ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 4, *pp.* 1183–1204

A computationally efficient correlated mixed probit model for credit risk inference

Elisa Tosetti and Veronica Vinciotti

Brunel University London, Uxbridge, UK

[Received August 2018. Revised March 2019]

Summary. Mixed probit models are widely applied in many fields where prediction of a binary response is of interest. Typically, the random effects are assumed to be independent but this is seldom so for many real applications. In the credit risk application that is considered in the paper, random effects are present at the level of industrial sectors and they are expected to be correlated because of interfirm credit links inducing dependences in the firms' risk to default. Unfortunately, existing inferential procedures for correlated mixed probit models are computationally very intensive already for a moderate number of effects. Borrowing from the literature on large network inference, we propose an efficient expectation–maximization algorithm for unconstrained and penalized likelihood estimation and derive the asymptotic standard errors of the estimates. An extensive simulation study shows that the approach proposed enjoys substantial computational gains relative to standard Monte Carlo approaches, while still providing accurate parameter estimates. Using data on nearly 64000 accounts for small and medium-sized enterprises in the UK in 2013 across 13 industrial sectors, we find that accounting for network effects via a correlated mixed probit model increases significantly the default prediction power of the model compared with conventional default prediction models, making efficient inferential procedures for these models particularly useful in this field.

Keywords: Credit risk modelling; Expectation–maximization algorithm; Graphical modelling; Mixed probit

1. Introduction

Discrete choice models with correlated group-specific random effects have wide applicability and practical importance in economics and the social sciences, as they can accommodate unobserved heterogeneity, overdispersion and intracluster as well as intercluster correlation across binary outcomes. In this paper, we consider the prediction of a firm's risk to default, whereby group random effects at the level of industrial sectors are to be expected, and, at the same time, dependences between and within the industrial sectors are also to be expected because of interfirm credit links.

Unfortunately, the presence of correlated random effects in mixed models poses substantial computational challenges, with maximum likelihood (ML) estimation typically requiring the evaluation of a high dimensional integral. To overcome these numerical difficulties, various methods have been proposed in the literature that approximate the likelihood by Gauss–Hermite quadrature or Monte Carlo integration and then maximize it by either Newton–Raphson or expectation–maximization (EM) algorithms (Breslow and Clayton, 1993; Schilling and Bock, 2005). Despite achieving a computational gain, these methods can still be applied only in the

Address for correspondence: Veronica Vinciotti, Department of Mathematics, Brunel University London, Uxbridge, UB8 3PH, UK.
E-mail: veronica.vinciotti@brunel.ac.uk

 0035–9254/19/681183
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

presence of a limited number of groups because the number of evaluation points in the Gauss–Hermite quadrature increases exponentially with the number of random effects. In addition, these approximate (ML) estimates have been proved to be inconsistent under various conditions, with an asymptotic bias that can be severe if the variance components are not small (Breslow and Lin, 1995).

An alternative, widely used, approach for estimating mixed models for binary variables combines Monte Carlo integration with various EM algorithms, leading to the so-called Monte Carlo EM algorithm (see, among others, Ashford and Sowden (1970), Chib and Greenberg (1998), McCulloch (1997) and Gueorguieva and Agresti (2001)). For the case of a mixed probit model with independent random effects, McCulloch (1994) proposed Monte Carlo versions of the EM algorithm for ML estimation based on Gibbs sampling. This approach has been extended by McCulloch (1997) to the more general case of generalized linear mixed models, by considering a Metropolis–Hastings algorithm at each E-step of the ML estimation. Similarly, for the case of a mixed probit model with correlated random effects, Chan and Kuk (1997) proposed an EM algorithm where the E-step is made feasible by Gibbs sampling. The approach proposed, however, is computationally very intensive, as it requires sampling from a multivariate truncated normal distribution. To deal with this problem, Tan *et al*. (2007) proposed a non-iterative importance sampling approach to evaluate the first- and the second-order moments of a truncated multivariate normal distribution associated with the Monte Carlo EM algorithm. An alternative, direct sampling-based, EM algorithm was advanced by An and Bentler (2012), who proposed to draw random samples from the prior Gaussian distribution of random effects. This is computationally easier than from the posterior distribution, but at the expense of a higher Monte Carlo error. One limitation of the above Monte Carlo EM algorithms is that, by combining Monte Carlo simulation with iterative procedures, they are still computationally very expensive. The estimation that is involved in the E-step of the Monte Carlo EM algorithm can require a prohibitively large amount of time for a large number of statistical units and already a moderate number of random effects.

In this paper, motivated by a large credit risk application, we propose an EM algorithm for estimation of a mixed probit model with correlated random effects that can be adopted for estimation and prediction from very large data sets and a large number of random effects. The algorithm relies on efficient approximations of conditional expectations that simplify the calculation of the moments of a truncated normal distribution and avoid computationally demanding sampling methods. Similar approximations have been adopted in the context of graphical models for ordinal (Guo *et al*., 2015; Behrouzi and Wit, 2019) and censored (Augugliaro *et al*., 2018) data but they have not been used in a regression context before. Similarly to those approaches, we also propose a penalized version of the likelihood estimator, by applying the graphical lasso approach (Friedman *et al*., 2008) within the proposed EM algorithm. Beyond pointwise estimation, standard errors of ML estimates in the context of mixed effects models are also typically obtained by time-consuming resampling approaches. In this paper we exploit the work of Louis (1982) to derive the observed Fisher information matrix of our proposed mixed probit model and thus to obtain the asymptotic standard errors of the estimates. In doing this, we adopt results by Horrace (2015) to calculate the third and fourth moments of univariate truncated normal distributions which appear in the observed Fisher information matrix. Our paper provides several contributions to the existing literature. First, we propose an extension of the literature on non-linear mixed models to the case of correlated random effects, offering an inferential procedure that enables estimation of unknown parameters and associated standard errors also in the presence of very large samples. In doing this, we investigate ways of overcoming serious computational difficulties that often arise in regression models with correlated binary responses. We

also show how penalized inferential procedures can be applied under this framework, enabling us to cover the case where the number of random effects exceeds the number of observations.

An extensive simulation study assessing the properties and computational efficiency of our inferential procedure shows good performance of the approach proposed compared with existing approaches. Using data on around 64000 accounts of unlisted small and medium-sized enterprises (SMEs) based in the UK and observed in the year 2013, we find that incorporating interfirm network dependences in the form of correlated random effects increases the default prediction power of the credit risk model compared with conventional models. The remainder of the paper is structured as follows. Section 2 describes the empirical application on credit risk which motivates this study. Section 3 introduces our mixed graphical probit model and describes the EM algorithm for parameter estimation, with the proposed efficient approximations of the conditional expectations, the inference under penalized likelihood and the derivation of asymptotic standard errors. Section 4 carries out an extensive simulation study on the method, and Section 5 describes the results of the proposed approach on real data. Finally, Section 6 provides some concluding remarks.

2. Motivating example: credit risk modelling of small and medium-sized enterprises

There is nowadays interest in creating default prediction models for SMEs. Academic research into failure prediction has focused almost exclusively on large companies, i.e. those which are listed on, and priced by, the market, proposing a wide range of models and methods to assess and quantify their risk of default. In contrast, there has been a relatively small number of prior academic studies examining default prediction and credit scoring models with reference to small, private, businesses, mostly because of the difficulty in obtaining sufficient and good quality data in these contexts. These models are likely to be different from those used for large corporates. For this reason, the recent Basel Accords are now directing the international credit system to pay closer attention to measuring and managing credit risk of SMEs (Sabato, 2010).

When modelling credit risk for SMEs, an important feature to be considered is the fact that companies are not simply independent agents competing for customers on markets. They are linked by supply–customer relationships. Firms interact with each other because they exchange items of value, such as information, goods, services and money. For example, the outputs of some firms (subcontractors) are input for some other firms. In addition, some firms may extend trade credit to other firms, thus creating some sort of interfirm credit links (Battiston *et al.*, 2007). Interdependence between firms' default can also arise because they share part of the management team and hence are subject to similar investment decisions, or because firms react similarly to external shocks such as a rise in the interest rate (Andrews, 2005). Under this framework, the failure of a firm is likely to increase the probability of failure of connected firms, giving rise to clustered fluctuations in the number of failed firms.

Despite the importance of interfirm links in determining firms' performances, only a few studies have looked at the role of interaction in determining firms' default and clusters of default, with the majority of these studies focusing on identifying the conditions under which local failures can result in bankruptcies across the network (Delli Gatti *et al.*, 2006), or exploring whether firms that issue more trade credit are more likely to experience debtor failure (Jacobson *et al.*, 2013). Yet fewer studies have considered incorporating information on firms' interdependences in a default prediction model. Among these, Barro and Basso (2010) have proposed a model of contagion that associates the economic relationship of sectors of the economy and the geographical proximity of each pair of firms in a network of firms, whereas Barreto and

Artes (2013) have developed a measure of local risk of default using ordinary kriging from data on 9 million Brazilian SMEs observed in 2010. After including this measure as an additional explanatory variable in a logistic credit scoring model, they showed that the performance of the model improved considerably.

It is well known that the financial performance of companies is in part driven by sector- and area-specific attributes, linked for example to heterogeneity across industries in accounting policies or local trends in demand (see, for example, Kukuk and Ronnberg (2013)). For this reason, mixed discrete choice models have been widely adopted to predict firm financial distress for large corporations (see, among others, Jones and Hensher (2004) and Kukuk and Ronnberg (2013)), with few studies also specific to SMEs (see, for example Alfo' *et al*. (2005)). Differently from this literature and considering the importance of interfirm dependences discussed above, in this paper, we allow group effects to be correlated by assigning them a non-diagonal covariance matrix. Under this framework, the dependence relationship of the binary outcomes (default) is induced by the underlying Gaussian graphical model on the random effects. In particular, we assume that the risk of default for one company follows a probit regression specification with correlated group random effects, where groups are given by all companies operating in the same sector of economic activity and located in the same region.

We exploit a rich data set from a large financial institution covering around 64000 accounts of unlisted firms based in the UK and observed in the year 2013. These are companies that have no more than 250 employees, a turnover of smaller than £25.9 million and a balance sheet total of no more than £12.9 million. In line with other studies, we define failure as entry into liquidation, administration or receivership. The accounts analysed for failed companies are the last set of accounts filed in the year before insolvency. The companies are spread over a total of 59 geographical areas, defined using the '*nomenclature des unités territoriales statistiques*', level 3, classification, and across 13 broad sectors (divisions) of economic activity. In our model, the sectors will appear as random effects, whereas the geographical areas as the sampling units.

The data set contains a set of financial variables extracted from the accounts of firms, as well as non-financial information, that are often included in conventional default prediction models (see, among others, Altman and Sabato (2007), Altman *et al*. (2010), Carling *et al*. (2007), Campbell *et al*. (2008) and Jacobson *et al*. (2013)). In terms of firm-specific financial variables, we include a set of financial ratios that cover the areas of profitability, liquidity, leverage, coverage and activity (Altman and Sabato, 2007). Profitability is the ability of the firm to generate sufficient profits or returns, liquidity measures the ability of the firm to meet its short-term obligations, leverage refers to the relative amount of debt and other obligations of the firm, coverage is the risk that is inherent in lending to the business in the long term and activity is the level of efficiency of a business. As for the non-financial indicators, we consider variables that are linked to the age and size of the companies. We expect a higher risk of default for newly formed companies that decreases with the age of the company, and that is particularly high in the years immediately after an initial 'honeymoon period' of around 2 years. Finally, we have matched information on the postal district of the trading address with data on latitude and longitude and other geographical information extracted from the UK Office for National Statistics, to calculate covariates at the aggregated level and to account for systematic risk. In particular, we include the '*nomenclature des unités territoriales statistiques*', level 3, gross domestic product, as a proxy for the economic conditions of the area where the company operates. Table 1 lists the financial ratios that are included in our analysis grouped according to the financial and the non-financial indicators, including company characteristics and aggregate variables.

Table 2 provides a set of descriptive statistics for the variables that are included in our model, for failed and non-failed companies. As expected, companies that failed have on average worse

Table 1. Credit risk data: definition of financial ratios, non-financial indicators and aggregate variables

Variable	*Accounting ratio category*
Financial indicators	
Total liabilities/total assets	Leverage
Net worth/total liabilities	Leverage
Cash/total assets	Liquidity
Current liabilities/current assets	Liquidity
Trade credit/total liabilities	Liquidity
Trade debt/total assets	Liquidity
Retained profits/total assets	Profitability
Account receivable/total liabilities	Activity
Non-financial characteristics	
Size	Total assets (logarithms)
Age (years)	Age from the date of incorporation (logarithms)
Age risk	1 if $3 \leqslant$ age $\leqslant 9$ years
Local gross domestic product	Gross domestic product in the *nomenclature des unités territoriales statistiques*, level 3

Table 2. Credit risk data: descriptive statistics for non-failed and failed companies on the training sample

Variable	*Results for non-failed companies*		*Results for failed companies*	
	Mean	*Standard error*	*Mean*	*Standard error*
Total liabilities/total assets	0.817	1.243	1.278	1.851
Net worth/total liabilities	6.315	22.461	3.155	15.173
Cash/total assets	0.333	0.348	0.377	0.380
Current liabilities/current assets	1.826	5.283	2.386	5.806
Trade credit/total liabilities	0.197	0.302	0.225	0.350
Trade debt/total assets	0.155	0.231	0.162	0.263
Retained profits/total assets	−0.030	0.594	−0.216	1.039
Account receivable/total liabilities	0.006	0.029	0.004	0.025
Size	12.311	2.899	10.489	2.484
Age (years)	2.382	0.927	1.757	0.873
Age risk	0.346	0.476	0.445	0.497
Local gross domestic product	10.229	0.447	10.213	0.433

leverage and liquidity indicators than firms that did not fail; they are smaller in size and younger and more frequently fall in the age risk group. It is interesting to observe that both trade debt and trade credit ratios have higher values for defaulted companies. This result is supported by the literature on trade credit which shows evidence that financially distressed small companies not only have higher levels of trade debt supplied to customers but also of trade credit obtained from suppliers (Carbó-Valverde *et al.*, 2016).

In the next section, we formalize the proposed mixed probit model with correlated random effects and describe an inferential procedure that is computationally efficient for data such as those described in this section, for which existing mixed probit models are prohibitively slow.

3. Efficient mixed probit model with correlated random effects

3.1. The model

Consider a sample of N_r companies in region r, with $r = 1, 2, \ldots, R$. Let y_{ir} be the dichotomous variable equal to 1 when company i in region r defaults. Let G be the number of industrial sectors. Using the latent response model, we assume that y_{ir} is generated by thresholding the latent variable y_{ir}^* that follows the Gaussian mixed model:

$$\begin{aligned} y_{ir}^* &= \boldsymbol{\beta}' \mathbf{x}_{ir} + \mathbf{z}_{ir}' \mathbf{u}_r + \varepsilon_{ir}, \\ y_{ir} = 1 \qquad & \text{if } y_{ir}^* \geqslant 0, \ 0 \text{ otherwise}, \end{aligned} \tag{3.1}$$

where $\mathbf{x}_{ir}$ is a K-dimensional vector of explanatory variables, $\boldsymbol{\beta}$ is a K-dimensional vector of unknown parameters, $\mathbf{u}_r = (u_{1r}, u_{2r}, \ldots, u_{Gr})'$ is a G-dimensional vector of Gaussian random errors with $\mathbf{z}_{ir}$ being a G-dimensional vector of (known) loadings, with all entries equal to 0 except for a 1 for the entry corresponding to the sector that is associated with observation i, and ε_{ir} are Gaussian random errors. We assume that $\mathbf{u}_r$ and ε_{ir} satisfy the following conditions for all r:

$$\begin{gathered} E(\varepsilon_{ir}) = 0, \qquad E(\varepsilon_{ir}^2) = 1, \qquad \text{for } i = 1, 2, \ldots, N_r, \\ E(\varepsilon_{ir}\, \varepsilon_{js}) = 0, \qquad \text{for } i \neq j = 1, 2, \ldots, N_r, \quad r, s = 1, 2, \ldots, R, \\ E(\mathbf{u}_r\, \mathbf{u}_r') = \boldsymbol{\Sigma}_G, \\ E(\mathbf{u}_r\, \mathbf{u}_s') = \mathbf{0}, \qquad \text{for } r \neq s, \\ E(\mathbf{u}_r\, \varepsilon_{is}) = \mathbf{0} \qquad \text{for } r, s = 1, 2, \ldots, R, \end{gathered}$$

where $\boldsymbol{\Sigma}_G$ is a positive definite matrix with σ_{gh} the (g, h) off-diagonal element and σ_g^2 the gth diagonal element. In stacked form model (3.1) can be written as

$$\mathbf{y}_r^* = \mathbf{X}_r \boldsymbol{\beta} + \mathbf{Z}_r \mathbf{u}_r + \boldsymbol{\varepsilon}_r,$$

where $\mathbf{y}_r^* = (y_{1r}^*, y_{2r}^*, \ldots, y_{N_r,r}^*)'$, $\mathbf{X}_r = (\mathbf{x}_{1r}, \mathbf{x}_{2r}, \ldots, \mathbf{x}_{N_r,r})'$, $\boldsymbol{\varepsilon}_r = (\varepsilon_{1r}, \varepsilon_{2r}, \ldots, \varepsilon_{N_r,r})'$ and $\mathbf{Z}_r$ is an $N_r \times G$ matrix. In addition, $\mathbf{y}_r^*$ has covariance

$$\boldsymbol{\Sigma}_r = \mathbf{Z}_r \boldsymbol{\Sigma}_G \mathbf{Z}_r' + \mathbf{I}_{N_r}. \tag{3.2}$$

The model above allows for group effects that vary across R and G, although the dependences are allowed only across the G-dimension.

3.2. Inference

The interest is in estimating the regression parameters $\boldsymbol{\beta}$, as well as the dependence structure among the G groups, given by the elements of the precision matrix $\boldsymbol{\Phi}_G = \boldsymbol{\Sigma}_G^{-1}$. As also remarked in the graphical modelling literature, estimating the elements of the precision matrix enables us to assess whether any two units are conditionally independent given all other units (Lauritzen, 1996), thus providing a network of dependences at the level of random effects. Accordingly, let $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \text{vech}(\boldsymbol{\Phi}_G))$ be the vector of unknown parameters in the above model, and note that the observed data $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_R')'$ are a function of the unobserved variables $\mathbf{y}^* = (\mathbf{y}_1^{*\prime}, \mathbf{y}_2^{*\prime}, \ldots, \mathbf{y}_R^{*\prime})'$ and $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_R')'$. The log-likelihood of the observed data is given by

$$l(\boldsymbol{\vartheta}) = \log \left\{ \int f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y}, \mathbf{y}^*, \mathbf{u} | \boldsymbol{\vartheta}) \mathrm{d}\mathbf{y}^* \mathrm{d}\mathbf{u} \right\}. \tag{3.3}$$

The integral in equation (3.3) makes it difficult to maximize $l(\boldsymbol{\vartheta})$ directly, but an EM algorithm for computing ML estimates can be adopted, by maximizing the conditional expectation of the log-likelihood function for the complete data given the observed data $\mathbf{y}$. Treating $\mathbf{y}$, $\mathbf{y}^*$ and $\mathbf{u}$ as the complete data, and $\mathbf{y}$ as the incomplete data, we have

$$l(\boldsymbol{\vartheta}) = \log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\boldsymbol{\vartheta})\} - \log\{f_{\mathbf{y}^*,\mathbf{u}|\mathbf{y}}(\mathbf{y}^*,\mathbf{u}|\mathbf{y},\boldsymbol{\vartheta})\}, \tag{3.4}$$

where $\log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\boldsymbol{\vartheta})\}$ is the log-likelihood function for the complete data, namely

$$\begin{aligned}\log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\boldsymbol{\vartheta})\} &= \log\{f(\mathbf{u})\, f(\mathbf{y}^*,\mathbf{y}|\mathbf{u})\} \\ &\approx \frac{R}{2}\ln|\boldsymbol{\Phi}_G| - \frac{1}{2}\sum_{r=1}^{R}\mathbf{u}_r'\boldsymbol{\Phi}_G\mathbf{u}_r \\ &\quad - \frac{1}{2}\sum_{r=1}^{R}(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta} - \mathbf{Z}_r\mathbf{u}_r)'(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta} - \mathbf{Z}_r\mathbf{u}_r).\end{aligned}$$

Taking conditional expectations given $\mathbf{y}$ on both sides of equation (3.4) yields

$$\begin{aligned}l(\boldsymbol{\vartheta}) &= E[\log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\boldsymbol{\vartheta})\}|\mathbf{y}] - E[\log\{f_{\mathbf{y}^*,\mathbf{u}|\mathbf{y}}(\mathbf{y}^*,\mathbf{u}|\mathbf{y},\boldsymbol{\vartheta})\}|\mathbf{y}] \\ &= Q(\boldsymbol{\vartheta}) - H(\boldsymbol{\vartheta}),\end{aligned} \tag{3.5}$$

where

$$\begin{aligned}Q(\boldsymbol{\vartheta}) \approx \frac{R}{2}\ln|\boldsymbol{\Phi}_G| - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Phi}_G\frac{1}{R}\sum_{r=1}^{R}E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)\right\} \\ - \frac{1}{2}\sum_{r=1}^{R}E\{(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta} - \mathbf{Z}_r\mathbf{u}_r)'(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta} - \mathbf{Z}_r\mathbf{u}_r)|\mathbf{y}_r\}.\end{aligned} \tag{3.6}$$

The Q-function is the main ingredient of the EM algorithm. Let $\hat{\boldsymbol{\vartheta}}^{(m)}$ denote the estimate of $\boldsymbol{\Theta}$ after the mth iteration. Then the E- and M-steps of the $(m+1)$th iteration are respectively given by

(a) compute $Q(\boldsymbol{\vartheta}|\hat{\boldsymbol{\vartheta}}^{(m)}) = E[\log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\hat{\boldsymbol{\vartheta}}^{(m)})\}|\mathbf{y}]$,
(b) compute $\hat{\boldsymbol{\vartheta}}^{(m+1)} = \arg\max Q(\boldsymbol{\vartheta}|\hat{\boldsymbol{\vartheta}}^{(m)})$.

These steps are iterated until convergence is achieved. Looking further at the optimization in the M-step, the first-order conditions for $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}_G$ are given by

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \left(\sum_{r=1}^{R}\mathbf{X}_r'\mathbf{X}_r\right)^{-1}\sum_{r=1}^{R}\mathbf{X}_r'\left\{E(\mathbf{y}_r^*|\mathbf{y}_r) - \mathbf{Z}_r E(\mathbf{u}_r|\mathbf{y}_r)\right\}, \tag{3.7}$$

$$\boldsymbol{\Phi}_G^{(m+1)} = \left\{\frac{1}{R}\sum_{r=1}^{R}E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)\right\}^{-1}. \tag{3.8}$$

Hence, the M-step alternates between estimation of $\boldsymbol{\beta}$ by using equation (3.7) and estimation of $\boldsymbol{\Phi}_G$ by using equation (3.8). At each step, the new estimate of $\boldsymbol{\Phi}_G$ uses the previous value of $\hat{\boldsymbol{\beta}}$ and the new value of $\hat{\boldsymbol{\Phi}}_G$ is used to update $\hat{\boldsymbol{\beta}}$. Meng and Rubin (1993) showed that iterating between these two equations in the EM algorithm provides convergence to the ML estimates. However, the above expressions depend on the unknown quantities $E(\mathbf{u}_r|\mathbf{y}_r)$ and $E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)$. In what follows, we propose an approximation of conditional expectations $E(\mathbf{u}_r|\mathbf{y}_r)$ and $E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)$ and show how this can be adopted to simplify the EM algorithm.

3.3. Approximating conditional expectations

Using the law of iterated expectations and the theorem on conditional normal distributions, $E(\mathbf{u}_r|\mathbf{y}_r)$ and $E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)$ are typically calculated by

$$E(\mathbf{u}_r|\mathbf{y}_r) = \boldsymbol{\Sigma}_G \mathbf{Z}_r' \boldsymbol{\Sigma}_r^{-1} \{E(\mathbf{y}_r^*|\mathbf{y}_r) - \mathbf{X}_r\boldsymbol{\beta}\}, \tag{3.9}$$

$$\begin{aligned} E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r) = {} & \boldsymbol{\Sigma}_G \mathbf{Z}_r' \boldsymbol{\Sigma}_r^{-1} E\{(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta})(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta})'|\mathbf{y}_r\} \boldsymbol{\Sigma}_r^{-1} \mathbf{Z}_r \boldsymbol{\Sigma}_G \\ & + \boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_G \mathbf{Z}_r' \boldsymbol{\Sigma}_r^{-1} \mathbf{Z}_r \boldsymbol{\Sigma}_G, \end{aligned} \tag{3.10}$$

following appendix B and Chan and Kuk (1997).

From these expressions it is clear that $E(\mathbf{u}_r|\mathbf{y}_r)$ and $E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)$ depend on the first two moments of a multivariate truncated normal distribution, namely $E(\mathbf{y}_r^*|\mathbf{y}_r)$ and

$$E\{(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta})(\mathbf{y}_r^* - \mathbf{X}_r\boldsymbol{\beta})'|\mathbf{y}_r\}.$$

Some researchers have proposed algorithms for direct estimation or approximation of moments of multivariate truncated normal distributions (see, among others, Tallis (1961), Lee (1979) and Leppard and Tallis (1989)). Others have proposed a Markov chain Monte Carlo approach that consists of randomly generating a sequence of samples from the multivariate truncated normal distribution and then approximating the first two moments by the empirical conditional moments from these samples (Kotecha and Djuric, 1999; Chan and Kuk, 1997; Chib and Greenberg, 1998; Abegaz and Wit, 2015). Although this method is faster than direct estimation of the moments, it is still computationally very demanding for large-scale problems. A recent strand of literature has proposed approximating the first and second moments of a multivariate truncated normal distribution through an iterative procedure within the M-step (Guo *et al.*, 2015; Behrouzi and Wit, 2019; Augugliaro *et al.*, 2018), leading to a computationally much faster approach than any previous methods. Exploiting this literature, we consider a mean field approximation of the second moments, namely, for $i \neq j$ and for all $r = 1, 2, \ldots, R$,

$$E\{(y_{ir}^* - \boldsymbol{\beta}'\mathbf{x}_{ir})(y_{jr}^* - \boldsymbol{\beta}'\mathbf{x}_{jr})|\mathbf{y}_r\} \approx E\{(y_{ir}^* - \boldsymbol{\beta}'\mathbf{x}_{ir})|\mathbf{y}_r\} E\{(y_{jr}^* - \boldsymbol{\beta}'\mathbf{x}_{jr})|\mathbf{y}_r\}. \tag{3.11}$$

Hence, once controlled for the observed values in $\mathbf{y}_r$ and the regressors $\mathbf{X}_r$, y_{ir}^* and y_{jr}^* become decoupled. In Section 4 we shall show good properties of our proposed estimator with that based on the slower Monte Carlo EM procedures, that do not make the above approximation. Under approximation (3.11), to compute equations (3.9) and (3.10), we only need to find $E(y_{ir}^*|\mathbf{y}_r)$ and $E(y_{ir}^{*2}|\mathbf{y}_r)$. For this, first write the first and second conditional moments as

$$E(y_{ir}^*|\mathbf{y}_r) = E\{E(y_{ir}^*|\mathbf{y}_{-i,r}^*, y_{ir})|\mathbf{y}_r\}, \tag{3.12}$$

$$E(y_{ir}^{*2}|\mathbf{y}_r) = E\{E(y_{ir}^{*2}|\mathbf{y}_{-i,r}^*, y_{ir})|\mathbf{y}_r\}, \tag{3.13}$$

where $\mathbf{y}_{-i,r}^* = (y_{1r}^*, y_{2r}^*, \ldots, y_{i-1,r}^*, y_{i+1,r}^*, \ldots, y_{N_r r}^*)'$. Noting that $\mathbf{y}_r^*$ is a vector of jointly normal variables with mean 0 and covariance $\boldsymbol{\Sigma}_r$, and exploiting the theorem on conditional normal distributions, we obtain that the conditional distribution of y_{ir}^* given $\mathbf{y}_{-i,r}^*$ has mean and variance respectively given by

$$\begin{aligned} \tilde{\mu}_{ir} &= \boldsymbol{\beta}'\mathbf{x}_{ir} + \boldsymbol{\Sigma}_{r,i,-i} \boldsymbol{\Sigma}_{r,-i,-i}^{-1} (\mathbf{y}_{-i,r}^* - \mathbf{X}_{-i,r}\boldsymbol{\beta}), \\ \tilde{\sigma}_{ir}^2 &= \sigma_{ir}^2 - \boldsymbol{\Sigma}_{r,i,-i} \boldsymbol{\Sigma}_{r,-i,-i}^{-1} \boldsymbol{\Sigma}_{r,-i,i}, \end{aligned}$$

where σ_{ir}^2 is the (i,i)th element of $\boldsymbol{\Sigma}_r$. Replacing these expressions in the equation for the mean and second moment of truncated normal distributions (see Appendix A) we obtain the

following expressions for the first conditional moment (3.12) and the second conditional moment (3.13):

$$E(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir}|\mathbf{y}_r) = \boldsymbol{\Sigma}_{r,i,-i}\boldsymbol{\Sigma}^{-1}_{r,-i,-i}E(\mathbf{y}^*_{-i,r} - \mathbf{X}_{-i,r}\boldsymbol{\beta}|\mathbf{y}_r) + \rho_{1,ir}\tilde{\sigma}_{ir}, \tag{3.14}$$

$$\begin{aligned} E\{(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir})^2|\mathbf{y}_r\} = {} & \boldsymbol{\Sigma}_{r,i,-i}\boldsymbol{\Sigma}^{-1}_{r,-i,-i}E\{(\mathbf{y}^*_{-i,r} - \mathbf{X}_{-i,r}\boldsymbol{\beta})(\mathbf{y}^*_{-i,r} - \mathbf{X}_{-i,r}\boldsymbol{\beta})'|\mathbf{y}_r\}\boldsymbol{\Sigma}^{-1}_{r,-i,-i}\boldsymbol{\Sigma}_{r,i,-i} \\ & + \tilde{\sigma}^2_{ir} + 2\rho_{1,ir}\tilde{\sigma}_{ir}\boldsymbol{\Sigma}_{r,i,-i}\boldsymbol{\Sigma}^{-1}_{r,-i,-i}E(\mathbf{y}^*_{-i,r} - \mathbf{X}_{-i,r}\boldsymbol{\beta}|\mathbf{y}_r) + \rho_{2,ir}\tilde{\sigma}^2_{ir} \\ & + (\boldsymbol{\beta}'\mathbf{x}_{ir})^2 - 2\boldsymbol{\beta}'\mathbf{x}_{ir}E(y^*_{ir}|\mathbf{y}_r), \end{aligned} \tag{3.15}$$

where $\rho_{1,ir}$ and $\rho_{2,ir}$ are defined in Appendix A. These equations show that there is a recursive relationship between the elements in $E(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir}|\mathbf{y}_r)$ and $E\{(\mathbf{y}^*_r - \mathbf{X}_r\boldsymbol{\beta})(\mathbf{y}^*_r - \mathbf{X}_r\boldsymbol{\beta})'|\mathbf{y}_r\}$ and offer an iterative procedure for estimating these quantities. More specifically, let $E(y^*_{jr} - \boldsymbol{\beta}'\mathbf{x}_{jr}|\mathbf{y}_r)^{(h)}$ and $E\{(y^*_{jr} - \boldsymbol{\beta}'\mathbf{x}_{jr})^2|\mathbf{y}_r\}^{(h)}$, for all j, be the estimates of $E(y^*_{jr} - \boldsymbol{\beta}'\mathbf{x}_{jr}|\mathbf{y}_r)$ and $E\{(y^*_{jr} - \boldsymbol{\beta}'\mathbf{x}_{jr})^2|\mathbf{y}_r\}$ respectively, at the hth stage in the M-step. We plug these into the right-hand side of equations (3.14) and (3.15) to compute new values of $E(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir}|\mathbf{y}_r)$ and $E\{(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir})^2|\mathbf{y}_r\}$ (inner iterations). After convergence has been reached, let $E(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir}|\mathbf{y}_r)^{(h)*}$ and $E\{(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir})^2|\mathbf{y}_r\}^{(h)*}$ be the final estimates. We plug these into equation (3.7) to obtain a new estimate of $\boldsymbol{\beta}$ and to compute equation (3.10) that enters equation (3.8) for estimation of $\boldsymbol{\Phi}_G$ (outer iterations). With the new $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}_G$, we recompute $E(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir}|\mathbf{y}_r)$ and $E\{(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir})^2|\mathbf{y}_r\}$ ready for another round of inner iterations. Note, however, that convergence for the inner iterations is not necessary; in fact, inner iterations can be reduced to a single round of computation.

According to the iterative procedure just described, the matrix inverse $\boldsymbol{\Sigma}^{-1}_{r,-i,-i}$, for $i = 1, 2, \ldots, N_r$, needs to be computed at each iteration of the EM procedure. Although the matrix can be rather large, given that it has size $(N_r - 1) \times (N_r - 1)$, a simplified expression can be obtained by noting that

$$\boldsymbol{\Sigma}_{r,-i,-i} = \mathbf{Z}_{r,-i}\boldsymbol{\Sigma}_G\mathbf{Z}'_{r,-i} + \mathbf{I}_{N_r-1},$$

and, using the matrix inversion lemma,

$$\boldsymbol{\Sigma}^{-1}_{r,-i,-i} = \mathbf{I}_{N_r-1} - \mathbf{Z}_{r,-i}(\boldsymbol{\Sigma}^{-1}_G + \mathbf{Z}'_{r,-i}\mathbf{Z}_{r,-i})^{-1}\mathbf{Z}'_{r,-i}.$$

Hence, $\boldsymbol{\Sigma}^{-1}_{r,-i,-i}$ involves computing only the inverse of G-dimensional matrices. This shows the power of using a mixed model approach, whereby dependences are captured at the lower dimensional space of the random effects.

In addition, when N_r is particularly large, such as in our real application, we found it computationally beneficial, and not detrimental to the resulting estimators, to replace the expectations (3.9)–(3.10) with the group averages of expectations of the latent variables, i.e.

$$E(u_{gr}|\mathbf{y}_r) \approx \frac{1}{m_{gr}}\sum_{i\in g}\{E(y^*_{ir}|\mathbf{y}_r) - \boldsymbol{\beta}'\mathbf{x}_{ir}\}, \tag{3.16}$$

$$E(u_{gr}u_{hr}|\mathbf{y}_r) \approx \frac{1}{m_{gr}m_{hr}}\sum_{i\in g; j\in h}E\{(y^*_{ir} - \boldsymbol{\beta}'\mathbf{x}_{ir})(y^*_{jr} - \boldsymbol{\beta}'\mathbf{x}_{jr})|\mathbf{y}_r\}, \tag{3.17}$$

where m_{gr} is the number of units belonging to group g and located in region r and $\Sigma_{i\in g}$ indicates the sum over all units belonging to group g and located in region r. This estimator is widely adopted to proxy random effects (Hsiao, 2003), also in the context of cross-sectionally dependent panels (Moscone *et al.*, 2017).

Finally, further computational efficiency can be achieved by applying penalized ML, as described in the next subsection.

3.4. Penalized maximum likelihood estimation

ML estimation of Φ_G is feasible for $R \geqslant G+K$, though it can become unstable for R approaching $G+K$. For $R < G+K$, ML estimation is not feasible and further computational challenges arise in the high dimensional case of $R \ll G$. In addition, the network of dependences is often expected to be sparse and the recovery of its structure is of interest. In our particular application, although the primary objective is the prediction of default, it is also of interest, and possibly motivating further decision making, to identify which sectors of the economy are mostly connected with each other in their risk of default.

To address these challenges, we add an L_1-norm penalty term to the log-likelihood of the model and optimize the penalized likelihood:

$$l_1(\vartheta) = \log\left\{\int f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\vartheta)\mathrm{d}\mathbf{y}^*\mathrm{d}\mathbf{u}\right\} - \rho_G \|\Phi_G\|_1,$$

where ρ_G is a tuning parameter controlling the degree of sparsity of the underlying network and '$\|.\|_1$' is the L_1-norm on the off-diagonal entries of the precision matrix. When ρ_G is sufficiently large, some coefficients in Φ_G are shrunk to 0, resulting in the removal of the corresponding links in the underlying network. Noting that the part of $\log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y},\mathbf{y}^*,\mathbf{u}|\vartheta)\}$ that depends on Σ_G^{-1} is the log-likelihood of a multivariate normal distribution,

$$Q_1(\vartheta|\hat{\vartheta}^{(m)}) = -\frac{R}{2}\ln|\Sigma_G| - \frac{1}{2}\mathrm{tr}\left\{\Sigma_G^{-1}\frac{1}{R}\sum_{r=1}^{R}E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)\right\},$$

and, following the same line of reasoning as in Section 3.2, we consider the penalized estimation problem for Φ_G within the M-step by optimizing

$$Q_{1,\mathrm{pen}}(\vartheta|\hat{\vartheta}^{(m)}) = \frac{R}{2}\ln|\Phi_G| - \frac{1}{2}\mathrm{tr}\left\{\Phi_G\frac{1}{R}\sum_{r=1}^{R}E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)\right\} - \rho_G\|\Phi_G\|_1. \tag{3.18}$$

Hence, we alternate between estimation of β by using equation (3.7) and estimation of Φ_G by using equation (3.18), for which efficient graphical lasso implementations can be used (Friedman *et al.*, 2008).

The regularization parameter ρ_G defines the level of sparsity of the associated network $\hat{\Phi}_G$. By tuning this parameter, we can explore the full path of solutions, from a disconnected network (large ρ_G, corresponding to a mixed model with uncorrelated random effects) to a fully connected network ($\rho_G = 0$, corresponding to the ML estimates). Various information criteria are available in the penalized likelihood literature for the selection of this parameter. These methods are based on the likelihood function of the observed data, which, for our model, is given by equation (3.5). Ibrahim *et al.* (2008), however, suggested the use of only the Q-function in approximation (3.6) for calculation of the likelihood. This is more efficient, as the Q-function is a direct output of the EM algorithm, whereas the H-function would need to be calculated separately. Augugliaro *et al.* (2018) showed how, for a similar model, this approximate information criterion behaves well when compared with that based on the full likelihood. In contrast, considering that prediction of default is a classification problem, other criteria can also be used that are more in line with this objective of the study. In the real application, we shall explore both with the selection of ρ_G that minimizes the extended Bayesian information criterion eBIC

(Foygel and Drton, 2010) and with the ρ_G that maximizes the area under the receiver operating characteristic (ROC) curve, AUC, on a test set.

3.5. Standard errors approximation

Calculating standard errors of estimates requires knowledge of the information matrix that is associated with the log-likelihood function of the observed data,which is known as the observed information matrix. However, this also involves computation of the H-function in equation (3.5), which is not a direct output of the EM iterations. Following Louis (1982), it is possible to compute the observed information matrix by exploiting the complete-data gradient and curvature. In particular, let $B(\mathbf{y}|\boldsymbol{\vartheta}) = \partial^2 l(\boldsymbol{\vartheta})/\partial\vartheta_i\,\partial\vartheta_j$ be the partial second derivatives of the observed data log-likelihood and

$$S(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta}) = \frac{\partial \log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})\}}{\partial\boldsymbol{\vartheta}}$$

and

$$B(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta}) = \frac{\partial^2 \log\{f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})\}}{\partial\vartheta_i\,\partial\vartheta_j}$$

be the gradient and second derivative of the complete-data log-likelihood respectively. It is possible to show that

$$\begin{aligned} B(\mathbf{y}|\boldsymbol{\vartheta}) = {} & E\{B(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})|\mathbf{y}\} + E\{S(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})S(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})'|\mathbf{y}\} \\ & - E\{S(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})|\mathbf{y}\}E\{S(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\boldsymbol{\vartheta})|\mathbf{y}\}'. \end{aligned} \tag{3.19}$$

Hence, by exploiting the law of iterated expectations as well as approximation (3.11), it is also possible to compute efficiently all terms appearing on the right-hand side of expression (3.19). In the on-line supplementary material, we provide finite expressions for the elements of $B(\mathbf{y}|\boldsymbol{\vartheta})$.

4. Simulation study

To assess the performance of our proposed approach, we consider a simulation study using the following data-generating process:

$$\begin{aligned} y^*_{ir} = \beta x_{ir} + \mathbf{z}'_{ir}\mathbf{u}_r + \varepsilon_{ir}, \qquad & i = 1, 2, \ldots, N_r, \quad r = 1, 2, \ldots, R, \\ y_{ir} = 1 \qquad & \text{if } y^*_{ir} \geqslant 0;\ 0 \text{ otherwise,} \end{aligned}$$

where we set $\beta = 1$, $\mathbf{x}_r = (x_{1r}, x_{1r}, \ldots, x_{N_r r}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$ and $\mathbf{u}_r \sim N(\mathbf{0}, \boldsymbol{\Sigma}_G)$. To generate $\boldsymbol{\Sigma}_G$, we start from $\boldsymbol{\Theta}_G = \boldsymbol{\Sigma}_G^{-1}$ and assume that $\theta_{gh,G} \sim \text{Bin}(1, 3/G)$ for $g = 1, \ldots, G$ and $h = g, \ldots, G$. We then let $\mathbf{D}$ be the Choleski decomposition of $\boldsymbol{\Sigma}_G$, namely $\boldsymbol{\Sigma}_G = \mathbf{D}\mathbf{D}'$, and we generate $\mathbf{u}_r = \mathbf{D}\boldsymbol{\epsilon}_r$, where $\boldsymbol{\epsilon}_r = (\epsilon_{1r}, \epsilon_{2r}, \ldots, \epsilon_{Gr})'$, with $\epsilon_{ir} \sim \text{IDN}(0, 1)$. We finally obtain $\boldsymbol{\Sigma}_r$ by applying formula (3.2). We generate $\boldsymbol{\Sigma}_X$ by following the same procedure. In the next subsections, we test the method under different scenarios and performance criteria.

4.1. Estimation of regression coefficients

In a first set of experiments, we assess the performance of our proposed method in estimating the regression coefficients and their standard errors. For this, we replicate 50 times the simulation that was described above and report the bias and root-mean-squared error (RMSE) of the slope parameter β, given by $(1/50)\Sigma_{s=1}^{50}\hat{\beta}_s - \beta$ and $\sqrt{\{(1/50)\Sigma_{s=1}^{50}(\hat{\beta}_s - \beta)^2\}}$ respectively. We carry out

Table 3. Simulated data from a mixed graphical probit: bias and RMSE of regression coefficients when estimated by (I) a standard probit with no random effects, (II) a mixed graphical probit using approximation (3.11), (III) a mixed graphical probit using approximation (3.11) and equations (3.16) and (3.17) in place of equations (3.9) and (3.10) and (IV) a mixed graphical probit with full Monte Carlo EM†

N	G	R	*(I) Results for probit*		*(II) Results for mixed graphical probit using equations (3.11), (3.16) and (3.17)*				*(III) Results for mixed graphical probit using equation (3.11)*			*(IV) Results for mixed graphical probit using full EM*		
			Bias	*RMSE*	*Bias*	*RMSE*	*Standard error (3.19)*	*Time (s)*	*Bias*	*RMSE*	*Time (s)*	*Bias*	*RMSE*	*Time (s)*
50	10	200	−0.2958	0.2961	−0.0021	0.0209	0.0170	3.0	−0.0080	0.0226	23.4	0.0266	0.0328	54.0
50	25	200	−0.2901	0.2906	−0.0771	0.0812	0.0136	4.1	−0.0939	0.0973	305.8	−0.0704	0.0758	46.8
100	10	200	−0.2963	0.2966	−0.0012	0.0149	0.0128	6.6	−0.0018	0.0149	13.5	0.0183	0.0247	149.7
100	25	200	−0.2932	0.2934	−0.0056	0.0169	0.0118	6.2	−0.0087	0.0181	153.8	0.0199	0.0253	128.3
250	10	200	−0.2929	0.2931	−0.0018	0.0108	0.0084	21.9	−0.0010	0.0107	25.9	0.0093	0.0130	1586.5
250	25	200	−0.2922	0.2923	0.0001	0.0115	0.0081	28.2	0.0054	0.0129	97.3	0.0198	0.0222	1665.2

†The average computational time for one estimation of the mixed graphical approaches is also reported, as well as the average standard error for (II) by using equation (3.19).

a comparison of the estimator that is obtained from the approximate EM algorithm proposed in this paper with that from the Monte Carlo EM estimator by Chan and Kuk (1997). Because of the high computational cost of the Monte Carlo EM approach, we select small values of G for the comparison ($G \leqslant 25$). We set $R = 200$ and conduct the comparison under ML estimation (i.e. setting $\rho_G = 0$) for varying values of $N_r = N$, the number of observations per regions. This comparison is important because the Monte Carlo EM estimator by Chan and Kuk (1997) does not rely on the conditional approximation (3.11). For the same combinations of N and R, we also compare the properties and computational time of the mixed graphical probit estimator using equations (3.9) and (3.10) with those of the same estimator based on their approximations (3.16)–(3.17).

Table 3 reports the results. These show that the three mixed graphical estimators have a small bias and RMSE, and that these decrease as N rises, whereas their performance slightly deteriorates as the number of groups, G, increases. In all cases, the estimators from the mixed graphical probit approaches are superior to a conventional probit model with no random effects (columns (I)). Comparing the results in columns (II) and (III), the computational time of the estimator based on equations (3.16) and (3.17) is significantly smaller than that of the estimator based on equations (3.9) and (3.10), thus supporting the use of group averages of conditional expectations to proxy random effects. The fact that the bias and RMSE of the estimators in columns (II) and (III) are of comparable size with that in column (IV) indicates that approximation (3.11), adopted both in columns (II) and in columns (III), does not significantly affect the properties of our estimators. This was found to be so also for the approximate standard errors calculated by using expression (3.19), whose averages across the 50 replications were found overall to be of comparable size with the RMSE, albeit with some discrepancy at small N. Whereas Table 3 shows little difference between the estimators in terms of bias and RMSE, the difference in computational time between the graphical mixed probit estimators in columns (II) and (III) and the full Monte Carlo EM estimator in columns (IV) is striking, with the mixed graphical probit carrying out one estimation in a few seconds across all experiments, against a computational time that can be as long as a few minutes in the case of the Monte Carlo EM algorithm.

4.2. *Recovery of the network of dependences under L_1-penalization*

In a second set of experiments, we assess the performance of our proposed method in recovering the underlying network of dependences. This network is constructed by drawing an edge in correspondence to each non-zero element of the estimated precision matrix $\mathbf{\Phi}_G$. For a range of values of N, G and R, we construct the ROC curve across the path of regularization parameters under L_1-penalization, i.e. across different levels of sparsity. Denoting by $\hat{\mathbf{\Phi}}_G^{\rho}$ the estimate of the precision matrix under the tuning parameter ρ, the curve plots the true positive rate, i.e. the percentage of true edges (non-zeros in the true $\mathbf{\Phi}_G$) correctly estimated as non-zeros in $\hat{\mathbf{\Phi}}_G^{\rho}$, against the false positive rate, i.e. the percentage of true missing edges (0s in $\mathbf{\Phi}_G$) incorrectly estimated as non-zeros in $\hat{\mathbf{\Phi}}_G^{\rho}$.

Fig. 1 plots these curves across the path of regularization parameters and averaged over 100 replications, for different configurations of N, G and R. As in the previous simulation, the performance of the mixed graphical probit estimator improves as N increases for fixed R and G (Fig. 1(a)), whereas it deteriorates as G rises, holding N and R constant (Fig. 1(b)). This result can be explained by looking at the main features of our model. In fact, as N increases there are increasingly more observations to estimate the unknown parameters β and $\mathbf{\Phi}_G$, whereas when G increases there are increasingly more parameters to estimate.

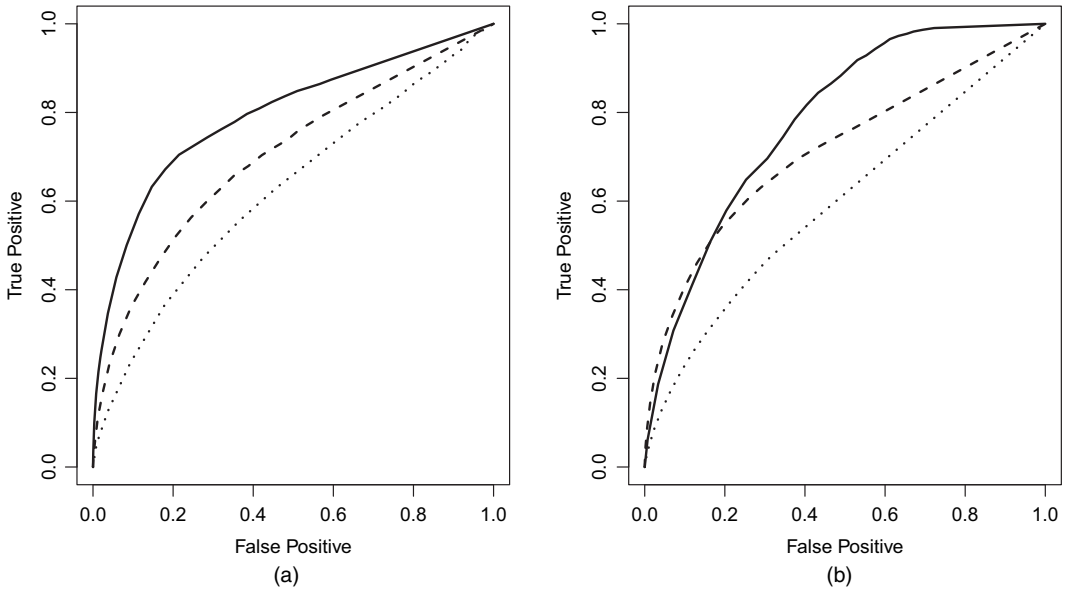

Fig. 1. Simulation study: averaged ROC curves of network recovery for varying N, G and R across the path of regularization parameters ρ_G, under penalized likelihood estimation: (a) $G = 25$, $R = 50$ (· · · · ·, $N = 50$; – – –, $N = 100$; ———, $N = 250$); (b) $N = 250$, $R = 50$ (———, $G = 5$; – – –, $G = 50$; · · · · ·, $G = 125$)

4.3. Prediction accuracy

In a final set of experiments, we assess the prediction accuracy of the proposed mixed graphical probit as a classification model. For this, we generate a testing sample with the same Monte Carlo design as above and employ the parameters that were estimated in the training sample to calculate the predicted probabilities

$$P(Y_{ir} = 1|\mathbf{x}_{ir}) = \Phi(\hat{\boldsymbol{\beta}}'\mathbf{x}_{ir} + \mathbf{z}'_{ir}\hat{\mathbf{u}}_r),$$

with Φ the standard normal cumulative distribution function and with $\hat{\mathbf{u}}_r$ the estimated group random effects calculated by using equation (3.9) (or for larger problems its approximation (3.16)).

We carry out two sets of experiments: one with $R = 200$ (the case of large R), where we compute the proposed ML estimator ($\rho_G = 0$) and one with $R = 50$ where we compute a penalized version of the estimator. In this case, we select the regularization parameter ρ_G with the value that is closest to the true sparsity level. This is possible only in a simulation setting and enables our results not to depend on the specific choice of model selection criterion. Fig. 2 shows the average ROC curves across 50 replications for varying configurations of N and G, under the large and small sample size cases respectively. The curve plots the percentage of non-zero outcomes that are correctly predicted as non-zero *versus* the percentage of 0s that are incorrectly predicted as non-zeros, as the classification threshold on the predicted probabilities varies between 0 and 1. Similarly to before, and as expected, we note an improvement in the performance of the mixed graphical probit as N increases and G decreases. The comparison also shows how the use of a densely estimated precision matrix ($\rho_G = 0$) under a setting of sparsity ($P(\theta_{jk} \neq 0) = 3/G$) does not hinder the performance of the classifier in terms of prediction accuracy. Indeed, the estimation of the regression parameters β is found to be quite stable across different levels of sparsity of the precision matrix, as noted also in the literature on correlated multivariate

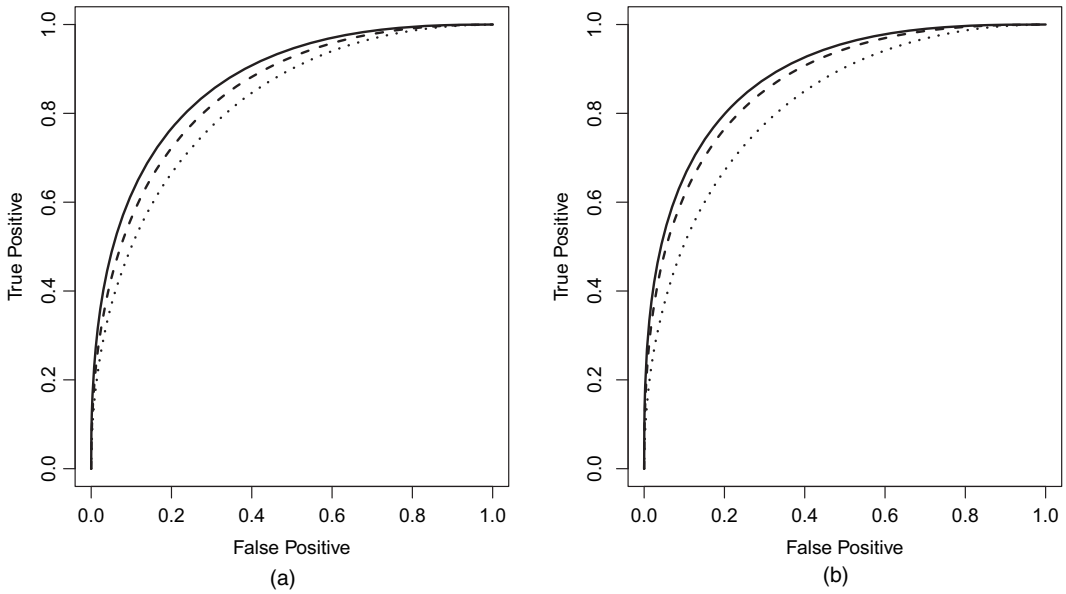

Fig. 2. Simulation study: average ROC curves on predicted outcomes on the test set for varying N, G and R, with model parameters estimated on the training set under (a) ML ($\rho_G = 0$) and (b) penalized likelihood (ρ_G set to the value closest to the true sparsity): (a) $G = 25$, $R = 200$ (· · · · · ·, $N = 50$; – – –, $N = 100$; ———, $N = 250$); (b) $N = 250$, $R = 50$ (———, $G = 5$; – – –, $G = 25$; · · · · · ·, $G = 125$)

probit models (Chib and Greemberg, 1998). In contrast with this, in all cases considered, the performance of the mixed graphical probit is far superior to that of a conventional probit model with no random effects, as shown in Fig. 3 for two representative cases. In the next section, we shall show how the mixed graphical model can be useful for credit risk prediction and we shall discuss more closely the aspect of model selection (selection of ρ_G) within that context, where recovering the underlying network of interfirm dependences is of particular interest.

5. Credit risk probit model with correlated effects

We employ the proposed approach to estimate a default prediction model for SMEs based on the data that were described in Section 2. To assess the performance of the classifier, we randomly split the sample into two groups: 40000 companies are used for estimation (the training sample) and the remaining accounts for testing the prediction accuracy of the model (the hold-out sample). We use the mixed graphical probit model in equation (3.1) and include network dependences at the level of the $G = 13$ industrial sectors, by exploiting the grouping of the data into $R = 59$ geographical regions. Thus, the model contains $59 \times 13 = 767$ correlated random effects from a total of 40000 observations in the training data. This means that ML estimation is feasible, but it is prohibitively slow by using standard implementations of mixed models; for example the R function `glmer` (Bates *et al.*, 2015) with uncorrelated random effects failed to converge. Thanks to the efficient implementation that is proposed in this paper, we can explore the full path of solutions from a fully connected network ($\rho_G = 0$; ML estimation) to a disconnected network (large ρ_G, corresponding to a mixed model with uncorrelated random effects with unequal variances). Fig. 4 shows two model selection criteria evaluated across the full path of solutions. In Fig. 4(a), we plot AUC of classification prediction on the test data for models fitted on the training data under various levels of sparsity. In Fig. 4(b),

we plot eBIC of the models across the path of solutions. Both plots show an optimal point somewhere in between a fully connected network (the rightmost value with 169 non-zeros in the precision matrix) and a disconnected network (the leftmost value with 13 non-zeros in the diagonal of the precision matrix). eBIC appears to favour a sparser model ($\rho_G = 0.005$, 41 non-zeros in the precision matrix), whereas AUC, although achieving a real maximum for the fully

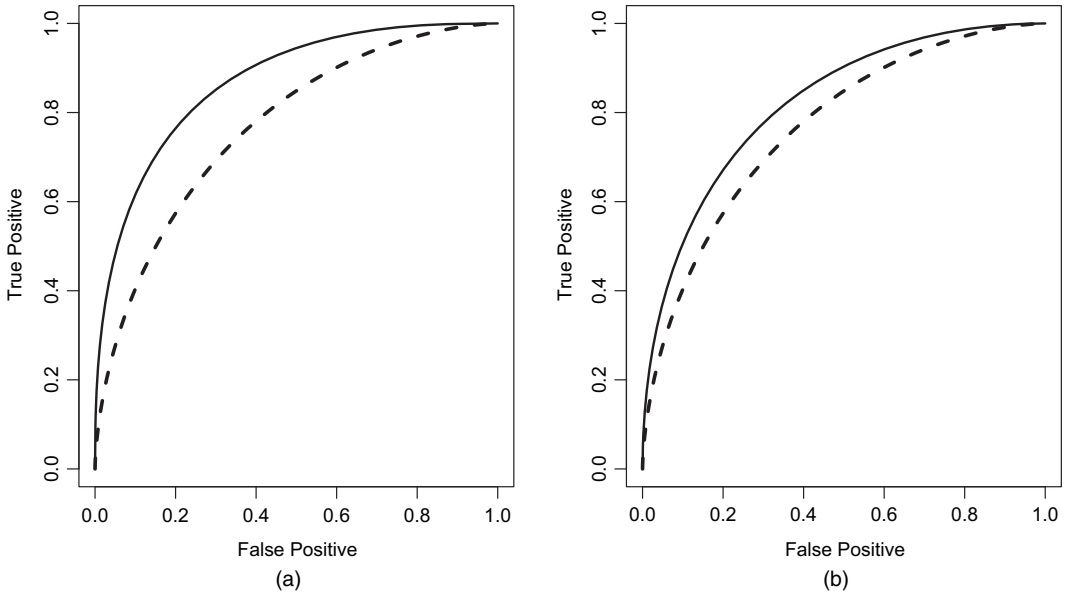

Fig. 3. Simulation study: average ROC curves on predicted outcomes on the test set, using the mixed graphical probit (———) under penalized estimation (with ρ_G set to the value closest to the true sparsity) and a conventional probit model (– – –) with no random effects: (a) $N = 250$, $G = 25$, $R = 50$; (b) $N = 250$, $G = 125$, $R = 50$

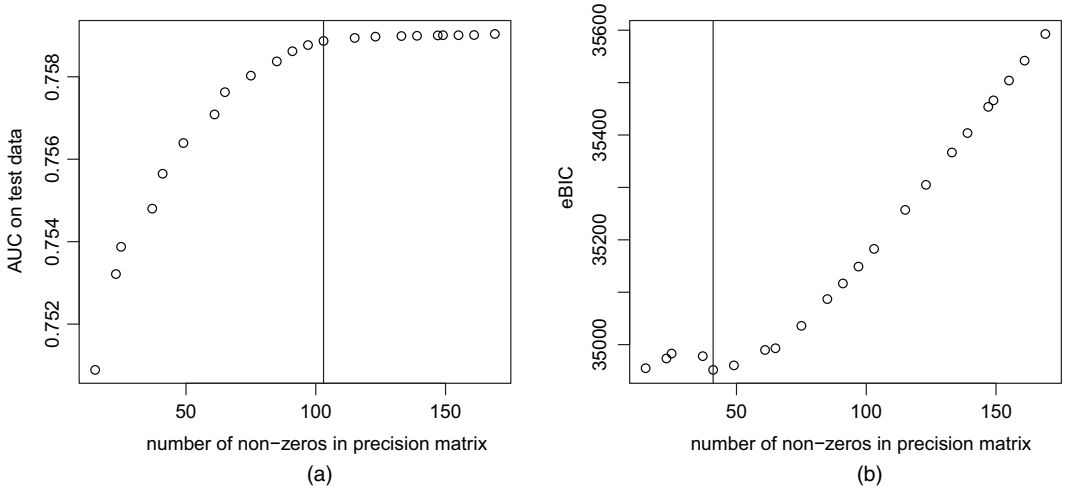

Fig. 4. Exploring the full path of solutions on the credit risk application by using two model selection criteria, (a) AUC on test data and (b) eBIC, for models fitted on the training data across a path of regularization parameters ρ_G

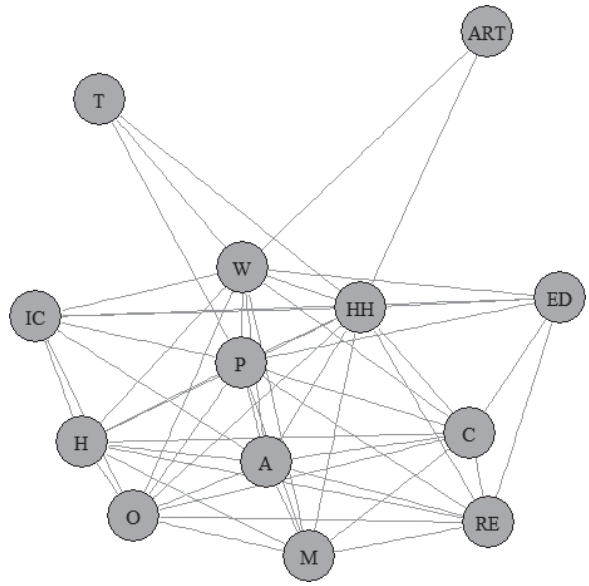

Fig. 5. Credit risk application: estimated network between sectors of economic activity by using L_1-penalization, choosing the ρ based on AUC on the test data ($\rho_G = 0.0006$): M, manufacturing; B, construction; W, wholesale and retail trade, repair of motor vehicles and motorcycles; T, transportation and storage; IC, information and communication; RE, real estate activities; P, professional, scientific and technical activities; A, administrative and support service activities; ED, education; H, human health and social work activities; ART, arts, entertainment and recreation; O, other service activities; HH, activities of households as employers

connected network, appears to decline significantly after fitting a model with a precision matrix with 103 non-zeros ($\rho_G = 0.0006$). We take the latter as the optimal model for subsequent analyses.

Firstly, we explore the estimated network, which gives an indication of the more connected sectors in the economy. This is plotted in Fig. 5, where links between any two sectors appear when there is a non-zero precision among them. It is interesting to see that the sectors that are more central to the network are those from real estate, manufacturing industry and the activities of households as employers, whereas we mostly find services activities sectors and, in particular, the sectors 'arts, entertainment and recreation' and 'transportation and storage' not highly connected.

Secondly, we consider the estimated regression coefficients from the fitted model. These are reported in Table 4 (column (I)), together with their standard errors which have been calculated by using the observed information matrix as described in Section 3.5 and further expanded on in the on-line supplementary material. In the remaining columns, we compare the estimates with those of a simpler mixed probit with uncorrelated random effects for sectors only (column (II), fitted with `glmer`) and of a conventional probit model without random effects, that ignores unobserved heterogeneity and that is often used in credit risk modelling (column (III)). Focusing on the estimates from the mixed graphical model (column (I)), the coefficient that is attached to cash over total assets is statistically significant with a negative sign, indicating that companies with higher cash reserves relative to current assets are less likely to default. The results also show a negative and statistically significant effect for the variable 'retained profits on total assets': the higher the net profits with respect to the investments made, the lower the probability that the firm will go bankrupt. The variable trade debt has a negative and significant coefficient, meaning

Table 4. Regression coefficients and standard errors estimated on the training sample of the credit risk application by using (I) the proposed credit risk mixed graphical model, (II) a mixed probit model with uncorrelated random effects per sector and (III) a conventional credit risk probit model

Variable	*(I) Results for mixed graphical probit*		*(II) Results for mixed probit (sectors)*		*(III) Results for conventional probit*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
Total liabilities/total assets	0.0201	0.0146	0.0218	0.0150	0.0360†	0.0146
Net worth/total liabilities	0.0005	0.0011	0.0001	0.0013	−0.0018	0.0013
Cash/total assets	−0.1115†	0.0359	−0.1157†	0.0409	−0.1018†	0.0396
Current liabilities/current assets	−0.0059	0.0088	−0.0063	0.0091	−0.0081	0.0089
Retained profits/total assets	−0.1420†	0.0248	−0.1428†	0.0250	−0.1465†	0.0254
Account receivable/total liabilities	−0.8015	1.3253	−0.7378	1.4844	−0.7570	1.4344
Trade credit/total liabilities	0.0132	0.0346	0.0354	0.0386	0.0437	0.0373
Trade debt/total assets	−0.2149†	0.0508	−0.2084†	0.0553	−0.1362†	0.0545
Size	−0.0855†	0.0039	−0.0807†	0.0053	−0.0616†	0.0050
Age	−0.1912†	0.0116	−0.1966†	0.0144	−0.2538†	0.0136
Age risk	0.0459	0.0246	0.0485	0.0252	0.0702†	0.0249
Regional gross domestic product	0.0146	0.0101	0.0131	0.0242	−0.0080	0.0234

†Significance at the 5% level.

Table 5. Classification performance of the credit risk models from Table 4 on the test sample

Model	*% correctly classified*	
	Non-failed	*Failed*
Mixed graphical probit	66.43	73.03
Mixed probit (sectors)	64.67	69.70
Conventional probit	62.99	66.54

that, the higher the money a company is expected to receive from other companies as a result of trade, the less likely the company is to default. Looking at the non-financial variables, the coefficients that are attached to size and age are significant and indicate that, as expected, larger and older companies have lower probabilities of default. Comparing with the results that are reported in columns (II) and (III), the inclusion of network effects in the probit model does not seem to change significantly the estimated coefficients, although the standard errors are slightly smaller for the method proposed. This has been observed also in related literature on multivariate probit models (Chib and Greenberg, 1998).

Finally, we compare the classification performance of the mixed graphical model with the same two models reported in Table 4. Table 5 reports the classification accuracy statistics on the hold-out sample. When adopting the mixed graphical probit, the overall classification accuracy is significantly improved. Given the high number of non-failed companies in the data, the mixed graphical probit is particularly good at identifying correctly companies that did not fail. This is confirmed also by the ROC curve in Fig. 6, where the ROC of the mixed graphical probit lies always above the ROC of the simpler mixed probit and of the conventional probit.

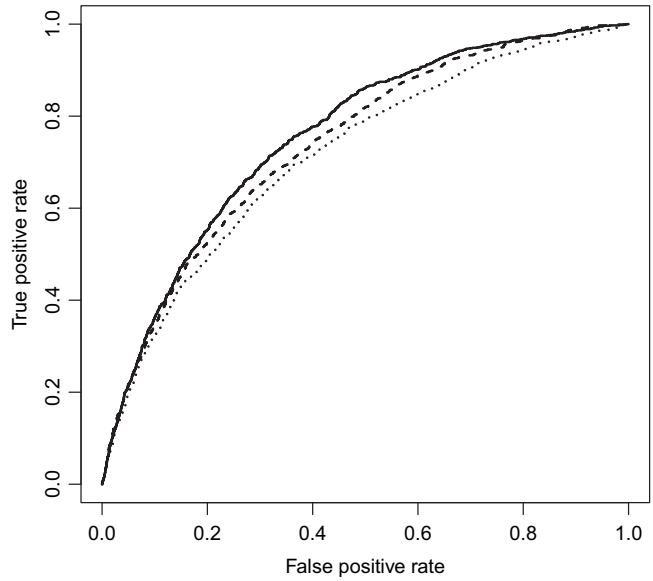

Fig. 6. ROC curves of predicted outcomes on the test sample: comparison between the mixed graphical probit (———), mixed probit (– – –) and conventional probit (· · · · · ·) on the credit risk application

6. Concluding remarks

In this paper we have proposed a computationally efficient EM algorithm for estimation of a mixed probit model with correlated group-specific effects and have shown its use in a credit risk application, for which existing approaches were prohibitively slow. We have proposed unconstrained and penalized likelihood estimation approaches for inference and have derived the observed information matrix and asymptotic standard errors of the estimates. The L_1-penalized approach is suitable for when the number of groups is large relative to the number of observations, for which ML fails, and/or when the recovery of the underlying network is of interest. If network recovery is not of interest but high dimensionality is present, other regularization methods can be used in the M-step of the EM algorithm proposed, such as by making use of a ridge penalty (Schäfer and Strimmer, 2005).

An extensive simulation study showed that our estimator has good finite sample properties and can be adopted for estimation and prediction using very large data sets, given its moderate computational costs. A large-scale credit risk application on a unique data set on SMEs, a setting in which credit risk modelling is currently underdeveloped, showed that accounting for network effects makes a significant contribution to increasing the default prediction power of risk models and therefore that efficient inferential procedures for these models are particularly useful in this field.

The R code to fit the mixed graphical probit that is described in this paper is available on GitHub (`https://github.com/veronicavinciotti/correlatedmixedprobit`).

Acknowledgements

The authors acknowledge financial support from the Engineering and Physical Sciences Research Council (EP/L021250/1). We thank the financial institution that provided the data,

George Foy for assisting with data retrieval and Francesco Moscone, Sergio di Cesare and Mark Lycett for helpful comments on the manuscript.

Appendix A: Moments of truncated normal distributions

We now provide the formulae for deriving the central and non-central moments of y_{ir}^* given $\mathbf{y}_{-i,r}^*, y_{ir}$. By the theorem on conditional normal distributions, y_{ir}^* given $\mathbf{y}_{-i,r}^*$ has a normal distribution with mean and variance

$$\tilde{\mu}_{ir} = \beta' \mathbf{x}_{ir} + \boldsymbol{\Sigma}_{r,i,-i} \boldsymbol{\Sigma}_{r,-i,-i}^{-1} (\mathbf{y}_{-i,r}^* - \mathbf{X}_{-i,r}\beta),$$
$$\tilde{\sigma}_{ir}^2 = \sigma_{ir}^2 - \boldsymbol{\Sigma}_{r,i,-i} \boldsymbol{\Sigma}_{r,-i,-i}^{-1} \boldsymbol{\Sigma}_{r,-i,i},$$

where σ_{ir}^2 is the (i,i)th element of $\boldsymbol{\Sigma}_r$. It follows that the conditional distribution of y_{ir}^* given $\mathbf{y}_{-i,r}^*, y_{ir}$ is a truncated normal distribution. Let $\xi_{ir,1} = (t_1 - \tilde{\mu}_{ir})/\tilde{\sigma}_{ir}$, $\xi_{ir,2} = (t_2 - \tilde{\mu}_{ir})/\tilde{\sigma}_{ir}$ and

$$\rho_{1,ir} = \frac{\phi(\xi_{ir,1}) - \phi(\xi_{ir,2})}{\Phi(\xi_{ir,2}) - \Phi(\xi_{ir,1})},$$

$$\rho_{2,ir} = \frac{\xi_{ir,2}\phi(\xi_{ir,1}) - \xi_{ir,1}\phi(\xi_{ir,2})}{\Phi(\xi_{ir,2}) - \Phi(\xi_{ir,1})}$$

with

$$t_1 = \begin{cases} 0, & \text{if } y_{ir} = 1, \\ -\infty, & \text{if } y_{ir} = 0, \end{cases}$$
$$t_2 = \begin{cases} \infty, & \text{if } y_{ir} = 1, \\ 0, & \text{if } y_{ir} = 0, \end{cases}$$

and ϕ and Φ are the density and cumulative distribution function respectively of a standard normal distribution. The first and second moments of y_{ir}^* given $\mathbf{y}_{-i,r}^*, y_{ir}$ are

$$\lambda_{i,1} = \tilde{\mu}_{ir} + \rho_{1,ir}\tilde{\sigma}_{ir},$$
$$\lambda_{i,2} = \tilde{\mu}_{ir}^2 + \tilde{\sigma}_{ir}^2 + 2\rho_{1,ir}\tilde{\sigma}_{ir}\tilde{\mu}_{ir} + \rho_{2,ir}\tilde{\sigma}_{ir}^2,$$

whereas the second, third and fourth central moments of y_{ir}^* given $\mathbf{y}_{-i,r}^*, y_{ir}$ are (see Horrace (2015))

$$\lambda_{i,2}^{\mathrm{c}} = \tilde{\sigma}_{ir}^2 - \tilde{\sigma}_{ir}\rho_{1,ir}\lambda_{i,1},$$
$$\lambda_{i,3}^{\mathrm{c}} = \tilde{\sigma}_{ir}\rho_{1,ir}(\lambda_{i,1}^2 - \lambda_{i,2}^{\mathrm{c}}),$$
$$\lambda_{i,4}^{\mathrm{c}} = 2\tilde{\sigma}_{ir}^4 - 3(\tilde{\sigma}_{ir}\rho_{1,ir}\lambda_{i,1}^{\mathrm{c}})^2 - \tilde{\sigma}_{ir}^{-1}\rho_{1,ir}\lambda_{i,3}^{\mathrm{c}} + \tilde{\mu}_{ir}^2\lambda_{i,2}^{\mathrm{c}}.$$

Appendix B: Conditional expectations

Using the law of iterated expectations we know that

$$E(\mathbf{u}_r|\mathbf{y}_r) = E\{E(\mathbf{u}_r|\mathbf{y}_r^*)|\mathbf{y}_r\},$$
$$E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r) = E\{E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r^*)|\mathbf{y}_r\}.$$

Noting that

$$\begin{pmatrix} \mathbf{u}_r \\ \mathbf{y}_r^* \end{pmatrix} \sim N\begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma}_G & \boldsymbol{\Sigma}_G\mathbf{Z}_r' \\ \mathbf{X}_r\beta & \mathbf{Z}_r\boldsymbol{\Sigma}_G & \boldsymbol{\Sigma}_r \end{pmatrix},$$

we can use the theorem on conditional normal distributions to obtain

$$E(\mathbf{u}_r|\mathbf{y}_r^*) = \boldsymbol{\Sigma}_G\mathbf{Z}_r'\boldsymbol{\Sigma}_r^{-1}(\mathbf{y}_r^* - \mathbf{X}_r\beta),$$

so that equation (3.9) holds. Similarly, focusing on $E(\mathbf{u}_r\mathbf{u}_r'|\mathbf{y}_r)$ and using again the theorem on conditional normal distributions we obtain equation (3.10).

References

Abegaz, F. and Wit, E. (2015) Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statist. Neerland.*, **69**, 419–441.

Alfo', M., Caiazza, S. and Trovato, G. (2005) Extending a logistic approach to risk modeling through semiparametric mixing. *J. Finan. Serv. Res.*, **28**, 163–176.

Altman, E. I. and Sabato, G. (2007) Modeling credit risk for SMEs: evidence from the US market. *Abacus*, **43**, 332–357.

Altman, E. I., Sabato, G. and Wilson, N. (2010) The value of non-financial information in small and medium-sized enterprise risk management. *J. Credt Risk*, **6**, 1–33.

An, X. and Bentler, P. M. (2012) Efficient direct sampling MCEM algorithm for latent variable models with binary responses. *Computnl Statist. Data Anal.*, **56**, 231–244.

Andrews, D. (2005) Cross section regression with common shocks. *Econometrica*, **73**, 1551–1585.

Ashford, J. R. and Sowden, R. R. (1970) Multivariate Probit analysis. *Biometrics*, **26**, 535–546.

Augugliaro, L., Abbruzzo, A. and Vinciotti, V. (2018) l_1-penalised censored Gaussian graphical model. *Biostatistics*, to be published.

Barreto, G. and Artes, F. R. (2013) Spatial correlation in credit risk and its improvement in credit scoring. *Working Paper WPE: 321/2013*. Instituto de Ensino e Pesquisa, São Paulo.

Barro, D. and Basso, A. (2010) Credit contagion in a network of firms with spatial interaction. *Eur. J. Oper. Res.*, **205**, 459–468.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *J. Statist. Softwr.*, **67**, 1–48.

Battiston, S., Delli Gatti, D., Gallegati, M., Greenwald, B. and Stiglitz, J. E. (2007) Credit chains and bankruptcy propagation in production networks. *J. Econ. Dynam. Control*, **31**, 2061–2084.

Behrouzi, P. and Wit, E. (2019) Detecting epistatic selection with partially observed genotype data by using copula graphical models. *Appl. Statist.*, **68**, 141–160.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Breslow, N. E. and Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Campbell, J. Y., Hilscher, J. and Szilagyi, J. (2008) In search of distress risk. *J. Finan.*, **63**, 2899–2939.

Carbó-Valverde, S., Rodríguez-Fernández, F. and Udell, G. (2016) Trade credit, the financial crisis, and SME access to finance. *J. Mony Credt Bankng*, **48**, 113–143.

Carling, K., Jacobson, T., Linde, J. and Roszbach, K. (2007) Corporate credit risk modelling and the macroeconomy. *J. Bankng Finan.*, **31**, 845–868.

Chan, J. S. K. and Kuk, A. Y. C. (1997) Maximum likelihood estimation for Probit-linear mixed models with correlated random effects. *Biometrics*, **53**, 86–97.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate Probit models. *Biometrika*, **85**, 347–361.

Delli Gatti, D., Gallegati, M., Greenwald, B., Russo, A. and Stiglitz, J. E. (2006) Business fluctuations in a credit-network economy. *Physica* A, **370**, 68–74.

Foygel, R. and Drton, M. (2010) Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 23* (eds J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel and A. Culott), pp. 604–612. Red Hook: Curran Associates.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Gueorguieva, R. V. and Agresti, A. (2001) A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Am. Statist. Ass.*, **96**, 1102–1112.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2015) Graphical models for ordinal data. *J. Computnl Graph. Statist.*, **24**, 183–204.

Horrace, W. C. (2015) Moments of the truncated normal distribution. *J. Productvty Anal.*, **43**, 133–138.

Hsiao, C. (2003) *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Ibrahim, J. G., Zhu, H. and Tang, N. (2008) Model selection criteria for missing-data problems using the EM algorithm. *J. Am. Statist. Ass.*, **103**, 1648–1658.

Jacobson, T., Lind, J. and Roszbach, K. (2013) Firm default and aggregate fluctuations. *J. Eur. Econ. Ass.*, **11**, 945–972.

Jones, S. and Hensher, D. A. (2004) Predicting firm financial distress: a mixed logit model. *Accountng Rev.*, **79**, 1011–1038.

Kotecha, J. and Djuric, P. (1999) Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. *IEEE Comput. Soc.*, **3**, 1757–1760.

Kukuk, M. and Ronnberg, M. (2013) Corporate credit default models: a mixed logit approach. *Rev. Quant. Finan. Accountng*, **40**, 467–483.

Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Oxford University Press.

Lee, L. (1979) On the first and second moments of the truncated multi-normal distribution and a simple estimator. *Econ. Lett.*, **3**, 165–169.

Leppard, P. and Tallis, G. M. (1989) Algorithm AS 249: Evaluation of the mean and covariance of the truncated multinormal distribution. *Appl. Statist.*, **38**, 543–553.
Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.
McCulloch, C. (1994) Maximum likelihood variance components estimation for binary data. *J. Am. Statist. Ass.*, **89**, 330–335.
McCulloch, C. (1997) Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–170.
Meng, X. L. and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.
Moscone, F., Tosetti, E. and Vinciotti, V. (2017) Sparse estimation of huge networks with a block-wise structure. *Econmetr. J.*, **20**, S61–S85.
Sabato, G. (2010) Credit scoring. In *Encyclopedia of Quantitative Finance*. Chichester: Wiley.
Schäfer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Molec. Biol.*, **4**, no. 1.
Schilling, S. and Bock, R. D. (2005) High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, **70**, 533–555.
Tallis, G. M. (1961) The moment generating function of the truncated multi-normal distribution. *J. R. Statist. Soc.* B, **23**, 223–229.
Tan, M., Tian, G. and Fang, H. (2007) An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data. *J. Statist. Computn Simuln*, **77**, 929–943.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: A computationally efficient correlated mixed probit model for credit risk inference'.