Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica
UNIVERSITÀ POLITECNICA DELLE MARCHE
Dipartimento di Scienze Economiche e Sociali DISES
LIUC Università Cattaneo
LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB
Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

Copyright © 2023
PUBLISHED BY PEARSON
WWW.PEARSON.COM
ISBN 9788891935618AAVV

Contents

Preface **XXII**

1 Plenary Sessions **1**

Inequality indices: accurate simulation-based inference 2
Maria-Pia Victoria-Feser

Examples from the Interface of Neural Models and Spatio-Temporal Statistics in Environmental Applications 7
Christopher K. Wikle, Likun Zhang, Myungsoo Yoo and Xiaoyu Ma

Demographic change and sustainability: novel approaches from digital and computational demography n.a.
Emilio Zagheni

2 Invited Sessions **14**

Machine learning in the design, analysis and integration of sample surveys

Causal Discovery for complex survey data 15
Paola Vicard

Data Integration without conditional independence: a Bayesian Networks approach 21
Pier Luigi Conti, Paola Vicard and Vincenzina Vitale

Mass imputation through Machine Learning techniques in presence of multi-source data 27
Fabrizio De Fausti, Marco Di Zio, Romina Filippini and Simona Toti

Machine learning: different uses and perspectives

Evaluation of pollution containment policies in the US and the role of machine learning algorithms 32
Marco Di Cataldo, Margherita Gerolimetto, Stefano Magrini and Alessandro Spiganti

Machine Learning for Official Statistics: An Application on External Trade n.a.
Mauro Bruno, Maria Serena Causo, Alessio Guandalini, Francesco Ortame and Silvia Russo

Machine learning, data quality and official statistics: challenges and opportunities n.a.
Stefano Menghinello

Statistical Machine Learning for environmental applications

Gaussian Processes and Deep Neural Networks for Spatial Prediction 38
Alex Cucco, Luigi Ippoliti, Nicola Pronello, Pasquale Valentini and Carlo Zaccardi

How can we explain Random Forests in a spatial framework? 42
Natalia Golini, Luca Patelli and Xavier Barber

Recent approaches in coupling deep learning methods with the statistical analysis of spatial point patterns 48
Jorge Mateu and Abdollah Jalilian

Statistical Process Monitoring for Complex Data in Industry 4.0

A Kernel-based Nonparametric Multivariate CUSUM for Location Shifts 53
Konstantinos Bourazas, Konstantinos Fokianos, Christos Panayiotou and Marios Polycarpou

An Approach for Profile Monitoring via Mixture Regression Models 58
Davide Forcina, Antonio Lepore and Biagio Palumbo

Anomaly Detection in Circular Data 63
Houyem Demni and Giovanni C. Porzio

Advances in Data Science and Statistical Learning [IMS Invited Session]

Empirical Bayes approximation of Bayesian learning: understanding a common practice n.a.
Sonia Petrone

Generalized Fiducial Inference on Differentiable Manifolds - a geometric perspective n.a.
Jan Hannig

Model-free bootstrap and conformal prediction in regression n.a.
Dimitris Politis

ENBIS Session: System Maintenance, Boosting algorithms for regression, and Research Excellence

Boosting Diversity in Regression Ensembles 69
Mathias Bourel, Jairo Cugliari, Yannig Goude and Jean-Michel Poggi

How ENBIS has contributed to the UK Universities Research Excellence Framework 71
Shirley Coleman

Maintenance of degrading systems by dynamic programming or reinforcement learning 75
Antonio Pievatolo

Population Dynamics, Climate Change and Sustainability

Climate change impacts on fertility in low- and middle-income countries: An analysis based on global sub-national data n.a.
Côme Cheritel, Roman Hoffmann and Raya Muttarak

Environmental Exposures and Under-5 Mortality in India: A Survival Analysis of DHS data 79
Vinod Joseph Kannankeril Joseph

The impact of temperature on expressed sentiment by migration status: Evidence from geo-located Twitter data 84
Risto Conte Keivabu and Jisu Kim

Statistical Learning for health research and omics data

An alternative to the Dirichlet-multinomial regression model for microbiome data analysis 95
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

Modelling ordinal response to treatment in a real-world cohort study 101
Marco Alfò, Maria Francesca Marino and Silvia D'Elia

On the application of the symmetric graphical lasso for paired data 105
Saverio Ranciati and Alberto Roverato

The Economic behaviour of Sustainability

Airports performances and sustainable practices. An empirical study on Italian data 110
Riccardo Gianluigi Serio, Maria Michela Dickson, Diego Giuliani and Giuseppe Espa

Sustainability: still an undefined concept for Italians 116
Raffaele Angelone and Andrea Marletta

Quasi-experimental evidence on COVID-19 lockdown effects on Italian household food shopping basket composition and its sustainability 122
Beatrice Biondi and Mario Mazzocchi

Advances in statistical methods for complex problems

Inferring multiple treatment effects from observational studies using confounder importance learning n.a.
Omiros Papaspiliopoulos

Path analysis in Ising models: an application to cyber-security risk assessment 127
Monia Lupparelli and Giovanni M. Marchetti

Causal Regularization n.a.
Lucas Kania and Ernst Wit

Explainable machine learning models

Enhancing Markowitz model: inspection of correlations and tail covariances 133
Gloria Polinesi

Objective and subjective dimension of economic well-being: an approach based on statistical matching 139
Daniela Marella, Vincenzina Vitale and Pierpaolo D'Urso

Sustainable, Accurate, Fair and Explainable Machine Learning Models n.a.
Paolo Giudici and Emanuela Raffinetti

Flexible Learning for Environmental Sustainability

Comparison of traffic flow data sources for air pollution modelling 145
Theresa Smith and Nick McCullen

Data analysis of photogrammetry-based mapping: the sea cucumbers in the Giglio Island as a case-study 150
Gianluca Mastrantonio, Daniele Ventura, Edoardo Casoli, Arnold Rakaj, Giovanna Jona Lasinio and Alessio Pollice

Understanding forest damage in Germany: Finding key drivers to help with future forest conversion of climate sensitive 156
Nicole Augustin, Heike Puhlmann and Simon Trust

Inequalities in higher education outcomes: learning from data

Inequalities in international students mobility 163
Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino

Uncovering the interplay of territorial, socioeconomic, and demographic factors in high school to university transition 169
Vincenzo Giuseppe Genova, Andrea Priulla and Martina Vittorietti

Statistical Learning of demographic and health dynamics

Estimating the impact of a vaccine mandate: the case of measles in Italy n.a.
Chiara Chiavenna

Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth n.a.
Andrea Nigri

Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data n.a.
Claire Dooley, Ridhi Kashyap, Douglas Leasure and Francesco Rampazzo

Challenges towards Fairness and Transparency for Data Processes, Algorithms and Decision-Support Models

Challenges on Ethics, and Privacy in AI Applications to Fintech 175
Catarina Silva, Joana Matos Dias and Bernardete Ribeiro

Uncertainty and fairness metrics 180
Anna Gottard

Educational Data mining: methods for complex data in students' assessment

Analysis of University Grades: An IRT Model for Responses and Response Times with Censoring 186
Michela Battauz

Predicting high schools' students performances with registry's data: a machine learning approach 191
Lidia Rossi, Marta Cannistrà and Tommaso Agasisti

Using response times to identify cheaters in CAT: A simulation study 195
Luca Bungaro, Bernard P. Veldkamp and Mariagiulia Matteucci

Spatial and Spatio-Temporal Modeling: Theory and Applications

A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy 200
Paolo Maranzano, Kelly McConville , Philipp Otto and Felicetta Carillo

Bayesian multi-species N-mixture models for large scale spatial data in community ecology 206
Michele Peruzzi

Minimum contrast for point processes' first-order intensity estimation 212
Nicoletta D'Angelo and Giada Adelfio

Statistical Framework for Measuring the Sustainability of Tourism

Data validity and statistical conformity with Benford's Law: the case of tourism in Sicily 217
Roy Cerqueti and Davide Provenzano

Exploring the level of digitalization of the Italian museums through a multilevel ordered logit model 228
Claudia Cappello, Sabrina Maggio and Sandra De Iaco

Functional Partial Least-Squares via Regression Splines. An application on Italian Sustainable Development Goals data 232
Ida Camminatiello, Rosaria Lombardo, Jean-Francois Durand and Leonardo S. Alaimo

Statistical learning for well-being analysis

Assessing multidimensional poverty of the Italian provinces during Covid-19: a small area estimation approach 238
Mariateresa Ciommi, Chiara Gigliarano, Francesca Mariani and Gloria Polinesi

The fuzzy set approach as statistical learning for the analysis of multidimensional well-being 244
Gianni Betti, Federico Crescenzi, Antonella D'Agostino and Laura Neri

What Makes a Satisfying Life? Prediction and Interpretation with Machine-Learning Algorithms n.a.
Conchita D'Ambrosio

Bayesian contributions to Statistical Learning

A Bayesian framework for early cancer screening 249
Sally Paganin and Jeff Miller

Imputing Synthetic Pseudo Data from Aggregate Data: Development and Validation for Precision Medicine n.a.
Cecilia Balocchi

Linear models with assumptions-free residuals: a Bayesian Nonparametric approach 254
Filippo Ascolani and Valentina Ghidini

Data Visualization for Smart Insights and Advanced Predictive Analytics

Applications of data visualization for industry 259
Martina Dossi, Stefano Sangaletti, Marilena Di Bari and Federica Bruschini

Some Notes on the Use of the Circular Boxplot n.a.
Giovanni Camillo Porzio and Davide Buttarazzi

TERRA: a smart visualization tool for international trade in goods statistics 265
Francesco Amato, Mauro Bruno and Maria Serena Causo

Methods for the analysis of distributional data

Clustering of Distributional Data based on LDQ transformation 271
Gianmarco Borrata and Rosanna Verde

Dynamic learning from data streams through the combined use of probability density functions and simplicial functional principal component analysis 276
Francesca Fortuna, Fabrizio Maturo and Tonio Di Battista

Multivariate Parametric Analysis of Distributional Data n.a.
Paula Brito

Migrants and Refugees in Europe: social, economic and health-related issues

Labor Market Return to Refugees' Human Capital Investment: A Natural Experiment in Sweden n.a.
Eleonora Mussino

Social networks and loneliness among older migrants in Italy 282
Viviana Amati, Eralba Cela and Elisa Barbiano di Belgiojoso

The Italian Decree on Security: An Analysis of the Impact on Asylum Applications 287
Giorgio Piccitto

Modelling and Forecasting High-dimensional time series

Adaptive combinations of tail-risk forecasts 293
Alessandra Amendola, Vincenzo Candila, Antonio Naimoli and Giuseppe Storti

Are Monetary Policy Announcements related to Volatility Jumps? 299
Giampiero Gallo, Demetrio Lacava and Edoardo Otranto

Regularized Estimation and Prediction of the El Nino/Southern Oscillation Cycle n.a.
Alessandro Giovannelli and Tommaso Proietti

3 Contributed Sessions 305

Bayesian nonparametric methods

Bayesian density estimation for modeling age-at-death distribution 306
Davide Agnoletto, Tommaso Rigon and Bruno Scarpa

Bayesian mixing distribution estimation in the Gaussian-smoothed 1-Wasserstein distance 311
Catia Scricciolo

Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models 316
Francesco Denti, Antonio Di Noia and Antonietta Mira

Bayesian nonparametric multiple change point detection for time series of compositional data 322
Edoardo Marchionni and Riccardo Corradin

Galton-Watson process: a non parametric prior for the offspring distribution 328
Massimo Cannas, Michele Guindani and Nicola Piras

Hierarchical processes in survival analysis 333
Riccardo Cogo, Federico Camerlenghi and Tommaso Rigon

Economics and Statistics

A regression analysis for count data to investigate the effectiveness of incentives on the adoption of 4.0 technologies 339
Stefano Bonnini and Michela Borghesi

Statistical analysis on SDGs indicators related to environmental sustainability 344
Najada Firza, Anisa Bakiu and Dante Mazzitelli

Empowering futures adopting a spatial convergence of opinions: a Real-Time Spatial Delphi approach 349
Yuri Calleo, Simone Di Zio and Francesco Pilla

Stocks price forecasts using Stochastic Differential Equations: an empirical assessment 355
Dario Frisardi and Matteo Spuri

The Added-Worker Effect within Italian Households 361
Donata Favaro and Anna Giraldo

Health statistics 1

A model for the natural history of breast cancer: application to a Norwegian screening dataset 365
Laura Bondi, Marco Bonetti and Solveig Hofvind

Generalized Bayesian Ensemble Survival Trees: an extension to categorical variables to apply it to real data 370
Elena Ballante

Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach 375
Chiara Masci, Marta Spreafico and Francesca Ieva

Mobility trends in Italy during the first wave of Covid-19 pandemic: analysis on Google data 381
Ilaria Bombelli and Daniele De Rocchi

Tracking attitudes towards COVID vaccines: A text mining analysis 387
Leonardo Scarso, Marco Novelli and Francesco Saverio Violante

Treatment effect assessment in observational studies with multi-level treatment and outcome 393
Federica Cugnata, Paola Vicard, Paola M.V. Rancoita, Fulvia Mecatti, Clelia Di Serio and Pier Luigi Conti

Indicators: composition, uses and limitations

Are European consumers willing to pay the true price for sustainable food? 399
Luca Secondi and Mengting Yu

Can the reliability of composite indexes be impacted by uncertainty of individual indicators? 406
Caterina Giusti, Stefano Marchetti and Vincenzo Mauro

Initial Coin Offerings and ESG: allies or enemies? 411
Alessandro Bitetto and Paola Cerchiello

On the impact of intraclass correlation in the ANVUR evaluation of academic departments 417
Giorgio Edoardo Montanari and Marco Doretti

Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes 422
Luigi Biggeri, Stefano Marchetti, Caterina Giusti, Monica Pratesi, Francesco Schirripa Spagnolo and Gaia Bertarelli

Smart Composite Indicators Measuring Corporate Sustainability: A Sensitivity Analysis 428
Camilla Salvatore, Annamaria Bianchi and Silvia Biffignandi

Multivariate data analysis 1

A note on most powerful tests for right censored survival data 434
Maria Veronica Vinattieri and Marco Bonetti

Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data 439
Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore

Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data 445
Farah Naz and Elena Ballante

ROBOUT: a multi-step methodology for conditional outlier detection 450
Matteo Farnè and Angelos Vouldis

Robustness of the Efficient Covariate-Adaptive Design for balancing covariates in comparative experiments 456
Rosamarie Frieri, Alessandro Baldi Antognini, Maroussa Zagoraiou, and Marco Novelli

Separation scores: a new statistical tool for scoring and ranking partially ordered data 462
Marco Fattore

Statistics in Society 1

Community detection analysis with robin on hashtag network 468
Valeria Policastro, Francesco Santelli and Giancarlo Ragozini

Film Tourism Motivation through the lens of Trip Advisor data 474
Nicolò Biasetton, Marta Disegna, Girish Prayag and Elena Barzizza

Life satisfaction and social activities in later life in Italy: a focus on the Internet use 480
Claudia Furlan and Silvia Meggiolaro

Social capital endowment's role in the intergenerational transmission of education 485
Alessandra Trimarchi, Maria Gabriella Campolo and Antonino Di Pino Incognito

Streaming Data from Social Networks to Track Political Trends 490
Emiliano del Gobbo and Barbara Cafarelli

The scientific production on gender dysphoria: a bibliometric analysis 495
Maria Gabriella Grassia, Marina Marino, Massimo Aria, Rocco Mazza,
Luca D'Aniello and Agostino Stavolo

Assessment and Education

A hierarchical modelling approach to explain differential functioning of mathematics items by student's gender 500
Clelia Cascella

A latent variable approach to Millennials' knowledge of green finance 506
Maria Iannario, Alessandra Tanda and Claudia Tarantola

Archetypal analysis and latent Markov models: A step-wise approach 512
Lucio Palazzo, Rosa Fabbricatore and Francesco Palumbo

From high school to university: academic intentions and enrolment of foreign students in Italy 518
Francesca Di Patrizio, Eleonora Trappolini and Cristina Giudici

Growth models for the progress test in Italian dentistry degree program 523
Giulio Biscardi, Leonardo Grilli, Carla Rampichini, Laura Antonucci and Corrado Crocetta

The COVID-19 pandemic and academic E-learning: Italian students and instructors' perceptions 527
Francesco Santelli, Teresa Gentile, Davide Bizjak and Lorenzo Fattori

Working Students and job market outcomes: Insights from the University of Florence 532
Gabriele Lombardi, Valentina Tocchioni and Alessandra Petrucci

Bayesian methods and applications 1

Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation 538
Elena Sabbioni, Enrico Bibbona, Gianluca Mastrantonio and Guido Sanguinetti

Approximate Bayesian Computation for Probabilistic Damage Identification 544
Cecilia Viscardi, Silvia Monchetti, Luisa Collodi, Gianni Bartoli, Michele Betti, Michele Boreale and Fabio Corradi

Estimation of scientific productivity with a hierarchical Bayesian model 550
Maura Mezzetti and Ilia Negri

Heat waves and free-knots splines 555
Gioia Di Credico and Francesco Pauli

The Hierarchical Beta-Bernoulli Process as Out-of-Scope Query Detector 560
Marco Dalla Pria and Silvia Montagna

Health and mortality

A novel definition of comorbidity based on the Global Burden of Diseases project weights 566
Angela Andreella, Lorenzo Monasta and Stefano Campostrini

An Age-Period-Cohort model of gender gap in youth mortality 572
Giacomo Lanfiuti Baldi and Andrea Nigri

Kinlessness in adult and old age across Europe 578
Marta Pittavino, Bruno Arpino and Elena Pirani

Parameter orthogonalization for Siler mortality model 584
Claudia Di Caterina and Lucia Zanotto

Pseudo-observations in survival analysis 590
Marta Cipriani, Alfonso Piciocchi, Valentina Arena and Marco Alfò

Sex Gap in Cancer-Free Life Expectancy: The Association with Smoking, Obesity and Physical Inactivity 595
Alessandro Feraldi, Cristina Giudici and Nicolas Brouard

Women's Exposure to HIV in Africa: the Role of Intimate Partner Violence 599
Micaela Arcaio and Anna Maria Parroco

Mixture Models

An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks 605
Dalila Failli, Maria Francesca Marino and Francesca Martella

Constrained Mixtures of Generalized Normal Distributions 611
Pierdomenico Duttilo, Alfred Kume and Stefano Antonio Gattone

Mixture-based clustering with covariates for ordinal responses 617
Kemmawadee Preedalikit, Daniel Fernàndez, Ivy Liuc, Louise McMillan, Marta Nai Ruscone and Roy Costilla

Partial membership models for soft clustering of multivariate count data 623
Emiliano Seri, Thomas Brendan Murphy and Roberto Rocci

Regression for mixture models for extremes 629
Viviana Carcaiso, Ilaria Prosdocimi and Isadora Antoniano-Villalobos

Robust matrix-variate mixtures of regressions 635
Salvatore Daniele Tomarchio and Michael P. B. Gallaugher

Sampling methods and analysis of survey data

On the use of auxiliary information to define the sampling design for large-scale geospatial data 641
Chiara Bocci and Emilia Rocco

Optimal joint inclusion probabilities for spatial sampling n.a.
Giuseppe Arbia, Piero Demetrio Falorsi and Vincenzo Nardelli

Robustness and Balance of Sampling or Experimental Designs and Mixture of Designs 647
Yves Tillé and Ejub Talovic

Robustness Bounds for Sampling and Experimental Designs 654
Ejub Talovic and Yves Tillé

Statistical Matching: Hotdeck or Propensity Score? 661
Elena Dalla Chiara, Marcello D'Orazio and Federico Perali

The Italian experience on register-based statistics considering measurement, coverage and sampling errors 667
Marco Di Zio, Romina Filippini and Simona Toti

Space-time statistics

A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis 673
Hiu Ching Yip, Gianluca Mastrantonio, Enrico Bibbona, Daria Valente and Marco Gamba

An approach to cluster time series extremes with spatial constraints 679
Alessia Benevento, Fabrizio Durante and Roberta Pappadà

An integrated space-time model to evaluate the innovation drivers in Italy 685
Emma Bruno, Rosalia Castellano and Gennaro Punzo

Revealing the dynamic relations between traffic and crowding using big data from mobile phone network 691
Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

SMaC: Spatial Matrix Completion method 697
Giulio Grossi, Alessandra Mattei and Georgia Papadogeorgou

The impact of traffic flow and road signs on road accidents: an approach based on spatiotemporal point pattern analysis on linear networks 702
Andrea Gilardi and Riccardo Borgoni

Clustering and classification 1

A clustering model for flow data: an application to international student mobility 708
Cinzia Di Nuzzo and Donatella Vicari

Contingency tables with structural zeros and discrete copulas 713
Roberto Fontana, Elisa Perrone and Fabio Rapallo

Levels Merging in the Latent Class Model 719
Christophe Biernacki

Model-based clustering of count processes with multiple change 725
Shuchismita Sarkar and Xuwen Zhu

Similarity Measures and Internal Evaluation Criteria in Hierarchical Clustering of Categorical Data 729
Jana Cibulková, Zdeněk Šulc, Hana Řezanková and Jaroslav Horníček

Spectral clustering of mixed data via association-based distance 735
Alfonso Iodice D'Enza, Francesco Palumbo and Cristina Tortora

Dynamic models and time series

A graph based convolution Neural Network approach for forecast reconciliation 741
Andrea Marcocchia and Pierpaolo Brutti

A multivariate hidden semi-Markov model for the analysis of multiple air pollutants 747
Marco Mingione, Pierfrancesco Alaimo Di Loro, Francesco Lagona and Antonello Maruotti

A smooth transition autoregressive model for matrix-variate time series 753
Andrea Bucci

Dynamic network models with time-varying nodes 759
Luca Gherardini, Mauro Bernardi and Monia Lupparelli

Time lapse analysis of nuclear calcium spiking in plant cells during symbiotic signaling 765
Ivan Sciascia, Andrea Crosino and Andrea Genre

Two-stage weighted least squares estimator of multivariate conditional mean observation-driven time series models 770
Mirko Armillotta

Environmental learning and indicators

Assessing the performance of nuclear norm-based matrix completion methods on CO_2 emissions data 776
Rodolfo Metulini, Francesco Biancalani, Giorgio Gnecco and Massimo Riccaboni

Deep Learning for smart and sustainable agriculture 782
Amalia Vanacore, Armando Ciardiello, Annalisa Izzo, Pierdomenico Zaffino, Carolina Vecchio, Gennaro Pio Auricchio and Luigi Uccelli

Do green transition, environmental taxes and renew-able energy promote ecological sustainability in G7 countries? Evidence from panel quantile regression 788
Aamir Javed, Agnese Rapposelli and Asif Javed

Doubly Robust DID for National Parks evaluation: “just” environmental benefits, or socioeconomics impacts as well? 795
Riccardo D’Alberto, Francesco Pagliacci and Matteo Zavalloni

On the gap between emitted and absorbed carbon dioxide. Are trees enough to save us? 801
Lorenzo Mori and Maria Rosaria Ferrante

Small scale analysis of energy vulnerability in the municipality of Palermo 806
Giuliana La Mantia

Health statistics 2

A test for non-differential misclassification error in database epidemiological studies 812
Giorgio Limoncella, Leonardo Grilli, Emanuela Dreassi, Carla Rampichini, Robert Platt and Rosa Gini

Is the COVID-19 ‘color code’ of Italian regions subjected to political manipulation? 816
Giovanni Busetta and Fabio Fiorillo

Modelling multilevel ordinal response under endogeneity: application to DTC patients’ outcome 822
Silvia D’Elia

Monitoring drugs-based diagnostic therapeutic paths in heart failure patients using state-sequence analysis techniques 827
Nicole Fontana, Laura Savaré and Francesca Ieva

Optimal two-stage design based on error rates under a Bayesian perspective 833
Susanna Gentile and Valeria Sambucini

Migrants in Italy and return migration

Comparing migrant and “native” Italian adolescents in risky behaviours from FSS and SHARE Corona surveys n.a.
Daniela Foresta

EU-Border crisis on Twitter: sentiments and misinformation analysis 839
Elena Ambrosetti, Cecilia Fortunato and Sara Miccoli

Graduates' interregional migration in times of crisis: the Italian case 843
Thaís García-Pereiro, Ivano Dileo and Anna Paterno

Intentions to stay: The experience of return migrants in Albania 848
Maria Carella, Thaís García-Pereiro, Roberta Pace and Anna Paterno

Return migration to home country: a systematic literature review with text mining and topic modelling 853
Cecilia Fortunato, Andrea Iacobucci and Elena Ambrosetti

The allocation of time within native and foreign couples living in Italy 860
Giovanni Busetta, Maria Gabriella Campolo and Antonino Di Pino Incognito

Εἰλείϑυια comes from afar: The foreigners’ contribution to fertility by Italian provinces 866
Eleonora Miaci, Cristina Giudici, Eleonora Trappolini, Marina Attili, Cinzia Castagnaro and Antonella Guarneri

Sustainability assessment

ESG, sustainability and stock market risk 871
Michele Costa

Exploring the effect of consumer motivation and perception of sustainability on food choices with a Discrete Choice Experiment 875
Gloria Solano-Hermosilla, Jesus Barreiro-Hurle and Ilaria Amerise

Sustainability explained by ChatGPT artificial intelligence in a HITL perspective: innovative approaches 881
Vito Santarcangelo, Angelo Lamacchia, Emilio Massa, Saverio Gianluca Crisafulli, Massimiliano Giacalone and Vincenzo Basile

Measuring economic and ecological efficiency of urban waste systems in Italy: a comparison of SFA and DEA techniques 887
Massimo Gastaldi, Ginevra Virginia Lombardi, Agnese Rapposelli and Giulia Romano

Profile based latent distance association analysis for sparse tables. Application to the attitude of EU citizens towards sustainable tourism 893
Francesca Bassi, Josè Fernando Vera and Juan Antonio Marmolejo Martìn

Sustainable tourism: a survey on the propensity towards eco-friendly accommodations 899
Claudia Furlan and Giovanni Finocchiaro

Bayesian methods and applications 2

A comparison of computational approaches for posterior inference in Bayesian Poisson regression 903
Laura D’Angelo

Bias-reduction methods for Poisson regression models 908
Luca Presicce, Tommaso Rigon and Emanuele Aliverti

Finite Mixture Model for Multiple Sample Data 913
Alessandro Colombi, Raffaele Argiento, Federico Camerlenghi and Lucia Paci

On Bayesian power analysis in reliability 918
Fulvio De Santis, Stefania Gubbiotti and Francesco Mariani

Power priors elicitation through Bayes factors 923
Roberto Macrì Demartino, Leonardo Egidi and Nicola Torelli

Predictive Bayes factors 929
Leonardo Egidi and Ioannis Ntzoufras

Clustering and classification 2

A Clusterwise Regression Method for Distributional-Valued Data 935
Antonio Balzanella, Rosanna Verde and Francisco de A.T. de Carvalho

A novel statistical-significance based semi-parametric GLMM for clustering countries standing on their innumeracy levels 939
Alessandra Ragni, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Introducing a novel directional distribution depth function for supervised classification 945
Edoardo Redivo and Cinzia Viroli

Clustering alternatives in the preference-approval context 950
Alessandro Albano, José Luis Garcia-Lapresta , Mariangela Sciandra and Antonella Plaia

Computational assessment of k-means clustering on a Structural Equation Model based index 955
Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Handling missing data in complex phenomena: an ultrametric model-based approach for clustering 961
Francesca Greselin and Giorgia Zaccaria

Economics and labour markets

A multivariate ranking analysis on the employability of young adults 967
Rosa Arboretti, Elena Barzizza, Nicolo Biasetton, Riccardo Ceccato, Monica Fedeli and Concetta Tino

Analysis of the Gender Pay Gap in the Italian Labour Market 973
Giulia Cappelletti and Daniele Toninelli

Evaluating the effect of home-based working employing causal Bayesian networks and potential outcomes 979
Lorenzo Giammei

Patterns of flexible employment careers. Does measurement error matter? 985
Mauricio Garnier-Villarreal, Dimitris Pavlopoulos and Roberta Varriale

Staying or leaving? A nonlinear framework to explore the role of employee well-being on retention 991
Ulpiani Kocollari, Fabio Demaria and Maddalena Cavicchioli

The CAP instruments impact on GVA and employment: a multivalued treatment approach 997
Montezuma Dumangane and Marzia Freo

The determinants of leaving the parental home in Italy: 2012-18 1003
Ilaria Rocco and Gianpiero Dalla Zuanna

Environmental modeling

A Bayesian weather-driven spatio-temporal model for PM10 in Lombardy 1109
Michela Frigeri, Alessandra Guglielmi and Giovanni Lonati

A preliminary study on shape descriptors for the characterization of microplastics ingested by fish 1015
Greta Panunzi, Tommaso Valente, Marco Matiddi and Giovanna Jona Lasinio

Artificial neural network in predicting odour concentrations: a case study 1021
Veronica Distefano and Gideon Mazuruse

Bayesian analysis of PM10 concentration by spatio-temporal ARIMA and STS models 1026
Michela Frigeri and Ilenia Epifani

Functional ANOVA to monitor yearly Adriatic sea temperature variations 1032
Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Nicola Di Deo

New perspectives in the measurement of biodiversity 1038
Linda Altieri, Daniela Cocchi and Massimo Ventrucci

Multivariate data analysis 2

Feature Selection via anomaly detection autoencoders in radiogenomics studies 1044
Alessia Mapelli, Michela Carlotta Massi, Nicola Rares Franco, Francesca Ieva, Catharine West, Petra Seibold, Jenny Chang-Claude and the REQUITE and RADprecise Consortia

Further considerations on the Spectral Information Criterion 1050
Luca Martino

How to increase the power of the test in sparse contingency tables: a simulation study 1057
Federica Nicolussi and Manuela Cazzaro

Latent event history models for quasi-reaction systems 1063
Matteo Framba, VeronicaVinciotti and Ernst Wit

Quantile-based graphical models for continuous and discrete variables 1069
Luca Merlo, Marco Geraci and Lea Petrella

The logratio Student t distribution 1075
Gianna Monti and Gloria Mateu-Figueras

Statistics in Society 2

A decomposition of the changes in tourism demand in Tuscany over the 2019-2021 period 1079
Mauro Mussini

Bayesian networks as a territorial gender impact assessment tool 1084
Flaminia Musella, Lorenzo Giammei, Fulvia Mecatti and Paola Vicar

Can statistics be helpful in detecting electoral fraud? 1088
Massimo Attanasio, Vincenzo G. Genova and Michele Tumminello

Companies' sustainability disclosure and contrast to hunger: the role of social inclusion 1093
Chiara Di Maria and Rodolfo Damiano

Passing network-based performance indicator in football: evidence from UEFA Champions League 2016-2017 1099
Riccardo Ievoli, Lucio Palazzo and Giancarlo Ragozini

Topic Modeling for the travel and tourism industry: classical and innovative methods compared 1105
Fabrizio Di Mari

Bayesian methods and applications 3

An Importance Sampling Algorithm For Bayesian Logistic Regression with Independent Gaussian Scale Mixture Prior 1111
Paolo Onorati and Brunero Liseo

Bayesian analysis of Amazon's best-selling books via finite nested mixture model 1117
Laura D'Angelo and Francesco Denti

Binomial Extended Stochastic Block Model for Brain Networks 1121
Valentina Ghidini, Sirio Legramanti and Raffaele Argiento

Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures 1127
Giulia Capitoli, Simone Colombara, Alessia Cotroneo, Francesco De Caro, Riccardo Morandi, Chiara Schembri, Alfredo G. Zapiola and Francesco Denti

Efficient expectation propagation for high-dimensional probit models 1133
Augusto Fasano, Niccolo Anceschi, Beatrice Franzolini and Giovanni Rebaudo

Model-based clustering of non-stationary time series with common historical change times 1139
Riccardo Corradin, Luca Danese, Wasiur KhudaBukhsh and Andrea Ongaro

Functional Data Analysis

A functional Ground Motion Model for Italy built with a weighted analysis of reconstructed seismic curves 1145
Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio

Conditional Gaussian Graphical Models for Functional Variables whit Partial Separable Operators 1149
Rita Fici, Gianluca Sottile and Luigi Augugliaro

Does the Inflation Factor need tuning? Simulation-based adjustment for Outlier Detection via the Functional Boxplot 1155
Annachiara Rossi, Andrea Cappozzo and Francesca Ieva

Functional Graphical Models to map Brexit debate on Twitter 1160
Nicola Pronello, Emiliano del Gobbo, Lara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti and Sara Fontanella

Measuring Dependence in Multivariate Functional Datasets 1166
Francesca Ieva, Michael Ronzulli and Anna Maria Paganoni

Robust Statistical Process Monitoring of Multivariate Functional Data 1173
Christian Capezza, Fabio Centofanti, Antonio Lepore and Biagio Palumbo

The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021 1179
Veronica Mazzola, Giovanni Bonaccorsi, Piercesare Secchi and Francesca Ieva

Machine Learning and text mining

A vocabulary-based approach for risk detection in textual annotations of contracts of public procurement 1185
Giulio Giacomo Cantone, Simone Del Sarto and Michela Gnaldi

Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study 1191
Giovanni Bocchi, Alessandra Micheletti, Patrizio Frosini, Alessandro Pedretti, Andrea R. Beccari, Filippo Lunghini, Carmine Talarico and Carmen Gratteri

Hedging global currency risk with factorial machine learning models 1197
Paolo Pagnottoni and Alessandro Spelta

InstanceSHAP: An instance-based estimation approach for Shapley values 1203
Golnoosh Babaei and Paolo Giudici

Networks & Nature Based Solutions: an application for Milan hydric resources 1209
Alessia Forciniti and Emma Zavarrone

The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization 1215
Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Agostino Stavolo

Multivariate data analysis 3

A comparison of different techniques for handling missing covariate values in propensity score methods 1219
Anna Zanovello, Alessandra R. Brazzale and Omar Paccagnella

A New Penalized Estimator for Sparse Inference in Gaussian Graphical Models: An Adaptive Non-Convex Approach 1224
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

A tool for assessing weak identifiability of statistical models 1230
Antonio Di Noia, Francesco Denti and Antonietta Mira

Computing Highest Density Regions with Copulae 1235
Nina Deliu and Brunero Liseo

Parameter estimation via Indirect Inference for multivariate Wrapped Normal distributions 1241
Francesca Labanca and Anna Gottard

Sequential marginal likelihood selection for the estimation of sparse correlation matrices 1246
Claudia Di Caterina and Davide Ferrari

Nonparametric statistical methods

A Comparison of Distribution-Free Control Charts 1252
Michele Scagliarini

Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida 1257
Dafne Zorzetto

Comparing three robust procedures for CANDECOMP/PARAFAC estimation 1262
Valentin Todorov, Violetta Simonacci, Michele Gallo and Nikolay Trendafilov

How active is a genetic pathway? Comparative analysis of post-hoc permutation-based methods 1268
Anna Vesely and Angela Andreella

Non Parametric Combination methodology: a literature review on recent developments 1274
Elena Barzizza, Nicolò Biasetton and Riccardo Ceccato

Regression modeling

A Quantile Regression Model to Evaluate the Performance of the Italian Courts of Law 1280
Carlo Cusatelli, Massimiliano Giacalone and Eugenia Nissi

A variable selection procedure based on predictive ability: a preliminary study on logistic regression 1285
Rosaria Simone and Mariarosaria Coppola

Comparison of binary regressions with asymmetric link function for imbalanced data 1291
Michele La Rocca, Marcella Niglio and Marialuisa Restaino

New advances in Regression Forests 1297
Mila Andreani, Lea Petrella and Nicola Salvati

On the Optimal Non-Convexity of Penalty in Sparse Regression Models 1303
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

Using expectile regression with latent variables for digital assets 1309
Beatrice Foroni, Luca Merlo and Lea Petrella

4 Program 1315

Regression for mixture models for extremes

Viviana Carcaiso[a], Ilaria Prosdocimi[b], and Isadora Antoniano-Villalobos[b]

[a]Department of Statistical Sciences, University of Padova, Italy;
`viviana.carcaiso@phd.unipd.it`
[b]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice,
Italy; `ilaria.prosdocimi@unive.it`, `isadora.antoniano@unive.it`

Abstract

In many practical problems the assumption that the data are generated by a single process does not hold. Mixture models can deal with this issue and represent a useful resource also in the context of extreme values. A typical scenario with extreme events is that there are two processes underlying the data: one that occurs with a much higher frequency and another which takes place more rarely but can lead to stronger magnitudes. However, simulation studies and applications to real data show that the rare type may not always correspond to the most extreme events and that the tails of the type-specific data-generating processes are not always identified by this information. Therefore, we aim at creating a new regression model that exploits the type of event as a covariate, together with other variables of interest, such as spatial characteristics.

Keywords: extreme value theory, finite mixtures, environmental problems

1. Introduction

The typical assumption that the data are representative of a unique population of interest is often too restrictive in real world problems. This constitutes an issue that is interesting to tackle in the extreme value framework (1), where the focus is not as usual on the behaviour of the central part of the distribution of the process of interest, but on the tail behaviour, that corresponds to the rarely observed events. Extreme value theory aims indeed at quantifying the stochastic behaviour of a process at extremely large (or small) values on the basis of asymptotic results and provides approaches that are specifically designed for the analysis of this kind of data. In this context the observations, which consist in maximum values of the measured process in a block, are usually assumed to be i.i.d. draws from an appropriate long-tailed distribution, typically the Generalized Extreme Value (GEV) distribution, which is a limiting distribution supported by theoretical results, but other distributions have been proposed and exploited.

Examples of rare events where the data are actually generated by at least two distinct processes are rainfalls generated by typhoon and non-typhoon weather systems, floods originating from a mixture of rainfall and snowmelt or, in a non-environmental context, stock prices bursts linked to the impact of different economic cycles. In order to take into account the specific nature of the data and exploit appropriate methods to deal with variables whose distribution is determined by multiple components, it is possible to exploit finite mixture models (2; 3). These models are used to describe the data as being drawn from a density which is modelled as a convex combination of components, each with a specified parametric form. Hence, they typically involve more parameters than single population models, which results in a more complex estimation in favor of a gain in flexibility.

The research is driven by the application to hydrological area, where due to natural hazards the estimation of the frequency of the extreme events is crucial. In the literature there are numerous studies on the application of mixture distributions for analysing hydrological extremes. The first contribution in this direction is the one of Rossi et al. (5), who defined the two-component extreme value (TCEV) distribution, which assumes the individual floods to arise from a mixture of two Exponential distributions and the number of events in a year to be generated from two Poisson distributions. Hence they specified a four-parameter distribution, that can be considered a generalization of the Gumbel distribution, which is a special case of the GEV distribution obtained by setting the shape parameter equal to 0. In particular, the TCEV distribution has CDF which is exactly the product of two Gumbel CDF's. Among others, Kjeldsen et al. (4) proposed instead a Gumbel mixture distribution for modelling extreme events from two different phenomena. Considering the independent random variables X_1 and X_2 that correspond to the annual maxima of two different processes, the CDF of the annual maximum X is expressed in terms of conditional distributions as

$$F_X(x) = G_1(x)(1-\omega) + G_2(x)\omega,$$

where ω is the probability that the annual maximum value is a result of the process of type 2 and G_1 and G_2 are conditional distributions of the events of type 1 and type 2, respectively, that are two-parameter Gumbel distributions. This model, unlike the TCEV one, allows to deal with scenarios in which events of type 2 do not occur every year. It also exploits more information than the TCEV one, by taking into account the label that identifies the type of event associated to the annual maximum, but there is, however, one additional parameter to estimate. Furthermore, knowing a priori the population originating each event is often difficult in practice. Even when the data are labelled with the type of event, it is not always the case that these labels actually identify the various subgroups in the tail population. Indeed, the events that happen infrequently are not always the most extreme. In this research we aim at using finite mixture models incorporating the labels as a covariate of the model rather than a deterministic identifier.

2. Models for multi-component extremes

We are interested in finite mixture models for the analysis of series of maximum events originated from multiple populations, focusing on a flexible description of the tail of the distribution. We begin from models which assume that the labels are a known deterministic identifier, and extend the idea to allow for noisy labels.

2.1 Finite mixture models for fixed categories

Without loss of generality, we consider data from two different processes. In the TCEV model (5) the CDF of the annual maximum is written as the product of two Gumbel CDF's: one with location parameter $\mu = \theta_1 \log \lambda_1$ and scale parameter $\sigma = \theta_1$, and the other with location parameter $\mu = \theta_2 \log \lambda_2$ and scale parameter $\sigma = \theta_2$. Table 1 shows the average proportion of type 2 events in the 10 most extreme values in 500 samples of 1000 simulations from a TCEV model with different choices of ratios of parameters, accounting for the fact that by definition of the model $\lambda_1 > \lambda_2$. It is not straightforward to understand how the combination of parameters controls the prevalence in the tail, and it is difficult also

	$\lambda_1/\lambda_2 = 1.5$	$\lambda_1/\lambda_2 = 2$	$\lambda_1/\lambda_2 = 5$
$\theta_2/\theta_1 = 0.5$	0.0018 (0.2213)	0.0012 (0.1565)	0.0000 (0.0357)
$\theta_2/\theta_1 = 1$	0.4042 (0.3969)	0.3244 (0.3321)	0.1732 (0.1664)
$\theta_2/\theta_1 = 1.5$	0.9090 (0.5135)	0.8794 (0.4591)	0.7624 (0.2999)
$\theta_2/\theta_1 = 2$	0.9924 (0.5902)	0.9892 (0.5450)	0.9722 (0.4055)

Table 1: Average proportion of events of type 2 in the 10 most extreme ones obtained from 500 series of 1000 simulations of the TCEV model with the corresponding ratios of location and scale parameters. Average proportion of type 2 events in the whole sample in parenthesis. For reference $\lambda_2 = 2$ always and $\theta_1 = 1$ everywhere except in the first row, where it is 2.

to acknowledge the role in the parameters in determining the total proportion of type 2 events (displayed in parenthesis in the table). Indeed, for some choices of ratios, type 2 events are not the rarest ones, but are still the strongest in terms of magnitude.

The Mixture Gumbel model (4) defines the CDF of the annual maximum as

$$F_X(x) = (1-\omega)\exp\left\{-\exp\left[-\frac{x-\mu_1}{\sigma_1}\right]\right\} + \omega\exp\left\{-\exp\left[-\frac{x-\mu_2}{\sigma_2}\right]\right\}, \tag{1}$$

with $\omega < 0.5$ since, as in (5), events of type 1 are expected to occur more frequently. The parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$ are the location and scale parameters of the two Gumbel distributions of events from process 1 and 2, respectively. Figure 1 shows how the distribution of 1000 simulated data from this model (with $\omega = 0.2$) changes with different ratios between the location parameters and the scale parameters. The location parameter μ_1 is kept equal to 10, while the scale parameter σ_1 is put equal to 3, except in the case $\mu_2/\mu_1 = 1.5$ and $\sigma_2/\sigma_1 = 0.5$, when it is 4, and in the case $\mu_2/\mu_1 = 2$ and $\sigma_2/\sigma_1 = 0.5$, with $\sigma_1 = 5$. Every histogram displays the distribution of the simulated mixture model with corresponding choices of the parameters, and the proportion of density in each bin that comes from each process is identified by using two different colours, light for process 1 and dark for process 2.

It is possible to recognise that the conditional distribution dominating the tail of the mixture can change depending on the parameter combinations: with some choices of the parameters the rare events (type 2) do not prevail in the right tail and frequent events (type 1) also can correspond to the most extreme values. This is also noticeable in Table 2, which displays the proportion of type 2 events in the 10 most extreme values among 1000 simulations (top 1%) from the Mixture Gumbel model with the same ratios of scale and location parameters as Figure 1. As expected type 2 events are the majority when both μ_2/μ_1 and σ_2/σ_1 are large. These results indicate that the model is flexible and can represent multiple scenarios, which is appealing since real life data may also behave in this way. Therefore, it is preferred over the TCEV one as a starting point of our model.

	$\mu_2/\mu_1 = 1$	$\mu_2/\mu_1 = 1.5$	$\mu_2/\mu_1 = 2$
$\sigma_2/\sigma_1 = 0.5$	0.0062	0.0350	0.1030
$\sigma_2/\sigma_1 = 1$	0.2800	0.5408	0.8482
$\sigma_2/\sigma_1 = 1.5$	0.7254	0.9276	0.9632
$\sigma_2/\sigma_1 = 2$	0.9296	0.9742	0.9902

Table 2: Average proportion of events of type 2 in the 10 most extreme ones obtained from 500 series of 1000 simulations of the Mixture Gumbel model with the corresponding ratios of location and scale parameters. The mixing parameter ω is equal to 0.2.

2.2 Finite mixture models for uncertain categories

Although the division of the data point made using the labels may make sense from a physical point of view, it is not necessarily appropriate for describing the tails of the distribution, hence we want to use them to inform the inference but not completely condition it, using also other variables to enhance the model. Therefore, rather than having pre-fixed groups using the labels, we want to allow the data to be informative and to let the allocation be driven by them. Without loss of generality, we assume that the data come from two populations.

Considering the data $x = (x_1, \dots, x_n)$ and the latent allocation variables $z = (z_1, \dots, z_n)$, with z_i that identifies the mixture component x_i belongs to, and the denoting by $\ell = (\ell_1, \dots, \ell_n)$ an observed vector of binary labels such that ℓ_i indicates the type of event generating the data point x_i, for $i = 1 \dots, n$, we define the model

$$\begin{gathered} x_i \mid z_i = j \overset{ind}{\sim} \text{Gumbel}\,(x_i \mid \theta_j)\,; \quad \theta_j = (\mu_j, \sigma_j) \\ z_i \mid \ell_i \overset{ind}{\sim} \text{Bernoulli}\,(z_i \mid \omega_i)\,, \\ \omega_i = \frac{\exp(\beta_0 + \beta_1 \ell_i)}{1 + \exp(\beta_0 + \beta_1 \ell_i)}, \quad i = 1, \dots, n. \end{gathered} \tag{2}$$

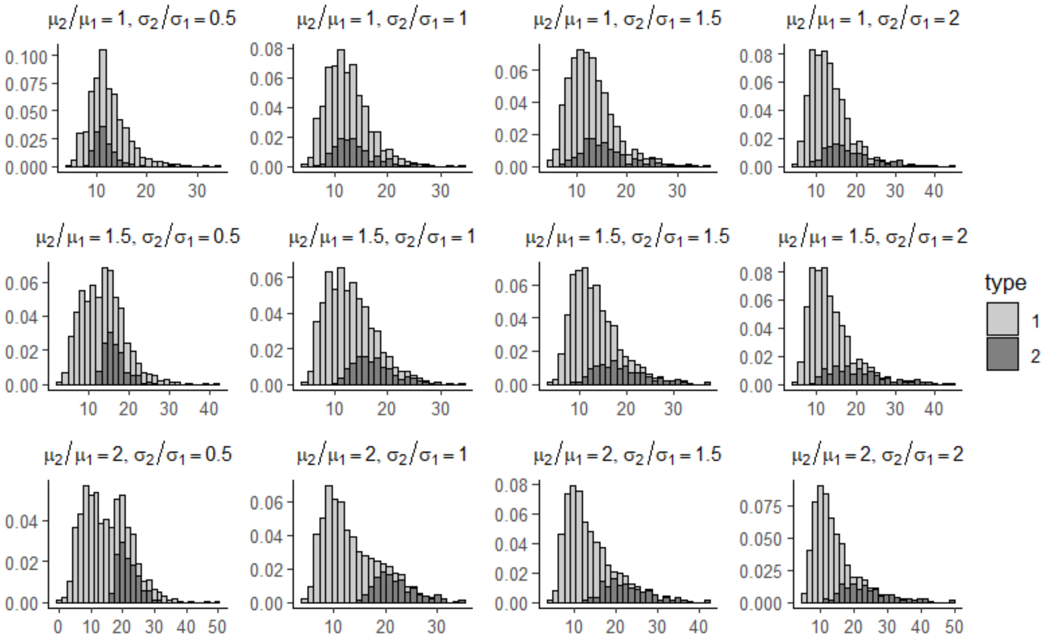

Figure 1: Histograms of the distribution of 1000 simulated data from the Mixture Gumbel model with different ratios of the location parameters and the scale parameters. The mixing parameter ω is always equal to 0.2. For each bin the area is coloured according to the proportion that is due to process 1 (light) and process 2 (dark).

Thus ω_i is derived from a logistic regression as a function of the labels, but this could be generalised to include additional covariates informing the data allocation in the tails. This hierarchical structure allows a more flexible modelling of extreme events that do not originate from a single population and, since the labelling is data-driven, it does not require the labels to be known.

3. Preliminary results

We carry out a simulation study to assess what happens when the labels are not a good representation of the mixture allocation for the tails, using both Model 1 and Model 2. Indeed, we explore how the results change and how robust they are by considering also a scenario where the labels are not the actual identifiers of the type of process, but a small percentage of them is swapped and therefore wrong. We produce 500 series of $n = 1000$ annual maximum events generated from a Mixture Gumbel model with parameters $\mu_1 = 10, \mu_2 = 20, \sigma_1 = 3, \sigma_2 = 5$ and $\omega = 0.2$, and we estimate the parameters by numerical maximum likelihood. For Model 2 we allocate to type 2 the data points with the estimated ω_i greater or equal to 0.5, and the other ones to type 1. For comparison with the Mixture Gumbel model, an estimate of a single ω is then computed as the number of units assigned to type 2 over the total. In Model 1 the parameter ω is simply estimated as the ratio between the number of observed type 2 events and the total number of events.

Tables 3 and 4 show the maximum likelihood estimates related to the two different likelihoods when the labels are the actual ones and when 10% of them are wrong, respectively. It is possible to notice that if the true labels are provided (so they actually identify the type of process) both models can accurately estimate the parameters, whereas when the information on the type of process is wrong, even by a small percentage, we get a more robust model by using the binary regression idea (Model 2). Similarly, the return level plots in 2 show that when the labels are not a good representation of the mixture allocation the model that solely relies on them for estimation (Model 1) is not able to correctly capture high quantiles while Model 2 does.

	Model with fixed ω				Model with varying ω_is			
	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
μ_1	9.985	10.06	10.06	10.13	9.957	10.03	10.03	10.11
σ_1	2.998	3.048	3.052	3.105	2.958	3.016	3.016	3.079
μ_2	20.12	20.37	20.39	20.63	20.12	20.38	20.40	20.65
σ_2	4.708	4.887	4.899	5.078	4.695	4.866	4.878	5.061
ω	0.193	0.200	0.200	0.208	0.193	0.200	0.200	0.208

Table 3: Summary of 500 maximum likelihood estimates of the parameters of a Mixture Gumbel model with $\mu_1 = 10, \mu_2 = 20, \sigma_1 = 3, \sigma_2 = 5, \omega = 0.2$. On the left the estimates assuming Model 1 and on the right the ones from Model 2.

	Model with fixed ω				Model with varying ω_is			
	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
μ_1	10.01	10.09	10.09	10.17	9.949	10.04	10.04	10.12
σ_1	3.027	3.088	3.091	3.152	2.958	3.023	3.025	3.102
μ_2	16.50	16.78	16.77	17.03	19.90	20.36	20.21	20.68
σ_2	6.477	6.649	6.655	6.828	4.695	4.914	4.979	5.185
ω	0.252	0.260	0.260	0.268	0.252	0.260	0.260	0.268

Table 4: Summary of 500 maximum likelihood estimates of the parameters of a Mixture Gumbel model with $\mu_1 = 10, \mu_2 = 20, \sigma_1 = 3, \sigma_2 = 5, \omega = 0.2$ and 10% of the labels swapped. On the left the estimates assuming Model 1 and on the right the ones from Model 2.

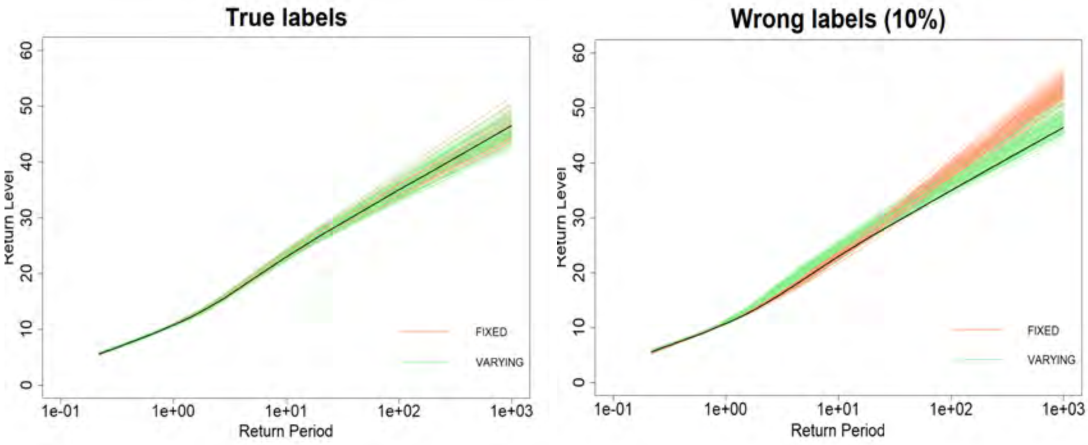

Figure 2: Return level plots corresponding to 100 different estimates of the parameters of the Mixture Gumbel model using Model 1 (pink) and Model 2 (green), when the labels are true on the left and when 10% are wrong on the right. In black the return level curve obtained with the true values of the parameters.

4. Discussion

When dealing with real world problems, it is reasonable to assume that extremes events originate from a number of different types of data-generating processes. However it is not easy to find data on extreme events that have information about the type of phenomenon which generated them. A further issue that is not very evident in the literature concerns the fact that even when this information is made available by domain experts, is not necessarily one that allows to discriminate between the multiple groups in the tail of the population. We define a model in which the allocation of the data points is not solely based on the knowledge on the type of process, i.e. the labels, but it exploits them to inform the inference, leading to results that are more robust to noise in the labels.

We are interested in enhancing the model by using Bayesian methods to exploit the ability of setting priors which encode the understanding we have of the problem, and to borrow information between the groups to estimate the model parameters. The Bayesian setting also allows to integrate in the estimation the uncertainty in the labels. Furthermore, we aim at exploring the results of the simulation study with real data applications, which are usually not characterised by such a high number of observations, with a focus on the problem of unknown categories.

References

[1] S. Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

[2] S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert. *Handbook of mixture analysis*. CRC press, 2019.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[4] T. R. Kjeldsen, H. Ahn, I. Prosdocimi, and J.-H. Heo. Mixture gumbel models for extreme series including infrequent phenomena. *Hydrological Sciences Journal*, 63(13-14):1927–1940, 2018.

[5] F. Rossi, M. Fiorentino, and P. Versace. Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, 20(7):847–856, 1984.