Dynamic modelling of corporate credit ratings and defaults

Vana Gür, Laura; Hornik, Kurt

Published in:
Statistical Modelling

DOI:
10.1177/1471082X211057610

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Statistical Modelling xxxx; **xx(x)**: 1–19

Dynamic modelling of corporate credit ratings and defaults

Laura Vana[1] and Kurt Hornik[1]
[1]Department of Finance, Accounting and Statistics, Institute for Statistics and Mathematics, Vienna University of Economics and Business, Austria

Abstract: In this article, we propose a longitudinal multivariate model for binary and ordinal outcomes to describe the dynamic relationship among firm defaults and credit ratings from various raters. The latent probability of default is modelled as a dynamic process which contains additive firm-specific effects, a latent systematic factor representing the business cycle and idiosyncratic observed and unobserved factors. The joint set-up also facilitates the estimation of a bias for each rater which captures changes in the rating standards of the rating agencies. Bayesian estimation techniques are employed to estimate the parameters of interest. Several models are compared based on their out-of-sample prediction ability and we find that the proposed model outperforms simpler specifications. The joint framework is illustrated on a sample of publicly traded US corporates which are rated by at least one of the credit rating agencies S&P, Moody's and Fitch during the period 1995–2014.

Key words: Bayesian inference; credit risk modelling, credit ratings, default probability, longitudinal model, multivariate ordinal data

Received August 2020; **revised** August 2021; **accepted** October 2021

1 Introduction

The last decades have witnessed an increased interest from practitioners in the financial industry, researchers and regulators alike in developing tools for appropriately measuring and modelling credit risk as well as developing and amending regulations which limit and monitor such risks. In this context, credit risk assessment typically relies on statistical models based on a historical database of defaults together with debtor-specific and market variables or on credit ratings which are forward-looking opinions of a debtor's creditworthiness and are assigned by external credit rating agencies (CRAs). This approach is also reflected in the regulations introduced in the Basel Accords I and II (Basel Committee on Banking Supervision, 2004, 2011). For example, under Pillar I of Basel II financial

Address for correspondence: Laura Vana, Department of Finance, Accounting and Statistics, Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, Building D4, 4th Floor, AT-1020, Vienna, Austria.
E-mail: laura.vana@wu.ac.at

10.1177/1471082X211057610

intermediaries can develop their own default prediction models, which, in case the history of defaults is limited and portfolio coverage is satisfactory, can be enhanced or replaced by models using external rating information. However, clear guidelines on how to integrate all available information sources into a proper modelling framework are scarce. More recently, the need of such an integrated approach has been stressed by both academia (e.g., Hilscher and Wilson, 2017; Hirk et al., 2020) and regulators with the IFRS 9 accounting standard issued by the International Accounting Standards Board (IASB, 2014) requiring banks to build provisions based on forward-looking expected loss models by considering 'all reasonable and supportable information, including forward-looking measures'.

In this article, we extend the work of Hirk et al. (2020) and propose a framework for modelling defaults and credit ratings from Standard and Poor's (S&P), Moody's and Fitch in a multivariate ordinal model, where the latent credit quality is modelled as a dynamic process. The dynamic modelling of credit risk allows for dependencies in the cross-section and over time to be accounted for by typically making a distinction between systematic and idiosyncratic risk, where the systematic risk relates to the relationship between credit risk and a business factor and is of prime importance in portfolio credit risk modelling (see Vasicek, 2002; Koopman and Lucas, 2005; McNeil and Wendin, 2007; Betz et al., 2018). We therefore build a longitudinal model of binary default events and ratings on an ordinal scale where a common latent variable which is a measure of credit quality is underlying the observations. In particular, we model the conditional distribution of these responses given a set of financial covariates known to be relevant for credit risk modelling by assuming that the latent variable corresponding to the credit quality, referred to in this article as a 'probability of default (PD) score', depends on unobserved firm-specific effects, and on a systematic as well as an idiosyncratic factor which both have an auto-regressive structure of order one. This approach allows us to disentangle differences in firms' credit quality due to idiosyncratic causes from the effects due to business conditions. In modelling the credit ratings we also consider several characteristics of the ratings market. First, CRAs claim to provide a forward-looking long-term measure of credit quality by employing a so-called 'through-the-cycle' (TTC) approach to ensure that their ratings are stable over the business cycle (as opposed to a 'point-in-time' (PIT) approach, which measures credit quality at a given point in time). Second, we do not disregard the criticism of the three big CRAs for their inability to assess risk accurately (e.g., Becker and Milbourn, 2011; Bolton et al., 2012; Bar-Isaac and Shapiro, 2013; Kashyap and Kovrijnykh, 2016) and, from the modelling perspective, we assume the ratings to be noisy observations of the underlying PD score by assuming a rater 'bias' for each of the CRAs which depends on covariates and has a time-varying component common to all the CRAs which captures yearly shifts in the rating standards of the rating agencies. Here, we resort to Bayesian estimation techniques implemented in the open-source software Stan (Stan Development Team, 2018) to estimate the parameters of interest and illustrate the dynamic framework on a subset of the COMPUSTAT-CRSP universe of publicly traded US firms which are rated by at least one of the big three CRAs over the period 1995–2014.

Over the last decade, joint modelling frameworks for credit risk measures have become more popular but it is still common practice in both industry and academia to model defaults and credit ratings separately. Statistical binary response (typically logit) models are often employed for predicting defaults (among the most prominent articles in the finance literature, for example Shumway, 2001; Campbell et al., 2008; Tian et al., 2015), while static ordinal or linear regression is used in modelling the credit ratings (e.g., Alp, 2013; Baghai et al., 2014). Ordered regression models with a dynamic specification have also been employed for modelling rating transitions (e.g., Malik and Thomas, 2012; Creal et al., 2014). Several articles have jointly investigated different credit risk measures, including credit ratings from possibly various raters for the purpose of credit risk measurement. For example, Hornik et al. (2010) propose a static parametric framework based on the existence of contemporaneous PD estimates for the same obligor provided by different rating sources for estimating consensus ratings as well as validate the different rating sources by analyzing the mean/variance structure of the rating errors. Grün et al. (2013) extend the analysis to a dynamic model and analyze a panel of ratings from the three big CRAs by first transforming ratings to PD estimates by using observed default rates. Creal et al. (2014) build a multivariate dynamic factor model for signal extraction and forecasting of macro, credit, and loss given default risk conditions in the US. Hilscher and Wilson (2017) provide an analysis of both ratings and defaults (even though not in a joint statistical model) where they investigate whether the measures have different information content. They conclude that, while the credit ratings are poor measures of raw default probabilities, they are strongly related to systematic risk. Hirk et al. (2018) build a joint ordinal model for the ratings from the big three CRAs while Hirk et al. (2020) propose a static framework for jointly modelling defaults and ratings using the class of multivariate ordinal regression models and show improved results in terms of default prediction conditional on the observed ratings.

The article is organized as follows: The joint modelling framework is introduced in Section 2 and details regarding the estimation and prior specifications are presented in Section 3. Section 4 introduces the data set used in the analysis and presents the results while Section 5 concludes the article.

2 Modelling framework

2.1 General set-up

Suppose that for each firm $i \in \{1, \dots, I\}$ in period $t \in T_i$, with T_i being the set of all available time points for firm i, we observe the corresponding $(P \times 1)$ vector of covariates $\boldsymbol{x}_i(t)$ measuring firm liquidity, profitability, indebtedness, a binary default indicator $y_i^{\text{def}}(t)$ which takes the value one if the firm defaulted between time t and $t+1$, and a vector of available credit ratings $\boldsymbol{y}_i^{\text{rat}}(t) = [y_{ij}^{\text{rat}}(t)]_{j \in J_i(t)}$ which are observed at the end of period t for a non-empty subset $J_i(t)$ of all available raters {S&P, Moody's, Fitch}.

The one-year probability of default of firm i at time t denoted by $\mathrm{PD}_i(t)$ is modelled as a random variable which follows a logistic regression model [Logistic regression is widely employed in the credit risk literature (e.g., Campbell et al., 2008; Tian et al., 2015).] and we assume that, conditional on $\mathrm{PD}_i(t)$, the default indicator $y_i^{\mathrm{def}}(t)$ follows a Bernoulli distribution:

$$\begin{aligned} &y_i^{\mathrm{def}}(t) \sim \mathrm{Bernoulli}(\mathrm{PD}_i(t)),\\ &\mathrm{PD}_i(t) = P(y_i^{\mathrm{def}}(t) = 1 | S_i(t)) = 1/(1 + \exp(-S_i(t))), \end{aligned}$$

where $S_i(t)$ is a real-valued score indicating the credit quality of firm i at time t with a high (low) value implying a low (high) credit quality. We refer to this quantity as the 'one-year PD score'.

For the credit ratings, we employ an ordinal regression model using the cumulative link approach by treating the ratings observed for the jth rater $y_{ij}^{\mathrm{rat}}(t)$ which can take one of C_j classes as a coarser version of an underlying latent variable $\widetilde{y}_{ij}^{\mathrm{rat}}(t)$, which can be interpreted as a real-valued rating score. A regression model is then assumed on the latent scale:

$$y_{ij}^{\mathrm{rat}}(t) = r \iff \theta_{j,r-1} < \widetilde{y}_{ij}^{\mathrm{rat}}(t) \leq \theta_{j,r}, \qquad \widetilde{y}_{ij}^{\mathrm{rat}}(t) = \underbrace{S_i(t) + \eta_{ij}(t)}_{\mu_{ij}(t)} + \varepsilon_{ij}(t), \tag{2.1}$$

where $\boldsymbol{\theta}_j = (\theta_{j,0}, \theta_{j,1}, \dots, \theta_{j,C_j})^\top$ is a set of rater-specific threshold parameters satisfying the order restriction $\theta_{j,0} = -\infty < \theta_{j,1} < \cdots < \theta_{j,C_j} = \infty$ which are used to slot the underlying variable into intervals corresponding to the non-default ordinal rating classes $r \in \{1, \dots, C_j\}$ (where 1 denotes the class with best creditworthiness); $\eta_{ij}(t)$ is a rater-specific bias term; $\varepsilon_{ij}(t) \overset{iid}{\sim} \mathcal{L}(0, 1)$ is a mean-zero noise term which follows a *standard* (to ensure identifiability) logistic distribution.

By allowing rater-specific thresholds we are able to capture the heterogeneity in the rating scale and methodology of the different CRAs. [The raters employ different coding for the rating classes: S&P and Fitch employ a rating scale with eight main non-default rating categories *AAA, AA, A, BBB, BB, B, CCC, CC* while Moody's scale is *Aaa, Aa, A, Baa, Ba, B, Caa, Ca*. Moreover, the CRAs claim to use different credit risk measures in their assessments: Moody's relies on loss given default while S&P and Fitch measure the relative likelihood of default.] Furthermore, in this application one may think of the term $\mu_{ij}(t)$ in Equation (2.1) as the 'expected rating score' assigned by rater j to firm i in year t, which the raters then transform to an ordinal scale using a suitable mapping. In the specification of expected rating score, we assume that the raters observe and/or produce noisy versions of the 'one-year PD score' $S_i(t)$ and that the CRAs' biases $\eta_{ij}(t)$ can be modelled additively on the scale of the underlying latent variables. The noisiness in the ratings can be motivated on the one hand by some sort of information asymmetry between firm owners and raters and on the other hand by the fact that the ratings are forward-looking measures of creditworthiness for horizons longer than one year which are assigned by the CRAs based on different metrics (see, e.g., Grün et al., 2013). On the other hand, the assumption of additivity on the rating score scale is in line with previous specifications of rating models (e.g., Alp, 2013; Baghai et al., 2014) but also by Merton-type models under partial information, where the error in the observation of the log firm value is additive on the scale of the log normal firm value process (first introduced in Duffie and Lando, 2001).

2.2 Model specification

2.2.1 Dynamic specification of the latent PD score $S_i(t)$

The basic framework underlying dynamic models of credit risk also adopted by regulators (see, e.g., the methodology underlying the CreditMetrics$^{\text{TM}}$ framework) is that credit risk depends on a systematic and an idiosyncratic component and a Gaussian single factor model is typically employed in the modelling process (Vasicek, 2002). Moreover, it is reasonable to assume that the PD scores are correlated over the time dimension owing to macroeconomic events such as recessions whose influence is not fully captured by the covariates or over the cross-section from direct effects of one corporate failure on other distressed corporations. We propose the following dynamic specification of the latent one-year PD score:

$$S_i(t) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i(t) + u_i(t), \tag{2.2}$$

$$u_i(t) = a_i - \omega b(t) + \epsilon_i(t), \tag{2.3}$$

$$a_i \overset{iid}{\sim} N(0, \tau_i^2), \qquad b(t) = \phi_b b(t-1) + \upsilon_b(t), \qquad \upsilon_b(t) \overset{iid}{\sim} N(0,1),$$

$$\epsilon_i(t) = \rho\epsilon_i(t-1) + \xi_i(t), \qquad \xi_i(t) \overset{iid}{\sim} N(0, \psi^2),$$

where β_0 is an intercept term, $\boldsymbol{\beta}$ is a $(P \times 1)$ vector of regression coefficients and $u_i(t)$ is a random effect which consists of a firm-specific effect a_i, a latent market factor $b(t)$ and idiosyncratic changes $\epsilon_i(t)$, and the loading ω measures the dependence of $S_i(t)$ on the latent market factor (a similar specification has been proposed in Grün et al., 2013). The effects a_i are firm-specific deviations from the overall intercept which can capture unobserved heterogeneity such as management ability, and are assumed to be normally distributed with a firm-specific variance. An auto-regressive structure of order one is assumed for the latent market factor $b(t)$ whereas the above sign convention implies that positive $b(t)$'s correspond to favourable market conditions. The modelling of the systematic factor as a latent quantity is rather standard in the literature, owing to the fact that the theory on which observed variables would be adequate as a proxy for systematic credit risk is rather scarce (see remarks in, e.g., Koopman and Lucas, 2005). By restricting ω to be constant for all firms and years we implicitly assume that $b(t)$ is indeed a common market factor impacting all firms equally, while the remaining unexplained variation can be captured by the idiosyncratic effect $\epsilon_i(t)$. [We also investigated whether the model improves when allowing the factor loading to depend on the industry in which firm i operates but find no compelling improvement in the performance.] The variance of $\upsilon_b(t)$ is fixed to one to ensure identifiability and we restrict $|\phi_b| < 1$ to ensure stationarity. Finally, the idiosyncratic disturbances $\epsilon_i(t)$ are independent over the firms but are serially correlated through an AR(1) process with $|\rho| < 1$ for ensuring stationarity. Due to the rather short history for most firms in the sample (see also Figure A.1 in the Supplementary Material) the data might not deliver enough information on identifying a firm-specific persistence or standard-deviation parameters, this is why we assume a constant ρ and ψ for all firms.

2.2.2 Dynamic specification of the rater bias $\eta_{ij}(t)$

The rater bias specification proposed in this article is given by:

$$\eta_{ij}(t) = \boldsymbol{\gamma}_j^\top \boldsymbol{x}_i(t) + \lambda_j \delta(t), \qquad \delta(t) = \phi_\delta \delta(t-1) + \upsilon_\delta(t), \qquad \upsilon_\delta(t) \overset{iid}{\sim} N(0,1),$$

where $\boldsymbol{\gamma}_j$ is a rater-specific $(P \times 1)$ vector of regression coefficients, $\delta(t)$ is a time-varying 'rater factor' which is modelled dynamically using an AR(1) process with $|\phi_\delta| < 1$ and λ_j is a rater-specific factor loading. The rater-specific bias $\eta_{ij}(t)$ depends linearly on the covariates $\boldsymbol{x}_i(t)$, which were also employed in the specification of the latent PD score $S_i(t)$ in Equation (2.2). This assumption is reasonable as these risk factors also affect the credit ratings, but to a different extent than they affect defaults. However, one could include the lagged ratings as a proxy for the 'stickiness' of credit ratings or the number of years that the firm has been rated by a CRA (having the firm as a client for a longer time period can potentially reduce information asymmetry) as potential covariates. We investigated whether these variables improve the model performance and find that, considering these additional covariates does not improve the results markedly. The specification above does not account for any dependence among the three raters, which might seem rather restrictive as lead-lag relationships among the raters have been found empirically (e.g., Güttler and Wahrenburg, 2007; Berwart et al., 2019). However, the rating adjustments are reported to follow within months of the lead rating change, so this effect can be expected to be negligible on an yearly basis.

The rater factor $\delta(t)$ is independent of observed covariates and of $S_i(t)$, and should thus pick up any time variation specific to the CRAs' behaviour and industry practices beyond that implied by the variation of the PD scores caused by the market factor. The reason for including $\delta(t)$ in the model lies mainly in previous results showing shifts in the rating standards over the sample period analyzed in this article (e.g., Alp, 2013). Moreover, given the low number of defaults in the sample, omitting the rater factor from the model specification might make the estimation of the business factor cumbersome given the strong (weak) signal in the data coming from the ratings (defaults). For the sake of parsimony, we do not employ one factor for each rating agency, as we expect the behaviour of the three CRAs to be similar given the oligopoly structure of the ratings' market and the high degree of agreement among the raters.

3 Bayesian inference

Several methods can be considered for the estimation of the proposed model, which contains effects at different levels of hierarchy. The non-nested firm and time effects, however, restrict the techniques which can be employed, as the estimation of ordinal models with non-nested effects is cumbersome due to the necessity to compute high-dimensional integrals. The maximum likelihood approach using the EM algorithm can be employed (e.g., Cagnone et al., 2009, employ the EM algorithm in estimating a multivariate ordinal model with item and time-specific random effects). McCulloch (1994) proposed the inclusion of a Metropolis Hastings step in the E-step of the EM algorithm so that the required expectation can be approximated by the average of Monte Carlo samples from the target distribution (approach used by, e.g., Xie et al., 2013, in a two-level model). Bellio and Varin (2005) tackle the dimensionality issue by maximizing the product of the pairwise marginal likelihoods and estimate the parameters of a two-level generalized linear mixed model (GLMM) with crossed random effects. The Bayesian framework is an attractive alternative for multi-level models with (crossed) effects at different levels of hierarchy. Through the specification of a prior distribution, Bayesian estimates of the effects can be obtained even in cases where there are few data points per group Moreover, the growing number of available open software tools for performing Bayesian inference make the implementation and estimation of such models more accessible.

3.1 Posterior

The posterior, which is proportional to the product of the likelihood of the four responses and the prior densities on all unobservables (i.e., parameters and latent quantities), is the object of interest in the analysis and inference relies on samples drawn from this posterior distribution. Samples from the posterior are drawn by using the package rstan (Stan Development Team, 2019) for R (R Core Team, 2020), which is an interface to the open-source software Stan. Stan is a probabilistic C++ library which provides full Bayesian inference through Markov chain Monte Carlo (MCMC) methods to obtain posterior simulations. In order to investigate how well the parameters of the proposed model can be estimated empirically, we conducted a simulation study which is presented in Section A.3 of the Supplementary Material.

Conditional on the latent PD scores $S_i(t)$, the responses are independent over all $i \in \{1, \dots, I\}$ and $t \in T_i$ and the joint likelihood is the product of the individual likelihoods of the responses.

$$p(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\zeta}) = \prod_{i=1}^{I}\prod_{t\in T_i}\left\{p(y_i^{\text{def}}(t)|\boldsymbol{x}_i(t), \boldsymbol{\zeta})\prod_{j\in J_i} p(Y_{ij}(t)|\boldsymbol{x}_i(t), \boldsymbol{\zeta})\right\}.$$

The term of the likelihood corresponding to the default indicator is given by the Bernoulli probability mass function $p(y_i^{\text{def}}(t)|S_i(t)) = \text{PD}_i(t)^{y_i^{\text{def}}(t)}(1 - \text{PD}_i(t))^{1-y_i^{\text{def}}(t)}$, while the likelihood of the ordinal responses is given by the probability of observing category $r_{ij}(t)$: $p(y^{\text{rat}}_{ij}(t)|\boldsymbol{x}_i(t), \boldsymbol{\zeta}) = \prod_{r=1}^{C_j} P(y^{\text{rat}}_{ij}(t) = r|\boldsymbol{x}_i(t), \boldsymbol{\zeta})^{\mathbb{1}\{r=r_{ij}(t)\}}$, where $\mathbb{1}\{.\}$ denotes the indicator function and for the sake of notational simplicity $\boldsymbol{\zeta}$ denotes unobservables. In the cumulative link logit model, the probability can be rewritten as $P(y^{\text{rat}}_{ij}(t) = r|\cdot) = P(y^{\text{rat}}_{ij}(t) \leq r|\cdot) - P(y^{\text{rat}}_{ij}(t) \leq r-1|\cdot)$ with $P(y^{\text{rat}}_{ij}(t) \leq r|\boldsymbol{x}_i(t), \boldsymbol{\zeta}) = \text{logit}^{-1}(\theta_{j,r} - S_i(t) - \eta_{ij}(t))$.

Priors are set on all model parameters. We find the results to be insensitive to the prior specified on the coefficients in the multivariate ordinal regression, which is to be expected given that the covariates were pre-selected based on their relevance in the existing literature. For the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$ of the standardized covariates, we proceed with the Student-t prior. For the threshold parameters, we employ a Dirichlet prior on the probability of the ordinal outcome falling in each of the C_j categories $\boldsymbol{\pi}_j \sim \text{Dirichlet}(\boldsymbol{\alpha}_j)$ and obtain the thresholds by the transformation $\theta_{j,r} = \text{logit}^{-1}\left(\sum_{i=1}^{r} \pi_{j,i}\right)$. For the firm-specific intercepts a_i a realistic assumption in our dataset is to expect the a_i's to have different variances $a_i \sim N(0, \tau_i^2)$ as we expect firms to have different variability in their creditworthiness. We separate the prior on the regression coefficients from the prior on the random firm-specific intercepts a_i, especially due to information imbalance (the number of observations available for identifying a_i is given by the number of years each firm spends in the sample, which for the analyzed sample ranges from 1 to 20, as can be seen in Figure A.1 in the Supplementary Material). Individual shrinkage of the firm effects is achieved by imposing a shrinkage prior which has a hierarchical representation. We consider here a non-Gaussian shrinkage prior on a_i with $\tau_i \sim N(0, q^2)$, where we treat q^2 as a hyper-parameter of the shrinkage prior above with an inverse Gamma distribution $q^2 \sim \mathcal{G}^{-1}(c_0, C_0)$ (for more details see, e.g., Frühwirth-Schnatter and Wagner, 2011). The prior on the persistence parameters ϕ_b, ϕ_δ and ρ is chosen as a scaled beta distribution $\frac{x+1}{2} \sim \text{Beta}(a^x, b^x)$, where the hyper-parameters are chosen to reflect prior information. For all other parameters, we use the improper prior $p(x) = 1/x$.

3.2 Model evaluation and out-of-time prediction

In order to evaluate the performance of the model in terms of out-of-sample prediction, we use two approaches. First, we employ approximate leave-one-out (LOO) cross-validation methods (as proposed in Vehtari et al., 2017). Second, we perform a K-fold cross-validation exercise adapted to the panel data structure. The difference between the two approaches is that the K-fold cross-validation requires refitting the model K times whereas approximate LOO methods require only one evaluation of the model. K-fold cross-validation has, however, the advantage that it allows to check the ability of the model to perform well out-of-time, in which case LOO methods are not suitable for assessing the performance on unseen time points (see, e.g., discussion in Vehtari and Ojanen, 2012). The details regarding the computation of the Bayesian LOO estimate of expected log pointwise predictive density employed here can be found in Vehtari et al. (2017). In the following, we proceed with the exposition of the out-of-time prediction exercise.

Assume we fit model M on the period up to a time t which implies we observe the ratings and the covariates up to time t and we also observe whether the company defaulted in the year following the rating observations. We denote the responses and the covariates observed up to time t by $\boldsymbol{Z}^{\text{o}}_{[1:t]} = \{(y^{\text{def}}_i(\tau), y^{\text{rat}}_{ij}(\tau)); \tau = 1, \ldots t, i \in I_t, j \in J\}$ and $\boldsymbol{X}_{[1:t]} = \{\boldsymbol{x}_i(\tau); \tau = 1, \ldots t, i \in I_t\}$, respectively, where I_t denotes the set of observed firms up to time t. We evaluate the predictive performance of M by making use of the posterior predictive density, which for a future data point containing the default and rating observations $\boldsymbol{z}_* = (y^{\text{def}}_*, \boldsymbol{y}^{\text{rat}\top}_*)^\top$ and a future set of covariates $\boldsymbol{x}_*$ is given by:

$$p(\boldsymbol{z}_*|\boldsymbol{x}_*, \boldsymbol{Z}^{\text{o}}_{[1:t]}, \boldsymbol{X}_{[1:t]}, M) = \int p(\boldsymbol{z}_*|\boldsymbol{x}_*, \boldsymbol{\zeta}, \boldsymbol{Z}^{\text{o}}_{[1:t]}, \boldsymbol{X}_{[1:t]}, M)p(\boldsymbol{\zeta}|\boldsymbol{Z}^{\text{o}}_{[1:t]}, \boldsymbol{X}_{[1:t]}, M)\text{d}\boldsymbol{\zeta}. \quad (3.1)$$

In addition to the joint posterior predictive densities, for the application case it is also relevant to assess whether adding the information about the ratings at the end of each year conditionally improves the prediction of the default component (see, e.g., Hirk et al., 2020). Hence, for a future default observation y^{def}_* we compute the conditional default probability implied by the model, that is, the one-year probability of default conditional on the corresponding ratings $\boldsymbol{y}^{\text{rat}}_*$ taking values $r_1, \ldots r_J$:

$$\begin{aligned}&P(y^{\text{def}}_* = 1|y^{\text{rat}}_{*,1} = r_1, \ldots, y^{\text{rat}}_{*,J} = r_J, \boldsymbol{x}^*, \boldsymbol{Z}^{\text{o}}_{[1:t]}, \boldsymbol{X}_{[1:t]}, M) = \\ &\qquad \frac{P(y^{\text{def}}_* = 1, y^{\text{rat}}_{*,1} = r_1, \ldots, y^{\text{rat}}_{*,J} = r_J|\boldsymbol{x}^*, \boldsymbol{Z}^{\text{o}}_{[1:t]}, \boldsymbol{X}_{[1:t]}, M)}{P(y^{\text{rat}}_{*,1} = r_1, \ldots, y^{\text{rat}}_{*,J} = r_J|\boldsymbol{x}^*, \boldsymbol{Z}^{\text{o}}_{[1:t]}, \boldsymbol{X}_{[1:t]}, M)},\end{aligned}$$

where both the denominator and the numerator can be rewritten as integrals similar to the one in Equation (3.1).

The predictive performance of M can then be measured by evaluating the above quantities at specific observations. For this purpose, we employ Bayesian cross-validation adapted to account for the time dimension of the data at hand. More specifically, we employ a one-step-ahead out-of-time exercise on an expanding window basis where, for increasing t, we repeatedly partition the data into a training set $(\boldsymbol{Z}^{\text{train}}_{[1:t]}, \boldsymbol{X}^{\text{train}}_{[1:t]})$ and a one-step-ahead test set $(\boldsymbol{Z}^{\text{test}}_{t+1}, \boldsymbol{X}^{\text{test}}_{t+1})$. After fitting the model M to the training set, given the resulting posterior distribution $p(\boldsymbol{\zeta}|\boldsymbol{Z}^{\text{train}}_{[1:t]}, \boldsymbol{X}^{\text{train}}_{[1:t]}, M)$ we evaluate the fit by evaluating for each firm the predictive densities or conditional probabilities at the test data.

Given that the integrals such as the one in (3.1) cannot be solved analytically, they may, however, be approximated through Monte Carlo integration. Assuming the posterior distribution $p(\boldsymbol{\zeta}|\boldsymbol{Z}^{\text{train}}_{[1:t]}, \boldsymbol{X}^{\text{train}}_{[1:t]}, M)$ can be summarized by S simulation draws $\boldsymbol{\zeta}^s$, we calculate the following (one-step-ahead) predictive measures:

- $\log p(z^{\text{test}}_{i,t+1}|\boldsymbol{x}^{\text{test}}_{i,t+1}, \boldsymbol{Z}^{\text{train}}_{[1:t]}, \boldsymbol{X}^{\text{train}}_{[1:t]}, M)$, the joint log predictive density of firm i in the test sample at $t+1$, which is approximated by the Monte Carlo estimate $\text{LPD}_{i,t+1}(M) \approx \log \frac{1}{S}\sum_{s=1}^{S} p(z^{\text{test}}_{i,t+1}|\boldsymbol{\zeta}^{s}, \boldsymbol{x}^{\text{test}}_{i,t+1}, M)$; then, for each test sample, we summarize the results by computing the average joint log predictive density $\text{LPD}_{t+1}(M)$ over all firms in the test sample.
- the posterior mean conditional default probability for firm i in $t+1$:

$$\text{PD}^{\text{cond}}_{i,t+1}(M) = \frac{\frac{1}{S}\sum_{s=1}^{S} P(y^{\text{def,test}}_{i,t+1} = 1|\boldsymbol{x}^{\text{test}}_{i,t+1}, \boldsymbol{\zeta}^{s}, M) p(\boldsymbol{y}^{\text{rat,test}}_{i,t+1}|\boldsymbol{\zeta}^{s}, \boldsymbol{x}^{\text{test}}_{i,t+1}, M)}{\frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}^{\text{rat,test}}_{i,t+1}|\boldsymbol{\zeta}^{s}, \boldsymbol{x}^{\text{test}}_{i,t+1}, M)}.$$

From these conditional probabilities, we compute for each test sample a measure of calibration, namely the (square root of) Brier score (Brier, 1950), which is mean the squared error between the conditional PDs and the observed binary default indicator. Finally, for evaluating the discriminatory power we compute the area under the precision-recall curve (Davis and Goadrich, 2006) [Using the precision-recall curve is more desirable than employing the receiver operating curve given the imbalance in the default indicator.].

4 Empirical application

In this section, we introduce the dataset used in the analysis and proceed with a discussion of the results obtained from the proposed dynamic framework as well a with comparison to benchmark models in terms of the predictive performance.

Throughout the analysis, the hyper-parameters for the dynamic specification are kept constant. We set $c_0 = C_0 = 0.5$ for the distribution of q^2, which implies a median of one for the variance τ_i^2 of the random effects and allows for heavy tails to accommodate for extreme observations; the parameters for the Student-t priors for the regression coefficients are fixed to four degrees of freedom, mean zero and unit variance; for the intercept β_0 we employ the same prior with a variance of two; for the threshold parameters, the ‘concentration’ hyper-parameter of the Dirichlet distribution is chosen in a data-dependent fashion by setting $\alpha_{j,r}$ as the number of ratings in class r assigned by rater j. For the persistence parameters, we choose $a^{\phi_b} = 20$ and $b^{\phi_b} = 2.5$ which translates into a prior mean of roughly 0.8 and a prior standard deviation of 0.12, motivated by previous results which find the market factor to be persistent. We choose the same hyper-parameters for the prior on ϕ_δ. For the persistence of the idiosyncratic effects we set $a^{\rho} = 20$ and $b^{\rho} = 5$ which has a larger variation, that is, is vaguer and has a mean around 0.60. For all models, five chains of length 2 000 were randomly initialized and run in parallel. The first 1000 MCMC iterations of each chain were discarded as burn-in. Trace plots and density plots show satisfactory convergence of all chains in each model.

4.1 Data

The empirical analysis is performed on a dataset containing credit ratings from S&P, Moody's and Fitch, default data and firm-level information for a sample of publicly traded and rated US corporates, excluding financial and utilities companies over the period 1995–2014. Long-term issuer credit ratings of S&P were downloaded by the S&P Capital IQ's COMPUSTAT North America Ratings file. The ratings from Moody's and Fitch were purchased by the research institution directly from the CRAs. Data on corporate defaults and bankruptcies were obtained from the UCLA-LoPucki Bankruptcy Research Database and the Mergent issuer default file. Firm-level data was downloaded from the merged COMPUSTAT/CRSP database.

We perform the analysis on a calendar year basis and match the latest available firm-level information with all available end-of-year ratings. The default indicator is set to one whenever we observe a bankruptcy filing under Chapter 7 (liquidation) or Chapter 11 (reorganization) of the US bankruptcy code or if a default rating [Default ratings assigned by a CRA include in distressed exchanges or missed interest payments addition to bankruptcy filings.] is assigned by at least one CRA in the year following the rating observations. In all other instances, the default indicator is set to zero, including cases where the firm disappears from the dataset for some reason other than bankruptcy such as acquisition, delisting or if no longer rated. The firm-level information is used to construct covariates which have been identified as significant predictors of credit quality in the literature. In our analysis, we rely on the work of Tian et al. (2015), who employ model selection techniques to identify factors relevant for bankruptcy prediction: current liabilities/total assets (LCT/TA), total debt/total assets (F/TA), net income over market value of total assets (NI/MTA), annualized standard deviation of stock returns over a three month period (SIGMA), the logarithm of the end-of-year stock price, whereas the stock price is capped at USD 15 (PRICE) and gross excess log return over value-weighted S&P 500 return (EXRET). All variables are winsorized at 99% and 1% if negative values are allowed.

After eliminating the missing data in the covariates (which appears mainly due to different coverage of the data sources), the merged dataset contains 2528 firms and 19952 yearly observations for all firms (so-called firm-year observations), whereas the panel is highly unbalanced in the time dimension, with firms staying on average 7.89 years in the sample (see Figure A.1 in the Supplementary Material). The sample contains 375 defaults (1.88% sample default rate). Figure 1 shows the cyclical behaviour of the default rates from 1996 to 2014, with default rates increasing during and immediately after recessions.

There is a high rating agreement in the sample with Spearman's correlation higher than 90% for all pairs of raters. Not all ratings are observed for all firm-years, with 97.52% of the firm-years being rated by S&P, 58.67% by Moody's and 17.17% by Fitch. For all CRAs the number of ratings falling into the best and worst classes is small so for the analysis we use the following aggregated rating classes: *AAA/A, BBB, BB, B, CCC/C* and *Aaa/A, Baa, Ba, B, Caa/Ca*, respectively. The rating distribution for S&P and Moody's ratings is relatively stable over the sample period, while Fitch assigns more favourable ratings especially at the beginning

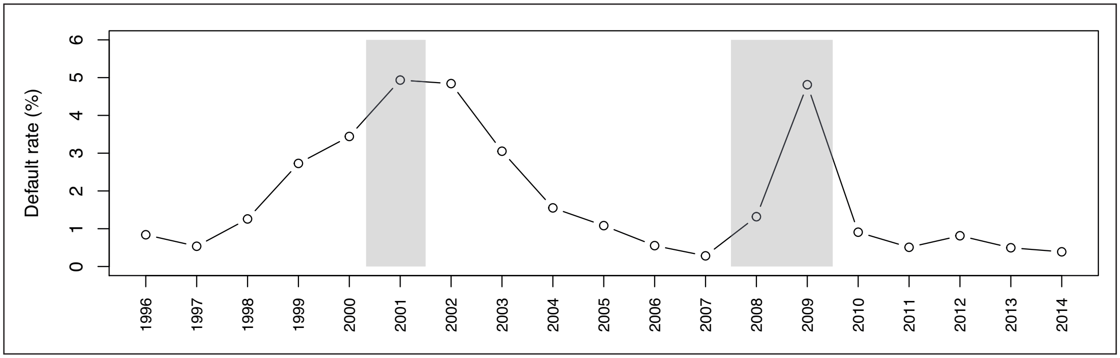

Figure 1 The figure illustrates the dynamics of the default rates from 1996 to 2014. The grey shaded areas represent recession periods as indicated by the NBER based Recession Indicators for the US and correspond to the burst of the dot-com bubble and the sub-prime mortgage crisis

of the period (Figure A.2 in the Supplementary Material shows the yearly distribution of the aggregated rating grades). Finally, the aggregated ratings change rarely: the S&P rating changes on average 0.712 times per firm, 60.4% of which are upgrades; the Moody's rating changes on average 0.747 times, 43.4% of which are upgrades; the Fitch rating changes on average 0.106 times, 62.1% of which are upgrades.

In the sample, we also observe that defaulted firms exhibit on average higher liabilities ratios (LCT/TA, F/TA, TL/MTA), higher stock price volatility, lower stock prices as well as negative income ratios and negative excess returns. The summary statistics of the covariates for both the entire and the defaulted sample are presented in Table A.1 in the Supplementary Material.

4.2 Results of the proposed model

Table 1 shows the posterior mean and posterior standard deviation for the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$ corresponding to the standardized covariates (the posterior distribution of these coefficients is illustrated in Figure A.3 of the Supplementary Material). We observe that the coefficients $\boldsymbol{\beta}$ of the latent PD score all have the expected signs. Firms with higher current liabilities, debt or total liabilities ratios have a higher likelihood to default. Similarly, a higher volatility of the stock price is associated with higher PD scores. On the other hand, higher profitability ratios, higher stock prices and higher excess returns lead to lower PD scores and improved creditworthiness. When looking at the rater biases, we observe no marked differences among the three raters. The quantity $\boldsymbol{\beta} + \boldsymbol{\gamma}_j$ corresponds to the rater-specific coefficients and, while caution should be employed when comparing the magnitude of these coefficients among raters (as absolute scale of the underlying variables in ordinal models is unidentifiable and assumed to be equal to one, see, e.g., Kern and Stein, 2015), insight can be gained from looking at the signs of these coefficients. It is worth noting that the resulting rater-specific coefficients for the covariates change sign for the covariates LCT/TA and EXRET. This has been

Table 1 Posterior mean and posterior standard deviation of the regression coefficients β and of the coefficients $\gamma_{\text{S\&P}}$, $\gamma_{\text{Moody's}}$ and γ_{Fitch} of the rater bias

	β		$\gamma_{\text{S\&P}}$		$\gamma_{\text{Moody's}}$		γ_{Fitch}	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
LCT/TA	0.362	0.087	−0.884	0.078	−0.873	0.082	−1.074	0.108
F/TA	0.826	0.105	0.518	0.091	0.467	0.097	0.621	0.139
NI/MTA	−0.067	0.076	−0.133	0.078	−0.136	0.086	−0.230	0.131
TL/MTA	2.087	0.172	−1.653	0.156	−1.188	0.162	−0.976	0.197
PRICE	−0.250	0.096	−0.692	0.087	−0.548	0.093	−0.474	0.145
SIGMA	0.171	0.108	1.153	0.110	1.109	0.117	0.664	0.167
EXRET	−1.050	0.079	1.294	0.082	1.408	0.087	1.356	0.119

previously observed empirically (e.g., Alp, 2013; Hirk et al., 2018) and is likely due to the ratings being forward-looking measures of creditworthiness over periods extending beyond one year. In this setting, the resulting negative coefficient of LCT/TA can be explained by the fact that short-term liabilities, that is, liabilities due within one year, which are problematic for firms close to default, in the longer run are less risky than long-term liabilities. Similarly, the positive coefficient of EXRET indicates that firms with higher excess returns, while more likely to avoid an imminent default, are considered typically to be riskier.

The posterior mean and the 80% symmetric credible intervals for the market factor $b(t)$ are shown in the left-most panel of Figure 2. We observe the drops in the estimated posterior means of the market factor during the burst of the dot-com bubble and the financial crisis 2007–2009. We also investigate the time-varying intercept in the rater bias which captures changes in the CRAs' behaviour. Figure 2 illustrates the posterior

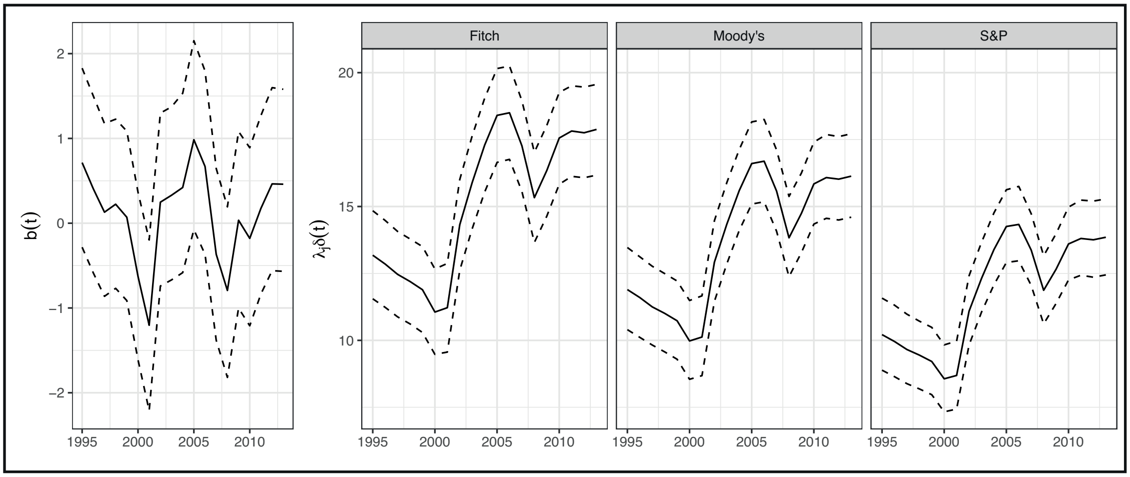

Figure 2 This figure shows the posterior mean and the 80% symmetric credible intervals for the latent systematic factor $b(t)$ and the scaled rater factor $\lambda_j\delta(t)$ at the end of each year over the period 1995 to 2013

means and the 80% credible intervals of the $\lambda_j\delta(t)$ term for each CRA, (note that in the analysis we fix $\lambda_{\text{S\&P}} = 1$ to reduce the parameter space and to ensure a better convergence of the model). We observe a rather counter-cyclical behaviour, where the drops (spikes) represent a relaxation (tightening) of the rating standards. The factor loading λ_{Fitch} is larger than the one for the other two CRAs, probably owing to Fitch being the youngest of the CRAs in the US market and to the least ratings being observed for Fitch in the sample. The tightening and relaxation of the rating standards in a counter-cyclical fashion could potentially be explained by the TTC approach employed by the raters, who adapt their standards to counterbalance the business cycle in order to keep the rating distribution constant. This hints towards the fact that the rating standards become stricter after economic downturns, behaviour not completely explained by the TTC approach (in line with e.g., Alp, 2013; Bar-Isaac and Shapiro, 2013).

The posterior summaries of the other parameters are presented in Table A.2 of the Supplementary Material.

4.3 Comparison with benchmark models

For comparison purposes, we formulate several alternative model specifications which differ from the proposed modelling framework in the specification of the random effect $u_i(t)$ and in the specification of the average rater bias $\eta_{ij}(t)$. We consider two static benchmark models which only contain one level of hierarchy in the random effects specification after accounting for the covariates, that is, the idiosyncratic term is normally distributed term $u_i(t) \sim N(0, \psi^2)$. In model *(S1)* we assume no rater bias $\eta_{ij}(t) = 0$ for all raters while in model *(S2)* we allow for covariate dependent rater bias $\eta_{ij}(t) = \boldsymbol{\gamma}_j^\top \boldsymbol{x}_i(t)$ for all raters. The other benchmark models considered share the dynamic specification of $u_i(t)$ introduced in Equation (2.3) but differ in the specification of the rater bias. Model *(D1)* assumes $\eta_{ij}(t) = 0$, while model *(D2)* assumes that the rater bias is a linear combination of the covariates $\eta_{ij}(t) = \boldsymbol{\gamma}_j^\top \boldsymbol{x}_i(t)$. The proposed model is denoted by *(PM)*. For a motivation on the choice for the rater bias specification in model *(PM)*, we direct the reader to Section A.4 of the Supplementary Material for an exploratory residual analysis.

We consider for model comparison purposes the widely used Bayesian leave-one-out estimate of the expected log pointwise predictive (ELPD LOO). Table 2 contains the difference in ELPD LOO relative to the best estimated model, that is, the model with highest ELPD LOO, together with an estimate for the standard error of the differences. We observe a value of zero for the proposed model and notice that the differences for all other models are more than two standard deviations away from zero, confirming the superior performance of Model *(PM)*.

In order to compare the proposed joint model with the above models in terms of out-of-time performance, we set up a cross-validation exercise as described in Section 3.2 and start by training the model on the period 1995–2006 and then sequentially add one sample year of data to the training set. This results in seven

Table 2 This table presents for five models considered the difference in ELPD LOO relative to the best performing model, the corresponding standard error, and the measures of the out-of-time exercise: one-step-ahead joint log predictive density, the square root of the Brier score and the area under the precision-recall curve averaged over 2007–2013 (whole) and 2007–2009 (crisis) test period. The best values are marked in bold

Measure	Period	*(PM)*^{a,d}	*(S1)*^{c,e}	*(S2)*^{b,e}	*(D1)*^{c,d}	*(D2)*^{b,d}
ELPD LOO (diff)		**0.000**	−12955.267	−12617.137	−579.408	−249.818
SE (diff)		0.000	106.838	101.362	52.115	28.989
$\overline{\text{LPD}}_{t+1}$	whole	**−1.188**	−1.765	−1.750	−1.214	−1.200
	crisis	−**1.212**	−1.790	−1.765	−1.240	−1.223
$\overline{\text{sqrtBrier}}_{t+1}$	whole	**0.097**	0.118	0.115	0.121	0.116
	crisis	**0.131**	0.159	0.148	0.161	0.151
$\overline{\text{AUPRC}}_{t+1}$	whole	0.367	0.284	0.365	0.342	**0.389**
	crisis	0.279	0.308	**0.351**	0.262	0.307

Notes: $^{a}\eta_{ij}(t) = \lambda_j \delta(t) + \boldsymbol{\gamma}_j^{\top} \boldsymbol{x}_i(t)$,
$^{b}\eta_{ij}(t) = \boldsymbol{\gamma}_j^{\top} \boldsymbol{x}_i(t)$,
$^{c}\eta_{ij}(t) = 0$,
$^{d}u_i(t)$ as in Eq. 2.3,
$^{e}u_i(t) \sim N(0, \psi^2)$.

training and test samples. Table 2 presents the one-step-ahead joint and marginal LPD averaged over all test samples (i.e., over the period 2007–2013) and over test samples covering the financial crisis 2007–2009, respectively. Similarly, we report the average one-step-ahead square root of the Brier score and area under the precision-recall curve based on the conditional probabilities of default. We observe that the proposed model *(PM)* performs best out-of-time in terms of the joint log predictive density scores, and more generally, the models with a dynamic specification in the PD score are better in terms of predictive performance than the static models *(S1)* and *(S2)*. Model *(PM)* also performs best in terms of calibration, as it achieves the lowest Brier score on average over all samples and over the crisis period. In terms of discriminatory power, the static model *(S2)* performs best for the crisis years, while dynamic model *(D2)* performs best for the whole test period. This suggests that including a time-varying rater factor in the rater bias specification does not necessarily improve the ability of the model to discriminate between defaults and non-defaults.

5 Concluding remarks

In this article we present a joint analysis of defaults and credit ratings for a sample of US publicly traded corporates by considering a multidimensional longitudinal model of multivariate ordinal data. We integrate both default and forward-looking credit rating data in a joint statistical model and employ a dynamic specification in the latent creditworthiness equation and in the rater bias. Bayesian methods implemented in the R package rstan are used for estimation. To examine the empirical performance of the posterior estimates under the proposed joint model, we conducted a simulation study which is presented in Section A.3 of the Supplementary Material. Conditional on well-established covariates in the bankruptcy prediction literature, we allow the latent PD score to depend on a dynamic unobservable market factor common to all firms, as well as on idiosyncratic effects with a dynamic specification. Moreover, the ratings from the big three CRAs are assumed to be noisy observations of the latent PD score, where a rater bias which depends on covariates and has a time-varying component is specified for each CRA. When comparing the predictive performance of the proposed framework to benchmark models we find that the proposed model has superior overall predictive performance.

Future research avenues include the incorporation of a wide range of covariates in the model and tackling the issue of variable selection to allow a data-driven identification of relevant factors for both the latent PD score and the rater biases as well as the exploration of more flexible mixed-effect specifications for the latent PD score, for example, more general parameterizations for the factor loading ω capturing the dependence between the latent PD score and the latent market factor as well as for the persistence ρ and standard deviation ψ of the idiosyncratic effects by allowing, for example, industry-specific parameters. Finally, the potential of the proposed joint model in serving as a framework for estimating and validating TTC versus PIT PDs and for measuring the degree to which rating systems are employing the TTC approach (topics of renewed interest mainly in the context of IFRS 9), could be further investigated.

Acknowledgements

The authors would like to thank the anonymous reviewers for their suggestions and comments that have improved the presentation of the manuscript.
Computational details have been performed on 25 IBM dx360M3 nodes within a cluster of workstations. The R codes for reproducing the simulation study in the Supplementary Material and the analysis presented in this article for a synthetic data set are available at the corresponding author's github repository.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This research was funded in part by the Austrian Science Fund (FWF) [Grant id ZK35]. For the purpose of open access,

Supplementary materials

Supplementary materials for this article are available from http://www.statmod.org/smij/archive.html and contain additional figures and tables, the results of a simulation study and a residual analysis for the empirical application presented in this manuscript.

References

Alp A (2013) Structural shifts in credit rating standards. *The Journal of Finance*, **68**, 2435–70. doi: 10.1111/jofi.12070.

Baghai RP, Servaes H and Tamayo A (2014) Have rating agencies become more conservative? Implications for capital structure and debt pricing. *The Journal of Finance*, **69**, 1961–2005. doi: 10.1111/jofi.12153.

Bar-Isaac H and Shapiro J (2013) Ratings quality over the business cycle. *Journal of Financial Economics*, **108**, 62–78. doi: 10.1016/j.jfineco.2012.11.004.

Basel Committee on Banking Supervision (2004) *Basel II: International convergence of capital measurement and capital standards: A revised framework* (Technical report). Basel: Bank of International Settlements. URL http://www.bis. org/publ/bcbs107.htm (last accessed 25 October 2021).

Basel Committee on Banking Supervision (2011) *Basel III: A global regulatory framework for more resilient banks and banking systems* (Technical report). Basel: Bank of International Settlements. URL https://www.bis.org/publ/bcbs189.htm (last accessed 25 October 2021).

Becker B and Milbourn T (2011) How did increased competition affect credit ratings? *Journal of Financial Economics*, **101**, 493–514. doi: 10.1016/j.jfineco.2011.03.012.

Bellio R and Varin C (2005) A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, **5**, 217–27. doi: 10.1191/1471082X05st095oa.

Berwart E, Guidolin M and Milidonis A (2019) An empirical analysis of changes in the relative timeliness of issuer-paid vs. investor-paid ratings. *Journal of Corporate Finance*, **59**, 88–118. doi: 10.1016/j.jcorpfin.2016.10.011.

Betz J, Kellner R and Rösch D. (2018) Systematic effects among loss given defaults and their implications on downturn estimation. *European Journal of Operational Research*, **271**, 1113–44. doi: 10.1016/j.ejor.2018.05.059.

Bolton P, Freixas X and Shapiro J (2012) The credit ratings game. *The Journal of Finance*, **67**, 85–111. doi: 10.1111/j.1540-6261.2011.01708.x.

Brier G. W (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Cagnone S, Moustaki I and Vasdekis V (2009) Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*, **62**, 401–15. doi: 10.1348/000711008X320134.

Campbell JY, Hilscher J and Szilagyi J (2008) In search of distress risk. *The Journal of Finance*, **63**, 2899–2939. doi: 10.1111/j.1540-6261.2008.01416.x.

Creal D, Schwaab B, Koopman SJ and Lucas A (2014) Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, **96**, 898–915. doi: 10.1162/REST a 00393.

Davis J and Goadrich M (2006) The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML ï¿½06, pages 233–40, New York, NY, USA. Association for Computing Machinery. doi: 10.1145/1143844.1143874.

Duffie D and Lando D (2001) Term structures of credit spreads with incomplete accounting information. *Econometrica*, **69**, 633–64. doi: 10.1111/1468-0262.00208.

Frühwirth-Schnatter S and Wagner H. (2011) Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In Bayesian Statistics 9, edited by J. Bernardo, MJ Bayarri, James O Berger, AP Dawid, D Heckerman, Adrian FM Smith and Mike West, pages 165–200. Oxford: Oxford University Press. URL 10.1093/acprof:oso/9780199694587.003.0006 (last accessed 25 October 2021).

Gruñ B, Hofmarcher P, Hornik K, Leitner C and Pichler S (2013) Deriving consensus ratings of the big three rating agencies. *Journal of Credit Risk*, **9**, 75–98. doi: 10.21314/JCR.2013.156.

Güttler A and Wahrenburg M. (2007) The adjustment of credit ratings in advance of defaults. *Journal of Banking & Finance*, **31**, 751–67. doi: 10.1016/j.jbankfin.2006.05.014.

Hilscher J and Wilson M (2017) Credit ratings and credit risk: Is one measure enough? *Management Science*, **63**, 3414–37. doi: 10.1287/mnsc.2016.2514.

Hirk R, Hornik K and Vana L (2018) Multivariate ordinal regression models: An analysis of corporate credit ratings. *Statistical Methods & Applications*. doi: 10.1007/s10260-018-00437-7.

Hirk R, Vana L, Pichler S and Hornik K (2020) A joint model of failures and credit ratings. *Journal of Credit Risk*. doi: 10.21314/JCR.2020.264.

Hornik K, Jankowitsch R, Leitner C, Lingo M, Pichler S and Winkler G (2010) A latent variable approach to validate credit rating systems. In *Model Risk in Financial Crises*, edited by D Rï¿½osch and H Scheule, pages 277–96. London: Risk Books.

IASB (2014) *IFRS 9 financial instruments*. URL https://www.iasplus.com/en/standards/ifrs/ifrs9#: :text=IFRS%209%20Financial%20Instruments%20issued,derecognition%20and%20general%20hedge%20accounting (last accessed 25 October 2021).

Kashyap AK and Kovrijnykh N (2016) Who should pay for credit ratings and how? *The Review of Financial Studies*, **29**, 420–56. doi: 10.1093/rfs/hhv127.

Kern C and Stein P (2015) Comparing coefficients of nonlinear multivariate regression models between equations. *Survey Research Methods*, **9**, 159–67. doi: 10.18148/srm/2015.v9i3.6211.

Koopman SJ and Lucas A (2005) Business and default cycles for credit risk. *Journal of Applied Econometrics*, **20**, 311–23. doi: 10.1002/jae.833.

Malik M and Thomas LC (2012) Transition matrix models of consumer credit ratings. *International Journal of Forecasting*, **28**, 261–72. doi: 10.1016/j.ijforecast.2011.01.007.

McCulloch CE (1994) Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330–35. doi: 10.1080/01621459.1994.10476474.

McNeil AJ and Wendin JP (2007) Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, **14**, 131–49. doi: 10.1016/j.jempfin.2006.05.002.

R Core Team (2020) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL https://www.r-project.org/about.html (last accessed 25 October 2021)

Shumway T (2001) Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, **74**, 101–24. doi: 10.1086/209665.

Stan Development Team (2018) *Stan modeling language users guide and reference manual*.

Version 2.18.0. URL http://mc-stan.org (last accessed 25 October 2021).

Stan Development Team (2019) RStan: *The R interface to Stan*. R Package Version 2.19.2. URL http://mc-stan.org/ (last accessed 25 October 2021).

Tian S, Yu Y and Guo H (2015) Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, **52**, 89–100. doi: 10.1016/j.jbankfin.2014.12.003.

Vasicek O (2002) The distribution of loan portfolio value. *Risk*, **15**, 160–62.

Vehtari A and Ojanen J (2012) A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6**, 142–228. doi: 10.1214/12-SS102.

Vehtari A, Gelman A and Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413–32. doi: 10.1007/s11222-016-9696-4.

Xie Y, Chen Z and Albert PS (2013) A crossed random effects modeling approach for estimating diagnostic accuracy from ordinal ratings without a gold standard. *Statistics in Medicine*, **32**, 3472–85. doi: 10.1002/sim.5784.