeri coinvolti per comprendere meglio il fenomeno, anche in vista di una regolamentazione più precisa. Poiché le statistiche ufficiali sono carenti, abbiamo cercato di utilizzare, quali fonti informative alternative, due tra le principali piattaforme di intermediazione web (AirBnB e Booking) per cercare di valutare le dimensioni del fenomeno a Firenze.*

Key words: Tourism, accommodation facilities, short-term rentals, web platforms.

[1] Silvia Bacci, University of Florence; email: silvia.bacci@unifi.it
Bruno Bertaccini, University of Florence; email: bruno.bertaccini@unifi.it
Gianni Dugheri, Florence Municipality; email: gianni.dugheri@comune.fi.it
Paolo Galli, University of Florence; email: paolo.galli@stud.unifi.it
Antonio Giusti, University of Florence; email: antonio.giusti@unifi.it
Veronica Sula, University of Florence; email: veronica.sula@stud.unifi.it

1 Introduction

Short-term rents of rooms or apartments were always existed. In recent years, however, the phenomenon has developed considerably both due to the increasing spread of hit and run tourism and for the emergence of internet platforms (OTA: online travelling agency), such as Airbnb, Booking, etc., which allow easy and rapid intermediation (Griswold, 2016).

At first glance, this under-regulated development may seem a positive fact, at least from an economic point of view, but it has also many drawbacks. First of all, the competition with the traditional accommodation businesses whose managers consider private rentals as unfair competition, given the lower costs and regulatory facilities faced by this type of offer. In addition, there is the issue of the increasing tourist pressure, especially in the central districts of the most important cities, the reduction of the availability of housing for long-term rentals for residential purposes with the consequent depopulation of urban centres, and the increase in the costs of municipalities for services such as urban garbage, water supply, surveillance, etc. (Nieuwland and van Melik, 2018).

To date, these types of accommodation now form a significant part of tourist presence. This requires an estimate of the number of rentals and beds for rent to better understand the real accommodation capacity of a generic tourist location also in view of its more precise regulation. Unfortunately, official statistics are rather lacking in this area; so we decided to use the data highlighted on the sites of two of the most important brokerage platforms to get a more precise idea of the true level of this phenomenon.

Referring to the city of Florence, to collect data from brokerage platforms we used two different strategies. A first method was to consult the Booking website (2020) to collect, in addition to the reviews of the customers on stays, information on the location and other details of the apartments used and the main data about the stay. A different strategy was used for AirBnB website (2020) for which there is a site, Insideairbnb (2020), that collects in-depth information on the whole rental listing published on Airbnb website.

With reference to the month of November 2019, we collected 2585 announcements from Booking on 1 or more apartments (although the vast majority of registrations concern only one apartment), while from Insideairbnb we found 8136 listings. In this case, each registration concerns only one apartment.

The two data sources are partially overlapping, but for the moment it has not been possible to determine the common cases to create a single archive, both because the presentation criteria of the accommodation facilities are very different between the two sources and because the geographical coordinates present in the files are partially altered. For this reason, we used the two data sources separately in our analyses.

2 The accommodation capacity of the Florence municipality

According to official sources, “Tourism statistics” period July-September 2019 (Metropolitan City of Florence, 2020), the accommodation offer in the Florence municipality is made up of 20.9% of hotel facilities and the remainder 79.1% of extra-hotel facilities. However, hotels collect more than 60% of rooms (63.7%) and beds (63.5%). Table 1 provides details about the “official” accommodation offer disentangled by hotel category and type of extra-hotel facilities.

The short-term rent offer represented by guesthouses and non-professional guesthouses captures the 64.5% of the offer of the extra-hotel facilities, rising to 69.9% with historical residences, corresponding to the 40.3% (47.3% with historical residences) of rooms and 36.0% (44.3% with historical residences) of beds. Overall, guesthouses and non-professional guesthouses represent the 51.0% (55.0% with historical residences) of the overall accommodation structures (i.e., hotel and extra-hotel facilities), with only the 14.6% of rooms and 13.1% of beds.

Table 1: Hotel and extra-hotel facilities in the Florence municipality, by number of structures, number of rooms, number of beds, number of bathrooms, period July-September 2019 (Source: Tourism statistics - Metropolitan City of Florence)

Type of structure	***Structures***	***Rooms***	***Beds***	***Bathrooms***
Hotel facilities	390	14853	32926	15640
1 star	*39*	*402*	*845*	*292*
2 stars	*69*	*1060*	*2246*	*1083*
3 stars	*146*	*4077*	*9201*	*4269*
4 stars	*115*	*8089*	*17375*	*8620*
5 stars	*20*	*1213*	*3240*	*1363*
Hotel-tourism residences	*1*	*12*	*19*	*13*
Extra-hotel facilities	1477	8469	18900	7641
Farmhouses	*8*	*44*	*85*	*36*
Guesthouses	*701*	*2839*	*5641*	*2848*
Holiday homes	*37*	*1008*	*1738*	*948*
Holiday dwellings (rented)	*353*	*1312*	*3106*	*1284*
Tourist camp-sites	*3*	*909*	*2290*	*307*
Youth hostels	*18*	*607*	*2034*	*553*
Residence	*25*	*586*	*1276*	*549*
Non-professional guest houses	*252*	*576*	*1160*	*484*
Historical residence	*80*	*588*	*1570*	*632*
Total	**1867**	**23322**	**51826**	**23281**

Although these surveys at provincial level contribute to the production process of the statistical data that ISTAT disseminates at national level, some categories that are part of the "extra-hotel facilities” (i.e., other collective accommodation establishments) are not taken into consideration at this last level (I.Stat, 2020).

However, looking at data about short-term rents coming from Airbnb, the overall situation seems clearly underestimated. In Figure 1 we show the distribution of the short-term rent facilities across the Florence municipality, by census area: the top

panel displays data coming from Booking, whereas the bottom panel refers to data coming from Insideairbnb. The census areas not marked on the maps are sparsely populated non-urban areas.

The two maps show a certain level of overlapping, as already mentioned in the introduction, however data coming from Insideairbnb include a higher number of facilities. In more detail, Insideairbnb took over 8136 active listings that offer more than 18000 beds all around the Florence municipality. Further details about the territorial distribution of the facilities disentangled by census area are provided in Table 2.

The short-term rental facilities are distributed in a very heterogeneous way across the territory (Figure 1 and Table 2). The highest density of structures is observed in the UNESCO area, corresponding to census areas 12, 14, and 13. In particular, census area 14 (where the most famous historical monuments are concentrated, such as Brunelleschi's Dome and Palazzo della Signoria) collects 1241.2 structures per km^2, corresponding to just less than 3000 beds per km^2. Census area 12 follows with a density of 866 structures per km^2 and more than 2000 beds per km^2. Overall, the short-term rental facilities are quite small, with an average of 2.2 beds per structure.

Table 2: Short-term rental facilities, by census area: density of facilities (per km^2), density of beds (per km^2), ratio beds/facilities; period November 2019 (Source: Insideairbnb)

Census area	***Density of facilities***	***Density of beds***	***Ratio beds/facility***
11	105.3	214.1	2.0
12	866.0	2061.4	2.4
13	370.1	815.7	2.2
14	1241.2	2937.1	2.4
21	150.6	299.1	2.0
22	57.0	120.4	2.1
23	39.8	91.1	2.3
24	57.9	122.9	2.1
25	32.3	58.8	1.8
31	30.5	61.4	2.0
32	20.3	49.0	2.4
41	13.5	28.3	2.1
42	11.1	29.8	2.7
43	61.5	146.6	2.4
44	84.7	191.2	2.3
51	26.0	55.5	2.1
52	156.3	335.9	2.1
53	14.1	28.1	2.0
54	9.0	19.4	2.2
55	53.4	110.3	2.1
56	26.2	58.8	2.2
Total	**163.2**	**373.1**	**2.2**

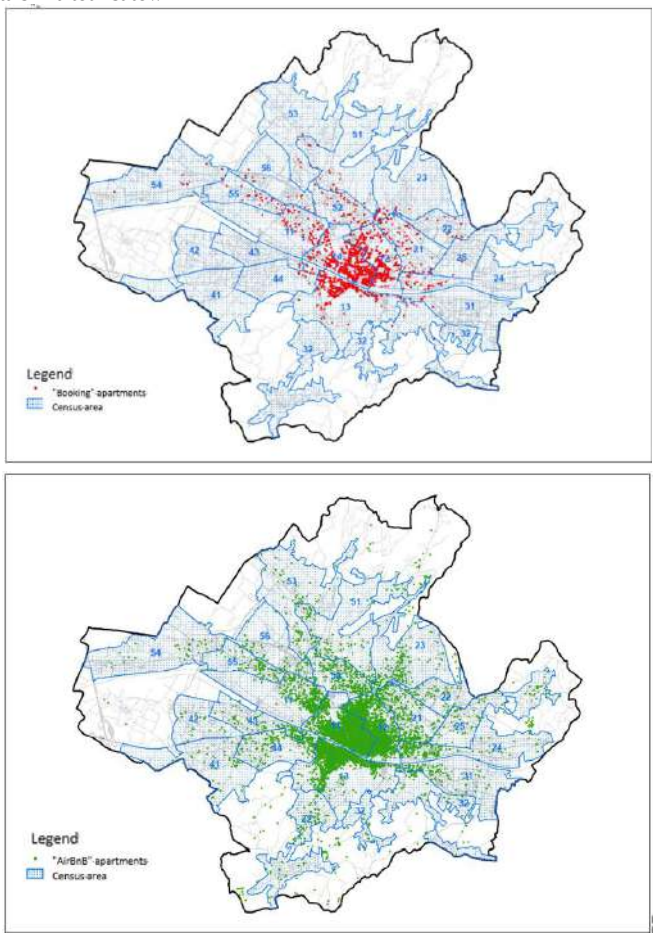


Figure 1: Map of the short-term rental facilities in the Florence municipality: source Booking (top) and Insideairbnb (bottom).

3 Some final remarks

This work is ongoing, but the relevance of the problem and its consequences in terms of political issues in the administration of a tourist destination is quite clear, especially when this tourist destination is a municipality with peculiar characteristics that require specific regulations.

Basically, according to data discussed in the previous section, beds surveyed by Insideairbnb correspond to all those surveyed by the Metropolitan City of Florence survey for the entire sector of extra-hotel facilities (i.e., more than 18000). Moreover, structures surveyed are 1477 according to the Metropolitan City of Florence (Table 1), 1139 according to ISTAT (data not shown here), and 8136 according to Insideairbnb. These figures mean that official statistical sources cover only a small percentage of the entire phenomenon of short-term rental facilities. It is true that these values are not exactly homogeneous neither by date nor by categories included, but the underestimation is so high that it cannot be further neglected. In particular, ISTAT does not cover at all the categories of those who do not carry out these activities in an entrepreneurial way, while the Metropolitan Area data are limited part of the facilities.

For the future work, we intend to perform a comparison between data coming from Insideairbnb and official data related with the city tax of the Municipality of Florence. We also intend to estimate in an accurate way the overlapping among Booking and Airbnb listings.

In perspective, our approach aims at experimenting an integrated system of sources that allows the public administration to follow this important phenomenon, with the least possible delay, to allow the study and definition of new regulations such as to reconcile the needs of the parties involved and, above all, the interest of citizens.

4. References

1. Airbnb: https://www.airbnb.it (accessed 4 March 2020).
2. Booking: https://www.booking.com/holiday-homes/index.it.html (accessed 4 March 2020).
3. Griswold, A.: It's time for hotels to really, truly worry about Airbnb. Quartz (2016) Available at: https://qz.com/729878/its-time-for-hotels-to-really-truly-worry-about-airbnb (accessed 4 March 2020).
4. Insideairbnb: http://insideairbnb.com (accessed 4 March 2020).
5. I.Stat: http://dati.istat.it/?lang=en&SubSessionId=65d172e9-a7e7-426a-b6df-287f5e2296bf# (accessed 4 March 2020).
6. Metropolitan City of Florence: http://www.cittametropolitana.fi.it/turismo/statistica-del-turismo/movimenti-turistici-e-consistenza-delle-strutture-ricettive (accessed 4 March 2020).
7. Shirley Nieuwland & Rianne van Melik: Regulating Airbnb: how cities deal with perceived negative externalities of short-term rentals. Current Issues in Tourism (2018) doi: 10.1080/13683500.2018.1504899.

SportIstat: a playful activity to developing statistical literacy

La promozione della cultura statistica attraverso un'attività ludico sportiva: il caso di SportIstat

Alessandro Valentini, Francesca Paradisi

Abstract SportIstat is the acronym of "Sport with Istat", a playful activity proposed in the occasion of scientific dissemination events involving predominantly young public. This activity is part of the context of gamification. The objective is to capture the attention of youth out of the schooling system, to introduce them to the use of quantitative measurement tools, and to develop their critical sense, using the expedient of sport. SportIstat is a combination of sports and statistics, a path divided into stages. The performance reached in the various stages are compared between the different players. Results are illustrated using interactive statistical tools. The activity is modular and can be re-engineered in various contexts with different levels of complexity.

Abstract *SportIstat è l'acronimo di "Sport con Istat", un'attività ludica proposta in occasione di eventi di divulgazione scientifica che coinvolgono un vasto pubblico prevalentemente giovane. Tale attività rientra nel contesto della gamification in quanto ha l'obiettivo di cogliere l'attenzione dei ragazzi in ambito extrascolastico per introdurli all'uso di strumenti di misurazione quantitativi, ed allo sviluppo del senso critico, utilizzando l'espediente dello sport. SportIstat è un percorso combinato di sport e statistica, articolato in tappe. Viene misurata la performance di ogni tappa e illustrato sia il posizionamento specifico che il rendimento complessivo utilizzando strumenti statistici interattivi quali indicatori e grafici. L'attività si sviluppa in maniera modulare ed è reingegnerizzabile in vari contesti.*

Key words: statistical literacy, developing of critical sense, gamification, youth

[1] Alessandro Valentini, Istat; alvalent@istat.it
Francesca Paradisi, Istat; paradisi@istat.it

1 Introduction

The role of Istat is to serve the community through the production and communication of high-quality statistical information, analysis and forecasts. To achieve this goal, one of the key factors is represented by investments for developing statistical literacy, that is intended as (Wallman, 1993) the ability to understand and interpret data and statistics; the capacity to develop and understand data-driven reasoning.
The large amount of statistical information freely accessible and the worsening of the so-called problem fake news (Corselli-Nordblad and Gauckler, 2018) made the need for a strong investment in this area even more urgent.

1.1 Methods for developing statistical literacy

The approach adopted by Istat for the development of statistical culture is broad-spectrum. The method is consistent with the guidelines defined by UNECE for Statistical Institutes (UNECE, 2014), as well as with the international literature on the subject (see for example Ferligoj, 2015; Gal and Ograjensek, 2017). The reference subjects for Istat's initiatives are varied and potentially include all citizens. They range from researchers to policy makers, from respondents to statistical surveys, to journalists, without neglecting the general public. A significant target is represented by students, differentiated according to the school level, from primary schools to university. Istat designs a series of products and initiatives, also in collaboration with scientific companies, such as SIS and with Community (Eurostat) or international institutions. In particular, it promotes participation in the Statistical Olympiad (Pollice e Barbieri, 2019) and in the international competition of ISLP statistical posters. Local activities are also boosted thanks to the Territorial Network for the development of statistical literacy, a network of specialists in the design and implementation of statistical culture promotion activities that operate at the Institute's territorial offices.
The focus of this paper is the use of a non curricular instrument to catch the attention of young people. Methodology is that of gamification.

1.2 Gamification

A recent approach to developing statistical literacy is that of gamification. The general objective of gamification is to foster the active interest of users, that is, their engagement, to modify their behaviour, to convey information in the desired direction. There are many studies concerning the effectiveness of gamification in the field of learning (Mohamad, Sazali and Sallek 2018). The comforting results allowed the development of this approach in the context of statistical literacy

(Zhang. Fang, 2019; Legaki, Karpouzis and Assimakopoulos, 2019). Gamification is also one of the key points of the DIGICOM project, devoted to modernizing the communication and dissemination of European statistics.
Istat also reserves a relevant attention to gamification activities by making available on its institutional website quizzes, puzzles and games. SportIstat is part of this new approach.

2 SportIstat: approach and model

The activity of *SportIstat* (acronym for "Sport with Istat") is designed for developing statistical literacy during scientific events involving a large audience. The model is based on the expedient of sport to attract young people and to introduce them to the use of quantitative tools, and to the development of critical sense (Gal, 2002; Watson and Callingham, 2003). This approach was tested in two occasions: a) during *The European Researchers' Night*, a set of public events dedicated to bringing researchers closer to the public (Perugia and Siena, 27th Sept, 2019); b) during *The Einsein's Island* (Perugia, 30th Aug – 1st Sep, 2019), an international festival of science show.

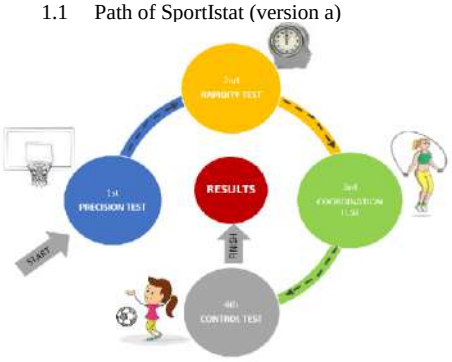
2.1 Materials and methods

SportIstat is an outdoor path inserted in the context of a wider scientific event. A particular care is devoted to the scenario, fundamental to attract the kids. This scenario is a sort of sport village divided into stages, each of which simulating an aspect of sports competition. The path is covered by participants one to one; each one accompanied by an Istat expert with the scope to measure the sport performance and to illustrate results giving statistical messages. Tests are projected to measure the presence of various skills useful for the most part of sport disciplines: precision, rapidity, coordination and control.
The number and typologies of competitions may change depending on the target audience and the space available. For instance, the version a) is modelled in four stages (Figure 1 and Table 1). In version b), in which the available space is very large, the model is slightly different: the total number of stages is 5; a specific stage is dedicated to a statistical test; control test is realized using bowling pins instead of dribbles. In both cases a specific care was dedicated to the output in terms of construction of graphics through the interaction of the participants in order to highlight the individual contribution to the construction of the collective phenomenon. A qualifying point was the use of building bricks to illustrate the content and meaning of the graphs and to make people understand the meaning of the correlation between variables. The prototype of chart made with bricks is illustrated in Figure 2.

Table 1: Detail of the path (model a)

Cod	*Name*	*Objective*	*Score*	*Range*
1	Precision test	To make the most baskets on three throws available. The participant positions himself about 3 meters away from the basket.	0 goal=0; 1 goal =8; 2 goal=16; 3 goal =25	0 – 25
2	Rapidity test	To grasp a falling graduated rod with your hand as quickly as possible. The rod is placed vertically with the zero at the bottom near the participant's fingers.	The interval level (x) ranges from 17 to 31 hundredths of a second. Score is calculated as: 25/14 * (31-x)	0 – 25
3	Coordination Test	To do as many jumps as possible in 20 seconds. Jumps need not be continuous.	One point for each jump up to a maximum of 25 points	0 – 25
4	Control Test	To do as many consecutive dribbles on the ground as possible with one hand, up to a maximum of 25.	One point for each dribble up to a maximum of 25 points	0 – 25

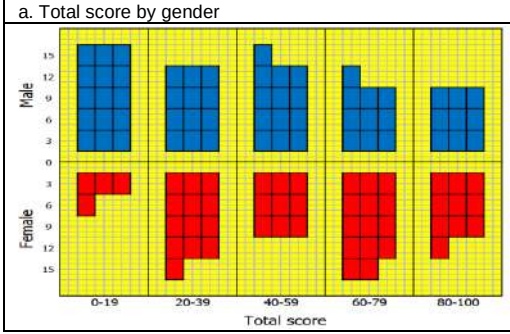
1.1 Path of SportIstat (version a)



1.2 A detail of third step: coordination test

Figure 1: Structure of the path SportIstat and a detail of third step

a. Total score by gender



b. Correlation between Age and reaction time

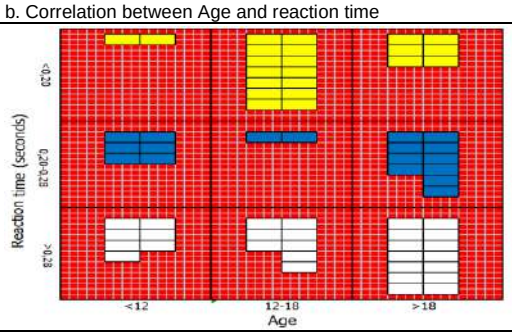


Figure 2: Prototypal chart made with bricks

3 Results

Activities carried out have shown that the SportIstat approach is able to involve a large audience of young people. Youth are attracted by the games and the sports competition. They are enthusiastic both in order to test their individual performance and to compare their scores with other participants having the same covariates (in terms of gender and age).

The path of SportIstat was followed by a total of 650 people: 348 attended event a), 302 event b). Demographics characteristics of participants are quite similar between the two cases in terms of gender: 50% of male in case a); 55% in case b). To be noted a difference by age: the quota of pupils in the ages corresponding to primary and secondary grades is 83% in the first case (a) and 67% in the second (b).

The experience of SportIstat is interesting not only for the project of promoting statistical literacy, but also for the unpublished results on individual performances.

In Table 2 the main results of case a). The total score (which ranges from 0 to 100) is on average 56.3 point. There are significant differences according to demographic characteristics. The score for Male (58.9) is higher than the score for Female (53.7); youth over 15 years (and adults) have a performance (71.3) significantly higher than the ones under 15 (53.5). Another variable that affects the total score is the frequency of sport (in terms of number of time for a week). The 56 participants that do sport more than 4 times a week on average reach a score (62.6) higher than others (55.1). The test with the higher average score is that of control (21.2 in the range 0-25), the score is particularly high (24.9, near to the top) for the 47 participants older than 15 years. Symmetrically the test with the lower score is that of precision (9.5): for this topic the more significant spread is that between male (11.2) and female (7.8). Among other tests, a particularly interesting difference concerns coordination, and is ascribed in the relevant range between younger (11.4) and older (20.4) participants. This should mean that pupils aren't longer able to jump!

Table 2: Average score by sports practice

	Cases	***Average score*** *(range 0-25)*				***Average score Total*** *(range 0-100)*
		Precision	***Rapidity***	***Coordination***	***Control***	
Total	348	9.5	13.0	12.6	21.2	56.3
Gender						
Male	173	11.2	13.4	11.6	22.6	58.9
Female	175	7.8	12.5	13.6	19.8	53.7
Age (years)						
Less than 15	301	9.3	12.6	11.4	20.7	54.0
More than 15	47	10.4	15.6	20.4	24.9	71.3
Weekly sports practice (times)						
Less than 4	291	9.5	12.7	11.8	21.0	55.1
More than 4	56	9.2	14.3	16.6	22.5	62.6

4 Future remarks

The approach of SportIstat is able to give a series of wide-ranging benefits. The expedient of sport, in fact, allows to capture the attention of young people for a few minutes through an activity of gamification, interesting for them and their respective families and friends. This is the occasion to launch a series of important statistical messages that can be transposed unconsciously to the public: the concept of statistical collective, average, correlation and so on; the interpretation of data; the figures. The graphics made with building bricks are very passionate for children and also appreciated by adults for their role of didactic sensory material. Older kids, using IT tools, are also able to create dynamic and interactive graphics: they can thus visualize how the insertion of the data contributes to the continuous evolution of the graphic representation. A good memory of the moment spent with statistics will allow young people to assimilate curiosity towards this world and to take in mind the main concepts, those are the main goals attended.

Finally, SportIstat is a very flexible extracurricular tool that can be exported in various forms and in many contexts thanks to: the possibility to change the number of steps and the disciplines, ductility of materials, soft approach, presence of specialists. For the future it is expected that this instrument will be used in wider areas and with even wider audiences.

References

1. Corselli-Nordblad L, Gauckler B.: New tools to improve statistical literacy – developments and projects. Paper prepared for the 16th Conference of the International Association of Official Statisticians (IAOS). Paris, France, 19-21 (2018)
2. Ferligoj A.: How to Improve Statistical Literacy? Metodološki zvezki, Vol. 12, No. 1, 2015, pp. 1-10 (2015).
3. Gal I.: Adults' statistical literacy Meanings, components, responsibilities, International Statistical Review, 70, pp. 1-51 (2002).
4. Gal I. and Ograjensek I..: Official Statistics and Statistics Education: Bridging the Gap. Journal of Official Statistics, Vol. 33, No. 1, 2017, pp-79-100 (2017).
5. Legaki Z.N., Karpouzis K, Assimakopoulos A.: Using gamification to teach forecasting in a business school setting, paper presented for GamiFIN Conference 2019, Levi, Finland, April 8-10, (2019)
6. Mohamad, S. N. M., Sazali N.S.S., Salleh M.A.M.: Gamification approach in education to increase learning engagement. International Journal of Humanities, Arts and Social Sciences, Volume 4 issue 1 pp. 22-32 (2018)
7. Pollice A. e Barbieri M.M.: Le Olimpiadi italiane di statistica. Statistica e Società, n. 2 (2019)
8. Unece: Making data meaningful. Part 4: A guide for statistical organizations, Geneva: ed. United Nation Economic Commission for Europe (2014)
9. Wallman, K. K.:. Enhancing Statistical literacy: Enriching Our Society, Journal of the American Statistical Association, Vol 88, no 421 (1993)
10. Watson J. M. and Callingham R.A.: Statistical literacy: A complex hierarchical construct. Statistics Education Research Journal, 2(2), 3-46 (2003)
11. Zhang H., Fang L.: Project-Based Learning for Statistical Literacy: A Gamification Approach. In: Väljataga T., Laanpere M. (eds) Digital Turn in Schools—Research, Policy, Practice. Lecture Notes in Educational Technology. Springer, Singapore (2019)

Statistical modeling for some features of Airbnb activity

Un modello statistico per spiegare le caratteristiche dell'attività di Airbnb

Giulia Contu and Luca Frigau

Abstract Airbnb is a platform where to rent tourist accommodations. The Airbnb activity is based on hosts. Every three months, Airbnb awards the best hosts with the badge of Superhost taking into account specific aspects such as the number of reservations, the cancellation policy, the response rate and the overall rating expressed by the guests. It is not clear what can really affect host performance and the probability of becoming a Superhost. The aim of this study is to comprehend which aspects can impact host activity and the achievement of Superhost badge. The analysis is realized considering Milan, which is one of the Italian most visited cities. Bivariate odds model and logistic model have been carried out to data. The first results show the impact on host activity of specific aspects as, for instance, the number of reviews, the services provided and the role of the Superhost. Moreover, they evidence the relevance of the overall rating on the probability of becoming a Superhost.

Abstract *Airbnb è una piattaforma online in cui è possibile affittare alloggi turistici. L'elemento base dell'attività di Airbnb è l'host e la sua attività. Ogni tre mesi, Airbnb assegna ai migliori host l'etichetta di Superhost considerando differenti aspetti quali, ad esempio, il numero di prenotazioni, le politiche di cancellazione delle prenotazioni, il tasso di risposta degli host e il giudizio complessivo espresso dagli ospiti. Non è chiaro cosa possa realmente influenzare le performance dell'host e la probabilità di diventare Superhost. Lo scopo di questo studio è comprendere quali aspetti possano influire sull'attività dell'host e sull'ottenimento del badge di Superhost. Si è deciso di analizzare gli host operanti a Milano, una delle città italiane più visitate, e di utilizzare due differenti metodologie: il bivariate odd ratio model e il modello logistico. I primi risultati mostrano l'impatto di variabili quali il numero di*

Giulia Contu
University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: giulia.contu@unica.it

Luca Frigau
University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: frigau@unica.it

recensioni e i servizi offerti sulle performance degli host; e l'impatto della variabile giudizio complessivo sulla probabilità di divenire Superhost.

Key words: Airbnb, hosts, superhost badge, VGAM, logistic model.

1 Introduction

Airbnb has been founded in 2008 in San Francisco and in the years has grown significantly. It can be defined as an online platform where it is possible to book single rooms or whole apartments for one or more days [2, 4]. Actually, it operates in more than 65,000 cities and 191 countries and it sells millions of room nights for tourists around the globe [1]. The core of Airbnb activity is constituted by hosts and their activity. Every three months, Airbnb awards the best hosts with the status of Superhost. To become a Superhost is necessary for the previous 12 months: to host at least 10 stays; to honor every reservation; to respond within 24 hours at least 90% of the time; to achieve a 4.8+ overall rating. However, it is not clear if the four variables that Airbnb declares to use in attributing the badge have a direct impact on the probability of becoming a Superhost and on the host activity in general.
In literature, few articles analyzed the role and features of the hosts and Superhosts (see: [5, 6]). Particularly interesting is the study realized by Gunter [3]. His goal was to to comprehend if the variables that Airbnb uses to define the Superhost status can really impact the probability to become a Superhost.The results have shown the relevant impact of overall rating and the existence of a specific rank of variables. In this rank, the first position is occupied by overall rating followed by the cancellation policy, the response rate and the number of bookings.
The aim of this paper is to comprehend which variables can directly influence the activity of hosts and Superhosts. To reach this aim, different kinds of variables will be considered. Specifically, they are the Airbnb variables, the managerial aspects and the characteristics of the accommodation. The analysis will be focused on Milan, one of the Italian cities with the highest number of official tourists. Two different models will be used: the bivariate odds model and the logistic model. The innovative aspects of this study are, firstly, the focus on hosts and their activities and, secondly, the use of Vector Generalized Linear Model (VGLM) in the analysis of tourism data. The main findings evidence the impact of different aspects on the host and Superhost activity.
Five sections comprise this study. The second and third Sections will be related to data and methodologies. The results and implications will be explained in the fourth part. Finally, the last Section will be focused on the concluding remarks, limitation and future developments.

2 Data

The study has been realized using a dataset provided by *Airdna*, a company that manages Airbnb data. Three main groups of variables have been used, as shown in Table 1. The first group includes the variables that Airbnb declares to use to attribute the status of Superhost. The second group is composed of the managerial variables, aspects that can be directly chosen by hosts. The third group includes the variables related to the characteristics of accommodation.

Table 1: Variables of the analysis

Group	Variable	Description
Airbnb	Cancellation policy	Flexible, moderate, and strict
	Number of Bookings	Number of reservations
	Response Rate	% of time a host responds to guests within 24 hours
	Overall Rating	Score related to whole Airbnb experience
Managerial	Max Guests	Maximum number of guests in an accommodation
	Response Time	Amount of time to obtain an answer to their questions
	Extra People fee	Price to add one or more guests
	Minimum Stay	Minimum number of days that must be booked
	Security Deposit	Payment (or not) of the deposit
	Business Ready	Services that can support business travelers
	Instantbook Enabled	To book accommodation without an explicit host approval
	Superhost	The host is or not a Superhost
	Occupancy rate	Reservation Days/(Reservation Days + Available Days)
	Number of reviews	Amount of reviews
Accommodation	Property Type	25 different types of accommodation
	Listing Type	Entire home, private room, and shared room
	Bedrooms	Number of bedrooms
	Bathrooms	Number of bathrooms

3 Models

Two different models have been used. The choice is affected by [3] where a logistic model is carried out to estimate the probability to become a Superhost. The model of this study is the logistic model applied with different level of complexity: the logistic model with one response variable and the bivariate odds ratio model with two response variables. Specifically, a logistic model is defined as:

$$F(\mathbf{X}'\boldsymbol{\beta}) = \frac{e^{\sum_{i=0}^{p}\beta_i X_i}}{1+e^{\sum_{i=0}^{p}\beta_i X_i}} \tag{1}$$

It is estimated by Maximum likelihood estimation. The likelihood function is

$$\mathscr{L}(\boldsymbol{\beta}|Y,\mathbf{X}) = \prod_{i=1}^{n} \left(F(\mathbf{X}'\boldsymbol{\beta})^{y_i} [1 - F(\mathbf{X}'\boldsymbol{\beta})]^{(1-y_i)} \right) \tag{2}$$

To reach the aim of the paper, the variable Superhost has been chosen as response variable; the Airbnb, the managerial and accommodation variables as covariates.
The bivariate odds ratio model is a part of a Vector Generalized Linear Model family functions describe by Yee [7]. It can be defined as a logistic regression model with two different response variables, Y_1 and Y_2. Moreover, it models the marginal distributions of two Y_j and the odds ratio. Specifically, it is defined through:

$$\text{logit } p_j(x) = \eta_j(x), \quad j = 1,2 \tag{3}$$

$$\log \psi(x) = \eta_3(x) \tag{4}$$

where ψ identifies the odds ratio and describes the association between the two responses. It is calculated as:

$$\psi(x) = \frac{p_{00}(x)p_{11}(x)}{p_{01}(x)p_{10}(x)} = \frac{P(Y_1 = 0, Y_2 = 0|X)P(Y_1 = 1, Y_2 = 1|X)}{P(Y_1 = 0, Y_2 = 1|X)P(Y_1 = 1, Y_2 = 0|X)} \tag{5}$$

For the application of this model, occupancy rate and overall rating have been chosen as response variables, whilst managerial and accommodation variables as covariates.

4 Results

The study starts with the identification of the variables that can directly influence the probability to obtain good results in Airbnb activity through the bivariate logistic model. The impact on occupancy rate and on overall rating of the managerial and accommodation variables have been analyzed. Tables 2 and 3 show that in Milan managerial variables and characteristics of the accommodation impact significantly on the probability to reach high results in terms of Overall rating and Occupancy rate. Some variables, as Instant book and Response rate, have a significative impact not only on the two response variables, but also on the odds ratio.
Then, through a logit model, it is investigated which variables can influence the probability to obtain the status of Superhost. The results show how in Milan all Airbnb and managerial variables influence the probability to obtain the badge of Superhost, as shown in Tables 4 and 5. On the contrary, only the accommodation variables have not a significant impact on this probability, as shown in Table 6.

Table 2: Impact of the managerial variables on occupancy rate (Occ.) and overall rating (Over.)

	Occ.	Over.	(Occ.= 1,Over.= 1)
Intercept	−1.80 ***	−0.22	0.73 ·
Max Guests	−0.07 ***	−0.09 ***	0.05 ·
Response Time min	−0.00 ***	0.00	−0.00
Security Deposit (Yes)	−0.07 ·	0.23 ***	−0.02
Extra People Fee (Yes)	0.09 ·	0.03	0.12
Business Ready (Yes)	−0.06	0.41 ***	0.15
Instantbook Enabled (Yes)	0.81 ***	−0.47 ***	−0.28 *
Number of Reviews	0.04 ***	−0.00	−0.00
Response Rate	0.014 ***	0.01 ***	−0.01 *
Superhost (=1)	0.62 ***	2.87 ***	0.39
Cancellation Policy (=1)	0.08 ·	0.02	-0.09

Table 3: Impact of the accommodation variables on occupancy rate (Occ.) and overall rating (Over.)

	Occ.	Over.	(Occ.= 1,Over.= 1)
Intercept	-0.09	0.18 *	0.07
Listing Type (1)	0.31 ***	-0.07	-0.01
Property Type (1)	0.16 **	0.15 *	0.10
Bathrooms	-0.15 **	0.08 ·	0.05
Bedrooms	-0.20 ***	-0.02	0.00

Table 4: Impact of Airbnb variables on the probability to be a Superhost

	Coef.	*pvalue*	AME
Intercept	-32.05	***	
Cancellation.Policy (1)	0.24	**	0.00
Number of Bookings	0.02	***	0.00
Response Rate	0.08	***	0.01
Overall Rating	4.56	***	0.39

5 Conclusions

The analysis highlights how the best results in terms of occupancy rate and overall rating were obtained with the highest scores in business ready, number of reviews, instant book, response rate, listing type and property type. Moreover, the results suggest that the Airbnb variables, the managerial variables and the accommodation variables influence the achievement of badge. In general, the findings evidence the necessity for hosts to offer high-quality services, to support guests before and during the experience, to enhance writing reviews.

Table 5: Impact of managerial variables on the probability to be a Superhost

	Coef.	*pvalue*	AME
Intercept	-2.5752	***	
Occupancy Rate	1.2972	***	0.12
Max Guests	-0.0605	**	-0.01
Response Time min	-0.0015	***	-0.00
Security Deposit (Yes)	0.2778	***	0.03
Extra.People.Fee (Yes)	0.2598	***	0.02
Business.Ready (Yes)	0.5985	***	0.06
Instantbook Enabled (Yes)	-0.8048	***	-0.06
Number of Reviews	0.0076	***	0.00

Table 6: Impact of accommodation variables on the probability to be a Superhost

	Coef.	*pvalue*	AME
Intercept	-2.2046	***	
Listing Type (1)	-0.0060		-0.00
Property Type (1)	0.1845	*	0.02
Bedrooms	-0.0666		-0.01
Bathrooms	0.0344		0.00

The study has the limitation that is focused only on Milan. In further researches more destinations and more variables will be considered.

References

1. Aznar, J. P., Sayeras, J. M., Rocafort, A., Galiana, J.: The irruption of Airbnb and its effects on hotel profitability: An analysis of Barcelonas hotel sector. Intangible Capital, **13**, 147–159 (2017)
2. Choi, K.H., Jung, J.H., Ryu, S.Y., Do Kim, S., Yoon, S.M. : The relationship between airbnb and the hotel revenue: in the case of korea. Indian Journal of Science and Technology **8**, (2015)
3. Gunter, U.: What makes an airbnb host a superhost? Empirical evidence from San Francisco and the bay area. Tourism Management **66**, 26–37 (2018)
4. Guttentag, D., Smith, S., Potwarka, L., Havitz, M. : Why tourists choose airbnb: A motivation-based segmentation study. Journal of Travel Research **57**, 342359 (2018)
5. Liang, S., Schuckert, M., Law, R., Chen, C.C. : Be a superhost: The importance of badge systems for peer-to-peer rental accommodations. Tourism management **60**, 454–465 (2017)
6. Ma, X., Hancock, J.T., Mingjie, K.L., Naaman, M. : Self-disclosure and perceived trustworthiness of airbnb host profiles., in: CSCW, 2397–2409 (2017)
7. Yee, T. W. : Vector generalized linear and additive models: with an implementation in R. Springer (2015)

Tertiary students with migrant background: evidence from a cohort enrolled at Sapienza University

Background migratorio e percorsi universitari: osservazioni su una coorte di studenti iscritti a Sapienza Università di Roma

Giudici Cristina, Vicari Donatella, Trappolini Eleonora

Abstract The participation of foreign and international students in the Italian tertiary education shows a strong positive trend in the last decades, and has recently drawn attention in the literature. The aim of this study is to examine disparities in timing of graduation of a cohort of full time students which have been enrolled at Sapienza University in the a.y. 2012/2013 and graduated within twice the theoretical duration of the programme of study, considering both their citizenship and country of prior education. Although migrant background is not necessarily indicative of a disadvantage, the analysis suggests the existence of a link between migrant background and longer time to complete the programme of study, mostly in Bachelor's programmes.

Abstract *La presenza di studenti stranieri ed internazionali nel sistema universitario italiano ha conosciuto nel tempo una tendenza ampiamente positiva, con un interesse crescente da parte della letteratura. Il presente studio è volto ad analizzare i tempi di conseguimento della laurea per una coorte di studenti iscritti alla Sapienza nell'a.a. 2012/13 e laureati entro il doppio della durata legale del corso, distinguendo per cittadinanza (italiana e straniera) e paese di conseguimento del titolo di accesso all'Università (italiano o estero) e stratificando per classi di laurea. L'analisi suggerisce l'esistenza di un legame tra background migratorio e tempi di conseguimento del diploma, in particolare nelle lauree triennali.*

Key words: University students, migrant background

[1] Giudici Cristina, Sapienza University; email: cristina.giudici@uniroma1.it
Vicari Donatella, Sapienza University; email: donatella.vicari@uniroma1.it
Trappolini Eleonora, Sapienza University; email: eleonora.trappolini@uniroma1.it

1 Introduction

The Organization for Economic Co-operation and Development has recently pointed out the massive expansion of the number of mobile students enrolled in tertiary education programmes worldwide, going from 2 million in 1998 to 5.3 million in 2017, with an annual increase of 5%. English-speaking countries are the most attractive destinations, and Asian students constitute the largest group of international students across the OECD countries (OECD, 2019).

The increase of international enrolment is particularly high at master and doctoral level, driven by a variety of push and pull factors. Actually, in many countries, the ever-growing demand for tertiary education may be associated with low education capacity in the origin country, and a growing number of students may look for educational opportunities abroad. On the other hand, social and economic factors are contributing to make international student mobility more affordable than in the past: the presence of an immigrant community already present in the country of destination, the existence of previous colonial ties, the use of a common working and teaching language, the job prospects after graduation are among the most motivating factors for studying abroad, together with the increasing investment by Universities in the quality of education (Norton and Fatigante, 2018).

In the literature, international students are generally assumed to be people moving to a country for the purpose of study and accessing University programmes with a foreign prior degree which would formally entitle him/her to be enrolled.

The term foreign student is often used as an approximation of international student, but they should be distinguished by taking into account also the country of citizenship, because foreign citizens may enter university with a national educational background and, conversely, international students may be national citizens.

The participation of students with a migrant background in the Italian Higher Education Institutions has been progressively growing and shows a strong positive trend in the last ten years. According to national statistics, 55% of the students enrolled as foreign citizens (almost 80,000 in 2017) is already resident in Italy and obtained the last educational degree from an Italian institution: this is the case, for instance, of the children of immigrants. The remaining 45% are international students, who accessed the Italian university system with a diploma obtained abroad (MIUR, 2018).

To our knowledge, in the Italian context, no studies exist on tertiary education, where international and second-generation students are distinguished. The few studies focusing on the first/second generation (Lagomarsino and Ravecca, 2014, Vaccarelli, 2016, Bertozzi, 2018) and on international students (Norton and Fatigante, 2018 and references therein) in the Italian University, underline the need of investigating student pathways in terms of access, academic pathways, success (or dropout), integration (well-being) and university-work transition.

Sapienza University of Rome is considered the largest in Europe with over 100,000 students and it is characterised by the largest number of both international and foreign students in Italy. Thus, it may be considered as a representative case for the analysis of tertiary students with migrant background in Italy.

The aim of the study is to analyse the timing of graduation for a cohort of graduate students, distinguishing by migrant and educational background and stratifying by medical, social-humanistic and technical-scientific programmes of studies.

2 Data and methods

Data come from Sapienza University administrative archives and are collected by Infosapienza on 13,061 full time students enrolled during the a.y. 2012/2013 in Bachelor's (7,873) and Master's (5,188) degrees and graduated within twice the theoretical duration of their programme of study, following the approach suggested by OECD (2019).
For comparability reasons, students enrolled in single-cycle Master's programmes with a legal duration of more than two years (Laurea Magistrale a ciclo unico) have not been included in this study.

Student demographic and socio-economic characteristics include age, gender, country of birth, citizenship, scholarships, graduation date and broad classes of study programme (medical, social-humanistic and technical-scientific.) A dummy variable also provides information on the country (Italy/abroad) where the student has obtained the *prior education,* i.e., the qualification (high school diploma or bachelor degree) which formally entitles him/her to be enrolled at Sapienza (as Bachelor or Master student, respectively). An international student is defined *stricto sensu* as a student who obtained his/her prior education abroad.

Table 1 outlines four categories of students based on their citizenship and country of prior education.

Table 1: *Categorization of students according to their citizenship and country of prior education*

Country of prior education	***Italian Citizenship***	***Foreign Citizenship***
Italy (non-international student)	*Italian Student* (*IS*)	*Foreign Student with Italian Educational Background* (*FS-IEB*)
Abroad (international student)	*Italian Student with Foreign Educational Background* (*IS-FEB*)	*Foreign Student with Foreign Educational Background* (*FS-FEB*)

International students include both Italian and foreign students who obtained their prior education abroad; analogously, non-international students include both Italian and foreign students who obtained their prior education in Italy. In addition to *Italian Students* (IS), i.e., Italian citizens with Italian prior education, the other three groups include *Foreign Students with Italian Educational Background* (FS-IEB), *Italian Student with Foreign Educational Background* (IS-FEB), and *Foreign*

Students with Foreign Educational Background (FS-FEB). FS-FEB should be distinguished by FS-IEB because the latter are supposed to better handle the Italian education system and to master the Italian language. Note that FS-IEB may also include long term residents or even students born in Italy. Analogously, IS-FEB moved from another country to study in Italy, but they may be supposed to know the Italian language and to be well integrated in the Italian education system.

Globally, our data include 12,564 Italian students, 192 FS-FEB, 275 FS-IEB and 30 IS-FEB. Table 2 shows the percentages of Bachelor and Master students by citizenship and country of prior education: 1.8% and 1.5% of Bachelor and Master students, respectively, obtained their prior education abroad. In both Bachelor's and Master's programmes the majority of foreign citizens have Italian educational backgrounds.

The cohort has been followed for twice the theoretical duration of the programme of study, i.e., 3+3 years for Bachelor's and 2+2 years for Master's, and a regression analysis of survival data has been performed, based on the Cox proportional hazards model. In the following survival analyses, Italian Students with Foreign Educational Background (IS-FEB) have not been included because of their limited number.

Table 2: *Percentages of Bachelor and Master students by citizenship and country of prior education*

Country of prior education	***Bachelor's***			***Master's***		
	Italian Citizens	***Foreign Citizens***	***Total***	***Italian Citizens***	***Foreign Citizens***	***Total***
Italy	*99.63%*	*58.33%*	*98.2%*	*99.96%*	*59.69%*	*98.5%*
Abroad	*0.37%*	*41.67%*	*1.8%*	*0.04%*	*40.31%*	*1.5%*
Total	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*

3 Preliminary results

Figure 1 shows survival functions for the cohort of students enrolled in 2012/13 and graduated within twice the theoretical duration of their programme of study, for both Bachelor's and Master's programmes, by country of prior education.

Three years after entering university, more than 50% of bachelor students are still in education and the percentage increases for international students. For Master's programmes no substantial differences emerge in the survival patterns between Italian and international students. The analysis by citizenship (Figure 2) shows similar results, with a lower gap.

Table 3 shows hazard ratios estimated from the Cox regression model for bachelor students by programme of study adjusted for age, gender and categories. Male gender and higher age at the enrolment are generally associated with longer time of graduation. Looking at categories, students with foreign citizenship generally take longer to graduate than Italians (IS), and this is more pronounced when students have a foreign educational background (FS-FEB) than an Italian one (FS-IEB). Foreign educational background is confirmed to be slightly associated to

the graduation delay for medical and social-humanistic degrees. Results for Master students are not reported for space reasons; although globally the analysis confirms what was found for Bachelor students, the results are less evident, probably for the greater heterogeneity of the student population.

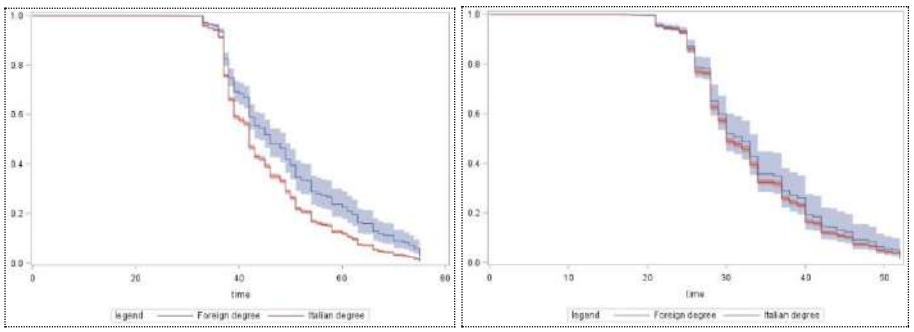


Figure 1: *Survival function (in months) for a cohort of full time students enrolled in 2012/13 in Bachelor's (on the left) and Master's (on the right) programmes and graduated within twice the theoretical duration of the programme of study, by country of prior education (95% CI).*

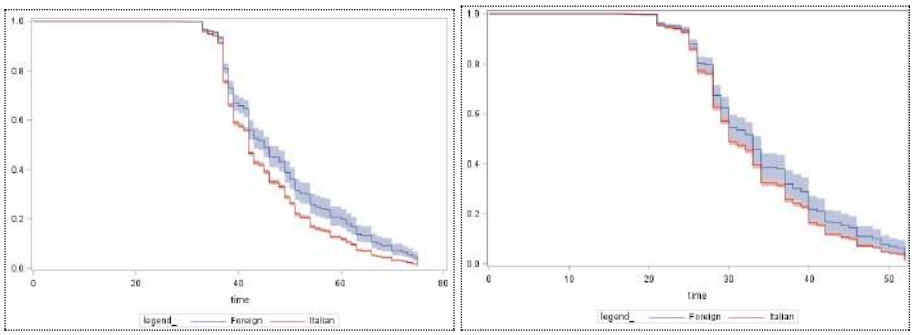


Figure 2: *Survival function (in months) for a cohort of full time students enrolled in 2012/13 in Bachelor's (on the left) and Master's (on the right) programmes and graduated within twice the theoretical duration of the programme of study, by citizenship (95% CI).*

Table 3: *Hazard ratios estimated from the adjusted Cox regression model - Bachelor's degree*

		Total	***MD***	***TSD***	***SHD***
Gender	Female		*Ref.*		
	Male	-0.11***	-0.13*	-0.01	-0.10*
Age	<20		*Ref.*		
	20-21	-0.08*	-0.09º	-0.20*	-0.21***
	22-23	-0.03	-0.09	-0.24	-0.30**
	24-25	0.06	-0.09	-0.07	-0.19
	>26	0.15º	0.05	-0.35	-0.20º
Categories	IS		*Ref.*		
	FS-FEB	-0.46***	-0.49º	-0.29	-0,24º
	FS-IEB	-0.20º	-0.19	-0.11	-0.18
N		*7845*	*2029*	*2032*	*3433*

*Levels of significance: ***p<0.0001; **p < 0.001; *p<0.01; º p<0.1*
MD: Medical Degrees; TSD: Technical and Scientific Degrees; SHD: Social and Humanistic Degrees
IS: Italian Students; FS: Foreign Students with Foreign Educational Background; FS(IEB): Foreign Students with Italian Educational Background.

4 Concluding remarks

Although migrant background is not necessarily indicative of a disadvantage, the analysis suggests the existence of a link between migration and lower student performance in the Academic context, which need to be further analysed. Notably, we need to examine academic paths of those students who have not graduated within a reasonable period. As pointed out by OECD (2019), delay graduation or dropping out are not necessarily symptoms of student or institutional failure. A strong labour market demand may lead a student to start working before attaining his degree. Nonetheless, specific difficulties associated with displacement, such as language barriers, socio-economic disadvantage, or lack of social integration, should not be underestimated.

References

1. Bertozzi, R.: University Students with Migrant Background in Italy. Which Factors Affect Opportunities? It J Soc Ed, 10(1), 23-42 (2018)
2. Lagomarsino F., Ravecca, A., Il passo seguente. I giovani di origine straniera all'università, Franco Angeli (2016)
3. MIUR, Anagrafe nazionale degli studenti https://anagrafe.miur.it/index.php (accessed in July 2018).
4. Norton, L., Fatigante, M., Being international students in a large Italian university: Orientation strategies and the construction of social identity in the host context. Rassegna di Psicologia, Vol 35, N. 3 (2018)
5. OECD: *Education at a Glance 2019: OECD Indicators*, OECD Publishing, Paris (2019).
6. Vaccarelli, A.: Studiare in Italia. Intercultura e inclusione all'Università. Franco Angeli (2016)

The Causal Effect of Immigration Policies on Income Inequality

L'effetto causale delle politiche migratorie sulla disuguaglianza del reddito

Irene Crimaldi, Laura Forastiere, Fabrizia Mealli and Costanza Tortù

Abstract This work evaluates the effect of immigration policies on the income inequality. The employed methodology has been recently introduced in [10]. It takes into account network interference among units, in the presence of a multi-valued individual treatment. The estimation strategy is based on the usage of a Joint Multiple Generalized Propensity Score, which properly guarantees balance with respect to both individual and network characteristics. Results suggest that a highly restrictive political approach leads to a higher income inequality index.

Abstract Questo lavoro valuta l'effetto delle politiche di immigrazione sulla disuguaglianza del reddito. La metodologia impiegata è stata recentemente introdotta da [10]: tiene conto della interferenza tra le unità, in presenza di un trattamento individuale multidimensionale. La procedura di stima è basata sul Joint Multiple Generalized Propensity Score, che consente di ottenere un adeguato bilanciamento rispetto a covariate sia individuali sia di rete. I risultati mostrano che politiche migratorie più restrittive portano ad un maggiore coefficiente di disuguaglianza del reddito.

Key words: causal inference, interference, complex networks, multi-valued treatment, multiple network exposure, immigration policies

---

Irene Crimaldi
IMT School for Advanced Studies Lucca, Piazza San Ponziano 6, 55100, Lucca (IT) e-mail: irene.crimaldi@imtlucca.it

Laura Forastiere
Yale University, School of Public Health 350 George St, New Haven (CT) e-mail: laura.forastiere@yale.edu

Fabrizia Mealli
University of Florence, Viale Morgagni 59, 50100, Florence (IT) e-mail: mealli@disia.unifi.it

Costanza Tortù
IMT School for Advanced Studies Lucca, Piazza San Ponziano 6, 55100, Lucca (IT) e-mail: costanza.tortu@imtlucca.it

1 Introduction

Policy evaluation studies intend to assess the effect of an intervention. This goal often results to be challenging, since most empirical scenarios involve complex treatments, which have not been randomly assigned over the population and, moreover, the analysis might be also affected by the presence of interference among units. This last phenomenon arises when the treatment of one unit has an effect on the response of another unit [2]. For instance, the political attitude of a country towards a relevant issue may influence the neighboring countries. Furthermore, evaluating the impact of a policy often requires the definition of a multi-valued treatment: a complex policy might be multi-faceted and, therefore, vary over multiple dimensions.

This work evaluates the impact of governments' stance towards immigration on the Gini index for income inequality. The employed methodology is the one proposed in [10], which in turn finds its roots in [4]. These works both refer to the branch of the growing statistical literature about causal inference under network interference, which extends the Rubin Causal Model for policy evaluation studies admitting first-order spillover effects. Specifically, the model proposed and studied in [10] allows to account for interference between units, in observational studies where the treatment varies over multiple categories. We here define a multi-valued characterization of an immigration policy looking at the restrictiveness of the national regulations, which are explicitly addressed to migrants, and of the control mechanisms, which monitor adherence to the regulations. Furthermore, we model interference between countries according to a specific indicator, that we call *Interference Compound Index*.

2 The Model

Let us consider a population $\mathscr{N}$, with N units. Denoted as K the number of treatment levels, let $Z_i \in \{1,\dots,K\}$ be a categorical variable representing the treatment assigned to unit i and Y_i^{obs} the observed outcome for the same unit. By $\mathbf{Z}$ and $\mathbf{Y}^{obs}$ we denote the corresponding vectors of the whole sample $\mathscr{N}$. Moreover, $\mathbf{X}_i$ denotes a vector of P covariates (or pre-treatment variables). We assume that the relations between units are fully described by the observed *weighted* undirected network $\mathscr{G} = (\mathscr{N}, \mathbb{E}, \mathbb{W})$, where $\mathscr{N}$ is the set of nodes (the units), $\mathbb{E}$ is the set of edges, indicating links between the units, and $\mathbb{W}$ is the set of the weights on the edges, that is $\mathbb{W}$ collects the quantities I_{ij}, each representing the weight of the link between i and j. We denote as $\mathscr{N}_i$ the set that includes all the nodes j which are in the neighborhood of i, that is all the units j such that exists a link between j and i. In this framework, the standard SUTVA assumption, which includes the no-interference hypothesis, is replaced by an alternative assumption, known as Stable Unit Treatment on Neighborhood Value Assumption (SUTNVA), which grants first order spillover effects:

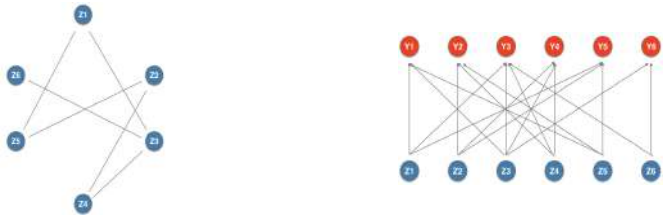


Fig. 1 Relations between units and the corresponding interference structure

Assumption 1 (SUTNVA).

1. No Multiple Versions of Treatment (Consistency)*: $Y_i(\mathbf{Z}) = Y_i(\mathbf{Z}^1)\ \ \forall \mathbf{Z}, \mathbf{Z}^1$ such that $\mathbf{Z} = \mathbf{Z}^1$, that is, the mechanism used to assign the treatments does not matter and an alternative assignment does not constitute a different treatment.*
2. Neighborhood Interference*: There exists a function g_i such that, for all $\mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}^1_{\mathcal{N}_{-i}}$ and $\mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}^1_{\mathcal{N}_i}$ with $g_i(\mathbf{Z}_{\mathcal{N}_i}) = g_i(\mathbf{Z}^1_{\mathcal{N}_i})$, we have*

$$Y_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = Y_i(Z_i, \mathbf{Z}^1_{\mathcal{N}_i}, \mathbf{Z}^1_{\mathcal{N}_{-i}}).$$

We denote as $\mathbf{G}_i$ the variable which results from applying the function g_i on the i's neighbors' treatment vector: $\mathbf{G}_i = g_i(Z_{\mathcal{N}_i})$. This variable, known as *neighborhood treatment*, measures the unit's indirect exposure to the treatment and represents a numerical synthesis of the treatment vector which characterizes the neighborhood of i. Figure 2 provides a graphical intuition of the interference mechanism, in a toy example: it shows the observed network and the corresponding spillover mechanism that the network links generate. Note that when the individual treatment is multi-valued, the neighborhood treatment is multivariate (specifically, it covers K dimensions) and it quantifies the individual indirect exposure to all treatment levels. Therefore, we can map the indirect exposure by means of the *Neighborhood Treatments Exposure Matrix* (NTEM) $\mathbf{G}$, which is defined as a $N \times K$ matrix that collects the individual neighborhood exposure to all the treatment levels. In particular, the generic element $G_{i,z} \in \mathbb{G} \subseteq \mathbb{R}$ indicates the indirect exposure of unit i to the treatment level z. Consequently, each row of $\mathbf{G}$ represents the neighborhood treatment of the generic unit i, $\mathbf{G}_i \in \mathbb{G}^K$. For instance, we can define the element $G_{i,z}$ as

$$G_{i,z} = \sum_{j \in \mathcal{N}_i} I_{ij} \delta_{zj}, \tag{1}$$

where δ_{zj} is a dummy variable that equals 1 if $Z_j = z$ and 0 otherwise. In this setting, each unit i is exposed to a *joint treatment* $(Z_i, \mathbf{G}_i)$: the *individual treatment* Z_i, which is a categorical variable with K levels, and the *neighborhood multi-treatment* $\mathbf{G}_i$, which is a K-variate variable. Hence, potential outcomes, for each unit i, are indexed by the joint treatment: $Y_i(Z_i, \mathbf{G}_i) = Y_i(Z_i = z, \mathbf{G}_i = \mathbf{g})$. As a consequence, the propensity score has to be defined with respect to this particular joint treatment.

Definition 1 (Joint Multiple Generalized Propensity Score (JMGPS)). The JMGPS is the probability of being jointly exposed to a K-variate individual treatment equal to z and to a K-dimensional neighborhood treatment equal to $\mathbf{g}$, conditioning on baseline covariates. That is,

$$\begin{aligned}\psi(z,\mathbf{g};\mathbf{x}) &= P(Z_i = z, \mathbf{G}_i = \mathbf{g}|\mathbf{X}_i = \mathbf{x}) \\ &= P(\mathbf{G}_i = \mathbf{g}|Z_i = z, \mathbf{X}_i^g = \mathbf{x}^g)P(Z_i = z|\mathbf{X}_i^z = \mathbf{x}^z) \\ &= \lambda(\mathbf{g};z,\mathbf{x}^g)\phi(z;\mathbf{x}^z),\end{aligned} \tag{2}$$

where $\lambda(\mathbf{g};z,\mathbf{x}^g)$ is the neighborhood propensity score and $\phi(z;\mathbf{x}^z)$ is the individual propensity score. $\mathbf{X}_i^z$ and $\mathbf{X}_i^g$ are vectors collecting covariates that affect the individual and the neighborhood treatment, respectively.

The estimation strategy is based of the usage of JMGPS and proposes a parametric approach for imputing missing potential outcomes, for the possible characterizations of the joint treatment. Please refer to [10] for a more detailed description of it.

3 Empirical Analysis

We make use of the above methodology in order to assess the causal effect of immigration policies on the income inequality. Specifically, we focus on a subset of 22 OECD countries, which are located in Europe, and we assess their political attitude towards migrants, evaluating its effect on the (lagged) Gini index for income inequality. We deal with country-year observations $i = (c,t)$, $i \in \mathcal{N}$, where c refers to a generic country and t indicates a given year, $t \in \{1980,\dots,2010\}$. We merge three main sources of data: (i) The IMPIC (Immigration Policies in Comparison) Dataset, which properly measures the restrictiveness of many specific immigration policies implemented by OECD countries between 1980 and 2010 [6]; (ii) The CEPII Gravity Dataset, which provides information about similarity and proximity measures involving country-year profiles [5]; (iii) The World Development Indicators Dataset, that includes country-year indexes [1]. In this empirical setting, SUTVA may be violated: the political receipt that each country implements with respect to immigrants has an impact on the international equilibrium and the global social system. The idea is that immigrants avoid highly restrictive countries and decide to go in places that are similar to their first choice with respect to some characteristics, but more politically welcoming. We assume that the extent to which each country can be influenced by another country is modeled by a suitable indicator, that we call Interference Compound Index (ICI) and that summarizes the macro-components which reasonably rule the spillover mechanism:

Definition 2 (Interference Compound Index (ICI)).

$$ICI_{cc',t} = \alpha \times IG_{cc'} + \beta \times IC_{cc',t}$$

where $IG_{cc'} \in [0,1]$ is the *geographic proximity indicator* which measures the geographical proximity between country c and country c' and $IC_{cc',t} \in [0,1]$ is the *cultural similarity indicator* that measures how much each pair of countries are culturally similar at time t. The constants α and β, with $\alpha+\beta=1$, are the weights that determine the extent of which each component contributes to the global index.

We vary the interference input weights so to check the robustness of our results, under different assumptions about the network structure.

We now define the treatment variable, that "quantifies" the immigration policy, which has been implemented by national governments. The IMPIC Dataset provides indicators which measure the country-year restrictiveness towards migrants with respect to *regulations* and *control* mechanisms. Let reg_i be the reported value of the restrictiveness in terms of regulations of the generic country-year profile $i=(c,t)$ and $cont_i$ represent the corresponding value in terms of control. Denoting as med_{Reg} and med_{Cont} the medians of the regulations and control indicator, respectively, we classify the national political approach towards immigration with respect to the distribution of the two indicators:

Definition 3 (Nominal Treatment Categories).

- Z_i=A if $\text{reg}_i \leq \text{med}_{Reg}$ and $\text{cont}_i \leq \text{med}_{Cont}$: this category identifies profiles that are barely restrictive with respect to the two mechanisms.
- Z_i=B1 if $\text{reg}_i > \text{med}_{Reg}$ and $\text{cont}_i \leq \text{med}_{Cont}$: this category detects profiles which implement restrictive regulations but weak control strategies.
- Z_i=B2 if $\text{reg}_i \leq \text{med}_{Reg}$ and $\text{cont}_i > \text{med}_{Cont}$: this category indicates a welcoming attitude in terms of regulations but intense control protocols.
- Z_i=C if $\text{reg}_i \geq \text{med}_{Reg}$ and $\text{cont}_i \geq \text{med}_{Cont}$: this category denotes an highly restrictive policy towards migrants with respect to both regulations and control.

With respect to this definition of the treatment variable, and given the ICI formulation, we can define the neighborhood treatment as $\mathbf{G}_{ct}$ as

$$\mathbf{G}_{ct} = \begin{pmatrix} G_{ctA} \\ G_{ctB1} \\ G_{ctB2} \\ G_{ctC} \end{pmatrix} = \begin{pmatrix} \sum_{c' \in \mathcal{N}_{ct}} ICI_{cc',t}\,\delta_{Ac't} \\ \sum_{c' \in \mathcal{N}_{ct}} ICI_{cc',t}\,\delta_{B1c't} \\ \sum_{c' \in \mathcal{N}_{ct}} ICI_{cc',t}\,\delta_{B2c't} \\ \sum_{c' \in \mathcal{N}_{ct}} ICI_{cc',t}\,\delta_{Cc't} \end{pmatrix},$$

where $\delta_{Ac't}, \delta_{B1c't}, \delta_{B2c't}, , \delta_{TCc't}$ are dummy variables such that $\delta_{Ac't}=1$ if $Z_{c',t}=A$ and 0 otherwise; $\delta_{B1c't}=1$ if $Z_{c',t}=B1$ and 0 otherwise; $\delta_{B2c't}=1$ if $Z_{c',t}=B2$ and 0 otherwise; $\delta_{Cc't}=1$ if $Z_{c't}=C$ and 0 otherwise. Consequently, the joint potential outcomes are defined as $Y_{ct}(Z_{ct},\mathbf{G}_{ct})$. As said before, the outcome variable of interest, Y_i, is the Gini index for income inequality: it is bounded between 0 and 1 and higher values are associated to more unequal societies.

The estimation strategy is based on the Joint Multiple Generalized Propensity Score. We separately estimate the two propensity scores, relying on the factorized nature of the JMGPS: we estimate the individual propensity score, referred to the

categorical variable Z_i, that is $\phi(z;\mathbf{x}^z) = P(Z_i = z|\mathbf{X}_i^z = \mathbf{x}^z)$, fitting a standard Multinomial Logit Model; moreover, we estimate the neighborhood propensity score following the approach which has been proposed by [11]. Under continuous treatments (in our setting the neighborhood treatment is a multivariate continuous variable), we can estimate the propensity score assuming a functional form for the treatment of interest, such that the functional parameters depend on individual covariates. In particular, we assume that the (transformed [1]) neighborhood treatment $\mathbf{G}_i^*$ is normally distributed as

$$\mathbf{G}_i^* \sim MN(\mu_{\mathbf{G}_i^*}, \Sigma_{\mathbf{G}^*}) \tag{3}$$

where $\mu_{G_{i,z}^*} = \alpha_{G_z^*} + \beta_{G_z^*}^T \mathbf{X}_i^g + \beta_{G_z^*}^T Z_i$.

4 Key Results

We evaluate the impact of immigration policies on the Gini index for income inequality, dealing with a treatment of four categories and exploiting the pairwise comparisons between the treatment levels. Note that, for dealing with possible reverse causality issues, we introduce a one-year lag process in the analysis: that is, baseline covariates of a given country at time t affect its own joint treatment exposure at time $t+1$, which in turn affects the outcome at time $t+2$. Furthermore, we include in the propensity score estimation the no-lagged value of income inequality (evaluated at time t), so to control for the country-specific baseline outcome and more accurately identify the effect of the treatment. We take into consideration four scenarios depicting the interference mechanism to check the robustness of results with respect to different *a-priori* assumptions about the dependency structure. In particular, we check the following configurations of the influence weights: i) $\alpha = \beta = \frac{1}{2}$, *(gc)*: both geographical proximity and cultural similarity affects the interference between units, and contribute in determining the global interference compound index with equal weight; ii) $\alpha = 1, \beta = 0$, *(g)*: only *geographical proximity* drives interference; iii) $\alpha = 0, \beta = 1$, *(c)*: interference is prompted by cultural similarity only and iv) $\alpha = 0, \beta = 0$, *(noint)*: *no interference* mechanism comes into play. Table 1 reports the estimated treatment effects and their 95% Confidence Interval (computed using bootstrap), while Figure 2 represents the box plots and forest plots, referred to the distribution of the estimated effects.

We can see that highly restrictive political approaches cause significantly higher income inequality indices. In particular, severe control mechanisms prompt inequality (B2-A,B2-B1). Observing the results, we can also state that ignoring the possible spillover mechanism leads to a downward bias in the estimates, thus relying on the SUTVA makes the size of the effects becomes smaller (*noint*). This conclusion is stable in all the contrasts of interest. Allowing for the presence of interfer-

[1] We have applied a preliminary transformation over each component of the neighborhood treatment so that the normality assumption is plausible.

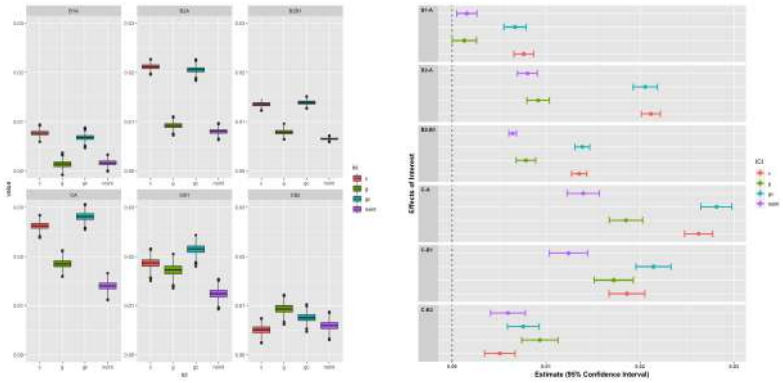


Fig. 2 The causal effect of immigration policies on the Gini index for income inequality: Colors signal the different assumption about interference: *gc*(lightblue), *g*(green), *c*(red), *noint*(purple)

ence increases the magnitude of the effects, and their robustness, with respect to the no-interference scenario. In particular, introducing the cultural similarity in the mechanism of dependencies points the way to an enhancement in terms of effects' intensities (c) (unless in the C-B2 comparison). Geographical proximity instead, mitigates the impact of interference on the results (g). These considerations hold in all the considered comparisons.

Table 1 The causal effect of immigration policies on the Gini index for income inequality: key results

ICI weights (α,β)	B1-A Est (95% CI)	B2-A Est (95% CI)	C-A Est (95% CI)	B2-B1 Est (95% CI)	C-B1 Est (95% CI)	C-B2 Est (95% CI)
$(\frac{1}{2},\frac{1}{2})$	0.00669 ***	0.02055 ***	0.02812 ***	0.01386***	0.02143***	0.00757***
	(0.00545;0.00787)	(0.01919;0.02174)	(0.02653;0.02974)	(0.01307;0.01465)	(0.01917;0.02306)	(0.00578;0.00921)
$(1,0)$	0.0013 *	0.00916 ***	0.0185 ***	0.00785 ***	0.01719 ***	0.00934 ***
	(-3e-05;0.00254)	(0.00782;0.01027)	(0.01679;0.02026)	(0.00681;0.00891)	(0.01462;0.01898)	(0.00736;0.01125)
$(0,1)$	0.00763 ***	0.02115 ***	0.02623 ***	0.01352 ***	0.0186 ***	0.00508 ***
	(0.00652;0.00868)	(0.0201;0.0221)	(0.02484;0.02775)	(0.01278;0.01434)	(0.01641;0.02033)	(0.00336;0.00665)
$(0,,0)$	0.00157 ***	0.00801 ***	0.01396 ***	0.00645 ***	0.01239 **	0.00594 ***
	(0.00049;0.00263)	(0.0069;0.00902)	(0.01234;0.01565)	(0.00604;0.00685)	(0.01024;0.01442)	(0.00399;0.00777)

5 Concluding Remarks

This work presents an innovative approach for exploring the causal relationship between the national immigration policies and income inequality. The employed methodology proposes a multi-valued characterization of the treatment variable and accounts for first-order spillover mechanisms. Results suggest that implementing severe policies has a negative impact on the income inequality, intensifying the dis-

crepancies between individuals. Specifically, enacting severe control mechanisms on migrants leads to a higher Gini index for income inequality.

Acknowledgements Irene Crimaldi, Fabrizia Mealli and Costanza Tortù are members of the Italian Group "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni" of the Italian Institute "Istituto Nazionale di Alta Matematica".

References

1. Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Skaaning, S.-E.,Teorell, J., Altman, D., Bernhard, M., Cornell, A., Fish, M. S., et al.: V-dem codebook v8 (2018). https://www.v-dem.net/media/filer_public/e0/7f/e07f672b-b91e-4e98-b9a378f8cd4de696/v-dem_codebook_v8.pdf
2. Cox, D. R.: Planning of experiments. Springer, Cox(1958)
3. Del Prete, D., Forastiere, L., Leone Sciabolazza, V.: Causal inference on networks under continuous treatment interference: an application to trade distortions in agricultural markets (2019). Available at SSRN 3363173. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3363173
4. Forastiere, L., Airoldi, E. M., Mealli, F: Identification and estimation of treatment and interference effects in observational studies on networks (2016). Available on arXiv: preprint arXiv:1609.06245. https://arxiv.org/pdf/1609.06245.pdf Forthcoming in *Journal of the American Statistical Association*.
5. Fouquin, M., Hugot, J., et al., 2016. Two centuries of bilateral trade and gravity data: 1827-2014. Tech. rep., Universidad Javeriana-Bogotá (2016) .
6. Helbling, M., Bjerre, L., Römer, F., Zobel, M.: Measuring immigration policies: The impic database. European Political Science **16**, 79–98 (2017).
7. Lopez, M. J., Gutman, R., et al: Estimation of causal effects with multiple treatments: a review and new ideas. Statistical Science **32 (3)**, 432–454 (2017)
8. Rubin, D. B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology **66 (5)**, 688.
9. Rubin, D. B.: Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association **75 (371)**, 591–593 (1980.
10. Tortù, C., Forastiere, L., Crimaldi, I., Mealli, F.: Modelling Network Interference with Multivalued Treatments: the Causal Effect of Immigration Policies on Crime Rates (2020). Available via arXIV, arXiv preprint arXiv:submit/3064478.
11. Wu, X., Mealli, F., Kioumourtzoglou, M.-A., Dominici, F., Braun, D.: Matching on generalized propensity scores with continuous exposures (2018).Available via arXIV, arXiv preprint arXiv:1812.06575. https://arxiv.org/abs/1812.06575

The job condition of academic graduates: a joint longitudinal analysis of AlmaLaurea and Mandatory Notices of the Ministry of Labour

Le carriere lavorative dei laureati: un'analisi longitudinale dei dati AlmaLaurea e delle comunicazioni obbligatorie del Ministero del Lavoro

Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini, Alessandra Petrucci

Abstract Education is one of the fundamental factors of development [1]. The evaluation of students' occupational status at different times -from six months to many years after graduation- is important to assess the university system's performance and contribute to its progressive improvement [2,3]. In the twenty-first century the Italian university system has experienced an expansion of enrolments benefiting from the Bologna Process. The expansion of higher education in Italy reduced the selectivity of courses increasing the heterogeneity of graduates. Thus, an academic degree may no longer be sufficient to access the most prestigious and remunerative occupational positions [9].In this perspective, the growth of the education system has led to an oversupply of graduates, especially in some fields, worsening the employment and working conditions of degree holders [14,5]. Relying of various sources of data, i.e. AlmaLaurea, the Mandatory Notices of the Ministry of Labour datasets and the administrative database of University of Florence (UNIFI) students, this work aims to provide an overview of UNIFI graduates' employment and labour market participation. A preliminary analysis is provided.

Abstract *L'istruzione rappresenta uno degli elementi cardine dello sviluppo economico. La valutazione degli sbocchi professionali dei laureati rappresenta un utile strumento per la definizione dell'efficacia del sistema universitario e per il suo conseguente miglioramento. Il trattato di Bologna ha favorito un importante*

[1] Maria Veronica Dorgali, University of Florence; email: mariaveronica.dorgali@unifi.it;
Silvia Bacci, University of Florence; email: silvia.bacci@unifi.it;
Bruno Bertaccini, University of Florence; email: bruno.bertaccini@unifi.it;
Alessandra Petrucci, University of Florence; email: alessandra.petrucci@unifi.it;

incremento delle iscrizioni all'interno degli atenei italiani, riducendo la selettività dei corsi e aumentando l'eterogeneità dei laureati. In questo modo, non solo un titolo accademico è diventato insufficiente a garantire un'occupazione remunerativa e prestigiosa, ma il crescente numero di laureati ha peggiorato fortemente le condizioni lavorative preesistenti. Utilizzando diverse fonti di dati, i.e l'indagine AlmaLaurea, le comunicazioni obbligatorie del ministero del lavoro e i dati amministrativi dell'Università di Firenze lo scopo del presente lavoro è quello di analizzare le condizioni lavorative dei laureati UNIFI al fine di valutare l'efficacia dei corsi universitari intrapresi dagli stessi. In questo estratto è presentata un'analisi preliminare dei risultati ottenuti.

Key words: Field of education; occupational outcome; university graduates; mixture cure models; Italy.

1 Introduction

Education is one of the fundamental factors of development [1]. Education human capital contributes to the economic growth in both developing and developed countries and has a positive and significant effect on labour productivity [2,3]. Thus, graduates' professional careers represent a tool to measure quality and effectiveness of university courses by providing an overview of the relationship between higher education and subsequent graduate employment [4,5]. The evaluation of students' occupational status at different times -from six months to many years after graduation- is important to assess the system's performance and contribute to its progressive improvement [6,7].

In the twenty-first century the Italian university system has experienced an expansion of enrolments benefiting from the Bologna Process. Moreover, after the significant increase in the number of graduates over the period 2001-2007, the number of new students started to decline in 2008 with the beginning of the economic crisis. The decline continued till 2012, when the number of graduates began to increase again, especially in northern and central Italian universities [8,5]. The expansion of higher education in Italy reduced the selectivity of courses increasing the heterogeneity of graduates. Thus, an academic degree may no longer be sufficient to access the most prestigious and remunerative occupational positions [9]. According to Rostan and Stan [5] two important theories may explain the characteristics of Italian graduates' employment and work conditions [10,8]. Firstly, the Italian economy seems to lack the characteristics, such as firms' size, innovation capacity and technology, to valorise and reward qualified human capital [11,12,13,14,15, 4,5]. Secondly, the expansion of higher education in Italy is often not associated with the demand of highly qualified human capital. In this perspective, the growth of the education system has led to an over supply of graduates, especially in some fields, worsening the employment and working conditions of degree holders [14,5]. The deterioration of employment conditions has resulted in stronger competition among graduates entering the labour market, where

selection processes started requiring a number of additional criteria. These go beyond the minimum job-related qualifications and more often focus on the type of institution attended at secondary or tertiary [15].

Nevertheless, much of the variation among graduates in occupation and earnings outcomes are related to disciplinary areas, with graduates from humanities and social sciences often penalized in terms of employability [16,17]. In fact, some qualifications adapt more easily to the demands of the market and are focused to develop human capital skills, whereas others are less sensitive to professional practice [7]. In Italy, Ballarino and Bratti [17] showed that in the decade 1995-2004 graduates in the 'hard' sciences obtained better rewarded occupations in terms of earnings and occupational status and were less likely to be unemployed or overeducated compared to graduates in the 'soft' social sciences and humanities.

Another important indicator of the performance of higher education institutions is the job-education mismatch. According to the existing literature, there are different ways to measure study mismatch. In general, the most part of studies focused on the qualifications or skills mismatch, which happens when a job-seeker may have to downgrade to find a job, but often this is different from field of study mismatch, which occurs when workers educated in a particular field work in another [18,19,20,21,22]. According to the OECD [22] Italy is one of the countries where more than 45% of workers are mismatched with respect to their field of study (FoS).

Relying of various sources of data, i.e Almalaurea, the Mandatory Notices of the Ministry of Labour datasets and the administrative database of University of Florence (UNIFI) students, this work aims to provide an overview of UNIFI graduates' employment and labour market participation. In order to outline the employment situation of UNIFI graduates, we will focus on three main aspects of the job experience, that is, the graduates employment condition at one, three and five years after graduation, the most common types of contract, the time needed to obtain a permanent position and the education-job match. Every objective will be evaluated comparing graduates from different FoS.

2 Data and methods

Three data sources will be integrated in the analysis of UNIFI graduates: the Mandatory Notice of the Ministry of Labour, the administrative database on UNIFI students and the AlmaLaurea Survey on Occupational condition of UNIFI graduates. The first source is an administrative dataset provided by the Ministry of Labour that reports information on all job contracts signed by graduates (except self-employment spells) in the years after graduation. The administrative source provides different information related with the job activity, such as the type of contracts held (open-ended, fixed term, short term, permanent, etc.), the number of working days per contract, the contract effective date, graduate age and gender, economic sector, etc. The second source contains a wide range of student's information: enrolment data, graduation data, previous education experience, high school, the course of study description and field of study. The third source is the AlmaLaurea Survey on

Occupational condition that provides information on post-graduate education and working career path of UNIFI first- and second-level graduates and five-years masters. On annual basis, the AlmaLaurea survey ensures a more than significant photograph of the employment path of graduates but is often characterized by a high nonresponse rate.
For the analysis of employment conditions of graduates, contingency tables and general descriptive statistics will be considered in order to identify the most important determinants of degree holder job career. In addition, mixture cure models will be employed in order to analyse time needed to obtain a job position, explicitly modelling the survival function as a mixture of 2 types of graduates: those who obtained a permanent position and those who did not.

3 Preliminary results

All the preliminary results presented were obtained from the database of Mandatory Notice of the Ministry of Labour and the administrative database on UNIFI students. The datasets were merged by a probabilistic record linkage. The archive contains data on about 262250 contracts signed by 46931 UNIFI graduated from 1 January 2008 to 31 December 2016. All the information refers to UNIFI students that obtained their degree between 2008 and 2016.
Trying to catch a snapshot of occupational outcomes of UNIFI graduates, Table 1 shows the percentage of signed contracts from 2008 to 2016. Overall, more than 60% of contracts were signed after graduation, the 37.17% within 3 years from graduation and almost the 29% more than 3 years after graduation.

Table 1: Number of signed contracts between 2008 and 2016

Number of signed contracts	N	%
While studying and ended	78,369	29.88
While studying and not ended	11,170	4.26
Within 3 year after graduation(3 years since gradutation)	97,469	37.17
More than 3 years after graduation	75,242	28.69
Total	262,250	100

Graduate employment can be analysed according to the type of contract graduates have. Focusing on the contract signed after graduation and on those signed while studying (or during university) the most common contract among UNFI students (bachelor and master level graduates and five-years masters) was the temporary one (59.13%); only the 10.35% of contracts were permanents. The 19.81% of contracts belongs to the category “Others” that includes “atypical” or “non standard” employments. Looking at the details bachelor and master graduates permanent contracts was, respectively, the third (8.77%) and the fourth (8.59%) most common.

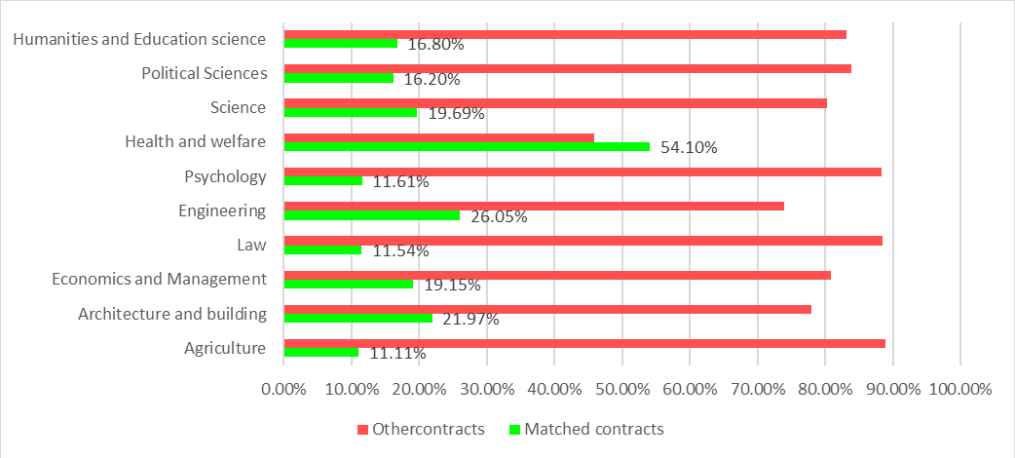
Looking at the difference by FoS, the average number of days of contract seems to differ across disciplinary areas and between bachelor and master students, but a high variability is observed for all the disciplinary areas. The average number of days increases for bachelor and master, except for Psychology that shows the lowest number of days for bachelor and master among all FoS. In the master group, Economics and Management, Health and Engineering show the highest average of days of contract.

Table 2: Average number of days of contracts and FoS (SD)

Field of study	Bachelor's degree	Master's degree and five-year master
Psychology	119.579(358.345)	118.39(336.86)
Humanities and Education science	122.104(312.233)	145.455(380.79)
Political Sciences	140.769 (378.678)	174.389(438.68)
Agriculture	142.23(301.814)	166.873 (309.85)
Science	153.767(314.434)	179.332(341.37)
Architecture and building	156.242(333.150)	113.45(340.22)
Economics and Management	167.152(341.708)	228.448(466.73)
Law	179.646(453.224)	164.977 (378.04)
Engineering	181.267(355.936)	259.618(495.81)
Health and welfare	213.424 (627.291)	228.811(852.59)

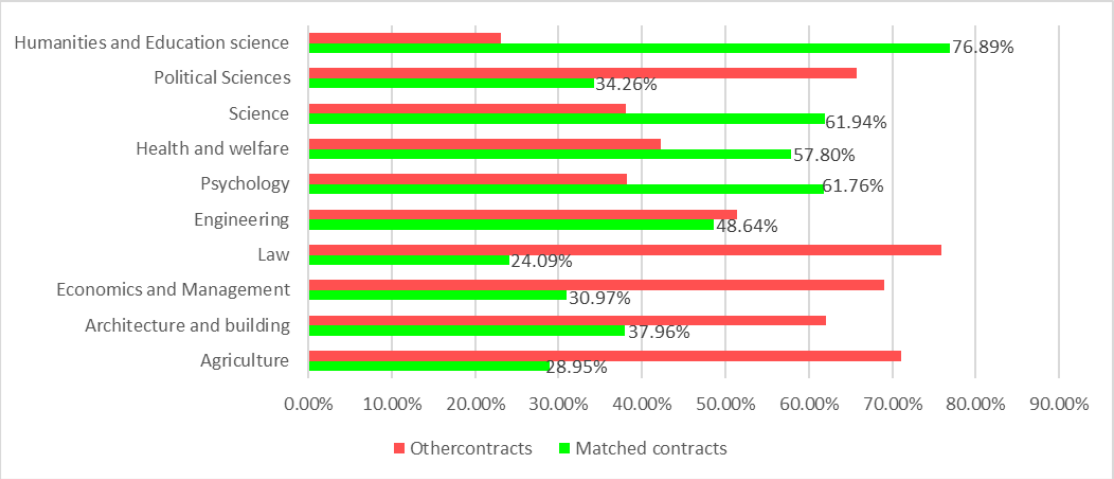
Qualification and skill mismatch was explored according to the ISTAT employment and work qualification list. In the list, employments are classified according to 9 main groups. Among bachelor graduates job-study mismatch can be identified looking at the percentage of contract belonging to the "TECHNICAL PROFESSIONS", whereas second cycle and single cycle are evaluated according to the "INTELLECTUAL, SCIENTIFIC AND HIGHLY SKILLED PROFESSIONS". As shown in Figures 1, among bachelor graduates only Health and Welfare seems to present the lowest percentage of mismatched contracts. In general, all FoS show a lower match between job and education attained in terms of skill and qualification (22.22%).

Figure 1: Study mismatch among first cycle graduates.



For master holders, Humanities and Education, Science, Health and Welfare, Psychology are less likely to be mismatched whereas the lowest percentage of matched contracts is showed by Law and Agriculture, as shown in Figure 2.

Figure 2: Study mismatch among secon and single cycle graduates.



4 Preliminary conclusions

Preliminary results revealed that graduates professional outcomes, in terms of average number of contract days, type of contracts and study mismatch seem to differ across disciplinary area. Further analysis are needed to consistently evaluate UNIFI graduates professional outcomes. Mixture cure model will be employed in order to analyse time needed to obtain a job position, explicitly modelling the survival function as a mixture of 2 types of graduates: those who obtained a permanent position and those who did not.

References

1. Ozturk, I.: The Role of Education in Economic Development: A Theoretical Perspective. Available at SSRN: https://ssrn.com/abstract=1137541.
2. Wang, Y. and Liu, S.: Education, Human Capital and Economic Growth: Empirical Research on 55 Countries and Regions (1960-2009). Theoretical Economics Letters, 6, 347-355. doi: 10.4236/tel.2016.62039 (2016). (https://www.researchgate.net/publication/301720577_Education_Human_Capital_and_Economic_Growth_Empirical_Research_on_55_Countries_and_Regions_1960-2009[accessed Mar 03 2020]).
3. Sonmez, F.D. and Sener, P.: Effects of Human Capital and Openness on Economic Growth of Developed and Developing Countries: A Panel Data Analysis. World Academy of Science, Engineering and Technology, No. 30, 1242- 1246 (2009).
4. TEICHLER, U.: Bologna – Motor or Stumbling Block for the Mobility and Employability of Graduates?, in H. Schomburg and U. Teichler (eds.), Employability and Mobility of Bachelor Graduates in Europe. Key Results of the Bologna Process, Rotterdam, Sense Publishers, pp. 3-41(2011).
5. Rostan, M., Stan., A.: Italian graduates' employability in times of economic crisis: overview, problems and possible solutions. Sociologica. SerieII (2017). doi: 10.4000/sociologico.1818.
6. Salas Velasco, M.: The transition from higher education to employment in Europe: the analysis of the time to obtain the first job. A: Higher Education, 54, pp.333-360 (2007).
7. Saurina,C., & Esperanca, V.: The match between university education and graduate labour market outcomes (education-job match): an analysis of three graduate cohorts in Catalonia. vols. Studies on higher education and graduate employment. Barcelona: Agencia per a la Qualitat del Sistema Universitari de Catalunya = Catalan University Quality Assurance Agency, 2010. [Online. Available: http://www.aqu.cat/doc/doc_12987231_1.pdf.]
8. Ballarino, G.: "Gli esiti occupazionali dei laureati", in P. Trivellato and M. Triventi (eds.), L'istruzione superiore: caratteristiche, funziona mento e risultati, Roma, Carocci, pp. 213-245 (2015).
9. Breen, R., & Goldthorpe, J. H.: Explaining educational differentials: Towards a formal rational action theory. Rationality and Society, 9(3), 275–305. (1997).
10. Ballarino, G.: Sistemi formativi e mercato del lavoro, in M. Regini (ed.), La sociologia economica contemporanea, Roma-Bari, Laterza, pp. 231-257 (2007).
11. Cipollone, P. e P. Sestito: Il capitale umano, Bologna, Il Mulino (2010).
12. VISCO, I.: Investire in conoscenza. Per la crescita economica, Bologna, Il Mulino (2009).
13. REYNERI, E.: Sociologia del mercato del lavoro. Vol. I, Il mercato del lavoro tra famiglia e welfare, Bologna, Il Mulino (2005).
14. Ballarino, G.: Le politiche per l'università, in U. Ascoli (ed.), Il welfare in Italia, Bologna, Il Mulino, pp. 197-223 (2011).

15. Macmillan, L., Tyler, C., & Vignoles, A. : Who gets the top jobs? The role of family background and networks in recent graduates' access to high-status professions. Journal of Social Policy, 44(3), 487–515 (2015).
16. Argentin, G., & Triventi, M. : Social inequality in higher education and labour market in a period of institutional reforms: Italy, 1992-2007. Higher Education, 61(3), 309–323 (2011). doi:10.1007/s10734-010-9379-6
17. Ballarino, G., & Bratti, M.: Field of study and university graduates' early employment outcomes in Italy during 1995-2004. Labour, 23(3), 421–457(2009). doi:10.1111/j.1467-9914.2009.00459.x.
18. Sloane, P.J.: Much Ado about Nothing? What Does the Overeducation Literature Really Tell Us, in F. Büchel, A. de Grip, Antje Mertens, Overeducation in Europe: Current Issues in Theory and Policy, pp. 11-45 (2003).
19. Robst, J.: Overeducation and College Major: Expanding the Definition of Mismatch between Schooling and Jobs, The Manchester School, 76(4), pp. 349-368 (2008).
20. Ortiz, L. and A. Kucel: Do Fields of Study Matter for Over-Education? The Cases of Spain and Germany, International Journal of Comparative Sociology, 49(4-5), pp. 305-327 (2008).
21. Quintini, G.: Over-Qualified or Under-Skilled: A Review of Existing Literature, OECD Social, Employment and Migration Working Papers, 121 (2011). doi:/10.1787/5kg58j9d7b6d-en.
22. OECD: The causes and consequences of field of-study mismatch: An analysis using PIAAC Social, Employment and Migration Working Papers No. 167 (2015). doi:10.1787/5jrxm4dhv9r2-en]

The joint effect of childcare services and flexible female employment on fertility rate in Europe

L'effetto congiunto dei servizi di assistenza per l'infanzia e della flessibilità lavorativa femminile sul tasso di fecondità in Europa

Viviana Cocuccio and Massimo Mucciardi

Abstract This research aims to deepen the effect of childcare services jointly to the flexible female employment on Total Fertility Rate (TFR) within the different European states contexts. In the 2002, the Barcelona European Council set objectives with regard to the availability of high quality and affordable childcare facilities for pre-school children, to improve the employment female rate and the TFR. Considering a spatial approach, the authors analysed how the current European situation in terms of childcare and employment policies are decisive on TFR levels

Abstract *Questo lavoro si propone di approfondire l'effetto dei servizi di assistenza per l'infanzia congiuntamente alla flessibilità del lavoro femminile sul Tasso di Fecondità Totale (TFT) nei differenti contesti nazionali europei. Considerando che già nel 2002 il Consiglio U.E. di Barcellona ha sancito gli obiettivi di elevate qualità e accessibilità dei servizi di assistenza alla prima infanzia al fine di migliorare il tasso di occupazione femminile e il TFT. Considerando un approccio spaziale, gli autori hanno analizzato quanto l'attuale situazione europea in termini di assistenza all'infanzia e di politiche occupazionali sia determinante sugli attuali livelli di TFT*

Keywords: Fertility, Childcare, Flexible Female Employment, Spatial Analysis

1 An overview of childcare services, flexible female employment and fertility in Europe

Many researchers have shown how new childcare policies are positively determinant on the female employment and, consequently, on the grown of Total fertility Rate (TFR) (Rindfuss et al., 2011). With "childcare policies", we mean

policies that support families like family allowances, public and private childcare services, parental leaves and tax exemptions. About the childcare services (CS) and the reproductive behaviour of the European regions, some researcher explain the evident stabilization of the phenomenon of fertility decline in Europe in the last decades (Lutz and Skirbekk, 2007). Besides the cultural causes of adaptation, however, there is the efficiency of early childhood care services on the growth of the fertility rate. According to this hypothesis, the surveys which were in the Eurobarometer 2006 (Testa, 2006), show that people who said they haven't fulfilled their desires on the childbearing, often said they didn't have the availability of a parental leaves and childcare, thus the presence of formal or informal childcare resources are determinant on the decision-making whether to have or not have a child. For instance, a cross-national study of 21 OECD countries shows a statistical positive correlation between the fertility behaviour of the European female employment and the presence of formal childcare services in their territory (Blum, 2012). After 1980 policies were introduced and directed at the reconciliation of work and family life as paid parental leaves and the TFR was grown. Let's briefly analyse some European cases. The Scandinavian countries represent an example of this good influence of CS on TFR, as it is proven by some recent literature. Rindfuss et al. (2007) with the use of comparative and spatial statistical models, has shown how in Norway there was a growth of childbirth rate for women of child-bearing age equal to 0.5 and 0.7 in correspondence with the implementation of child policies-care aimed at preschool children. That was a real welfare response favourable to the female worker and consequently to the family. Because of the aging of the population, Germany has decided to prevent the catastrophic socio-economic consequences of this phenomenon by implementing policies in Nordic style. One of these policies was the choice to realize the right of childcare for all preschool children by 2013. One of these consists to ensure that all children under the age of 3 have access to childcare services. A fertility study in Germany by Bauernschuster et al., (2012) shows that there is a correlation between the existence of more CS and the growth of local fertility rates, especially in the countries with a high level of education and less conservative cultural attitudes. Therefore, not only in the Scandinavian countries but also in countries which are similar to Italy but have culture and social conditions such as Germany, the strategy of implementation of early childcare policies present itself as a winning solution to the problem of low fertility. A recent paper by P. Baizàn (2009) supports the hypothesis of the positive correlation between early childhood care policies and growth in fertility rates, also in the Spanish case. This study highlighted important territorial differences growth: public CS increase more frequently in the regions of Spain where there was a higher rate of female employment, followed by a growth in fertility rates in these regions. In Russia there was an important demographic decline from 1990 to 2009, caused by emigrations of young aged people for the urban and rural unemployment and high death rates. Answering to this decline, Russia approved many new policies in the 2000s. New additional measures of public support for families with children had a positive impact on the TFR, through an increase of the per capita income, which allowed a decline in youth emigration. There were pro-natal measures and incentives for families with second and subsequent births since 2007, for the mothers

especially who are non-workers and have one or more children following the first. Following these policies to support the birth rate, Russia had experienced a “true demographic explosion. About the Russian case, the research of Miljkovica and Glazyrina (2008) on TFR in Russia shows us the unemployment phenomenon as a negative determinant on the percentage of fertility. Infact, as the level of unemployment increases, there is a decrease of the family income that make grown the perception of uncertainty that influence the procreative choice. At the contrary, France is an anomalous case given by the apparent similarity of the French context to neighbouring nations with low TFR. France recently has exceeded the TFR rates historically typical of the Scandinavian countries: in 2015, French’ TFR was equal to 1.96 while the Swedish one was 1.85. In France, policies in favour of the birth rate are not limited to classic monetary incentives such as the "baby bonus" but there are measures that counteract the progressive aging of the population through concrete services and work facilities that support families who embrace the choice of procreation. First of these policies was to guarantee ‘free kindergartens’ distributed throughout all the territory. Generating children in France is an event calculated in terms of merit for professional and pension purposes: for the calculation of the pension the period of maternity and leave for breastfeeding or other early childhood care is counted in favour of the female worker and contributions are paid by the state. This policy is the result of an ideological battle contaminated by pro-nativist values and gender equality supported by the state and the feminist movement. According to the study of Toulemon et al. (2008), the virtuous rate of fertility in France can be explained by the policies supporting fertility aimed directly at supporting the working mother and the not working one and many other political initiatives to combat youth unemployment. These can favourite the reconciliation of family and working time and should be replicable in other similar countries.

2 An attempt to estimate using a spatial approach

In order to identify the joint effect of CS and flexible female employment (FFE) on TFR, we use a geographically weighted regression (GWR) approach (Fotheringham et al., 2002). In the current state of research, the variables considered at member states level (NUTS0) are: average age of women at childbirth (year) (W_AGE_M); part-time female employment (W_E_PRT) (%); children aged 0-3 who use childcare services for less than 29 h a week (%) (CH_LESS). All the data is extracted from Eurostat dataset (2019) and are calculated considering the averages for the period 2014-2017. We analysed the results in two steps: first considering the global model (OLS regression); then considering the local model (GWR). To evaluate the spatial variability of the regression coefficients we refer to the Geographical Variation Test (Nakaya, 2015). As it is easy to see from the results obtained from the OLS model (table 1), the variable W_AGE_M is negatively correlated with the TFR. This correlation is probably due to the extension of the training education period and to the time taken to find employment, confirming the

Picchio's thesis (Picchio M. et al., 2018). Another classic thesis seems to be confirmed by our results. More specifically, the variable W_E_PRT is positively correlated with the fertility rate, thus the reproductive behaviour should be conditioned by the part-time female employment. This result probably explains that women need to feel more economic and social secureness to make this decision, but they prefer to spend time for family care too. Moreover, the prospection to have children is correlated with the use of CS too (see CH_LESS variable) which involves only less hours a week (less than 29 hours).

Table 1: Estimates and test for the OLS and GWR model[1]

Variable	OLS	Min	Lower quartile	Median	Upper quartile	Max
Intercept	3.54**	-1.315	2.549	2.808	3.135	8.089
W_AGE_M	-0.072**	-0.221	-0.059	-0.048	-0.039	0.101
W_E_PRT	0.008**	-0.006	0.006	0.007	0.008	0.068
CH_LESS	0.007*	-0.010	0.002	0.004	0.006	0.014

Global regression results (OLS): AICc=-20.66; Adj-R-square=0.25; *=p<0.05; **p<0.01
GWR results: AICc=-27.12; Adj-R-square=0.54; GWR ANOVA Test: F=3.86
GVT test results[#]: -88.14 (W_AGE_M); -8.56 (W_E_PRT); 3.37 (CH_LESS)
[#] Positive value of GVT suggests no spatial variability (Nakaya, 2015).

It is important to highlight how the variables W_AGE_M and W_E_PRT have a spatial variability respect to the variable CH_LESS which has a homogeneous impact on fertility on the European territory (see the positive GVT value in table 1). For the W_AGE_M and W_E_PRT variables, the GWR model detects a spatial variability of the coefficients (see the range of the coefficients from the min to the max value in table 1). Although there are some variabilities in the estimates, for the variable CH_LESS spatial model (GWR) does not reject the null hypothesis of absence of spatial variability. This should mean that the presence of CS is a crucial element on the reproductive choice, regardless of the territorial contest. Instead, as regarded the determinants W_AGE_M and W_E_PRT, the GWR model reveals a territorial influence of these regressors on the TFR. These variables affect the TFR differently according to the European local context.

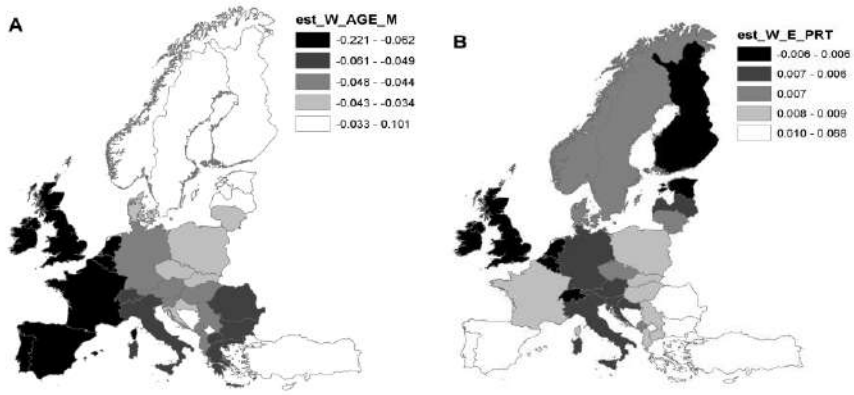
As it's known, analysing the spatial distribution of the TFR at European level[2], it's higher in Scandinavian countries, in Turkey and French, while the lower European TFR are in the Mediterranean part of Europe and in Poland, where the policies of childbirth and the FFE probably are not efficient to reconcile the work and family-caring time. From the territorial point of view, the FFE is present mostly in the Scandinavian states, UK and Netherland. It is less common in France, Turkey, Greece, German and Iberian and Baltic nations. Probably, this is just the useful time to help women to reconcile the part-time working and family care time and the

[1] The data was appropriately geocoded and processed using GWR 4.0 (estimates) and Arcgis 10.2 software (maps). Furthermore, for methodological reasons we exclude Iceland from the analysis.

[2] The maps of the TFR, W_AGE_M, W_E_PRT and CH_LESS variables are not shown for reasons of space but are available on request

services for childcare services for 0-3 year-old-children are used for few hours a week according to the part-time, seems to be more used in northern and west European states. Examining the childcare variable, it's possible to see how data confirms thesis of Sonja Blum's survey (Blum, 2012) that showed how the public childcare policies have almost everywhere an important role in family's decision-making on their demographic behaviour. Also other studies confirm our results: we refer, for example, at Baizàn's analyse (Baizàn, 2009), who underline the positive effect on fertility of formal childcare for children under the age of 3. However it will be necessary to investigate better, considering the joint effect between these two effects (CS and FFE). According to Gesano's study (Gesano, 2017) in a socio-political context that pushes women to extend the period of training to hyper-specialize and compete for managerial work positions, it becomes natural to postpone the age of the first birth. To limit this, the presence of efficient early CS becomes essential. In particular, Gesano's study shows how the average age of woman at the first birth is low in the provinces of Eastern European countries, while it is particularly advanced in Sardinia (Italy) and other regions of central and southern Italy and Spain. The results of ~~a~~ the research by Gustaffson and Kalwij (2006) consolidate this view: it seems that higher educational level of the woman makes lower chances of birth before the average age of woman at birth and at births subsequent to the first. This means a negative relationship between fertility and education level of woman. This dynamic reduces the demand for the quantity of children, which allows women to postpone childbirth and increase investment in human capital in a few children (i.e. higher quality demand). In conclusion, from our maps results (figure 1) it is possible to see the jointly influence of the regressors on TFR in some European countries. There are nations where some regressors have ~~a~~ stronger influence on TFR and other regressors that have a poor influence on the TFR. This spatial methodology allows us to show how different are the phenomenon as the CS and the FFE with the average age of giving birth, probably influence the reproductive choice in different local way. However a research on new models is currently in progress by authors on these topics to deepen the influence of the factors analysed by Gesano and Gustaffson and Kalwij, who are not considered on this European fertility research because of the lack of availability of territorial and/or temporal data. It would be interesting, for instance to correlate the average age of women at childbirth with the extension of the educational period on the female employment.

Figure 1: Local coefficient estimates of W_AGE_M (fig. 1A) and W_E_PRT (fig. 1B) by quintiles range



References

1. Baizàn, P., (2009): Regional childcare availability and fertility decisions in Spain, Demographic Research, volume 21, article 27.
2. Bauernschuster S., Rainer H., (2012) "Political regimes and the family: how sex-role attitudes continue to differ in reunified Germany", in J Popul Econ (2012) 25:5–27 DOI 10.1007/s00148-011-0370-z
3. Blum, S., (2012): Family policies and birth rate: evidence and challenges of European countries, Briefing Paper, Shanghai Coordination Office for International Cooperation.
4. EUROSTAT, (2019): https://ec.europa.eu/eurostat/data/database
5. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: the Analysis of Spatially Varying Relationships. Wiley.
6. Gesano G., (2019): "La riproduzione in Italia e nelle sue regioni nel quadro delle dinamiche demografiche in Europa",CNR, IRPPS e e-publishing, DOI: 10.14600/978-88-98822-15-7 Roma
7. Gustafsson S, Kalwij A. "Education and Postponement of Maternity: Economic Analyses for Industrialized Countries European Studies of Population". Vol. 15. Dordrecht: Kluwer Academic Publisher, Springer; 2006
8. Kalwij A., (2010): "The impact of Family Policy Expenditure on Fertility in western Europe" in Demography, May 2010; 47 (2): 503-519, PMCID: PMC3000017
9. Lutz W., Skirbekk V., Testa M. R., (2007) The Low-Fertility Trap Hypothesis: Forces that May Lead to Further Postponement and Fewer Births in Europe RP-07-001 March 2007
10. Miljkovica D., Glazyrina A., (2008) "The impact of socio-economic policy on total fertility rate in Russia", in Science Direct, Journal of Policy Modeling 37 (2015) 961–973
11. Nakaya, T. (2015). Semiparametric geographically weighted generalized linear modelling: the concept and implementation using GWR4. Brunsdon, C., & Singleton, A., eds. Geoco(Sage Publication), 201–220.
12. Picchio M., Pigini C., Staffolani S., Verashchagina A., (January 2018) "If Not Now, When? The Timing of Childbirth and Labour Market Outcomes" from IZA Institute of Labor Economics DP No. 11270
13. Toulemon L., Pailhé A., Rossier C., (2008) "France: High and stable fertility" in Demographic Research, Max Planck Institute for Demographic Research, 19 (16), pp.503-556. ff10.4054/DemRes.2008.19.16ff. ffhal-02081757f

The Left Behind Generation: How the current Early School Leavers affect tomorrow's NEETs?

La generazione degli emarginati: Quanti di coloro che abbandonano la scuola oggi si trasformano nei NEET di domani?

Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano, Antonella Rocca

Abstract The youth labour market in Italy is among the most problematic ones. Indeed, across European countries, Italy shows the highest share of young people (16-34 years), not in Employment, Education or Training (NEETs). Furthermore, while at European level, the share of high educated is one of the lowest ones, Italy is among countries with the highest share of Early School Leavers, which are those who leave education after the achievement of compulsory education or even less. In line with a great amount of literature on human capital, highlighting the relationship between the level of education attained and the performance in the labour of market, this paper aims at analyse the link between the decision to leave prematurely the studies and the future status of NEETs in Italy. At this end, the authors estimate a time-varying correlation model to a four-weekly time series ad hoc constructed by Labour Force Survey.

Abstract *A livello europeo, l'Italia è il paese con la più elevata percentuale di giovani NEET (giovani tra i 16 ed i 34 anni che non lavorano, non studiano né svolgono alcun tipo di attività di formazione). In Italia, inoltre, la quota di giovani laureati è la più bassa mentre la quota di coloro che smettono di studiare dopo il conseguimento della sola scuola dell'obbligo, o anche prima (Early School Leavers), è tra le più alte. In linea con un ampio filone della letteratura sul capitale umano, che ricollega le difficoltà di inserimento dei giovani nel mercato del lavoro al loro basso profilo educativo, questo lavoro analizza il legame tra la decisione di lasciare prematuramente gli studi ed il conseguente status di NEET di Italia. A tal fine, gli autori stimano un modello di time-varying correlation su una serie storica costruita sui dati dell'Indagine forze di lavoro per il periodo 2007-2017.*

[1] Giovanni De Luca, giovanni.deluca@uniparthenope.it;
Paolo Mazzocchi, paolo.mazzocchi@uniparthenope.it;
Claudio Quintano, claudio.quintano@uniparthenope.it;
Antonella Rocca, rocca@uniparthenope.it; Department of Management and Quantitative Studies, University of Naples "Parthenope", Naples, Italy.

Key words: Early school leavers, NEETs, Youth unemployment rate

1. Introduction

Young people suffer a particular condition of disadvantage in the labour market in comparison to their older peers. When they leave school, they meet many obstacles to find a job for the lack of job experience and in many cases for the absence or an inefficiency of the instruments regulating the School-to-Work transition. This latter refers to the time from the end of the studies until the achievement of a stable job. In Italy, this period is very long, and its causes cannot be ascribed only to the high levels of unemployment. Indeed, in Italy the ratio between the youth unemployment rate and the adult unemployment rate, defined as the young people relative disadvantage, is one of the highest ones.

As the difficulties met by young people in the labour market may induce, besides unemployment, inactivity, the most appropriate indicator in order to study the youth condition on the labour market is represented by the NEET rate. It expresses the percentage of young people Not in Employment, Education and Training (aged 16-34 years). In the last years, a great amount of literature has underlined the relevant role of education as main human capital factor in determining the employability chances of individuals. Looking at the level of education, Italy shows the lowest share of tertiary educated among young people while the share of who leave school after completing compulsory school or before (Early School Leavers, ESLs) is one of the highest one. In this paper, we try to verify to what extent the highest share of NEETs in Italian young population depends on the high share of ESLs.

The structure of the paper is as follows: Section 2 introduces the framework of analysis; Section 3 shows the methodology while Section 4 presents the main results. Some conclusive considerations follow.

2. The framework of analysis

The high share of NEETs among the young population is certainly due to the high vulnerability of young people in the labour market. However, this phenomenon shows a great variability across Europe. Indeed, in 2017, the share of unemployed in the age class 18-24 ranged from the 43.6% of Greece to the 6.8% of Germany. Italy with 34.7% was just overcome by Greece and Spain. Conversely, according to NEETs, Italy, with 25.5%, shows the highest share, against a EU-28 mean of only 14.7%. The causes of so high share of NEETs in countries like Italy are certainly due to the difficulties met by young people to enter the labour market. This depends on individual characteristics and macro-economic factors linked to the labour market and the institutions in force on it. However, as a great amount of literature

highlights, these difficulties are also due to the education system, its inability to attract and keep young people and to transmit the skills and competences requested by employers (Pastore, 2014). Indeed, among the factors strongly increasing the chance to be employed for an individual, the level of education plays a crucial role (Becker, 1967). It is therefore not surprisingly that the highest share of unemployed and inactive can be found among low educated. Within the objectives contained in the United Nations 2030 Agenda for Sustainable Development Goals (SDGs), the Goal 4, Quality of Education, refers at subpoint 4.3 to the reduction of ESLs to less than 10%. However, among EU countries, many of them are very far from this objective. In 2017 the share of ESLs ranged from the 18.3% of Spain to the 3.1% of Croatia. Italy, with a share of 14%, just follows to Spain, Malta and Romania, against a mean value at EU-28 level of 10.6%.

Repeated failures in job search may induce to two different conditions. The first one is unemployment and consequent risk of falling into a trap of low-paying, temporary, unstable, or part-time jobs (Caroleo et al. 2018; Pastore 2014). The second possible consequence consists in discouragement, which is when people do not look for a job simply because they believe the search will fail, provoking inactivity (Finegan 1978; Furlong 2006). Unemployed and inactive individuals acquire the NEET status which, therefore, better than the youth unemployment rate is useful to highlight the youth condition on the labour market.

Therefore, in order to investigate on the causes of the NEET status, it is crucial to identify the causes in terms of the different connected characteristics.

The left side of Fig.1 shows the share of NEETs and of ESLs who are inactive. In Italy inactive are at least the 50% of total NEETs among males while among females they reach almost the 70%. This outcome highlights the remarkable difference between Italy, on the one side, and the other countries like Greece and Spain, on the other side. These latter countries share with Italy the high levels of youth unemployment rate, but while in Spain and Greece the high share of NEETs is mainly the result of high levels of unemployment, in Italy the determinants are not only explainable in terms of unemployment.

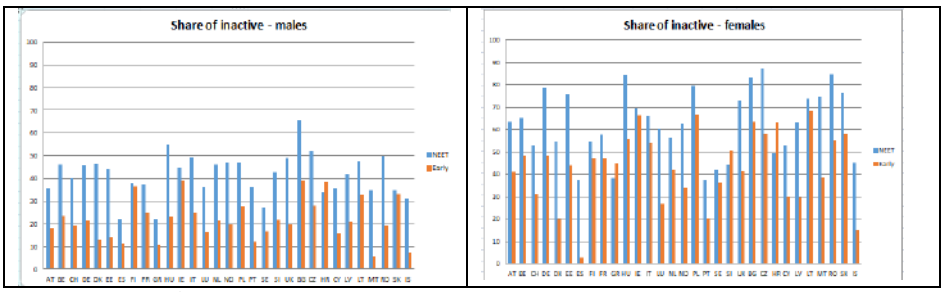


Figure 1: *Share of inactive on total NEETs by gender in EU countries in 2017.*
Source: Ad hoc elaboration on Labour Force Survey Data (2017).

Since the individuals' employability chances, and therefore the probability of becoming NEET, is affected by the level of education system, in Fig. 2 we have reported, on the left side, the share of ESLs who have become NEET and, on the

right side, the share of NEET who are ESLs. Even in this case, data highlight relevant differences across countries. The share of ESLs who have become NEET is very high in Italy, reaching the 45% among males and the 70% among females. Looking at the total of NEETs, those who are also ESLs are more than 40% among males and just a little bit less among women. Across EU countries, only Hungary and Denmark show highest percentages. Therefore, in the following we try to investigate on the relationship between NEET and ESL rates in terms of correlations on ad hoc constructed time series. The analysis is limited to the Italian data.

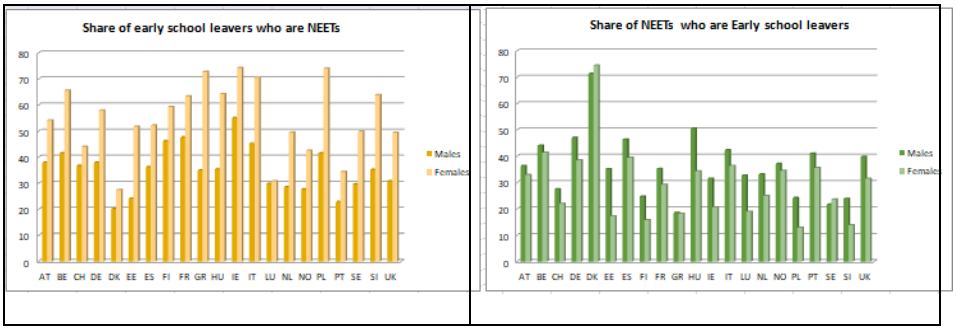


Figure 2: *Share of ESLs on total NEETs and share of NEETs on total ESLs by gender in EU countries in 2017.*
Source: Ad hoc elaboration on Labour Force Survey Data (2017).

3. The statistical methodology

In order to understand how the condition of ESL can lead to NEET status, starting from the Labour Force Survey, we have constructed an ad hoc time series based on four-weekly observations from 2007 to 2017, for a total number of 143 observations. Then we have investigated the dynamic relationship between NEET and ESL rates based on the estimation of a time-varying correlation. Indeed, given that in the life-cycle of an individual, after leaving school, the NEET status can occur immediately or after a certain temporal lag, we have estimated the correlation between $NEET$ and ESL, at time t and lag k

$$Corr_{t,k}(NEET, ESL) = \frac{Cov_{t,k}(NEET, ESL)}{\sqrt{Var_t(NEET)}\sqrt{Var_{t-k}(ESL)}}$$

with k = 1, 13. The two lags correspond, respectively, to a four-week and on-year periods.
The time-varying nature of the correlation has been captured using a moving-window with size s = 26, that is a two-year window, such that at time t

$$Cov_{t,k}(NEET, ESL) = \frac{1}{s-1}\sum_{i=t}^{t-s+1}(NEET_i - \overline{NEET}_{t:t-s+1})(ESL_{i-k} - \overline{ESL}_{t-k:t-k-s+1})$$

where $\overline{NEET}_{t:t-s+1}$ is the average of the 26 observations of the variable NEET from time t backward to $t-25$ and $\overline{ESL}_{t-k:t-k-s+1}$ is the average of the 26 observations of the variable ESL from time $t-k$ backward to $t-k-25$. Finally, $Var_t(NEET)$ and $Var_{t-k}(ESL)$ are the variance estimated using again 26 observations.

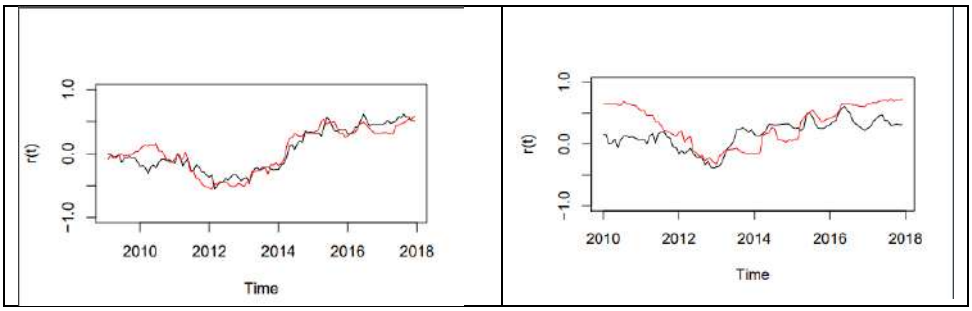


Figure 3: *Moving-window time-varying correlation based on lag=1 (left) and lag=13 (right). The red line is for females and the black one for males.*
Source: Ad hoc elaboration on Labour Force Survey Data (2017).

4. Results

Looking at the results of the rolling correlation, the lag 1 plot, on the left side of Fig.3, shows a null correlation at the beginning of the period in analysis, followed by a negative correlation in the years of the economic crisis, approximately between 2011 and 2014. Finally, starting from 2014, the association between the two variables is positive and increasing, in light of the low descending trend of ESLs and of the slightly decreasing pattern of NEETs of these last years. In any case, the NEET trend shows a persistence on higher levels than the situation prior to the crisis. No particular differences arise when we consider the gender perspective: at the end of the sample period, both for males and for females a positive and increasing dependence is detected. At lag 13, the relationship between NEET and one-year-lagged ESL does not show any relevant difference for male gender, even if the minimum and maximum values are less extreme. For women, there is evidence of a positive correlation at the beginning of the period, which decreases towards zero and finally becomes positive and increasing.

5. Conclusions

High share of NEETs represents a huge loss for a country economic growth. Among European countries, Italy shows the highest share of NEETs. One of the most important actions to take in order to reduce the high share of NEETs consists in stimulating young people to attain higher level of education, which can improve their skills and employability. A recent study demonstrated that the different capacity of the European countries to recover by the great recession can be explained through the strong role played by higher levels of education, which can reduce unemployment and inactivity (Pompei and Selezneva, 2018). In this paper, we tried to identify the main peculiarities of the condition of young people in Italy, where the share of NEETs is the highest one, despite the fact that the youth unemployment rate, even if high, is not maximum. The analysis shows that, besides the major propensity to inactivity, characterizing especially young women, the link between the NEET status and the low level of education is very strong. Compared to the other EU countries, in Italy the probability that an ESL becomes NEET is one of the highest ones. It is also higher than the Spanish one, where the share of ESLs is maximum and the levels of youth unemployment rate higher. After Denmark, Hungary and Spain, Italy shows also the highest share of low educated NEETs. The results of the rolling correlation add to this first results other relevant evidence. The connection in terms of correlation between the status of ESL and the status of NEET is very high and, especially for females, this relationship is even increased in these last years, when the economic crisis has been widely overcome. In a gender perspective, we can highlight that the differences between men and women, initially negligible, tend to amplify. Being young represents a factor of penalization on the labour market. However, when other conditions of vulnerability add, such as being low-educated, the difficulties dramatically arise. Finally, being female further increases the previous disadvantage, making still more difficult escape from a future of long-term unemployment or inactivity, that is from the NEET status.

References

1. Becker, G.: Human Capital and the Personal Distribution of Income: An Analytical Approach. University of Michigan Institute of Public Administration, Ann Arbor, MI (1967)
2. Caroleo, F.E., Demidova O., Marelli E. and Signorelli E. (eds.): Young people and the Labour Market. A comparative perspective. New York: Routledge Studies in Labour Economics (2018)
3. De Luca, G., Mazzocchi, P., Quintano, C. Rocca, A.: Italian NEETs in 2005–2016: have the Recent Labour Market Reforms Produced Any Effect? CESifo Economic Studies, 65(2), pp. 154–176 (2019)
4. Finegan, T.A.: Should Discouraged Workers Be Counted as Unemployed?, Challenge. 21(5), pp. 20-25 (1978)
5. Furlong, A.: Not a very NEET Solution: Representing Problematic Labour Market Transitions among Early School Leavers, Work. Employment and Society, 20(3), pp. 553–569 (2006)
6. Pastore, F.: The youth experience gap: Explaining national differences in the school-to-work transition, Berlin: Springer Verlag (2014)

The probability to be employed of young adults of foreign origin

La probabilità di essere occupati dei giovani adulti di origine straniera

Alessio Buonomo, Francesca Di Iorio and Salvatore Strozza

Abstract Our study focuses on a sub-sample of the 2011-2012 ISTAT survey on 'Social Condition and Integration of Foreign citizens' (SCIF). A Probit model is implemented in order to analyse the determinants of the probability to be employed of the young adult (aged 18-34) foreign population. Our findings prove the central role played by gender and citizenship at birth, without significative effects of migratory generation. The importance of the cultural integration emerges for women.

Abstract *Su un sotto-campione dell'indagine ISTAT del 2011-2012 su "Condizione e integrazione sociale dei cittadini stranieri" è stato implementato un modello Probit al fine di analizzare le determinanti della probabilità di essere occupati dei giovani adulti (18-34 anni) stranieri o di origine straniera. È confermato il ruolo centrale del genere e della cittadinanza di origine, mentre la generazione migratoria non appare significativa. Tra le donne emerge l'importanza dell'integrazione culturale.*

Key words: Employment, Foreigners, Migratory generations, Probit model, Italy

1 Introduction

Foreign immigration to Italy has a history that began more than 40 years ago (Strozza 2016). Since the end of the nineties, literature has focused attention on immigrants' descendants (Ambrosini e Molina 2004). Scholars proved that students with non-Italian citizenship are more likely to have lower levels of education, poorer

[1] Alessio Buonomo, University of Naples Federico II; email: alessio.buonomo@unina.it
Francesca Di Iorio, University of Naples Federico II; email: francesca.diiorio@unina.it
Salvatore Strozza, University of Naples Federico II; email: salvatore.strozza@unina.it

performances, and a greater probability of moving towards vocational schooling than Italian counterparts (Strozza 2008; Dalla Zuanna et al. 2009; Conti et al. 2013; Buonomo et al. 2018). The lack of fluency in the language of the majority population represents an important obstacle for children of immigrants to succeed in their schoolwork. These evidences imply that their insertion into the labour market will be more difficult compared to nationals (Meurs et al. 2008). Specifically, some studies proved that they have higher risk of unemployment and precarious employment (Kristen et al. 2008; Meurs, et al. 2008).

The aim of this contribution is to evaluate, through a Probit model, the relevance of different determinants on the probability to be employed among foreign and naturalized people aged 18-34, focusing our attention in assessing possible differences between first generation immigrants and children of immigrants born in Italy or immigrated in Italy before age 18 (hereafter 1.5 generation).

2 Data used, descriptive statistics and method of analysis

The survey on 'Social Condition and Integration of Foreign citizens' (SCIF) was conducted in 2011-2012 and released in 2016. It is a new and unique sample survey on this subject, designed by the Italian National Statistical Institute (ISTAT) in the system of multipurpose household surveys. The survey collects data on households with at least one foreign (or foreign origin) citizen and provides original information on foreigners living in Italy. In particular, many immigrants' characteristics and behaviours are considered: family, marriage, fertility, education, employment history, working conditions, religious affiliation, etc. This survey has a cross-sectional asset and it covers a sample of about 12,000 households. Our target sample refers to people aged 18-34 (6,342 unweighted observations).

We considered the weighted data in order to reproduce the main characteristics of the universe. Using a vector of weights resized to the total sample size, we observe that 58.5% of the target sample is employed with evident differences by gender (the percentage of employed is equal to 74.6% among men and only 44.6% among women) and migratory generation (62.6% among first generation immigrants against 48.6% among 1.5 generation). The difference in the employment rate by migratory generation could however depend on the different age structure of the two groups considered.

Many factors can be associated with employment probability. Considering the available information, we selected as possible explanatory factors 13 covariates grouped in two types of variables: individual characteristics (including intentions), and indicators of integration condition. The following individual covariates are included in the analysis (in brackets some weighted percentages): gender (males are 46.4%); age (continuous, we included the square term); area of residence (62.2% in the North, 23.9% in the Centre and 13.9 in the South of Italy); marital status (in three categories: single are 50.2%, married 43.7% and other conditions 6.1%); presence of children (56.3% without children); educational level (in three categories: 10.0% without education or with primary education, 52.3% with low or middle secondary education,

and 37.7% with complete secondary or tertiary education). We added a dummy variable that expresses the desire to have a child in the next 3 years (49.7% desire a child) in order to consider future intentions of the target sample. Migratory generation was summarized in another dummy variable that divides first generation of immigrants (70.9%) from children of immigrants born in Italy or immigrated before age 18 (1.5 generation). We used citizenship at birth, instead of the current one, in order to identify ethnic minorities even among the people who have acquired Italian citizenship. We considered three more numerous countries in our sample, while all the others were aggregated by geographical and economical macro-areas (table 1).

Table 1: Citizenship at birth by main countries and macro-areas. Absolute weighted values and percentages

Citizenship at birth	***N***	***%***
Albania (ALB)	755	11.91
Romania (ROM)	1,567	24.70
Morocco (MOR)	573	9.04
Rest of Eastern EU (EST_EU)	301	4.74
Rest of Eastern non EU Europe (EST_NEU)	747	11.77
Rest of Africa (AFR)	651	10.26
Asia (ASIA)	1,061	16.74
Latin America (LATAM)	530	8.36
More Developed Countries (MDCs)	157	2.48
Total	6,342	100.00

Following Blangiardo and co-authors (2013, 2018), we included three different dimensions of integration: cultural (knowledge and frequency of use of the Italian language, healthcare, eating habits, etc.), social (the active participation in social and public life), and political (the attention given to issues in Italian politics and their opinion regarding the importance of acquiring Italian citizenship). Each of the three indicators has a variation range between -1 (absence of integration) and +1 (maximum integration), with an average value of the total sample equal to 0.
Given the nature of the dependent variable, i.e. the probability to be employed, we choose to implement a Probit model (among others Hosmer and Lemeshow, 2000).

3 Results and discussion

Estimated coefficients are reported in table 2. Three different models are proposed. Model 1 include all the covariates described above excepted for variables on age and the dimensions of integration. It serves as the basis for the following models. The probability of be employed is higher for first generation immigrants compared to the 1.5 generation. As expected, females have negative effect on the estimated probability. Compared to single individuals, married have lower estimated probability to be employed while this is higher for separated or divorced individuals. Area of residence and education turn out not statistically significant at 5% level. The association between employment and the desire to have a child in the next three

years is positive. Considering the citizenship at birth (benchmark category is Albania), just Romania (positive), Morocco (negative) and others African nationalities (negative) significantly affect the response variable.

Table 2: Probit model coefficient estimates (in brackets benchmark level for categorical variable)

	Mod1		Mod2		Mod3		Mod3 Male		Mod3 Female	
Gender (bck: Male)										
female	-0.846	***	-0.782	***	-0.799	***				
First generation (bck: 1.5 gen.)	0.515	***	-0.158	**	-0.074		0.053		-0.111	
Residence area (bck: North)										
Center	0.093		0.093		0.084		0.116		0.071	
South	0.014		0.002		0.041		0.161	***	-0.065	
Education (bck: No or primary)										
Lower second. Educ.	0.049		0.114		0.068		-0.043		0.174	
High school	0.027		-0.011		-0.095		-0.267	*	0.067	
Marital status (bck: single)										
Married	-0.166	**	-0.380	***	-0.364	***	0.227		-0.643	***
Others status	0.470	***	0.130		0.138		0.004		0.293	*
Children	-0.064		-0.347	***	-0.336	***	-0.121		-0.497	***
Future child	0.251	***	0.199	***	0.196	***	0.124		0.172	**
Citizen at birth (bck: Albania)										
Romania	0.183	**	0.216	***	0.245	***	0.134		0.204	*
Morocco	-0.454	***	-0.503	***	-0.432	***	-0.432	**	-0.468	***
Rest of Eastern EU	0.137		-0.021		-0.011		-0.116		-0.026	
Rest of Eastern non-EU	-0.001		0.035		0.043		-0.084		0.001	
Rest of Africa	-0.362	***	-0.384	***	-0.313	***	-0.397	**	-0.228	
Asia	-0.009		-0.004		0.128		0.036		0.141	
Latin America	0.100		0.063		0.068		-0.242		0.232	
More Developed Countries	-0.065		-0.282		-0.247		-0.857	***	-0.109	
Age			0.804	***	0.806	***	0.845	***	0.826	***
Age squared			-0.013	***	-0.013	***	-0.014	***	-0.013	***
Cultural integration					0.408	***	-0.178		0.600	***
Social integration					0.157		0.388		0.093	
Political integration					0.109		0.220		-0.030	
Constant	0.247	**	-11.177	***	-11.175	***	-11.503	***	-11.992	***
Pseudo R^2	0.129		0.209		0.212		0.203		0.180	

Note: *** significant at 1%; ** significant at 5%; * significant at 10%

In Model 2, with the introduction of variables on age and its squared term, the coefficient associated to first generation assume a negative significant impact as well as the presence of children. This is the result of the age distribution, since the first generation is made up of older people and constitutes the greater part of the target sample. Moreover, age have expected positive (and age squared negative) effect on the estimated probability. The first generation coefficient become non-statistical significant in Model 3 with the introduction of variables on integration. Among them, only cultural integration has a significant (positive) effect. The remaining coefficients do not undergo significant changes compared to Model 2. This final model shows how the probability of be employed is linked mainly to gender, to the family context, the area of origin and the level of cultural integration,

and obviously to age. Education level and area of residence, as well as migratory generation, do not seem to play a significant role.
We repeated the analysis presented in Model 3 on males and females subsamples. The significant gender differences in the determinants of the probability of be employed are highlighted. The Probit model shows significant effect for variables associated to family context only for females: marital status (negative for married women), presence of children (negative for women having children) and desire to have a child (positive for women that desire a child). Citizenship at birth shows significant coefficients for the same categories in the three Models. At the same time the analysis carried on by gender using Model 3 highlights that several categories of citizenship at birth affects the response variable in different way in their respective sub-samples with respect to the reference group (Albanians). In particular, in the male subsample only Moroccans, other Africans and MCDs have significant negative effect. These evidences have a different meaning being probably due to high unemployment in the case of Africans and low activity rate in the case of citizens of MDCs. For the female subsample there is an advantage in terms of probability of be employed by the Romanians and a negative effect by Moroccan compared to the Albanians.

Figure 1: Estimated probability to be employed and 95% Confidence Interval by gender, migratory generation and citizenship at birth when all categorical variables are set to benchmark level and continuous variables at the mean.

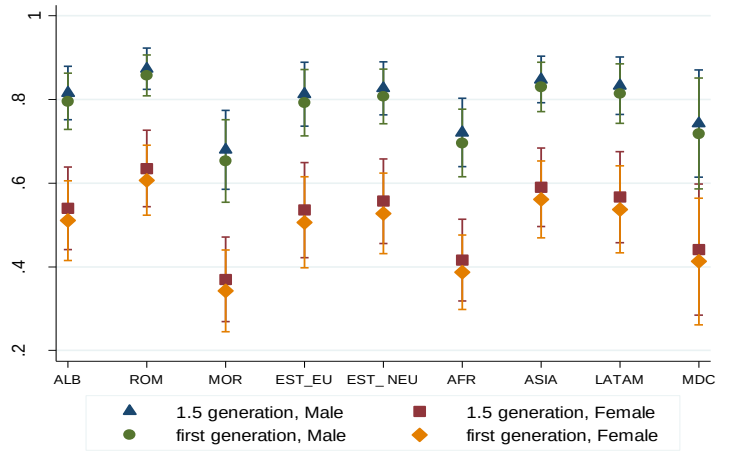


Figure 1 depict the estimated probability to be employed by gender, migratory generation and citizenship at birth, when the other dummies or polychotomous variables are at the reference level and the remaining variables at the mean. Gender differences are confirmed separately by citizenship at birth: for all citizenship at birth, except for MDCs, women always have values of the estimated probability of be employed clearly lower than those of the male counterpart. All other things being equal, it seems that there aren't significant differences in the possibility of access to employment for young adults of foreign origin distinguished between those of the first generation and those who are children of immigrants.

Described results gives a first insight on the probability to be employed of young-adults foreigners. It will be necessary to verify, for example, the quality of the occupational status level and the importance of over-education comparing the results among both migratory generations and citizenship at birth. This analysis is mainly devoted in evaluating the contribution of the SCIF survey, further investigation will require also the evaluation of a different source of data (for instance, Labor Force Survey). In any case, our contribution provides some policy implications. In particular, the results reveal the central role played by investing in (cultural) integration in order to favour the placement in the labour market of the host country of foreign (female) young-adults residents in Italy.

Acknowledgement: Grant from Ministry of Education, University and Research (MIUR), PRIN project 2017 titled "Immigration, integration, settlement. Italian-Style" (Grant No. 2017N9LCSC_004) is grateful acknowledged.

References

1. Ambrosini, M., Molina, S. (Eds.): Seconde Generazioni. Un'introduzione al futuro dell'immigrazione in Italia. Turin: Fondazione Giovanni Agnelli (2004)
2. Blangiardo, G. C., Perez, M., Quattrociocchi, L., Zizza, R.: Employment and economic Conditions. In: Ministero dell'Interno, Istat (Eds.) Integrazione. Conoscere, Misurare, Valutare pp. 29-47, International Conference, Rome (2013)
3. Blangiardo G., Mirabelli, S., M.: Misurare l'integrazione. In Perez M. (Ed.) Vita e percorsi di integrazione degli immigrati in Italia, pp. 361-382, Istituto nazionale di statistica, Roma (2018)
4. Buonomo, A., Strozza, S., Gabrielli, G.: Immigrant youths: Between early leaving and continue their studies. In: Merrill, B., Padilla Carmona, M. T., González Monteagudo, J. (Eds.) Higher Education, Employability and Transitions to the Labour Market, pp. 131-148. Sevilla, España: EMPLOY Project: Universidad de Sevilla (2018)
5. Conti, C., Di Bartolomeo, A., Rottino, F.M., Strozza, S.: Second Generation and Educational Attainment. In: Ministero dell'Interno e ISTAT (Eds.), Integration. Knowing, Measuring, Evaluating. Rome: 49-69 (2013)
6. Dalla Zuanna, G., Farina, P., Strozza, S.: Nuovi italiani. I giovani immigrati cambieranno il nostro paese?. Bologna: il Mulino (2009)
7. Hosmer, D. & Lemeshow, S Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc. (2000)
8. Kristen, C., Reimer, D., Kogan., I.: Higher Education Entry of Turkish Immigrant Youth in Germany. International Journal of Comparative Sociology 49 (2–3): 127–151 (2008)
9. Meurs, D., Pailhé, A., Simon, P.: Discrimination despite integration: Immigrants and the second generation in education and the labour market in France. In: Bonifazi, C., Okólski, M., Schoorl, J., Simon, P. (Eds.) International Migration in Europe. New Trends and New Methods of Analysis, pp. 247-270. Amsterdam: IMISCOE Research Amsterdam University Press (2008)
10. Strozza, S.: Partecipazione e ritardo scolastico dei ragazzi stranieri e d'origine straniera. Studi Emigrazione 171: 699-722 (2008)
11. Strozza, S.: Foreign immigration in Italy: a forty-year-old history. Proceedings of the 48th Scientific Meeting of the Italian Statistical Society. Università di Salerno 8-10th June (2016)
12. Strozza, S., De Santis, G., (Eds.): Rapporto sulla popolazione. Le molte facce della presenza straniera in Italia. Il Mulino, Bologna (2017)

The risk of inappropriateness in geriatric wards: a comparison among the Italian regions.

Rischio di inappropriatezza in ambito geriatrico: un confronto tra le regioni italiane.

Paolo Mariani, Andrea Marletta, Marcella Mazzoleni and Mariangela Zenga

Abstract Italy could be considered one among the oldest countries in Europe and this situation is becoming increasingly relevant in the future decades. This scenario will lead the Italian Health System to cope with a significant increase in healthcare consumption. This work will analyse the ordinary acute admissions in the geriatric wards of the Italian hospitals using the Hospital Discharge Data. Specifically, the aim is to describe the risk of inappropriateness of the hospitalizations related to chronic diseases among the Italian regions. A risk of inappropriateness index will be proposed to compare the Italian regions into the period 2012-2015.

Abstract *L'Italia può essere considerate una tra le nazioni più anziane in Europa, e questa situazione non può che aumentare nelle prossime decadi. Questo scenario porta il Sistema Sanitario Nazionale italiano verso un maggiore consumo di risorse per la salute. Questo lavoro si propone di analizzare i ricoveri ordinari in acuto nei reparti di geriatria degli ospedali italiani attraverso i dati delle dimissioni ospedaliere. In particolare, lo scopo è quello di descrivere il rischio di inappropriatezza per i ricoveri per malattie croniche tra le regioni italiane. Verrà proposto un indice di rischio di inappropriatezza per comparare le regioni italiane nel periodo 2012-2015.*

Key words: Italian Geriatric Wards, Chronic diseases, Inappropriateness risk.

[1]Paolo Mariani, DEMS, UNIMIB; email: paolo.mariani@unimib.it; Andrea Marletta, DEMS, UNIMIB; email: andrea.marletta@unimib.it; Marcella Mazzoleni, DISMEQ UNIMIB; email: marcella.mazzoleni@unimib.it; Mariangela Zenga, DISMEQ UNIMIB; email: mariangela.zenga@unimib.it.

Introduction

In the last decades, the proportion of elderly population has increased across all European countries, showing the highest growth rate amongst groups of all ages. Italy could be considered one among the oldest countries in Europe: in fact, the population aged 65 and over is 22.6% of the Italian population with an aging index of 168.7% as stated by ISTAT in 2018. Moreover, a high percentage (49.6%) of elderly people shows at least one of chronic/chronic degenerative disease. Moreover, in 2012 the Italian regions have been undergone to spending review in relations to the high cost of healthcare review (Ministero della salute, 2012). Given by the Italian spending review in healthcare, the acute hospitalizations in geriatric wards decreased by 3.9% from 2012 to 2015.
This work proposes a synthetic index to measure the risk of inappropriate hospitalization that is applied to the Italian regions, so they are classified into three groups based on the distance from the national average behaviour.

Definition of inappropriate hospitalization

In this work, we pay attention on the measure of inappropriateness in the hospitalizations for elderly patients admitted in the Italian geriatric wards. From a healthcare point of view, several authors (Tsang and Severs, 1995; Inglis et al., 1995; Houghton et al., 1996; Beringer and Flanagan, 1999; Leah and Adams, 2010; Mayo and Allen, 2010) proposed to use a definition based on patients' clinical details by a team of doctors. Glasby and Littlechild (2001) proposed to use also the patients' opinion.
In this case, we adopt the definition of inappropriateness hospitalization as stated by the Italian Healthcare Ministry. In 2001, the Essential Assistance Levels (LEAs) were introduced with the aim to rationalise the use of resources in the healthcare sector and they were related to the Diagnosis-Related Group (DRG). The concept of appropriateness was defined with the aim to improve the quality of provided health services and the proper use of resources. An indication of potential inappropriateness of the healthcare was represented by a high variability in the demand and supply of healthcare services at a regional level. In this way, 108 DRGs with a high risk of inappropriateness for hospitalization of acute diseases were defined to monitor and verify effective delivery of performance in the national territory. The 2010-2012 Pact for Health (Patto per la Salute, 2009) has been drawn up. According to this reform, the aspect of appropriateness plays a major role in hospital care. This reform defines a set of indicators to be used to control the achievement of an appropriate service delivery and performance in healthcare system. The indicators system affords to the regions to define standards of organisational appropriateness for the purpose of self-assessment in order to optimise the delivery of healthcare. Some standards have been defined at regional level indicating the threshold values within which to permit

inpatient admissions. A comparison of the regional values of the indicators can identify regional critical areas and it could help the central level to intervene.

Data and Methodological Approach

The data used in this paper consists of the ordinary admissions of the patients aged 65 years or older to every geriatric ward in every acute care hospital operating in the 20 Italian Regions in 2012 and 2015.
The Italian Healthcare Ministry provided the data. Individual Hospital Discharge Charts (HDC) are reported in the dataset including patient information (gender, age, residence, etc.), the treatments received during hospitalization including information such as Disease Related Group (DRG), principal and secondary diagnoses and procedures, date of admission, and so on.
Patients are aged 84 years on average. Approximately 42% of patients were male. The ten percent of patients were admitted to the department of geriatric medicine through emergency admission. Approximately 57% of patients were admitted for surgery. Chronic patients represented approximately 33% of all patients. Moreover, 24% of the admitted patients had a principal diagnosis of circulatory system, 22% as respiratory system problems, and 16% problems of the nervous system. Approximately 73% of patients left the geriatric ward to return home, of whom about 3% were voluntary discharge; 13% died while in hospital and the remaining 14% transferred. The length of Stay is about 11 days. The Elixhauser comorbidity index diseases (Elixhauser et al., 1998) for the geriatric hospitalizations is about 1.3 comorbidity.
The Ministry of Health measures the rate of DRGs at Inappropriateness risk as the number of Inappropriate hospitalizations per Not Inappropriate hospitalizations (Ministero della Salute, 2010):

$$RI_{t,i} = \left(\frac{Number\ of\ Inappopriate\ Hospitalization}{Number\ of\ Not\ Inappopriate\ Hospitalization}\right)_{t,i}$$

for the *i-th* Italian region in *t-th* year. To compare the *i-th* region in t_1 and t_2 years, a number index is used:

$$DRI_{t_2,i/t_1,i} = \frac{RI_{t_2,i}}{RI_{t_1,i}}.$$

Moreover, considering Italy as a benchmark, the regions could be divided in three groups according to $I_i = (DRI_{t_2,i/t_1,i})/(DRI_{t_2,ITALY/t_1,ITALY}) \cdot 100$ if:

- $I_i \leq 100$ the *i-th* region shows a better performance in reduction of inappropriateness respect to Italy;
- $100 < I_i \leq 100 + std(I)$ the *i-th* region presents a limited worst performance in reduction of inappropriateness respect to Italy;

- $I_I > 100 + std(I)$ the *i-th* region presents a worst performance in reduction of inappropriateness respect to Italy.

Some evidences

Table 1 reports the rate of DRGs at Inappropriateness risk by the Italian regions in 2012 and 2015. In 2012 for Italy, for 100 Not Inappropriate Hospitalizations exist 11 Inappropriate Hospitalizations. Sardegna has the highest rate of DRGs at Inappropriateness risk (0.21), followed by Puglia (0.16) and Lombardia (0.15), while Valle d'Aosta is the region with the lowest rate of DRGs at Inappropriateness risk (0.04), followed by Piemonte (0.0593) and Marche (0.0599). In 2015, the Italian rate of DRGs at Inappropriateness risk decreased by 20%. Almost each region shows decreased rate of DRGs at Inappropriateness risk, except for Piemonte, Friuli-Venezia Giulia, Umbria and Molise. Again, Sardegna shows the higher rate of DRGs at Inappropriateness risk (0.18), followed by Lombardia (0.12) and Puglia (0.11). Calabria has the highest percentage of decreasing, going from 0.10 to 0.03 (-71.13%), while Friuli-Venezia Giulia has the highest percentage of increasing, going from 0.065 to 0.089 (+36.92%).
In the comparison in I_i, ($std(I)$*=29.57)* six regions (Group 1: Toscana, Marche, Campania, Puglia, Calabria and Sicilia) show better performance in reduction of inappropriateness respect to the Italian situation, even if different situations are present. For example, Marche shows lower rate of DRGs at Inappropriateness risk in 2012 and 2015 and the reduction in it is greater respect to Italy. On the contrary, Puglia and Campania have bad performances in terms of the rate of DRGs at Inappropriateness risk in 2012 and 2015, far below the Italian situation, but the reduction in inappropriateness results to be greater respect to Italy. The regions in the limited worst group (Group 2) are Valle d'Aosta, Lombardia, Trentino Alto Adige, Veneto, Liguria, Emilia Romagna, Lazio, Abruzzo, Basilicata and Sardegna. In the worst perfomance group (Group 3) there are Piemonte, Friuli-Venezia Giulia, Umbria and Molise.
This result is widely reflected on the Italian regional spending review (Ministero della salute, 2012).

Conclusions and future works

The study of the inappropriateness dynamics of the geriatric hospitalization recommends a more careful examination by temporal and spatial aspects. In this analysis, some regions showing better performance have been undergone heavy actions by the National Health Care System reducing costs. On the other hand, other regions showing undesirable values were subject to attention by the Italian government in these last years.

Future research is based on the investigation of the possible causes of inappropriateness that could be nested both in the comorbidity and in the different styles of health management that are currently in the hands of the individual regions.

Table 1: Rate of DRGs at Inappropriateness risk by Italian region in 2012 and 2015.

Region	RI_{2012}	RI_{2015}	$DRI_{2015/2012}$	I	Group
Piemonte	0.0593	0.0683	115.18	143.83	3
Valle Aosta	0.0424	0.0370	87.26	108.97	2
Lombardia	0.1460	0.1191	81.58	101.87	2
Trentino-Alto Adige	0.0976	0.0916	93.85	117.19	2
Veneto	0.0989	0.0834	84.33	105.30	2
Friuli-Venezia Giulia	0.0650	0.0890	136.92	170.98	3
Liguria	0.0972	0.0965	99.28	123.98	2
Emilia-Romagna	0.1117	0.0943	84.42	105.42	2
Toscana	0.0873	0.0565	64.72	80.81	1
Umbria	0.0926	0.0968	104.54	130.54	3
Marche	0.0599	0.0335	55.93	69.84	1
Lazio	0.0897	0.0732	81.61	101.90	2
Abruzzo	0.0687	0.0582	84.72	105.79	2
Molise	0.0992	0.1093	110.18	137.59	3
Campania	0.1311	0.0935	71.32	89.06	1
Puglia	0.1596	0.1100	68.92	86.07	1
Basilicata	0.1011	0.0877	86.75	108.32	2
Calabria	0.0963	0.0278	28.87	36.05	1
Sicilia	0.0694	0.0384	55.33	69.09	1
Sardegna	0.2107	0.1757	83.39	104.13	2
Italy	0.1059	0.0848	80.08	100	

Source: Elaboration on Individual Hospital Discharge Charts data.

References

1. Beringer, T., Flanagan, P.: Acute medical bed usage by nursing home residents. *Ulster Med J*. **68**(1), 27--29 (1999).
2. Elixhauser A, Steiner C, Harris DR, Coffey RM.: Comorbidity measures for use with administrative data. *Med Care*. **6**, 8--27 (1998)
3. Glasby, J., Littlechild, R., Pryce, K.: Show me the way to go home: a narrative review of the literature on delayed hospital discharges and older people. Brit J Soc Work. **34**, 1189--1197 (2004)
4. Houghton, A., Bowling, A., Jones, I., Clarke, K.: Appropriateness of admission and the last 24 hours of hospital care in medical wards in an east London teaching group hospital. *Int J Qual Health Care*. **8**, 543--553 (1996)
5. Inglis, A.L., Coast, J., Gray, S.F., Peters, T.J., Frankel, S.J.: Appropriateness of hospital utilization: the validity and reliability of the intensityseverity-discharge review system in a united kingdom acute hospital setting. *Med Care*. **33**(9), 952--957 (1995)
6. Leah, V., Adams, J.: Assessment of older adults in the emergency department. *Nurs Stand*. **24**(46): 42--45 (2010)

7. Mayo A, Allen A. Reducing admissions with social enterprises. *Emerg Nurse*. **18**(4):14--17 (2010)
8. Ministero della salute: Indicatori di Appropriatezza organizzativa. (2010) Available via http://www.salute.gov.it/imgs/C_17_pubblicazioni_1421_allegato.pdf cited on 10/02/2020.
9. Ministero della Salute Italiano: Decreto Legge n. 95, Disposizioni urgenti per la revisione della spesa pubblica con invarianza dei servizi ai cittadini. Supplemento Ordinario alla Gazzetta Ufficiale n 156 del 6 Luglio 2012- Serie Generale (2012)
10. Patto per la Salute: Provvedimento 03 dicembre 2009. Gazzetta Ufficiale, Serie Generale, n. 3 del 05 gennaio 2010 (2010)
11. Tsang, P., Severs, M.: A study of appropriateness of acute geriatric admissions and an assessment of the appropriateness evaluation protocol. *J R Coll Physicians Lond*. **29**, 311--314 (1995)

The role of the accumulation of poverty and unemployment for health disadvantages

Il ruolo delle traiettorie di povertà e occupazione negli svantaggi di salute

Annalisa Busetta, Daria Mendola, Emanuela Struffolino and Zachary Van Winkle

Abstract Health inequality is an important aspect of how advantages and disadvantages are distributed within societies. We extend previous research by considering how trajectories of poverty and employment affect self-reported health among young adults. We use data from the German Socio-Economic Panel, restricting the analytical sample to those who were 25 to 45 year-old in 2005. We calculated the indexes that account for persistence and intensity of poverty and employment on 10-year-long individual-level employment and household-level poverty trajectories. Ordinal logit regression models show that both long-lasting poverty and short periods out of employment are detrimental for men's health. In contrast, only more recent episodes in poverty have a negative effect on women's health.

Abstract *Le disuguaglianze di salute sono cruciali per la distribuzioni di vantaggi e svantaggi. Questo lavoro esamina come le traiettorie di povertà e occupazione influenzino la salute percepita tra i giovani adulti. Utilizzando il German Socio-Economic Panel e considerando un campione di 25-45enni al 2005, abbiamo misurato su 10 anni la persistenza e l'intensità della povertà e dell'occupazione. I modelli di regressione logistica ordinali mostrano che sia la povertà di lunga durata che i brevi periodi di disoccupazione sono dannosi per la salute degli uomini; mentre per le donne solo degli episodi di povertà più recenti mostrano un effetto negativo.*

Key words: in-work poverty, gender, health, life-course, index, persistence

[1] Annalisa Busetta, Università degli Studi di Palermo; annalisa.busetta@unipa.it

Daria Mendola, Università degli Studi di Palermo; daria.mendola@unipa.it

Emanuela Struffolino, Freie Universität Berlin and WZB Berlin; e.struffolino@fu-berlin.de

Zachary Van Winkle, University of Oxford & Nuffield College, zachary.vanwinkle@sociology.ox.ac.uk

Introduction

Health inequality is an important aspect of how advantages and disadvantages are distributed within societies. Individual health can be affected both positively and negatively by life course events in different domains, such as work and family (e.g., Marmot, 2015). Most mechanisms linking individual life courses to health outcomes depend on exposure to specific risks, require time to show up, and can have cumulative effects (Dannefer, 2003). Empirical research that adopts a longitudinal perspective shows that long-lasting exposure to poverty over the life course and changes in income and economic status are associated with poorer health (e.g., Gunasekara et al., 2011). Similarly, employment trajectories characterized by unemployment as well as precarious and poorly paid jobs are associated with poor health in mid-life (e.g., Cullati, 2014; Devillanova et al., 2019). Not only do differences in material standards of living, access to healthcare, and disposable financial resources account for the association between precarious life courses and health, but also the psychological and behavioural implications of unemployment and poverty.

Recent reviews on socioeconomic inequalities in health have identified limitations in the existing literature that can be addressed by adopting longitudinal analytical approaches and the life course perspective (Corna, 2013). Specifically, there has been so far little recognition of how the association between socio-economic position and health are—at least partly—constituted by labour market and family experiences across the life course. Additionally, a deep understanding of these dynamics requires to consider how gendered they are. We extend upon previous research and assess how longitudinal trajectories of poverty and labour market attachment independently and mutually affect self-perceived health among young adults. We devote particular attention to gender differences.

We focus on Germany, where the increase in in-work poverty between 1985 and 2017 was one of the largest in the EU (Eurostat, 2018) and for which high quality longitudinal data are available (SOEP-German Socio-economic Panel). At this stage, we will consider a cohort of young individuals whose life courses unfolded between 2005 and 2014. To assess how trajectories of poverty and employment impact health, we reconstruct separate sequences of household poverty and individual labour market attachment. We then adopt two recently developed indexes that capture cumulative effects of repeated poverty spells in terms of their severity and recentness (Busetta et al., 2019; Mendola et al., 2011). We therefore go beyond "point in time" approaches that consider how the transition from a permanent to a temporary job affects health and address how trajectories of poverty and labour market attachment affect health in a holistic manner.

Data and methods

Data. We used data from the German SOEP, a nationally representative household panel. We restrict the sample to those born between 1969 and 1980 who therefore were 25 to 34 year-old in 2005 (women=717, men=584). For each individual, we constructed an "employment trajectory" and a "poverty trajectory".

Dependent variable. Health is measured in 2015, that is at the end of the 10-year observational window, using a question that asks respondents to self-rate their own general health (herein SRH). We recoded SRH into four categories in each panel wave: very good (13.86% of the sample), good (49.97%), fair (25.89), and bad/very bad (bad, henceforth, 10.28%). SRH has been shown to be a valuable predictor of more detailed measures of health status (Jylhä, 2009).

Main dependent variables. Employment and poverty patterns are our main predictors for SRH. For employment trajectories, each time-point between the first observation in 2005 and the last in 2014 was coded with 0 if the individual was inactive/unemployed and 1 if he/she was employed. For poverty trajectories, each time-point where coded with 1 if the respondent's household income was below the poverty line (i.e. 60% of the median of equivalised household income distribution) for that year and with 0 otherwise. We then calculated the index of longitudinal poverty proposed by Mendola et al. (2011) on the poverty trajectories. The index takes into account the pattern and the severity of poverty experience, as it (i) gives the highest weight to each pair of consecutive years spent in poverty (and an increasingly higher weight to closer spells of poverty), i.e., they contribute more to increase the value of the overall measure) and (ii) puts more emphasis to poverty episodes experienced during years of low poverty incidence in the general population; (iii) puts more weight the more severe poverty spells are (i.e., when household income is far below the poverty line). To summarize employment trajectories to proxy individuals' labour market attachment, we propose an adaptation of the index proposed in Busetta et al. (2019) that considers individual and contextual labour market conditions simultaneously. By adopting this modified index, we account for the continuity of employment (vs. unemployment) and the "intensity" of employment, by weighting each employment episode by the numbers of hours worked (percentage of a full-time week, 40h/week). In addition, the employment index incorporates the probability of remaining in employment and the detrimental effect of years spent out of employment. Next to poverty and employment indexes, we calculated two ancillary indexes (so-called "closeness" effects) that account for how recent poverty and employment spells are to the measurement of self-rated health. Note that the income variable was lagged to preserve the correspondence between the employment status and income reference period.

Modelling strategy. We run a set of ordinal logit regression models for SRH, for men and women separately. Following a forward-stepwise criterion, they include: (1) the poverty index, (2) the poverty index and its closeness effect, (3) the labour market attachment index, (4) the labour market attachment index and its closeness

effect, (5) both the poverty and labour market attachment indexes and their closeness effect and the interaction between the indexes.

Control variables. All models are adjusted for a set of observable individual and household characteristics: age, age-squared, number of years in education (logarithm), being born in East Germany, and number of children in the household (age 18 and below) are measured at the onset of individuals' sequences. Another set of variables accounts for the change in the number of children in the household and for the share of years spent in singlehood and divorce across the 10-year observational window.

Main results

Table 1 shows the results (as betas) of a set of ordered logistic regressions for the probability of bad health for women and men separately.[1] Models 1a and 1b estimate the effect of the poverty persistence index: longer and closer periods spent in poverty increase the probability of reporting bad health for men. For them, the poverty index is no longer statistically significant after adjusting for the closeness effect of poverty (models 2a and 2b), which—by emphasising the more recent poverty episodes—captures all the effect of poverty patterns. The closeness effect of poverty is also statistically significant for women (model 1b). With regard to employment trajectories, stronger attachment decreases the likelihood of bad health for men but not for women (models 3b and 3a respectively). Similar to poverty, the effect of the overall trajectory of employment disappears once the effect of recent employment spells (model 4b) are taken into account. Finally, the interaction between the poverty and employment trajectories is not statistically significant for both for men and women, while closeness effect of poverty and of employment are significant for both groups (models 5a and 5b). Explanations for gender differentials in our relationships of interest are not straightforward. Although in Germany several steps towards gender equality have been made, the division of labour remain highly gendered, especially after childbirth. The so called "1.5-bread-winner model" prevails so that weak labour market attachment might affect men's health but not women's as for men the social norms associated with their primary role of breadwinner are stronger. However, when it comes to household poverty, the negative effect of the closeness of poverty affects women as well, as they are highly involved in the process of coping with poverty irrespective of their own position in the labour market and especially in presence of young children.

[1] As robustness checks, we replicate the models by (i) dichotomizing the outcome variable ("bad"/"very bad" vs. "very good"/"good"/"fair"), (ii) keeping the categories "bad" and "very bad" separated, (iii) estimating generalized ordinal logit models for the outcome with five categories; (iv) including interaction between closeness effects of poverty and employment. The results are highly stable and consistent with those presented in Table 1.

Conclusions and next steps

In this study, we investigate possible spill-over effects of poverty and employment trajectories on health. We adopt a longitudinal view that accounts for the persistence and intensity of poverty as well as of labour market attachment. Our preliminary results show considerable gender differences. For men having spent several years in poverty and out of the labour market has a negative effect on health, especially when these events occurred in the last few years before the self-reported assessment. However, we did not find an interaction effect between poverty and employment trajectories. For women, we only find a significant effect of poverty episodes experienced more recently in time. Possible explanations for the gendered differential effects of poverty and employment trajectories can be found in how the cultural and institutional contexts shape the role of men and women in the family and the labour market. Further, gender roles might influence men's and women's reporting behaviour also depending on how the employment status of partners affects respondents' perceived health (e.g., Oksuzyan et al., 2019). In the next steps, we will deepen our understanding of these dynamics by expanding the current analytical design to incorporate observations from older birth cohorts and consider how welfare state and labour market changes (such as the Hartz reforms started in 1998) strengthened or weakened the association between poverty and employment trajectories and health across social groups in East and West Germany.

References

Busetta, A., Mendola, D., & Vignoli, D. (2019). Persistent joblessness and fertility intentions. *Demographic Research*, *40*, 185–218.

Corna, L. M. (2013). A life course perspective on socioeconomic inequalities in health: A critical review of conceptual frameworks. *Advances in Life Course Research*, *18*(2), 150–159.

Cullati, S. (2014). The influence of work-family conflict trajectories on self-rated health trajectories in Switzerland: A life course approach. *Social Science & Medicine*, *113*, 23–33.

Dannefer, D. (2003). Cumulative advantage/disadvantage and the life course: Cross-fertilizing age and social science theory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *58*(6), 327–337.

Devillanova, C., Raitano, M., & Struffolino, E. (2019). Longitudinal employment trajectories and health in middle life: Insights from linked administrative and survey data. *Demographic Research*, *40*(47), 1375–1412.

Eurostat. (2018). *People at risk of poverty or social exclusion*.

Gunasekara, F. I., Carter, K., & Blakely, T. (2011). Change in income and change in self-rated health: Systematic review of studies using repeated measures to control for confounding bias. *Social Science & Medicine*, *72*(2), 193–201.

Jylhä, M. (2009). What is self-rated health and why does it predict mortality? Towards a unified conceptual model. *Social Science & Medicine*, *69*(3), 307–316.

Marmot, M. (2015). *The health gap: The challenge of an unequal world*. Bloomsbury Publishing.

Mendola, D., Busetta, A., & Milito, A. M. (2011). Combining the intensity and sequencing of the poverty experience: A class of longitudinal poverty indices. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(4), 953–973.

Oksuzyan, A., Dańko, M. J., Caputo, J., Jasilionis, D., & Shkolnikov, V. M. (2019). Is the story about sensitive women and stoical men true? Gender differences in health after adjustment for reporting behavior—ScienceDirect. *Social Science & Medicine*, *228*, 41–50.

Table 1 Ordinal Logit Regressions for SRH- beta estimates for likelihood of bad health by gender

	Women					Men				
	(1a)	(2a)	(3a)	(4a)	(5a)	(1b)	(2b)	(3b)	(4b)	(5b)
Poverty trajectory	0.208	-2.179			-2.280	1.409*	-1.389			-2.023
Closeness effect of poverty		1.433*			1.620*		1.639*			0.412
Employment trajectory			0.335	-0.0186	0.212			-2.153***	-0.858	-0.857
Closeness effect of employment				0.346	0.375				-1.385*	-1.734*
Poverty trajectory # Employment trajectory					2.008					3.621
Age	1.653***	1.666***	1.650***	1.638***	1.624***	-0.249	-0.169	0.110	0.056	0.082
Age-squared	-0.021***	-0.021***	-0.0201***	-0.020***	-0.020***	0.004	0.003	-0.001	-0.000	-0.000
ln no. of years of education at first wave	-1.305***	-1.188***	-1.395***	-1.389***	-1.241***	-1.059**	-0.952**	-1.047**	-0.976**	-0.941**
Born in East Germany (ref. West)	-0.365**	-0.391**	-0.379**	-0.375**	-0.436**	-0.0840	-0.0920	-0.0594	-0.071	-0.053
no. of years as divorcee	-0.047	-0.118	-0.036	-0.012	-0.147	-0.178	-0.242	-0.194	-0.261	-0.241
no. of years as single	-0.178	-0.211	-0.184	-0.172	-0.256	-0.296	-0.311	-0.381	-0.392	-0.397
no. of underage children in the hh. at w1	-0.240**	-0.250**	-0.205**	-0.213**	-0.218**	-0.230*	-0.238*	-0.185	-0.184	-0.199
no. of children in the hh. increased (ref. no change)	0.235	0.238	0.228	0.228	0.190	0.325	0.319	0.288	0.280	0.283
no. of children in the hh. decreased	-0.019	-0.054	0.034	0.028	0.016	-0.065	-0.114	-0.002	0.002	-0.025
N (individuals)	717	717	717	717	717	584	584	584	584	584
AIC	1665.06	1663.85	1664.56	1665.85	1667.03	1292.90	1291.58	1280.90	1280.09	1284.54
Cut point 1	26.76**	27.33**	26.59**	26.57**	26.77**	-9.405	-7.480	-3.091	-4.889	-4.563
Cut point 2	29.55***	30.12***	29.38***	29.36***	29.57***	-6.395	-4.461	-0.0544	-1.855	-1.525
Cut point 3	31.22***	31.79***	31.05***	31.03***	31.25***	-4.435	-2.491	1.937	0.153	0.487

SOEP data V.32. *** p<0.01, ** p<0.05, * p<0.10

Unemployment and fertility in Italy: A panel analysis.

Disoccupazione e fecondità in Italia: un'analisi panel.

Gabriele Ruiu and Marco Breschi

Abstract The effect of increasing unemployment on the decision to have children is not clear from both theoretical and empirical points of view. We focus our empirical analysis of the relationship between fertility and unemployment on a peculiar context represented by Italy. The demographic behaviour in this Country challenges the most influential theoretical predictions developed in social sciences. Implementing a panel ARDL analysis, our results indicate that a negative long run relationship exists between the two variables, but some surprising results are obtained when the short-run is considered.

Abstract *L'effetto della disoccupazione sulla fecondità non è chiaro sia dal punto di vista teorico sia da quello empirico. L'analisi della relazione tra queste due variabili svolta nel presente contributo è focalizzata su un contesto molto particolare come l'Italia. Il comportamento fecondo nel Paese rappresenta infatti la sfida per eccellenza nelle scienze sociali. I risultati di un'analisi panel ARDL indicano che tra le due variabili esiste una relazione negativa di lungo periodo, mentre nel breve periodo emerge una situazione più eterogenea tra le regioni.*

Keywords: Unemployment; Fertility; ARDL technique; panel analysis.

1 Introduction

The relationship between unemployment and fertility is widely debated in both economics and demography. In particular, the effect of increasing unemployment on the decision to have children is not completely clear, from both theoretical and

[1] Gabriele Ruiu, Department of Economics and Business, University of Sassari; gruiu@uniss.it

Marco Breschi, Department of Economics and Business, University of Sassari; breschi@uniss.it

empirical points of view. For men, unemployment implies not being able to access previously accessible goods due to the reduction of income. Therefore, if we are willing to accept the idea that having a child may be modelled as an economic decision, the prediction is inevitably that a sharp reduction in the demand for "children" will be observed. For women, in addition to this income effect, another force pushes in the opposite direction. Indeed, unemployment implies that the opportunity cost of having a child (i.e., the wage that one has to give up net of the possible maternity allowance) suffers a drastic decline, thus favouring fertility choices (substitution effect). Therefore, only when the latter force compensates the income effect, there will be an overall increase of fertility in times of crisis, while the opposite will happen when the substitution effect is overwhelmed by the income effect. However, Regions with low female participation rates experience a higher incidence of households in a situation of zero earnings when men lose their jobs. Therefore, the positive substitution effects should be negligible in these regions with upsetting effects on fertility. Pro-cyclical relationships have been confirmed by several empirical studies that use macro-level data [1, 8]. However, there are also empirical works, especially those analysing data before the late 1980s, which have found a counter-cyclical tendency in fertility [6, 5].

Considering the Italian case, [4] have shown that the direction of the relationship between unemployment and general fertility rates in Northern Italy is different from that of Southern regions. Specifically, they found that in Northern and Central Italy, the effect of unemployment is negative, while the opposite is true for Southern Italy (although in the latter case the results were weakly statistically significant). They conclude that this difference may be due to regional heterogeneity in the pervasiveness of the underground economy. In particular, it is well known that Southern Italy is characterised by the highest level of underground economic activity. This means that the level of unemployment, as measured by official statistics, may not reflect the real economic situation of this part of the Country. It should be noted that, the empirical results of [4] are obtained by pooling together all the regions belonging to the same macro-area and estimating a dynamic regression model for each area. In other words, this approach does not consider regional heterogeneity. We believe that Italy represents a peculiar context for analysis for several reasons. First of all, the country is (together with Spain, Portugal, Greece, Cyprus, and Malta) characterised by the lowest total fertility rate (TFR) in the World. In addition, given that female participation in the labour market in Italy is among the lowest in Europe, the positive relationship between unemployment and fertility challenges the predictions made by neoclassical microeconomic models. Furthermore, Italy has often been depicted as the least secularised among the developed countries. Hence, the Second Demographic Transition theory cannot help explain the low fertility rate. In the following section we will present data and methodology. In the third section, we will show and comment the obtained results-

2 Data and methods

We calculated the time series of quarterly GFR for each Italian region in the period 1992 (fourth quarter)-2017 (fourth quarter) using official data produced by ISTAT. The log of GFR is used as a dependent variable in an ARDL panel estimation where both the log of the male unemployment rate (hereafter *unem_male*) and the log female unemployment rate (hereafter *unem_female*) are the main explicative variables. In figure 1, it is reported the evolution of both the general fertility rate and of the total unemployment rate in the considered period.

Going into greater detail on ARDL panel techniques, we implement both the Mean Group (MG) and Pooled Mean Group (PMG) estimator, proposed respectively by [9] and [10]. In addition to distinguishing between short-run and long-run effects, these techniques make it possible to deal with time series that have different orders of integration; however, the series must not be second order stationary. Our battery of stationarity tests (not reported here) shows indeed that male unemployment is integrated of order one, while the other time series are integrated of order zero. The PMG model enables to estimate regional specific short-run coefficients and heterogeneous (region by region) speed of adjustment to the long-run equilibrium values, allowing error variances to be heterogeneous across regions, while the long-run slope coefficients are restricted to be homogeneous across regions. With $LGFR_{it}$ indicating the log of general fertility rate in region i at time t and with x the log of the unemployment rates in region i at time t-j, the PMG could be written as follows:

$$LGFR_{it} = \sum_{j=1}^{p} \lambda_{ij} LGFR_{i,t-j} + \sum_{j=3}^{q} \eta_{ij} x_{i,t-j} + m_i + \varepsilon_{it} \qquad (1)$$

For i=Apulia, Basilicata, … and t=1992q4,…2017q4

m_i is a unit-specific effect. In equation 1, j=3 is exemplificative of a model in which we assume that the short-run effects on fertility produced by the increase in unemployment are exerted with a lag of three quarters. This has been done to consider the potential effect of economic uncertainty on delaying conception to better times. Thus, this should have a negative effect on fertility, at least nine months later. However, we estimated the model also allowing j=4, j=5, and j=6. ε_{it} is a white noise error term. p and q (i.e. the number of lags of both dependent and independent variables) are usually decided by adopting the AIC criteria.

The model can be reparametrized as a VECM system:

$$\Delta LGFR_{it} = \theta_i \left(y_{i,t-1} - \beta' x_{i,t-4} \right) + \sum_{j=1}^{p-1} \lambda_{ij} \Delta GFR_{i,t-j} + \sum_{j=3}^{q-1} \eta_{ij} \Delta x_{i,t-j} + m_i + \varepsilon_{it} \qquad (2)$$

Where β is the long-run parameter and assumed to be equal across i. θ_i are the error correction parameters, which can be interpreted as the speed of adjustment when a shock perturbs the long-run equilibrium. As usual in econometric literature, the operator Δ indicates that the variables are differenced. The main difference between PGM and MG is that the latter does not impose long-run homogeneity. In particular, the MG estimator obtains the long-run parameters from autoregressive distribution lag models estimated separately for each unit i, while for the whole panel the long-run parameter is computed as the arithmetic mean of the individual β_i. MG is less efficient than PMG when the assumption of homogeneity holds. In the latter case we have also that the PMG is both consistent and efficient and thus

should be preferred to MG. We will implement both estimations and use the Hausman to decide between the MG and PMG estimators.

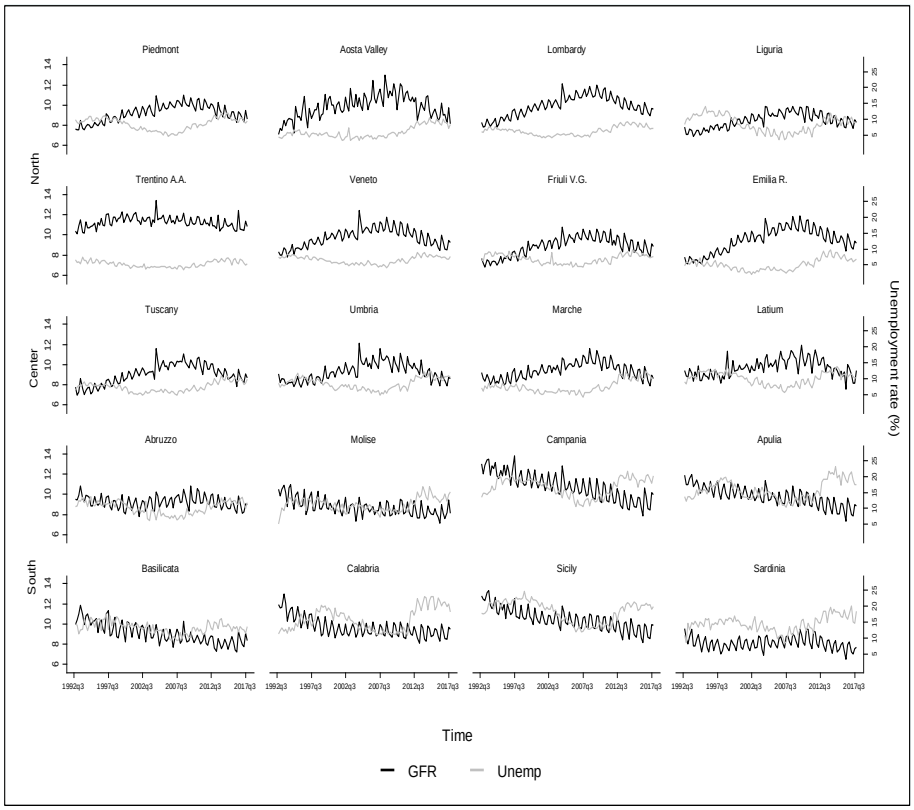


Figure 1: General fertility rates and unemployment rates in Italian Regions, 1992-2017

3 Results

Table 1 shows the results associated with long-run coefficients of both the PMG and MG model and the averaged error correction term.

Table 1: The relationship between fertility and unemployment

	(1)$\Delta LGFR_{it}$) - PMG	(2) $\Delta LGFR_{it}$ - MG
Long-run equilibrium		
Log(*unemp_male*	-0.132 (0.026)***	-0.096 (0.175)
Log(*unemp_female*)	-0.089 (0.033)***	-0.036 (0.170)
Error Correction Term	-0.149 (0.021)***	-0.205 (0.027)***

Nr of lags selected	(5,1,1)	(5,1,1)
Hausman test	Chi-square st. 4.20 sig:0.012	

Lags were selected by minimizing the AIC indicator. Max lags allowed: 8. Standard errors in parentheses; Sig: *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

According to the Hausman test (reported at the end of the table), PMG should be preferred to MG. It is worthwhile noting that only the long-run coefficient associated with male unemployment is larger than that associated with female unemployment. This evidence gives some support to the idea that Italy is a country where the traditional male breadwinner model is still strong, and is coherent with the findings of [3,4] regarding the stronger correlation for male unemployment and fertility with respect to the same correlation calculated for female unemployment.

For a 1% increase in the male unemployment rate we have a decrease in the GFR of about 0.13%. Considering that Southern Italy tends to be characterised by an unemployment rate of more than the 10%, this suggests that the reduction of unemployment in these regions may produce a considerable gain in fertility in the long run. The error correction term is negative and highly statistically significant. If this term is not different from 0, then one may conclude that there is no evidence of a long-run relationship. Accordingly, our result suggests that when the long-run equilibrium is disturbed, about 15% of the shock is absorbed in the following quarter. Table 2 reports the short-run coefficients associated with the unemployment rates for each region.

Table 2: Short-run coefficients associated with unemployment and speed of adjustment

Northern Italy				
	Piedmont	**Aosta V.**	**Lombardy**	**Liguria**
ΔLog(*unemp_female*)	-0.019***	0.007***	-0.020***	-0.031***
ΔLog(*unemp_male*)	0.039***	-0.019***	0.012***	0.015***
	Veneto	**Friuli**	**Emilia**	**TTA**
ΔLog(*unemp_female*)	-0.047***	-0.013***	-0.035***	0.030***
ΔLog(*unemp_male*)	0.016***	0.021***	-0.001*	0.002***
Central Italy				
	Tuscany	**Umbria**	**Marche**	**Latium**
ΔLog(*unemp_female*)	-0.040***	0.027***	-0.016***	-0.030***
ΔLog(*unemp_male*)	-0.004***	0.003***	0.049***	0.040***
Southern Italy				
	Abruzzo	**Molise**	**Campania**	**Apulia**
ΔLog(*unemp_female*)	0.028***	0.027***	0.052***	0.087***
ΔLog(*unemp_male*)	0.029***	-0.100***	0.002	0.030***
	Basilicata	**Calabria**	**Sicily**	**Sardinia**
ΔLog(*unemp_female*)	-0.004**	-0.033***	0.100***	0.032***
ΔLog(*unemp_male*)	0.063***	-0.062***	-0.048***	-0.011***

Sig: *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The short-run specification also includes dummies for the quarters and a dummy for the recession period 2008-2013 (not reported to save space). The results depict a very heterogenous situation across Italian regions. We are able to confirm the results obtained by [4] about the positive relationship between male and female unemployment rates and fertility only for Abruzzo and Apulia. However, the same result has been found in Trentino Alto Adige (Northern Italy) and in Umbria (Central Italy). In the remaining Northern Regions, the relationship is negative. This excludes the hypothesis suggested by [4] about the possible role of the shadow economy, given that according to the estimation carried out by [7], Trentino Alto Adige is characterised by the lowest incidence of shadow economy in Italy. Therefore, at least for this region, a possible explanation could be that when unemployment is expected not to be persistent, couples who have postponed their first birth, are at risk of not having children and this may induce them to lower the desired "quality" of the child [2], thus partly counteracting the negative income effect. Perhaps, the depicted mechanism is weaker in Southern regions where female participation in the labour market is lower and unemployment tends to be longer. Therefore, when men lose their job, the zero earning situation is more frequent, also causing an immediate negative effect. This explanation seems to be consistent also with the findings obtained by [3]. According to these results, the economic uncertainty that derives from the persistence of joblessness negatively impacts fertility intentions with an effect that is stronger for men than for women (even in dual earner family). It remains to establish why this is not true for Abruzzo and Apulia. Perhaps, an investigation at individual level may help to better understand the obtained result.

References

1. Ahn, N. and Mira, P.: A note on the changing relationship between fertility and female employment rates in developed countries. J Popul Econ 15(4): 667–682 (2002).
2. Becker, G.S.: An economic analysis of fertility: Demographic and economic change in developed countries. Princeton, NBER (1960).
3. Busetta, A., Mendola, D. and Vignoli Daniele: Persistent joblessness and fertility intentions. Dem Res, 40, 185-218 (2019).
4. Cazzola, A., Pasquini, L. and Angeli, A.: The relationship between unemployment and fertility in Italy. Dem Res, 34: 1-38 (2016),.
5. Comolli, C. L.: Finnish fertility: Pro - or counter-cyclical?. Research on Finnish Society, 11: 58-64 (2018).
6. Ermisch, J.: Econometric analysis of birth rate dynamics in Britain. J Hum Resour 23(4),563-576 (1998).
7. Istat: Conti Economici Territoriali (2020). Available https://www.istat.it/it/files/2020/01/Conti-economici-territoriali.pdf. Accessed 20 February 2020.
8. Matysiak, A., Sobotka, T. and Vignoli, D.: The great recession and fertility in Europe: A Sub-National Analysis. Eur J Popul (2020) doi: 10.1007/s10680-020-09556-y.
9. Pesaran, H. and Smith, R.: Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. J Econom, 68(1), 79-113 (1995),.
10. Pesaran, M.H., Shin, Y. and Smith, R.P.: Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. J Am Stat Assoc , 94(446), 621-634 (1999).

University drop out and mobility in Italy. First evidences on first level degrees

Abbandono degli studi universitari e mobilità in Italia.

Prime evidenze sulle lauree di primo livello

Nicola Tedesco and Luisa Salaris

Abstract Italian university system presents a widespread inefficiency in the management of training processes which has – among others - a significant impact on the ratio between enrolled and graduated students. Dropout is an important issue which could be considered as an indicator of criticality of the training system. Using data from MIUR "Anagrafe Nazionale Studenti" from MIUR (ANS), the present work aims to measure the levels of dropout in Italy among a cohort of students (2011-2012) enrolled in a bachelor course (Level I) and followed for a period of 5 years. The analysis is carried out by means of Logistic 2-levels Model, to investigate the possible the determinants of the probability of student dropout, together with the role played by student mobility.

Abstract *Il sistema universitario italiano presenta una diffusa inefficienza nella gestione dei processi formativi che ha, tra gli altri, un impatto significativo sul rapporto tra gli studenti iscritti e laureati. Importante risulta il tema degli abbandoni che possono essere considerati come un indicatore della presenza di criticità nel sistema formativo. Grazie all'uso dei dati provenienti dall'Anagrafe Nazionale Studenti del MIUR (ANS), il presente lavoro ambisce a misurare i livelli di abbandono in Italia nella coorte degli studenti (2011-12) iscritti ad un corso di laurea triennale (Livello 1) seguita per un periodo di 5 anni. L'analisi è stata condotta attraverso l'uso di un modello logistico a 2 Livelli, per investigare le possibili determinanti dell'abbandono, assieme al ruolo giocato dalla mobilità.*

[1] Nicola Tedesco, Dipartimento di Scienze Politiche e Sociali, Università di Cagliari; email: tedesco@unica.it

Luisa Salaris, Dipartimento di Scienze Politiche e Sociali, Università di Cagliari; email: salaris@unica.it

This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P. CUP: B78D19000180001

Key words: dropout, mobility, university, cohort, multilevel model

1 Introduction

That of University dropout is a relevant issue not only for student, but it is also represent a relevant topic for the evaluation of higher education system performances. Its incidence is highly variable and depends on the different methods of teaching delivery and the time needed to complete the studies. These aspects make it an issue that all European countries, despite with some differences in changing contexts. In Italy, high levels of university student dropout are generally observed, and this is partially related to the fact that Italian university system suffers from a low level of efficiency. Despite an important number of students every year decide to enrol at the University, there is a low number of graduates. Another relevant peculiarity of Italian system is that an enrolled student can indefinitely remain in this "enrolled" status and this occurs even if he/she does not achieve CFU (University Training Credits). As it is easy to understand this condition poses many problems for the estimation of dropout.

Prior studies have reported that the determinants of dropout are different: economic, need to get a job, wrong choices (see [1], [2], [3], [4]); but the possible role of student mobility has been never considered as a push/break factor.

The primary aim of this paper is measure student dropout in Italy, while instigating the possible determinants of this phenomenon based on the data provided by the MIUR student registry (Anagrafe Nazionale Studenti, ANS). Moreover, according to the fact that in Italy a consistent level of mobility is observed, in particular from Southern regions to Northern ones, in this work the possible role of mobility within the dropout dynamics is explored and considered into the analysis.
In section 2 the used data and methods are described together with the definition of key concepts for analysis, i.e. that of dropout and of mobility. In section 3 presents the results of data analysis according to the following two steps: firstly, the events in the career trajectories of students occurred during a period of 5 years are reported; secondly, thank to the application of Logistic 2-level Models the possible determinants of student dropout are investigated, where respectively region of residence and region of university of enrolment were considered as Second Level variables.

2 Data and methods

The data comes from the MIUR student registry ANS. The studied population is the cohort of students that during academic year 2011-2012 enrolled in a bachelor

course (Level I). The university career of these student cohorts was followed for a period of 5 years. We observed outcomes at the first level degree and in the possible access to the second level. The ANS data set contains the micro-data on the individual students enrolled in all Italian universities, on their university careers and the events occurred in their administrative career. In this study, the following individual information were considered: gender, region of residence and type of diploma. The cross-checking information on municipality of residence of the student and on municipality of the university site led to a classification of enrolled students as "non-mover" or "mover", where students belonging to the last group are those that are enrolled in University that is located in a different region to that of their residence. This variable allows to include in the analysis the possible role of mobility and to explore if the status of "mover" has an increasing/decreasing effect on student dropout. The analysis considers regional level mobility as it proves to be a more stable movement compared to mobility occurring within provinces which could be affected by commuting phenomenon. The analysis longitudinally considers the changes in the student status at every academic year. The considered student statuses are the following: stay enrolled, formal drop out, re-enrolment in another course of study, do not confirm the enrolment.

Data were analysed in two steps. Firstly, we observed the status of each student every a.y. to measure real drop out (net of re-enrolment); secondly, we modelled the probability of dropping out as success event (formal drop out and do not confirm the enrolment) vs failure (take the degree and enrol to a course of the II level, re-enrolment in another course of study and stay enrolled). Every year a student can change his/her status. All status that represent the permanence in the university system are aggregated as "enrol", the others as "drop out" to obtain a dichotomous dependent variable (drop out vs enrol).

3 Data analysis

The first step of the analysis regards the reconstruction of student trajectories over a period of 5 academic years. Table 1 firstly reports the outcomes of cohort of 2011 Italian students (269,309 subjects). During the first year of observation 41,180 dropout are recorded. The number of dropouts in subsequent year increases up to 58,110 in 2014. 2015 data was not considered in the analysis (see note).

Table 1 also shows the behaviour of the cohort of dropping out student in 2011 in subsequent years. In 2012, a great number of students that dropped out in 2011 (18,529, 45.0%) started a new university career. This behaviour persists in subsequent years, but with a low magnitude. So, it is evident that drop out in 2011 and 2012 are clearly a repositioning in the choice of university course.

In Table 2 the evolution of dropping out considering the student status of mover/not-mover is considered, focusing on regional movers, who are those students that are resident in a different region other than that of the chosen university. Region dimension was preferred to provincial one as the latter could reduce the estimation of real number of movers. In many geographic areas of Italy, it is easy to observe

mover-students between provinces, but a large proportion of this mobility is generated by commuter students, because their residence is not so far from the university but is still in another province. Accordingly, despite a regional approach might be reductive, as it does not consider spatial variability due to internal regional movements, it still represents a first useful attempt to investigate the possible role of mobility and to think about alternative measure that will be implemented in future steps of this research (i.e. a proximity indicator (according to distance and time to reach university).

Table 1: Enrolled students and dropout in Italy in 2011-2015 a.y.

Outcomes	***2011***	***2012***	***2013***	***2014***	***2015***
Enrolled	228,096	-	-	-	-
Drop out (registered every year)	41,180	48,463	52,742	58,110	175,735*
Dropping-out cohort of 2011					
Drop out	41,180	22,651	18,368	16,789	15,575
Re-enrolled		18,529	3,847	1,579	1,214

*The data of 2015 is extraordinarily high because it contains first-level graduates who decided not to enrol into a II level degree in 2015.

Table 2: 2011 cohort student enrolled in 2011 according to mobility status

Outcomes	***2011***	***2012***	***2013***	***2014***	***2015****
Mover	50,787	82,980	92,353	126,657	144,770
of which drop out	6,912	38,839	48,068	54,492	100,613
Not Mover	218,522	186,329	176,956	142,652	124,539
of which drop out	34,268	9,624	4,674	3,618	75,122
Dropping-out cohort of 2011					
Mover	6,912	20,396	17,945	16,624	15,721
Not Mover	34,268	1,819	423	165	34

*The 2015 data consider I level graduates who decide not to enrol in the II level degree in 2015.

The number of movers has increased every year. In 2011, 50.787 student begin their academic career moving to another region. In 2012 this number increase significantly as in 2014, when the first graduates at I level enrols at the second level degree. It is interesting to observe how the number of mover student dropout is greater only after one a.y. (38.839), while between not-mover students is higher in the first year and then decrease consistently.

The analysis of the cohort of dropouts shows that among not-mover students most of them leaves studies in the first year (2011) to reorient their own academic career (as shown in Table 1). If a student is a mover, the great part of dropout delays by one year (in 2012 20.396), probably because his own condition of mover encourages the choice to stay enrolled also for parental influence [3].

Table 3: 2-levels logistic model for the probability to drop out according to sex, mobility status, type of diploma and region of residence (2^ level units). Cohort 2011.

Covariates model 1	***OR***	***SE***	***p***	***95% Confidence Interval***	
Sex (base=Female)					
Male	1.14	0.013	0.000	1.12	1.17
Regional mover (Base=No)					
Yes	0.83	0.012	0.000	0.81	0.86
High school diploma (Base = Scientific High School)					
Other High School	1.37	0.023	0.000	1.33	1.42
Classical High School	1.02	0.018	0.160	0.99	1.06
Abroad High School	1.13	0.042	0.000	1.06	1.22
Professional High School	1.81	0.042	0.000	1.73	1.90
Technical High School	1.54	0.021	0.000	1.50	1.59
Variance (regions)	0.1566	0.051		0.08	0.30

Table 4: 2-levels logistic model for the probability to drop out according to sex, mobility status, type of diploma and university of enrolment (2^ level units). Cohort 2011.

Covariates model 2	***OR***	***SE***	***p***	***95% Confidence Interval***	
Sex (base=Female)					
Male	1.15	0.013	0.000	1.13	1.18
Regional mover (Base=No)					
Yes	0.98	0.015	0.149	0.95	1.01
High school diploma (Base = Scientific High School)					
Other High School	1.39	0.024	0.000	1.34	1.44
Classical High School	1.05	0.018	0.011	1.01	1.08
Abroad High School	1.19	0.044	0.000	1.11	1.28
Professional High School	1.81	0.042	0.000	1.73	1.89
Technical High School	1.52	0.021	0.000	1.47	1.56
Variance (universities)	0.4831	0.0092		0.10	0.13

To preliminary investigate the role of some important covariates, we build a 2-level logistic regression model. The dependent variable is the probability to drop out. The considered covariates are sex (male, female), mobility status (yes, no) at the first observation year of the cohort 2011, type of high school diploma. The mobility status is a condition that can change every year. But to simplify data analysis, we decided to consider only the initial mobility status.

In model 1 (Table 3) the region of residence of the student was introduced as second level variable. In this way we can measure the variability of the dependent variable due to the geographical breakdown. The variability explained by region of residence is quite important. Among covariates male has a greater probability to dropout (OR=1.14), the opposite for movers (OR=0.83). Except for student with classical high school, for the others is higher the probability of dropout.

In model 2 (Table 4) we substitute the 2-level variable and we introduce the university of enrolment. The parameters estimated does not change compared to model 1, but it grows considerably the variance component (0.4831).

4 Conclusion and further developments

The cohort analysis (Tables 1) showed that dropout is a relevant problem of Italian university system. However, in the first year, the possibility of not confirming their enrolment is clearly an opportunistic behaviour adopted by an important share of students. In fact, if a student drops out he/she does not have to pay enrolment fees as he/she can sign up again to another course. The analysis of dropping-out cohort shows it clearly (Table 1).

Regional mobility influences the propensity to drop out (Table 2). If a student is a mover, he decides to drop out after one a.y. (probably because the status of mover delays his decision to drop out). If a student is not mover the decision to drop out is immediately, probably to re-orient his own choice.

Model results suggest that if we consider universities as 2-level units, we can better explain the variability of the probability of drop out. Probably the different offer of university courses, the different organization, the reputation of some specific university locations creates differ conditions that significantly influence the experience of university students.

These are preliminary results. The university career of a student is characterized by different events. Each student in fact can enrol to a first level course, drop out, enrolment in another course, take a degree. Later the student can enrol to second level course, drop out, to enrol in another course, take a second level degree. All these possible events suggest to analyze data with Multilevel Multistate Models. This will be our next step.

References

1. AA.VV., Drop out and Completion in Higher Education in Europe, Main Report European Commission Education and Culture (2015) ISBN: 978-92-79-52352-6 doi: 10.2766/826962
2. Rodríguez-Gómez, D., Feixas, M., Gairín, J., Muñoz, J.L., Understanding Catalan university dropout from a cross-national approach, Studies in HigherEducation, 40:4, 690-703, (2015) doi: 10.1080/03075079.2013.842966
3. Donggeun, K., Seoyong, K., Sustainable Education: Analyzing the Determinants of University Student Dropout by Nonlinear Panel Data Models, Sustainable, 10:4, (2018) https://doi.org/10.3390/su10040954
4. Aina, C.: Parental background and university dropout in Italy, High Educ (2013) 65:437–456 doi 10.1007/s10734-012-9554-z
5. Tedesco, N., Puggioni, G., (2002), Il rischio di abbandono degli studi universitari. Problemi di rilevazione e di misura. In: Fabbris, L., (eds.) Indicatori e metodi per l'analisi dei percorsi universitari e post-universitari collana "Determinazione e previstone di rischi sociali e sanitari", pp. 89-100. CLEUP (2002). ISBN: 88-7178-593-2

Worthiness-based Scale Quantifying
Re-interpretare scale ordinali equi-distanziate

Giulio D'Epifanio[1]

Abstract The construction of an ordinal scale Y, to be associated to a performance index in evaluating social-agents, is outlined which is quantified, adopting an "intrinsic worthiness" criterion, standardized on a chosen reference-agent, eg that which is intended as representing a (actual or perhaps hypothetical) "best practice". The usual practice of using the equispaced scale is re-interpreted. The ordinal levels of Y are identified by design through scheduling a hierarchical sequence of increasingly stringent "goals to be achieved". The index, to be associated to Y, borrows the structure of the Yaari-Quiggin functional, from the RDEU theory. But, the concept of "value increases", in advancing on the scheduled goals, is meaningfully re-interpreted herein, besides utility-based meaning, as "social worthiness". These "value increases" may be extracted, fully normalized on the data-behavior of the chosen "reference social-agent", upon a probabilistic formal setting using data-analysis (perhaps pseudo-Bayesian) tools. Thus, the ordinal levels of Y remain quantified, alternatively to various other approaches.

Abstract Si delinea la costruzione di una scala ordinale (da associare ad un indice di prestazione), quantificata adottando un principio di "merito intrinseco", normalizzata sulla scelta di un agente di riferimento, ad es. quello il cui comportamento è da intendersi come una "buona pratica", reale o ipotetica che sia. I livelli ordinali of Y sono identificati, per disegno, attraverso la pianificazione di una seguenza di obiettivi da raggiungere, progressivamente più severi. L'indice, associato alla quantificazione di Y, è formalmente strutturato sul funzionale di Yaari-Quiggin, ripreso dalla teoria RDEU. Ma il concetto di "incremento di valore", nelle transizioni sulla catena di obiettivi, è qui reinterpretato come "merito sociale", piuttosto che utilità. Questi incrementi potranno essere operativamente estratti previa formalizzazione in un quadro probabilistico, attraverso strumenti di data-analysis.

Key words: indexing, social agent evaluation, scale quantifying, worthiness

University of the Study of Perugia, Department of Political Science, ggiuliodd@gmail.com

1 Introduction and methodological outline

[The higher level question] Some agent-schools $\{A_1, A_2, \ldots, A_q\}$ have to be benchmarked, from the view of the National Instruction Authority (the policy maker), with respect to *"the ability to address students*[1] *in achieving outcomes in learning"* on a certain outcome scale Y which classifies student-performance (eg by using a certain *learning test*) on the outcome-levels labeled as: "*very bad*" , "*bad*", "*almost enough*", "*sufficient*, "*more than sufficient*", "*good*", "*excellent*"). The policy-maker (PM) demands the agent-assessments be standardized on the behavior of a certain school A^* (the reference agent), perhaps chosen by experts to represent an actual instance of "*what should be intended as a, reasonable and desirable, best practice*". Performance-data are reported in table (1). More complex ex. are in [3].

level of performance Y:	I	II	III	V	V	VI	VII
actual agents:							
agent A1	0	16	24	31	12	0	4
agent A2	4	28	65	107	26	1	3
agent A3	4	42	71	102	33	0	2
agent A4	4	71	112	194	53	0	4

(a) Data by the agents to be benchmarked

Y:	I	II	III	IV	V	VI	VII
reference agent A^*	12	157	272	434	124	1	13

(b) Data by reference stand. agent A^*

Table 1: Example data

The PM[2] would need now an evaluation-machinery (the index) which, whenever applied to any agent A, it takes into input the performance-data of A to provide a certain performance-value, on a properly quantified ordinal scale Y, so that such a value is meaningful upon the conceptual framework which the PM has adopted, conditional on a set of design specifications. These specifications also including the choice of reference-agent A^* on the which the PM would normalize its evaluation-machinery.

[Ordinal scaling] In order to design an ordinal scale Y, on the which the evaluation-machine has to be constructed, we refer to the following scheme. Suppose that the PM, in pursuing its purposes, has been able to schedule a sequential hierarchy of, increasingly stringent, (Guttman-like ordered) goals[3]

[1] in a specified social domain $\mathcal{D}$, e.g. "18 year old female with a certain social background"

[2] To him, in setting value-levels, it seems be excessively "naive" considering equi-distanced scale; but, not clearly structured levels-score choices seem difficult to be justified in institutional benchmarking. On the other hand, economic utility-based interpretations seem lacking of meaning in social assessments (for an operative-research-based approach, see [2]). Worse still, merely data-analysis-criteria based methods (eg see [4]), "*di per se*" seem of little relevance at the PM's higher level question (eg see also [6]).

[3] Guttman order: $O_l \preceq O_{l+1} \Leftrightarrow$ "whenever goal O_{l+1} is achieved, also O_l has been achieved"

$$O_0 \preceq O_1 \preceq O_2 \preceq ... \preceq O_l \preceq ... \preceq \ldots \preceq O_{L-1} \preceq O_L := O_{Full}, \qquad (1)$$

Assume that a verbal ordinal scale for the outcome exists so that (unless of recoding it as a sequence of integers $0, 1, \ldots, l, \ldots, L$) the intrinsic meaning is established through the following identification[4]: $O_l \leftrightarrow (Y \geq l)$, $l := 0, \ldots, L$, where $O_0 := (Y \geq 0)$ represents the "tautological-goal" (ie the dummy goal always achieved by anyone). Therefore, an ordinal scale $Y \in \{0, 1, ..., l, ..., L\}$ remains identified with goals sequence (1) from the which it will inherit semantics, and vice versa.

[The formal evaluation machine] Suppose that the PM is able to assign, for any transitions in advancing sequence (1), the value-increase[5] $\omega_l := \Delta_{l-1}Val := Val(O_l) - Val(O_{l-1}) \geq 0$, $l := 1, \ldots, L$. Here, $Val(.)$ denotes a (non-negative, not decreasing in value by crossing goals-sequence (1)) value-function, which is initialized at $\omega_0 := Val(O_0) = 0$. The utility-value theory would provide the formal platform (e.g. see [1], pp. 559) from the which it could be inherited (as an instance of a more general class of Yaari-Quiggin-like value-functionals, eg see [1]) the following index-structure:

$$A \in \{A_1, .., A_q\} :\mapsto W[A] := \sum_{l:=1}^{L} \omega_l \cdot (1 - F_Y[p[A]](l)) = \sum_{l:=1}^{L} s(l; \omega) \cdot p_l[A] \quad (2)$$

unless of parameters list $\omega = (\omega_1, .., \omega_L)$ of value-increases on the transitions between adjacent levels of Y (ie in advancing performance on the scheduled ordered goals-sequence (1)). Here, $p[A] := (p_0, p_1, \ldots, p_L)[A]$ denotes the relative distribution which describes the behavior of agent A; $F_Y[p]$ the cumulative distribution such that $F_Y[p(A)](l) = p_0(A) + p_1(A) + \cdots + p_{l-1}(A)$; $s(l; \omega) := \omega_1 + \omega_2 + \cdots + \omega_l$ the Y-levels-quantifier.[6]

[The core methodological question] At the higher-level of the PM's decisions, the practical question now arises, which has an intrinsic methodological interest with regard to the operative way for specifying parameters ω_l, $l := 1, \ldots, L$, which will enter index structure (2) above, in a way that they will be actually meaningful and useful to the PM in evaluating social performance, besides formal utility-based settings.

[What it is proposed herein] Value-increases parameters ω_l on chain (1), which will enter (2), are re-interpreted with the meaning of "social worthiness-

[4] it is set here the logical identification of proposition *"goal O_l is achieved"* with that of "*the outcome-level of Y is at least l*"

[5] Here, $\omega_l := \Delta_{l-1}Val$ may be interpreted as the reference-value which would be gained by any social agent (due to its political activity in addressing the governed individuals) which is able to improve the condition-level of a certain "*standard individual*", from the current $(l-1)$th level to the next lth one

[6] Formally, $s(.; \omega)$ can be viewed as the quantification-function of the ordinal-levels $0, 1, \ldots, L$ of Y, which is obtained by (Choquet-)integrating value-increases ω_l, so that:$s(0; \omega) := Val(O_0) = 0 \leq s(1; \omega) := Val(O_1) = \omega_1 \leq s(2; \omega) := Val(O_2) = \omega_1 + \omega_2 \leq \ldots \leq s(L; \omega) := Val(O_L) = \omega_1 + \omega_2 + \cdots + \omega_L$

increases", by recalling a "principle of intrinsic worthiness". These increases are then specified upon a probabilistic formal setting, which may be also advanced enough to include complex contexts (eg see [3]). Hence, parameters ω_l could be numerically elicited from the reference-agent A^* data, conditionally on the PM's design-specifications, perhaps using pseudo-Bayesian data-analysis tools (eg as in [3]). The methodological point of arrival (unless of more complex extensions and advancing) is that an ordinal quantified scale Y (by Choquet-integrating increases ω_l) will remain operatively quantified, fully normalized on that specific "meaning of worthiness" which is intrinsic into the PM's choice of assuming, as reference-behavior among various alternatives, the behavior of agent A^* (eg a presumed "best-practice") .

2 Worthiness based indexing

Recalling (eg see [3]) the criterion of intrinsic worthiness[7], the "increases of worthiness"$\omega_l := \Delta_{l-1}Val(.)$ may be interpreted as follows. Let $\mathcal{P}^*$denote the population of the (real or perhaps virtual) individuals which are governed by the "reference agent" A^*.

> For any actual individual i (eg a student of a school in tab.1a), which has achieved goal O_{l-1} moving up goals-chain (1), *the higher* the $\mathcal{P}^*-$standardized risk of failing the next goal O_l, *the greater* the $\mathcal{P}^*-$standardized "increase of worthiness" $\Delta_{l-1}Val(.)$ which such an individual i gains whenever it also achieves goal O_l.

By adopting now a probabilistic interpretative setting, such a $\mathcal{P}^*-$standardized risk could be related to non-transition probability[8] $Pr\{Y = l-1|Y \geq l-1; \mathcal{P}^*\}$, up to some monotone transformation $\varphi_l(.)$ (eg see [3]). Thus, by setting value-increases as:

$$\Delta_{l-1}Val := \omega_l^* := \varphi_l(Pr\{Y = l-1|Y \geq l-1; \mathcal{P}^*\}),\ l := 1, \dots L \qquad (3)$$

[7] Consider hierarchical chain of goals (1). Given that a certain goal O_{l-1} has been achieved, the greater the resistance, with reference to the evaluation framework, to also achieve the next pursued goal O_l, by continuing to improve, the greater the increment of value $\omega_l := \Delta_{l-1}Val(.)$ due to the "intrinsic worthiness" of who, effectively, is able to achieve it.

[8] of course, $Pr\{Y=l-1|Y\geq l-1;\mathcal{P}^*\}=\frac{Pr\{Y=l-1;\mathcal{P}^*\}}{Pr\{Y\geq l-1;\mathcal{P}^*\}} = \frac{p^*_{l-1}}{p^*_{l-1}+p^*_l+\dots+p^*_L}$.

Example, by data-table (1b): $\omega_0 = Pr\{O_0\ fails\} = 0$, $\omega_1 = Pr\{O_1\ fails\} \simeq \frac{12}{12+157+272+434+124+1+13}$, $\omega_2 = Pr\{O_2\ fails|O_1\ achieved\} \simeq \frac{157}{157+272+434+124+1+13}$, $\omega_3 = Pr\{O_3\ fails|O_2\ achieved\} \simeq \frac{272}{272+434+124+1+13}$, $\omega_4 \simeq Pr\{O_4\ fails|O_3\ achieved\} = \frac{434}{434+124+1+13}$, $\omega_5 = Pr\{O_5\ fails|O_4\ achieved\} \simeq \frac{124}{124+1+13}$, $\omega_6 \simeq Pr\{O_6\ fails|O_5\ achieved\} = \frac{1}{1+13}$. Then (let here $\varphi_l(.)$ be the identity), by accumulating value-increases yields:

level-score of Y:	0	1	2	3	4	5	6
	0	0.011846	0.168689	0.490964	1.249705	2.148256	2.219685

it yields index structure (2) to be instanced as[9]

$$A :\longmapsto W[A;\, \omega^*] := \sum_{l:=1}^{L} \varphi_l(\frac{Pr\{Y = l-1; \mathcal{P}^*\}}{Pr\{Y \geq l-1; \mathcal{P}^*\}}) \cdot (1 - F_Y[p[A]](l)) \qquad (4)$$

From the which, the re-ranged on interval $[0, 100\%]$, normalized version[10]

$$A \in \{A_1, .., A_p\} :\longmapsto W^*[A;\, \omega^*] := \frac{W[A;\, \omega^*] - W[A_{worst};\, \omega^*]}{W[A_{best};\, \omega^*] - W[A_{worst};\, \omega^*]} \qquad (5)$$

3 Advancing and notes on equidistant scaling

[***Multi-domain indexing***] The PM might want consider worthiness-based performance difference, among the social agents, conditional on status $x := X \in \{x_1, \ldots, x_r, \ldots, x_R\}$ of the governed individuals, normalized on standard agent A^*. Here, $\{x_1, \ldots, x_r, \ldots, x_R\}$ represents a set of *"reference domains"*. A global worthiness-index, integrating on domains X from (4), could have the following structure:

$$A \longmapsto \sum_{r:=1}^{R} q_r \cdot W_{x_r}[p_r([A]; \omega_r(\mathcal{P}^*)] = \sum_{r:=1}^{R} q_r \cdot \{\sum_{l:=1}^{L} \varphi_l(\frac{\exp(\widehat{a}_l + \hat{b}_l x_r)}{1 + \exp(\hat{a}_l + \hat{b}_l x_r)}) \cdot (1 - F_{Y|x_r}[p[A]](l))\}$$

where parameters $\hat{a}$ and $\widehat{b}$ were determined, using a sequence of logistic models to model value-increases (3) conditional on status x, given reference-population $\mathcal{P}^*$associated to A^*. Here, $q_r \geq 0$ weights[11] ($\sum_{i:=1}^{R} q_r = 1$) the reference domains. A more complex example is presented in [3].

[***Equi-distanced quantifying***] Process any individual, governed by standard-agent A^* in a certain condition-domain $\mathcal{D}$ (ie in reference-population $\mathcal{P}^*_{|\mathcal{D}}$), sequentially against goal-achievement detectors associated to goals-hierarchy

[9] Here, a continuous monotone functions $\varphi_l(.)$ could be chosen for specifying some types of design-requirements (e.g. the additivity) on the worthiness-scale.

[10] Here, the performance of agent $A \in \{A_1, .., A_p\}$ is graduated on the percentage of gained worthiness, in advancing from the complete social-failure (represented by distribution $p[A_{worst}] := (1, 0, \ldots, 0)$ associated to the "worst-virtual agent" named A_{worst}) toward the full achievement of the social overall-goal (represented by distribution $p[A_{best}] := (0, 0, \ldots, 0, 1)$ associated to the best virtual agent named A_{best})

[11] the weights should represent the political relevancy of the "social reference domains" to the overall purpose of the PM

(1). Then[12], recalling worthiness-increments (3), the value-increases will be provided (for simplicity use identity $\varphi_l(t) = t$) by:

$\omega_l^* = \frac{q_1 q_2 \cdots q_l (1-q_{l+1})}{q_1 q_2 \cdots q_l (1-q_{l+1}) + q_1 q_2 \cdots q_l \cdot q_{l+1}(1-q_{l+2}) + \ldots + q_1 q_2 \cdots q_l \cdot q_{l+1} \cdots q_L} = \frac{(1-q_{l+1})}{(1-q_{l+1}) + q_{l+1}(1-q_{l+2}) + \ldots + q_{l+1} \cdots q_L}$

Suppose now that, in particular, A^* has been chosen (to be put into example-table 1b) such that transition probabilities $q_l := Pr\{Y > l-1 | Y \geq l-1; \mathcal{P}^*\} = q,\ l := 1, \ldots L$ ($0 < q < 1$ constant) were invariant through chain (1). Then, it will happen[13] that worthiness increases $\omega_l^* = 1-q,\ l := 1, \ldots L$ remain constant. Thus, normalized on A^*, it will be induced an equi-distanced scale, with some curious interpretation[14]with respect to the PM's choice of considering A^* as a "best practice".

References

1. Chateauneuf A., Cohen M., Meilijson I. (2004), Four Notions of Mean-preserving Increase in Risk Attitudes and Applications to the Rank-dependent Expected Utility Model, *Journal of Mathematical Economics* 40, 547-571 1
2. D'Epifanio G. (2009), Implicit Social Scaling. From an institutional perspective. *Social Indicator Research 94: 203-212* 2
3. D'Epifanio G. (2018), *Indexing the Normalized Worthiness of Social Agents*, in Springer Proceedings in Mathematics & Statistics 227, C. Perna et al. (eds.), pp 263-274, Studies in Theoretical and Applied Statistics, Springer International Publishing AG, Cham, Switzerland ISBN 978-3-319-73905-2, ISSN 2194-1009 (eBook) https://doi.org/10.1007/978-3-319-73906-9 Pre-print in:pre-print 1, 1, 2, 3
4. Jöreskog K.K, Moustaki I. (2011), A Comparison of Three Approaches, *Multivariate Behavioral Research*, 36 (3), 347-387 2
5. Kampen J., Swyngedouw M. (2000), The ordinal controversy revisited, *Quality and Quantity*, Volume 34, Number 1
6. Journal Royal Statistical Society (2005) Performance indicators: good, bad, and ugly. *J. R. Statist. Soc. A, 168, Part 1, 1–27*

[12] Let $q_l := Pr\{Y > l-1 | Y \geq l-1; \mathcal{P}^*\}$, $l := 1, \ldots, L$, denote the "transition probability" which an individual, once arrived at level $(l-1)$th of Y, is also able to pass over beside. Of course, the expected "social behavior" of A^*, with respect to outcome-classifier $Y \in \{0, 1, \ldots, L\}$, is represented by parameters $p[A^*] := (p_0, p_1, \ldots, p_L)[A^*]$, where $p_0 := 1-q_1$, $p_1 := q_1(1-q_2)$, ..., $p_l := q_1 q_2 \cdots q_l(1-q_{l+1})$,.., $p_L := q_1 q_2 \cdots q_l \cdot q_{l+1} \cdots q_L$; here, $p_l[A^*]$ represents the probability of definitively remaining at level lth of Y.

[13] $\omega_l^* = \frac{1-q}{(1-q)+q(1-q)+q^2(1-q)+\ldots+q^{L-1}(1-q)+q^L} = \frac{1-q}{1-q+q-q^2+q^2-q^3+\ldots+q^{L-1}-q^L+q^L} = 1-q$

[14] Vice versa, the choice of using an equi-distanced ordinal scale for Y would be equivalent, in our interpretative setting in schools evaluations, to choose as reference standard-agent A^* that "virtual" school (the distribution $p[A^*] := (1-q, q(1-q), \ldots, q^{L-1}(1-q), q^L)[A^*]$ to be put into example-table 1b) whose students constitute that reference population $\mathcal{P}^*$ where everyone, subjected to a sequence of k-outcome-formatted learning-tests (intended as goal-achievement detectors on goals-hierarchy (1)) until a goal is failed, tries to guess, among the k (k constant positive integer) proposed answers, the correct one. That is to say that, in particular using a sequence of $k := 2$ outcome-formatted learning-tests, it would be "*as if*" any student in $\mathcal{P}^*$ would respond according to the random outcome from tossing a certain coin.

Young people in Southern Italy and the phenomenon of immigration: what is their perception?

I giovani del Sud Italia e il fenomeno dell'immigrazione: quale percezione?

Nunziata Ribecco, Angela Maria D'Uggento and Angela Labarile

Abstract In Italy, the issues of immigration, integration and perception of the phenomenon are extremely topical and have prompted a rather heated debate. Lacking official data on the citizens' "sentiment", a survey was carried out through a questionnaire addressed to about 1,200 students attending the Apulian high schools. The aims are mainly two: to ascertain possible misperceptions or information gaps and to better understand young people's opinions about the positive or negative effects of the immigrants' presence in our communities, and their behaviour to facilitate the integration process. Some relevant misperceptions are highlighted, and the respondents' perceptions are detected by means of exploratory analyses, heterogeneity indexes and multiple correspondence analysis.

Abstract *In Italia, i temi dell'immigrazione, integrazione e percezione del fenomeno risultano estremamente attuali e originano accesi dibattiti. In mancanza di dati ufficiali sul "sentiment" dei cittadini, è stata svolta un'indagine attraverso la somministrazione di un questionario che ha coinvolto circa 1200 studenti delle scuole superiori della Puglia. L'obiettivo è duplice: individuare eventuali bias informativi, nonché conoscere la percezione degli intervistati circa la presenza di immigrati nella propria realtà territoriale e gli atteggiamenti per favorire il processo di integrazione. In particolare, la ricerca evidenzia la presenza di bias rilevanti; la percezione degli intervistati è stata analizzata dal punto di vista della eterogeneità mentre l'analisi delle corrispondenze multiple delinea tre profili.*

Key words: immigration, misperception, heterogeneity indexes, multiple correspondence analysis.

[1] Nunziata Ribecco, Università degli Studi di Bari; nunziata.ribecco@uniba.it
Angela Maria D'Uggento, Università degli Studi di Bari; angelamaria.duggento@uniba.it
Angela Labarile, Università degli Studi di Bari Aldo Moro, angelalabarile30@gmail.com

1 Introduction

Perceptions represent the acquisition of awareness of the external context through a series of sensory *stimuli*, filtered through intuitive, psychic and intellectual processes. It is the way in which the sensory-intellectual apparatus perceives external reality in relation to the ability of the individual response. By transposing perceptions in the field of information, it is understood how fundamental becomes to be exposed to external stimuli, and to frequent information flows, whose sources have undergone a profound change in recent years. The massive spread of social media alongside traditional news channels (TV, newspapers, books, internet) has profoundly affected the mechanisms of formation of collective and individual opinions and perceptions. The paper aims at analysing young people's perceptions of the phenomenon of foreign immigration affecting Italy and Puglia, as one of the main migration routes towards Europe. Although immigration is a widely discussed issue, some misperceptions with respect to crucial aspects have emerged. The survey was carried out in February 2019, when Italy was the country of first reception in a European emergency phase, and the phenomenon has also been a battleground in the political and social debate. Public opinion was split up between those who supported the policies of hospitality and their opposites, who leveraged on suspicion and reduction of the sense of security due to immigrants' presence. Then, the research "Immigration and integration: the reality and perception of young people" was carried out among Apulian high schools' students. Besides investigating the perception of this phenomenon, a further issue is understanding the impact of media in conditioning people's perceptions, especially of that part of the population that has not yet developed an autonomous critical sense because of the young age and more likely to be influenced. In a globalised society, reality becomes increasingly difficult to understand, and therefore public perception risks to move further and further away from real data. This distance is defined misperception or cognitive bias and refers to a systematic pattern of deviation from the norm or rationality in judgment. In Psychology, it indicates a tendency to create one's own subjective reality, not necessarily corresponding to the evidence, developed on the basis of the interpretation of the information hold, which, therefore, may lead to an error of judgment or lack of objectivity in many fields [2,3]. Then, by means of some specific questions, the existence and magnitude of cognitive biases on the phenomenon of immigration is verified by comparing the sample's responses with the official information. The paper is organized as follows: section 2 deals with a description of the survey's and the questionnaire, section 3 with the statistical procedures and the discussion on the main findings about misperception measurement, exploratory and multiple correspondence analyses; section 4 briefly concludes.

2 The survey and the data

The survey involved a representative sample of high schools in Puglia; students interviewed were given a questionnaire of 32 questions divided into four sections: I. Knowledge of quantitative data of the phenomenon and comparison to real data; II. Perception of the phenomenon and integration; III. Impact of immigration on the country's living system; IV. Initiatives to share national political choices. Students were asked to express their level of agreement/disagreement with a set of 10 statements, in Section III, related to the problem being investigated, using a five-point Likert scale [4] in which 1 indicates "not at all agree" and 5 "completely agree". The statements are the following and the acronyms of the corresponding variables are also specified:

1. Immigrants contribute to the cultural development of our country (CULTDEV);
2. Immigrants contribute to the economic development of our country (ECONDEV);
3. Immigrants are necessary to compensate for the demographic decline caused by the progressive ageing of the Italian population (DEMODECL);
4. Immigrants spread behaviours contrary to our traditions and put Italian cultural identity at risk (IDENTITY);
5. Immigrants introduce and spread diseases (DISEASE);
6. Immigrants introduce dangerous ideologies into the host country (IDEOLOGY);
7. Immigrants exacerbate public order problems (PUBORDER);
8. The presence of immigrants is a danger to social security (SOCIALSEC);
9. Immigrants compete with the Italian workforce, thus causing job losses and lower wages (JOBLOSSES);
10. Immigrants accept illegal jobs and contribute to the spread of “black work” also for Italians (ILLEGALWORK).

We collected 1,222 online questionnaires, reduced to 1,196 after the phase of data cleaning. Data are analysed by means of explorative and multivariate techniques.

3. Main results and discussion

3.1 Informative misperception among young people

Italy, along with Germany, Great Britain, France and Spain is one of the five EU countries with the highest concentration of immigrants. To measure the main misperceptions, let us now compare some official data [5] with the students’ perceptions as collected by the sample (Tab.1). Most of them believe that immigrants come almost exclusively from Africa, then from Asia, Europe and America. The reality, however, is quite different, as 50.2% of immigrants come from European countries, namely Romania. It is the same if we consider the underestimation of the number of foreign people in Italy.

Another issue concerns the reasons for foreigners staying and, in this case, the perception of the sample does not differ much from the real situation, as the majority believes that the main reason is the family reunion, followed by humanitarian crises and asylum, and finally the working reason. Table 1 summarizes the main findings by showing the real data, the perceived data, and the misperception index, obtained as the difference between the two values.

Table 1: Misperception measurements on some information about immigration

Variables	***Real***	***Perceived***	***Misperception***
Number of immigrants (mean)	5,225,03	3,120,602	2,104,901
Countries (%)			
Europe	50.2	10.7	39.5
Africa	21.7	69.4	-47.7
Asia	20.8	18.1	2.7
America	7.2	1.8	5.4
Reasons (%)			
Family reunion	52.4	49.5	2.9
Asylum, Humanitarian crisis	41.6	41.2	0.4
Job	6.0	9.3	-3.3

3.2 Exploratory analyses

The main statistics synthetizing the responses given by the students to the ten statements about immigration, introduced in Section 2, are shown in Table 2. Since the variables are ordinal, the central tendency measures of interest are the median and the mode, but these indicators are not good at discriminating the different attitude of the interviewed in relation to the statements of the set, especially when they have similar distributions. Considering that the attitude is a continuum, on which individuals can be located according to their answers, each variable can be considered as continuous, so we can compute a numeric summary. This numeric summary is the mean value of the scores, which can be interpreted as a proxy of the average value of the continuous variable "attitude toward the phenomenon". To evaluate the accuracy of the answers collected, that is the reliability concerned with the study, some variability measures fitting ordinal variables need to be selected [1]. Then, we first consider Gini heterogeneity index in its standardised form and, in addition, two other indexes that can be obtained as monotonic functions of the index: Frosini index (F) and Laasko and Taagepera index (A). All the indexes assume values in [0,1] and the results are recapped in Table 2: among the three indexes, the one selected to complete the analysis is the Laasko and Taagepera index (A), because it has the widest range, so it is supposed to be more selective and able to discern better between heterogeneous situations. The following Fig. 1 displays the ten variables in relation to the mean and the A index values. It can be observed that all the mean values range from 2 to little more than 3: in particular, the value of 3 (which indicates the indifference state) is exceeded just by the variable "Illegal Work". Regarding dispersion, this appears to be rather high, since A index overtakes the value of 0.5 for all the variables; the highest measures are linked to the variables "Illegal Work" and "Demographic decline".

Table 2: Mean, Median, Mode and Variability indexes for the ten items

Variables	***Mean***	***Median***	***Mode***	***Gini Index***	***Frosini Index***	***Laasko Index***
Cultural development	2.758	3	3	0.948013	0.771993	0.784812
Economic development	2.471	2	2	0.941275	0.757668	0.762228
Demographic decline	2.817	3	3	0.981822	0.865171	0.915267

Identity	2.315	2	1	0.937928	0.750858	0.751372
Diseases	2.277	2	1	0.930207	0.735816	0,727194
Ideology	2.197	2	2	0.910427	0.700713	0.670275
Public order	2.415	2	2	0.938610	0.752223	0.753564
Social security	2.171	2	2	0.899530	0.683030	0.641661
Job losses	2.436	2	1	0.958447	0.796154	0.821845
Illegal work	3.009	3	3	0.992014	0.910635	0.961305

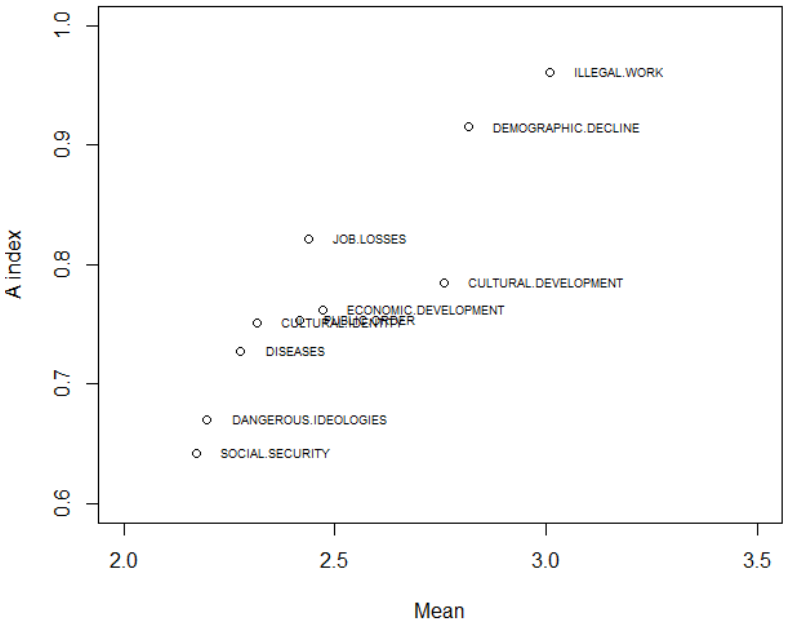


Figure 1: Relation between the mean values and the Laasko and Taagepera Index A

3.3 Multiple correspondence analysis

Another relevant research issue deals with the assessment of the real willingness to provide shelter, as a multidimensional latent construct to be observed trough some direct measurements. This latent variable investigates students' behaviours towards immigrants, beyond the simple statements of principle, mostly inspired by the "social desirability". The Multiple correspondence analysis highlights the proximity relationships among the responses collected with reference to the most representative five statements of the set (Fig. 2). A clear opposition among three groups is drawn: on the left, we find the students who consider immigration a resource for Italian economy, with a positive contribution to labour forces and no other problems for social security or Health. Near them, those who prefer to adopt a neutral behaviour, by selecting the intermediate point equivalent to 3 (labelled with DK). On the right, we find those that we may call the "conservatives", considering

the immigration almost a threat. The model shows a good fitting, explaining the 62.4% of the variance.

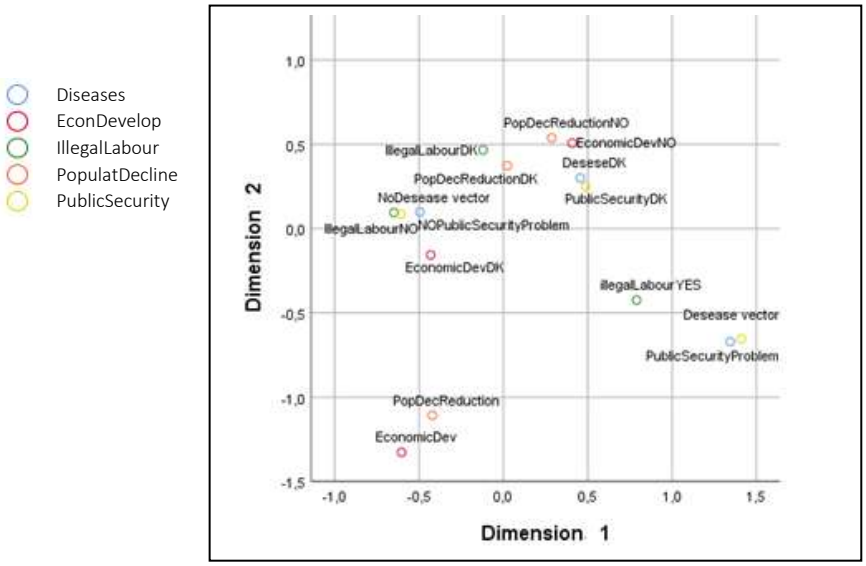


Figure 2: Conjoint plot of variables of Multiple correspondence analysis

4 Some conclusive remarks

The research on youths' perception of immigration, carried out as a part of the National Scientific Degree Plan-PLS project, showed interesting results both regarding the role of the media in affecting misperceptions, and the evidence that inaccurate information produce distrust and closure towards the phenomenon.
It is a fundamental task of educational institutions to support the development of a critical spirit in students, making them autonomous in judgment, and lay the foundation for the future society.

References

1. D'Elia A., Piccolo D., Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico. Quaderni di Statistica, vol. 7, 121-161, Liguori, Napoli (2005). ISBN: 9788820739508.
2. Flynn D.J., Nyhan B., Reifler J.: The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. Advances in Political Psychology, Vol.38, Suppl.1, 2017.
3. Kahneman D, Tversky A. On the study of statistical intuitions. Cognition, 11(1982) 123-141.
4. Likert R. A Technique for the Measurement of Attitudes. Archives of Psychology, 140, 1–55. (1932).
5. ISTAT, Cittadini non comunitari: presenza, nuovi ingressi e acquisizioni di cittadinanza - Anni 2017-2018.