Joint models of multivariate longitudinal outcomes and discrete survival data with INLA

European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Joint models of multivariate longitudinal outcomes and discrete survival data with INLA: An application to credit repayment behaviour

Victor Medina-Olivares[a,*], Finn Lindgren[b], Raffaella Calabrese[a], Jonathan Crook[a]

[a] Business School, University of Edinburgh, United Kingdom
[b] School of Mathematics, University of Edinburgh, United Kingdom

ARTICLE INFO

Article history:
Received 29 May 2021
Accepted 8 March 2023
Available online xxx

Keywords:
OR in banking
Bayesian joint models
Discrete time
Laplace approximation
Credit prepayment

ABSTRACT

Survival models with time-varying covariates (TVCs) are widely used in the literature on credit risk prediction. However, when these covariates are endogenous, the inclusion procedure has been limited to practices such as lagging these variables or treating them as exogenous. That leads to possible biased estimators (depending on the strength of the exogeneity assumption) and a lack of prediction framework that consolidates the joint evolution of the survival process and the endogenous TVCs. The use of joint models is a suitable approach for handling endogeneity, however, it comes at a high computational cost. We propose a joint model for bivariate endogenous TVCs and discrete survival data using integrated nested Laplace approximation (INLA). We illustrate the implementation via simulations and build a model for full-prepayment consumer loans. We also propose a methodology for individual survival prediction using the Laplace method that leads to more accurate approximations than comparable approaches. We evidence the superiority of joint models over the traditional survival approach for an out-of-sample and out-of-time analysis.

1. Introduction

The balance of a loan can be partially or fully paid before it is due, which results in unscheduled cash inflows and potential loss of interest for a bank. This prepayment risk requires banks to closely study their borrowers' payment behaviour so timely action can be taken (BCBS, 2019). We are interested in predicting when and which consumer loans will be paid off before the date agreed in the contract and in the presence of endogenous time-varying covariates.

Many approaches have been proposed to assess the credit risks in consumer loans (see Crook, Edelman, & Thomas, 2007). For instance, structural models are based on utility theory where the event of interest is assumed to occur when a latent stochastic process dips below a threshold. This threshold is commonly related to the difference between the consumer's assets and debts (see, for example, De Andrade & Thomas, 2007 in the context of consumer loans, and Campbell & Cocco, 2015; Foote, Gerardi, & Willen, 2008 in mortgage loans). Another approach uses mathematical programming to allocate the resources in a way that minimises a specific cost, for example, the cost of misclassifying applicants in the context of credit default (Shi, Peng, Xu, & Tang, 2002). In this work, we are interested in prediction-based approaches, conceivably the most used by organisations that wish to make commercial decisions, which are essentially statistical models that aim to predict an event of interest given the characteristics of the loan and the borrower (see, for example, Bhattacharya, Wilson, & Soyer, 2019; Crook & Bellotti, 2010; Fitzpatrick & Mues, 2016 for specific applications, and Thomas, Crook, & Edelman, 2017 for a thorough description of these techniques).

Survival models are a good approach to predict when an event, such as full-prepayment, will occur and have been widely used in the credit risk literature for the last 20 years (Banasik, Crook, & Thomas, 1999; Bellotti & Crook, 2009; Leow & Crook, 2016; Wang, Crook, & Andreeva, 2020). However, there is little work in this field that addresses how to include time-varying covariates (TVCs) that are endogenous to the subject, i.e. the path of the covariate influences the time to the event and the event influences the path of the covariate. In this context, for example, the movements in the balance of the loan can relate to how likely is the full-prepayment and, in turn, once the loan is paid off, the balance should be zero. The ubiquitous way to treat these TVCs in the literature is to either

* Corresponding author.
E-mail address: victor.medina@ed.ac.uk (V. Medina-Olivares).

https://doi.org/10.1016/j.ejor.2023.03.012

work with their lagged values[1] or simply assume they are exogenous, so partial likelihood estimation can be carried out. However, this latter procedure has two major problems. First, it can lead to biased estimators (Kalbfleisch & Prentice, 2002; Rizopoulos, 2012). Second, it lacks a prediction framework that could account for the mutual evolution of the survival process and the endogenous TVCs. In addition to correctly including endogenous variables in a survival model to reduce any bias in parameter estimates, we also wish to increase prediction accuracy given our application to prepayment risk.

A rapidly evolving field of medical research known as joint models for longitudinal and survival data (joint models hereafter) addresses the problem of endogeneity by jointly modelling the survival process and the endogenous TVCs (Henderson, Diggle, & Dobson, 2000; Rizopoulos, 2012; Tsiatis & Davidian, 2004; Wu & Carroll, 1988). In this context, we use the terms "endogenous TVCs" and "longitudinal outcomes" indistinguishably. These models are computationally expensive to estimate, cost that is further increased if more data and covariates are considered. For this reason and the lack of adequate software (Furgal, Sen, & Taylor, 2019), most of the literature focuses on the case with only one longitudinal outcome (Hickey, Philipson, Jorgensen, & Kolamunnage-Dona, 2016) and/or relatively small datasets (Brown, Ibrahim, & DeGruttola, 2005; Chi & Ibrahim, 2006; Rizopoulos & Ghosh, 2011). Other approaches to control for endogeneity in survival analysis have been proposed in the literature. For instance, Bartus (2017) showed how to estimate multilevel multiprocess models in Stata, including simultaneous hazard equations. However, this implementation seems to be restricted to the case where the observed endogenous covariate is qualitative as opposed to "continuous" or "numerical". As described in Section 2 and motivated by the application of Section 4, we are instead interested in the case where our observed endogenous covariates are numerical.

In credit analysis, the datasets are usually fairly large[2] and consist of many time-fixed covariates (e.g. amount of the loan, interest rate at origination, term, among others) and more than one endogenous TVC (e.g. balance of the loan, number of instalments in arrears, among others). Therefore, to study the possible advantages of joint models in credit applications it is relevant to explore fast inference methods that can handle more than one TVC (multivariate). Moreover, most joint models are implemented for continuous time but credit data is commonly delivered over discrete periods (e.g. monthly accounting data) and many ties occur between events making the discrete time version more appropriate (Bellotti & Crook, 2013; Djeundje & Crook, 2019).

We present two methodological and two empirical contributions to the literature. From the methodological point of view, first, we propose a joint model for bivariate longitudinal outcomes and discrete survival data using integrated nested Laplace approximations (INLA) (Rue, Martino, & Chopin, 2009), a deterministic algorithm for Bayesian inference. We extend Van Niekerk, Bakka, & Rue (2019) who use INLA to estimate a joint model for the univariate case with continuous time. By using this method, we suggest a faster estimation procedure that can easily scale to large datasets without compromising accuracy of the estimates and that otherwise would not be computationally feasible. We illustrate the implementation via simulation analysis which shows satisfactory recovery of the true parameter values. Second, we propose a methodology for individual survival predictions using the Laplace method (Tierney & Kadane, 1986) that leads to more accurate approximations than the empirical Bayes approach used in the joint model literature (Rizopoulos, 2012).

From the empirical perspective, our first contribution is to apply a multivariate joint model approach in the credit risk context for the first time, in particular, for predicting the probability of full-prepayment in a consumer loan portfolio. While Hu & Zhou (2019) use joint models to predict early mortgage loan repayment events and show performance improvements compared to survival models, the authors consider only the univariate case and time as continuous. Moreover, the estimation is done by the R package JM which uses the maximum likelihood approach and does not allow the multivariate case.

In addition, Medina-Olivares, Calabrese, Crook, & Lindgren (2022), focused on the probability of default rather than on the prepayment event, they proposed a discrete joint model with one endogenous covariate whose specification allows autoregressive terms to be included. This model was estimated with an MCMC sampling scheme, which does not scale appropriately for the multivariate case (see Appendix B for a computational comparison).

The second empirical contribution shows that the multivariate approaches result in better discrimination and calibration performances than the traditional survival models used in the literature (Thomas et al., 2017) for all evaluation times. This consistent result along different evaluation times was not demonstrated in the two previously cited papers that studied joint modelling with one longitudinal outcome with credit applications. This is evidence that the little-explored joint modelling framework in this context remains a promising research area.

The remainder of the article is organised as follows. In Section 2, we describe the joint model with multivariate longitudinal outcomes and discrete survival time. We also present the estimation procedure using INLA and how we compute the individual survival predictions. In Section 3, we perform a simulation analysis. In Section 4, we apply two multivariate joint models to empirical data and compare them with standard approaches in credit risk analysis. Section 5 concludes.

2. Methodology

In this section we describe the model (2.1), the estimation procedure (2.2), how the individual survival predictions are obtained (2.3) and the metrics to assess the models with censored cases and discrete time (2.4).

2.1. Joint model of multivariate longitudinal outcome and discrete survival data

We assume that the duration time T_i, for subject i ($i = 1, \ldots, N$), has a discrete domain. We model T_i in terms of time-fixed covariates, $\mathbf{z}_i$, and a set of M endogenous TVCs, $Y_{is}^{(m)}$ for covariate m ($m = 1, \ldots, M$), that are observed at time s. Also, we assume that the study lasts until time T and we observe subject i until $t = t_i$ (i.e. $t_i \leq T$ and $s \in \{0, \ldots, t_i - 1\}$), at which point either the event happens, or it is censored.[3] In principle, the number of observed measurements for the endogenous TVCs (longitudinal outcomes hereafter) can differ from the number of survival points. However, we assume that for each time t, the immediate previous realisations of the longitudinal outcomes are provided at time $t - 1$, so we can express $Y_{i,t-1}^{(m)}$ ($t = 1, \ldots, t_i$) (assuming no events happen at $t = 0$).

[1] This approach removes the first lagged observations and therefore assumes that there are no events in that period (induced bias). Furthermore, since the event at a particular time is related to the lagged observations, the latest information that may be relevant for the prediction may not be used. And finally, the lag is usually decided with respect to the prediction window, limiting the prediction for other time horizons (Bellotti & Crook, 2013).

[2] Around 30,000 cases in practical application. In our case, we have approximately 60,000 observations, see Section 4.

[3] In this context, censoring means that we observe subject i until time t_i but not later (Allison, 1982).

Following Allison (1982), we denote the sequence of survival points through a binary random variable X_{it} that takes the value 1 if subject i experiences the event at time t and 0 otherwise. Using this notation, the last observation of the sequence for subject i, i.e. x_{i,t_i} (we denote the realisations of a random variable in lowercase), is equal to the event indicator. We consider a logit link between the binary random variable X_{it} and the linear predictor η_{it}^S, then the discrete-time survival is

$$X_{it}|X_{i,t-1}=0,\eta_{it}^S \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{it}^S))$$
$$\eta_{it}^S = a_t^0 + \mathbf{z}_i^\intercal \boldsymbol{\gamma} + \sum_{m=1}^{M} \lambda^{(m)} g^{(m)}(\eta_{i,t-1}^{(m)}), \quad (1)$$

where a_t^0 is the baseline discrete event time distribution, usually represented by either fixed effects or spline models (Tutz, Schmid et al., 2016). Here, we follow the smoothing approach from Lindgren & Rue (2008) that is implemented in the INLA package (Rue et al., 2009).[4] We express a_t^0 as a discrete time second-order random walk model which is a discretization of a continuous time integrated Wiener process. This smoothing approach depends on one hyperparameter τ_0 that denotes the precision of the underlying Gaussian white noise. Furthermore, $\boldsymbol{\gamma}$ is the vector of coefficients for the time-fixed covariates $\mathbf{z}_i$. $\lambda^{(m)}$ is the association parameter that links the mth endogenous TVC with the event process. The association function $g^{(m)}(\cdot)$ takes as argument the mth latent linear predictor $\eta_{i,t-1}^{(m)}$ (described below) and returns some of its components. For example, a widely used version of the function $g^{(m)}$ is the identity function, i.e. $g^{(m)}(\eta_{i,t-1}^{(m)}) = \eta_{i,t-1}^{(m)}$ (see Hickey et al., 2016 for a comprehensive review of different association functions $g^{(m)}$).

With respect to the mth longitudinal outcome $Y_{i,t-1}^{(m)}$, we denote $\mathbf{w}_i^{(m)}$ as a vector of time-fixed covariates associated with subject i (it does not have to be equal to $\mathbf{z}_i$) and $\mathbf{q}_{i,t-1}^{(m)}$ as a vector of longitudinal covariates measured at time $t-1$. Then, the mth longitudinal outcome is assumed as a noisy version of an underlying latent predictor $\eta_{i,t-1}^{(m)}$ that can be decomposed into fixed effects, $\mathbf{w}_i^{(m)\intercal}\boldsymbol{\alpha}^{(m)} + \mathbf{q}_{i,t-1}^{(m)\intercal}\boldsymbol{\beta}^{(m)}$, and random effects, $\mathbf{d}_{i,t-1}^{(m)\intercal}\mathbf{U}_i$, where $\mathbf{d}_{i,t-1}^{(m)}$ is the design vector at time $t-1$. This leads to the following linear mixed-effect model (Laird & Ware, 1982)

$$Y_{i,t-1}^{(m)}|\eta_{i,t-1}^{(m)},\tau^{(m)} \sim \mathcal{N}(\eta_{i,t-1}^{(m)}, 1/\tau^{(m)}), \qquad m=1,\dots,M$$
$$\eta_{i,t-1}^{(m)} = \mathbf{w}_i^{(m)\intercal}\boldsymbol{\alpha}^{(m)} + \mathbf{q}_{i,t-1}^{(m)\intercal}\boldsymbol{\beta}^{(m)} + \mathbf{d}_{i,t-1}^{(m)\intercal}\mathbf{U}_i, \quad (2)$$

where the subject-level $\mathbf{U}_i$ are assumed as mutually independent and distributed as a zero-mean multivariate Gaussian distribution with $p \times p$ precision matrix $Q_{\mathbf{U}}$. Note that for the particular case of the mixed-effect "intercept-and-slope" model, we write $\mathbf{q}_{i,t-1}^{(m)\intercal}\boldsymbol{\beta}^{(m)} = \beta_0^{(m)} + \beta_1^{(m)}(t-1)$ and $\mathbf{d}_{i,t-1}^{(m)\intercal}\mathbf{U}_i = U_{0i}^{(m)} + U_{1i}^{(m)}(t-1)$ with $\mathbf{U}_i = [U_{0i}^{(1)}, U_{1i}^{(1)}, \dots, U_{0i}^{(M)}, U_{1i}^{(M)}]^\intercal$.

2.2. Estimation

The observed data are the longitudinal outcomes $\{\mathcal{Y}_i\}_{i=1,\dots,N}$ with $\mathcal{Y}_i = \{y_{i,t-1}^{(m)} : t=1,\dots,t_i; m=1,\dots,M\}$ and the survival data $\{\mathcal{X}_i\}_{i=1,\dots,N}$ with $\mathcal{X}_i = \{x_{it} : t=1,\dots,t_i\}$. We denote the complete observed data as $\mathcal{D} = (\{\mathcal{Y}_i\}_{i=1,\dots,N}, \{\mathcal{X}_i\}_{i=1,\dots,N})$. From Section 2.1, the parameters to estimate are $\{\{\boldsymbol{\alpha}^{(m)}\}, \{\boldsymbol{\beta}^{(m)}\}, \{\tau^{(m)}\}, Q_{\mathbf{U}}, \boldsymbol{\gamma}, \{\lambda^{(m)}\}, \tau_0\}$ in addition to baseline terms a_t^0 and the random effects $\mathbf{U}_i$. The main assumption in the joint model literature is that the longitudinal and survival processes are conditionally independent given the random effects $\mathbf{U}_i$, which simplifies the representation of the likelihood (see Henderson et al., 2000; Tsiatis & Davidian, 2004; Wu & Carroll, 1988; Wulfsohn & Tsiatis, 1997).

It is possible to implement this model via simulation-based MCMC schemes (see, for example, Andrinopoulou, Rizopoulos, Takkenberg, & Lesaffre, 2014). The main drawback of this strategy is that it is computationally expensive or even infeasible for some applications. A faster and accurate alternative is to use the INLA methodology proposed by Rue et al. (2009) and implemented in the INLA package for the R software. INLA approximates the Bayesian inference on the class of Latent Gaussian models (LGMs), as presented in Rue et al. (2009). This class comprises a large number of well-known statistical models (e.g. mixed-effects, dynamic, spatial-temporal models and more).

To briefly describe the INLA methodology and frame our model as an LGM, we define the following terms

$$\boldsymbol{\mu} = (\{\boldsymbol{\eta}_i^S\}, \{\boldsymbol{\eta}_i^{(m)}\}, \{a_t^0\}, \boldsymbol{\gamma}, \{\boldsymbol{\alpha}^{(m)}\}, \{\boldsymbol{\beta}^{(m)}\}, \{\mathbf{U}_i\})$$
$$\boldsymbol{\theta}_1 = (\tau_0, \theta_{\boldsymbol{\gamma}}, \theta_{\boldsymbol{\alpha}^{(m)}}, \theta_{\boldsymbol{\beta}^{(m)}}, Q_{\mathbf{U}}, \{\lambda^{(m)}\})$$
$$\boldsymbol{\theta}_2 = (\{\tau^{(m)}\})$$

where $\theta_{\boldsymbol{\gamma}}$, $\theta_{\boldsymbol{\alpha}^{(m)}}$ and $\theta_{\boldsymbol{\beta}^{(m)}}$ are hyperparameters (precision terms) for $\boldsymbol{\gamma}$, $\{\boldsymbol{\alpha}^{(m)}\}$ and $\{\boldsymbol{\beta}^{(m)}\}$, respectively, and defined by the user as detailed below. Note also that with this notation, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ correspond to the set of hyperparameters of the latent field $\boldsymbol{\mu}$ and observation density, respectively.

Given the conditional dependency assumed in Eqs. (1) and (2), the joint conditional density of $\mathcal{D}$ is $p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\theta}_2) = \prod_{j\in\mathcal{J}} p(\mathcal{D}_j|\mu_j,\boldsymbol{\theta}_2)$ with $\mathcal{J}$ the set of indices for all observed values in $\mathcal{D}$ and it is coded so that each observation is associated with its respective linear predictor η. We assume the density of $\boldsymbol{\mu}|\boldsymbol{\theta}_1$ as zero-mean Gaussian with precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$ and denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Therefore, the posterior follows

$$p(\boldsymbol{\mu},\boldsymbol{\theta}|\mathcal{D}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\mu}|\boldsymbol{\theta})\prod_{j\in\mathcal{J}} p(\mathcal{D}_j|\mu_j,\boldsymbol{\theta})$$
$$\propto p(\boldsymbol{\theta})|\mathbf{Q}(\boldsymbol{\theta})|^{1/2}\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\intercal\mathbf{Q}(\boldsymbol{\theta})\boldsymbol{\mu} + \sum_{j\in\mathcal{J}}\log\{p(\mathcal{D}_j|\mu_j,\boldsymbol{\theta})\}\right].$$

We are interested in the posterior marginals, $p(\mu_j|\mathcal{D})$ and $p(\theta_j|\mathcal{D})$, specified by

$$p(\mu_j|\mathcal{D}) = \int p(\mu_j|\boldsymbol{\theta},\mathcal{D})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$
$$p(\theta_j|\mathcal{D}) = \int p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}_{-j}. \quad (3)$$

The INLA methodology computes these marginals based on the Laplace approximation (Tierney & Kadane, 1986). For $p(\boldsymbol{\theta}|\mathcal{D})$ this follows

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \frac{p(\boldsymbol{\mu},\boldsymbol{\theta},\mathcal{D})}{p(\boldsymbol{\mu}|\boldsymbol{\theta},\mathcal{D})} \approx \frac{p(\boldsymbol{\mu},\boldsymbol{\theta},\mathcal{D})}{\tilde{p}_G(\boldsymbol{\mu}|\boldsymbol{\theta},\mathcal{D})}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}^*(\boldsymbol{\theta})} =: \tilde{p}(\boldsymbol{\theta}|\mathcal{D}),$$

where $\tilde{p}_G(\boldsymbol{\mu}|\boldsymbol{\theta},\mathcal{D})$ is the Gaussian approximation to the full conditional and $\boldsymbol{\mu}^*(\boldsymbol{\theta})$ its mode. A crucial step in the procedure is to further approximate the terms $p(\mu_j|\boldsymbol{\theta},\mathcal{D})$ by using the Laplace approximation one more time as

$$p(\mu_j|\boldsymbol{\theta},\mathcal{D}) \propto \frac{p(\boldsymbol{\mu},\boldsymbol{\theta},\mathcal{D})}{p(\boldsymbol{\mu}_{-j}|\mu_j,\boldsymbol{\theta},\mathcal{D})} \approx \frac{p(\boldsymbol{\mu},\boldsymbol{\theta},\mathcal{D})}{\tilde{p}_G(\boldsymbol{\mu}_{-j}|\mu_j,\boldsymbol{\theta},\mathcal{D})}\Big|_{\boldsymbol{\mu}_{-j}=\boldsymbol{\mu}^*_{-j}(\mu_j,\boldsymbol{\theta})}$$
$$=: \tilde{p}(\mu_j|\boldsymbol{\theta},\mathcal{D}).$$

Finally, the integrals in Eq. (3) are computed using numerical integration. Other computational efficient modifications for $\tilde{p}(\mu_j|\boldsymbol{\theta},\mathcal{D})$ are also implemented in the INLA package and detailed in Rue et al. (2009).

To fully specify estimation, we need to define the priors of the hyperparameters $\boldsymbol{\theta}$. In particular, we assume that the parameters $\boldsymbol{\gamma}$, $\{\boldsymbol{\alpha}^{(m)}\}$ and $\{\boldsymbol{\beta}^{(m)}\}$ have independent zero-mean Gaussian

[4] The package is hosted on http://www.r-inla.org/

priors with precision matrix of $\theta_f \mathbf{I}$, where $\mathbf{I}$ is the identity matrix with the corresponding dimension for each set of parameters and θ_f, a precision parameter, is equal to 0.01 (i.e. $\theta_{\boldsymbol{\gamma}}$, $\theta_{\boldsymbol{\alpha}^{(m)}}$ and $\theta_{\boldsymbol{\beta}^{(m)}}$ are all equal to θ_f). Moreover, for the log scale of $\{\tau^{(m)}\}$, the precision parameters of the error terms of the longitudinal outcomes, we assume weakly informative log-gamma prior distributions with shape and scale parameters of 1 and 5×10^{-5}, respectively. The prior of the $p \times p$ precision matrix $Q_{\mathbf{U}}$ is assumed as a Wishart distribution $\mathcal{W}_p(\mathbf{I}, p(p+1)/2+1)$ which shows sensible results on the simulation study (see Section 3). Finally, for the prior of τ_0, the precision parameter of the second-order random walk model, we assume a penalising complexity (PC) prior as described in Simpson, Rue, Riebler, Martins, & Sørbye (2017). The form of the prior is defined via the influence of the parameter on the latent process model, as measured by the deviation from a base model with zero variance. The prior is specified by choosing an upper α-quantile u for the standard deviation of the model, so that $P(\tau_0^{-1/2} = \sigma_0 > u) = \alpha$ for the choice of α and u. We use a weakly informative prior by choosing $P(\sigma_0 > 1) = 0.01$, indicating a small probability for large standard deviation.

Despite using weakly-informative priors, we studied the robustness of the results by estimating the model with even looser priors, without finding relevant differences (see Appendix E).

2.3. Individual survival predictions

We are interested in estimating the survival probability of a new subject k not originally included in the training data $\mathcal{D}$. Assume we have collected the M longitudinal outcomes for this subject up to time c and denote the set of these observations as $\mathcal{Y}_k = \{y_{k,t-1}^{(m)} : t = 1, \ldots, c+1; m = 1, \ldots, M\}$. Since we know that subject k has survived until at least c, we focus on the conditional probability of surviving at the time $c + \Delta c$, with $\Delta c \in \mathbb{Z}_+$, given that subject k has survived until time c. Formally,

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \mathcal{D}) = \int P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \Theta) p(\Theta|\mathcal{D}) \mathrm{d}\Theta, \tag{4}$$

where Θ represents all the parameters but the random effects $\mathbf{U}_k$ (which are not known), and $p(\Theta|\mathcal{D})$ is the posterior distribution of the parameters given the data $\mathcal{D}$.

Equation (4) does not have a closed form. In addition, in credit applications the interest is in estimating predictions for a large number of out-of-sample borrowers. Therefore, we propose the following procedure to approximate Eq. (4).

The first step considers that we have already estimated the posterior distribution of Θ and we can rely on a point estimate denoted by $\hat{\Theta}$ (we use the posterior mean). Thus, we can now concentrate on the expression $P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta})$ which can be formulated as

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) = \int P(T_k > c + \Delta c | \quad T_k > c, \mathbf{U}_k, \hat{\Theta}) \times p(\mathbf{U}_k | T_k > c, \mathcal{Y}_k, \hat{\Theta}) \mathrm{d}\mathbf{U}_k, \tag{5}$$

where $P(T_k > c + \Delta c | T_k > c, \mathbf{U}_k, \hat{\Theta})$ can be in turn written as

$$P(T_k > c + \Delta c | T_k > c, \mathbf{U}_k, \hat{\Theta}) = \frac{P(T_k > c + \Delta c | \mathbf{U}_k, \hat{\Theta})}{P(T_k > c | \mathbf{U}_k, \hat{\Theta})} = \frac{\prod_{t=1}^{c+\Delta c}(1 - p_{kt})}{\prod_{t=1}^{c}(1 - p_{kt})} = \prod_{t=c+1}^{c+\Delta c}(1 - p_{kt}),$$

with $p_{kt} = \text{logit}^{-1}(\eta_{kt}^S)$ and η_{kt}^S following Eq. (1).

A first order approximation of Eq. (5) is presented in Rizopoulos (2012) who uses the empirical Bayes estimates for the random effects $\mathbf{U}_k$ as follows

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) \approx \frac{P(T_k > c + \Delta c | \hat{\mathbf{U}}_k, \hat{\Theta})}{P(T_k > c | \hat{\mathbf{U}}_k, \hat{\Theta})}$$

where $\hat{\mathbf{U}}_k = \text{argmax}_{\mathbf{U}} \{\log P(T_k > c, \mathcal{Y}_k, \mathbf{U} | \hat{\Theta})\}$.

However, we can get a better approximation of the Eq. (5) by using the Laplace method introduced by Tierney & Kadane (1986). Consider $-c \cdot h_k(\mathbf{U}) = \log P(T_k > c, \mathcal{Y}_k, \mathbf{U} | \hat{\Theta})$ and $-c \cdot h_k^*(\mathbf{U}) = -c \cdot h_k(\mathbf{U}) + \log P(T_k > c + \Delta c | T_k > c, \mathbf{U}, \hat{\Theta})$ and denote the vectors that maximise $-h_k^*(\cdot)$ and $-h_k(\cdot)$ as $\mathbf{U}^*$ and $\hat{\mathbf{U}}$, respectively. Hence, Eq. (5) can be approximated as (see Appendix A for further details)

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) \approx \frac{|\Sigma^*|^{1/2} \exp\{-c \cdot h_k^*(\mathbf{U}^*)\}}{|\tilde{\Sigma}|^{1/2} \exp\{-c \cdot h_k(\hat{\mathbf{U}})\}} \tag{6}$$

where Σ^* and $\tilde{\Sigma}$ are the inverses of the Hessians of $h_k^*(\cdot)$ and $h_k(\cdot)$ evaluated in $\mathbf{U}^*$ and $\hat{\mathbf{U}}$, respectively. We use this procedure to calculate the conditional probabilities needed in the performance measures described below.

2.4. Performance measures

We measure the prediction and accuracy of the models using similar metrics and notation as in Medina-Olivares et al. (2022). In 2.4.1 and 2.4.2, we briefly describe the discrimination and calibration metrics, respectively, and refer the reader to the mentioned work for more details.

2.4.1. Discrimination

The AUC is a commonly used metric to evaluate a binary classifier's ability to differentiate between two classes at various thresholds (Hand, 2009; Thomas et al., 2017). It measures the area under the curve of correct predictions versus misclassifications. A value of 0.5 represents a random classifier, while 1 indicates a perfect one. Moreover, as shown by Hanley & McNeil (1982), the AUC can also be interpreted as for any random pair of subjects $\{i, j\}$ and evaluation times c and $c + \Delta c$ (with $\Delta c > 0$), such as $AUC_c^{\Delta c} = P(\pi_i(c + \Delta c|c) < \pi_j(c + \Delta c|c) | \{T_i \in (c, c + \Delta c]\} \cap \{T_j > c + \Delta c\})$, where $\pi_i(c + \Delta c|c)$ denotes the conditional survival probability $P(T_i > c + \Delta c | T_i > c, \mathcal{Y}_i, \mathcal{D})$. The latter follows Eq. (4) when subject i is not included in $\mathcal{D}$. Moreover, censoring is accounted by a model-based approach (Rizopoulos, Molenberghs, & Lesaffre, 2017), detailed in Medina-Olivares et al. (2022) for the discrete case.

Note that the AUC depends on the selected pair of time points c and $c + \Delta c$. Hence, to evaluate the overall performance, we use the concordance index (Harrell, Califf, Pryor, Lee, & Rosati, 1982) in the version introduced by Rizopoulos (2011) and adapted to the discrete form as

$$C_{AUC}^{\Delta c} = \sum_c AUC_c^{\Delta c} u(c), \tag{7}$$

where $u(c)$ is a weight function to account for not all time points contributing the same. Although the choice of $u(c)$ remains an open question, Rizopoulos (2011) suggests using $u(c) = P(T_i > c)/\sum_t P(T_i > t)$, where $P(T_i > c)$ represents the marginal survival probability, estimated using the Kaplan-Meier estimator (Kalbfleisch & Prentice, 2002). We also opted for this strategy.

2.4.2. Calibration

The commonly used measure to evaluate the accuracy of predictions in survival models is the expected error of predicting future events (PE) (Rizopoulos et al., 2017). It is defined as $PE(c + \Delta c|c) = \mathbb{E}(LI(T_i > c + \Delta c), \pi_i(c + \Delta c|c))$. L represents a loss function, such as logarithm, Brier score, absolute error, etc. A lower value of PE indicates a better-calibrated model.

To account for censoring, we follow Henderson, Diggle, & Dobson (2002) who propose the estimate $\widehat{PE}(c+\Delta c|c) = n(c)^{-1}\sum_{i:T_i>c}\{S_i(c+\Delta c|c)+E_i(c+\Delta c|c)+C_i(c+\Delta c|c)\}$, where $n(c)$ is the population at risk at time c. S_i, E_i, and C_i are the contribution to the loss function, evaluated between c and $c+\Delta c$, when i either survives, experiences the event or is censored (see Medina-Olivares et al., 2022, for further details).

Analogous to Eq. (7), we assess the overall calibration performance with

$$C_{PE}^{\Delta c} = \sum_{c} PE(c+\Delta c|c)u(c). \tag{8}$$

In this work, we use the logarithmic score (Good, 1952) as the loss function L instead of the widely used Brier score because it is consistent with the use of likelihoods (or log-likelihoods) to measure the models (Winkler, 1969).

3. Simulation

In this section we perform a simulation study of the discrete multivariate joint model with INLA presented in Section 2. The aim is to check how well the proposed implementation works under different sample sizes. The simulated data is inspired by the application described in Section 4 and generated from a joint model with two longitudinal outcomes, both of them with a fixed intercept plus random intercept and slope. The four random effects (two intercepts and two slopes) are assumed zero-mean multivariate Gaussian distributed. Moreover, the event process has an additional covariate, fixed in time, and the baseline hazard rate is drawn from a cubic polynomial function. Formally, the generated data for the longitudinal processes follows

$$Y_{i,t-1}^{(m)}|\eta_{i,t-1}^{(m)}, \tau^{(m)} \sim \mathcal{N}(\eta_{i,t-1}^{(m)}, 1/\tau^{(m)}), \qquad m = 1, 2,$$
$$\eta_{i,t-1}^{(m)} = \beta_0^{(m)} + U_{0i}^{(m)} + U_{1i}^{(m)}(t-1),$$
$$(U_{0i}^{(1)}, U_{1i}^{(1)}, U_{0i}^{(2)}, U_{1i}^{(2)})^{\intercal} \sim \mathcal{N}_4(\mathbf{0}, Q_{\mathbf{U}}^{-1}),$$

and the corresponding event process is

$$X_{it}|X_{i,t-1}=0, \eta_{it}^{S} \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{it}^{S})),$$
$$\eta_{it}^{S} = a_t^0 + \gamma z_i + \lambda^{(1)}\eta_{i,t-1}^{(1)} + \lambda^{(2)}\eta_{i,t-1}^{(2)},$$
$$a_t^0 = at^3 + bt^2 + ct + d.$$

We simulate this data considering a maximum of 36 periods for three different numbers of subjects (500, 1000 and 1500) which correspond to 15,183, 30,187 and 44,971 observations, respectively. Figure 1 shows the distribution of events versus time for the sample with 1500 subjects. Figure 2 shows both simulated longitudinal outcomes and, for visual purposes, we highlight ten subjects that experience the event (dashed line) and ten that are censored (dotted line).

Despite that the baseline hazard rate a_t^0 is originally generated from a cubic polynomial function, we estimate it under a second-order random walk model (Lindgren & Rue, 2008) to evaluate how suitable this assumption is in this context. The solid line in Fig. 3 shows the true baseline hazard rate (after the logistic transformation) versus the estimated 95% credible intervals for the three samples. We observe a good fit for all three samples and, as we increase the sample size, the interval narrows around the true value. The same behaviour is observed in Table 1, where we show the true values of all the parameters in the simulations versus those estimated under each setting. The covariance matrix $Q_{\mathbf{U}}^{-1}$ is parameterised via marginal precisions $\tau_{U_{01}}$, $\tau_{U_{11}}$, $\tau_{U_{02}}$, and $\tau_{U_{12}}$, and pairwise correlations ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, and ρ_{34}, giving

$$Q_{\mathbf{U}}^{-1} = \begin{pmatrix} 1/\tau_{U_{01}} & \rho_{12}/\sqrt{\tau_{U_{01}}\tau_{U_{11}}} & \rho_{13}/\sqrt{\tau_{U_{01}}\tau_{U_{02}}} & \rho_{14}/\sqrt{\tau_{U_{01}}\tau_{U_{12}}} \\ \rho_{12}/\sqrt{\tau_{U_{01}}\tau_{U_{11}}} & 1/\tau_{U_{11}} & \rho_{23}/\sqrt{\tau_{U_{11}}\tau_{U_{02}}} & \rho_{24}/\sqrt{\tau_{U_{11}}\tau_{U_{12}}} \\ \rho_{13}/\sqrt{\tau_{U_{01}}\tau_{U_{02}}} & \rho_{23}/\sqrt{\tau_{U_{11}}\tau_{U_{02}}} & 1/\tau_{U_{02}} & \rho_{34}/\sqrt{\tau_{U_{02}}\tau_{U_{12}}} \\ \rho_{14}/\sqrt{\tau_{U_{01}}\tau_{U_{12}}} & \rho_{24}/\sqrt{\tau_{U_{11}}\tau_{U_{12}}} & \rho_{34}/\sqrt{\tau_{U_{02}}\tau_{U_{12}}} & 1/\tau_{U_{12}} \end{pmatrix}.$$

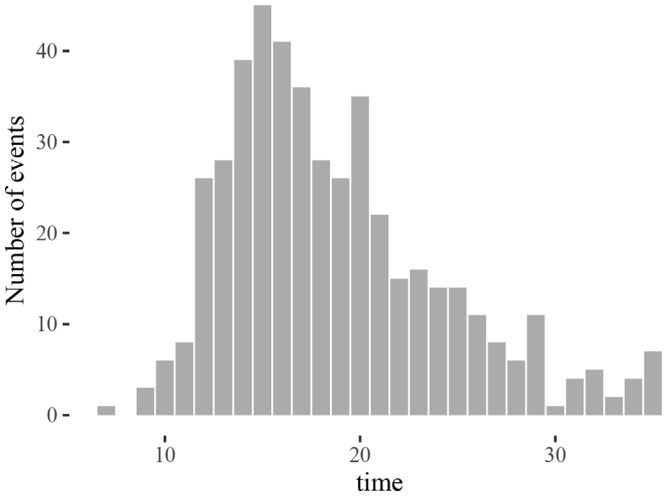

Fig. 1. Time-events distribution for the simulated sample of 1500 subjects.

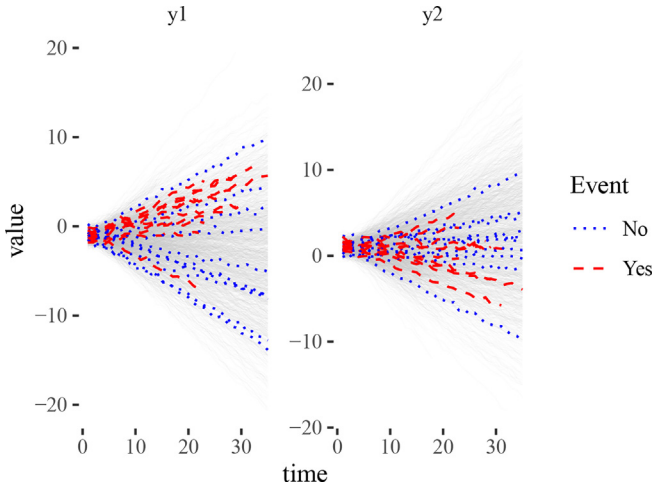

Fig. 2. Both longitudinal outcomes for the simulated sample of 1500 subjects. For visual purposes, we highlight ten subjects that experience the event (dashed line) and ten subjects that are censored (dotted line).

Finally, with the purpose to show how precise and fast is the INLA methodology for our model in comparison with an MCMC scheme, in Appendix B we carry out a comparative analysis, showing, for instance, that for the simulation with $N_{ids} = 500$ the inference takes more than 4 hours when is performed by MCMC sampling, whereas with INLA takes less than 3 minutes using exactly the same computational resources.

4. Application: Full-prepayment model

4.1. Data

The complete dataset is formed by two different cohorts of consumer loans granted by a bank. In the first cohort the loans were originated in April, 2012 and in the second the loans were originated in August, 2015. Each cohort has 40 consecutive months of performance. We use the first cohort as the training dataset and the second as the out-of-time dataset.

The training dataset corresponds to 2397 consumer loans with a total of 59,415 observations. The number of full-prepayment events is 470 and its distribution over time is shown in Fig. 4. The first longitudinal outcome is the cumulative sum of the ratio between the actual balance and the scheduled balance of each loan. This longitudinal outcome accounts for how different the bal-

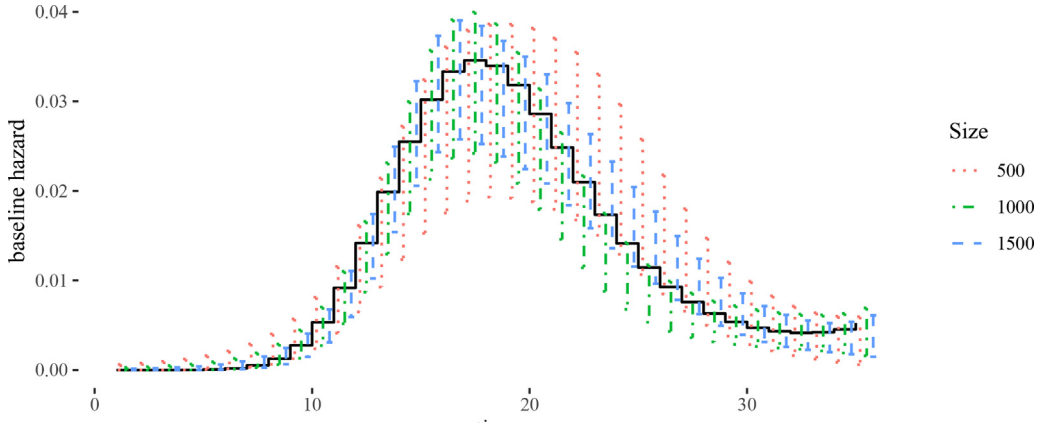

Fig. 3. Simulated baseline hazard (solid stepped line) and the estimated 95% credible intervals for the three sample sizes.

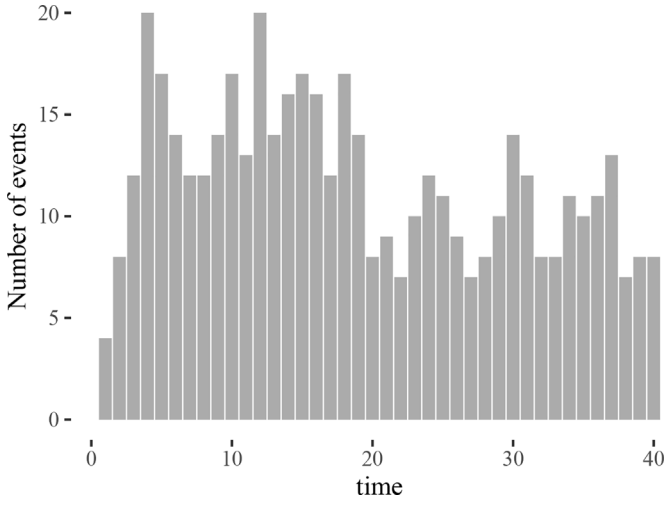

Fig. 4. Distribution of the full-prepayment events in time for the training dataset.

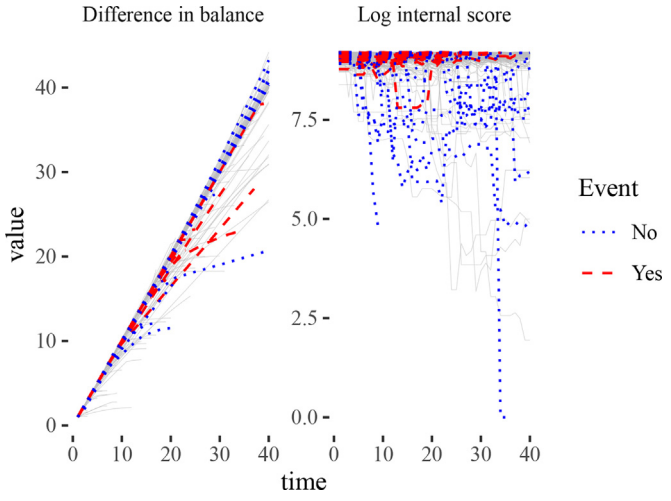

Fig. 5. Evolution of both longitudinal outcomes for the full-prepayment dataset. For visual purposes, we highlight borrowers that full-prepay the loan (dashed line) and borrowers that are censored (dotted line).

ance of the loan is from what was originally scheduled. This gives early signals, for example, if the loan is under or over paid. The second longitudinal outcome is an internal score calculated by the bank (in logarithmic scale) which measures the borrower's creditworthiness. Figure 5 shows the evolution of the two longitudinal outcomes and, analogously to Fig. 2, we have highlighted some borrowers that experience the full-prepayment event (dashed line) and borrowers that do not (dotted line). The rationale for choosing these two longitudinal outcomes is that we would expect that

Table 1
Estimations for the three simulation settings.

		$N_{id} = 500$				$N_{id} = 1000$				$N_{id} = 1500$			
	true	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%	mean	sd	2.5%	97.5%
$\beta_0^{(1)}$	−1.00	−1.02	0.02	−1.06	−0.98	−0.99	0.01	−1.02	−0.97	−1.00	0.01	−1.02	−0.98
$\beta_0^{(2)}$	1.00	0.98	0.02	0.94	1.02	0.99	0.01	0.96	1.02	0.99	0.01	0.97	1.01
γ	−0.50	−0.60	0.06	−0.71	−0.49	−0.49	0.04	−0.56	−0.41	−0.53	0.03	−0.58	−0.47
$\tau^{(1)}$	25.00	24.97	0.33	24.27	25.55	25.13	0.26	24.55	25.56	25.13	0.18	24.75	25.44
$\tau^{(2)}$	25.00	25.08	0.30	24.47	25.65	25.14	0.22	24.74	25.61	25.09	0.19	24.77	25.51
$\tau_{U_{01}}$	4.00	4.18	0.27	3.68	4.73	4.09	0.18	3.75	4.46	4.22	0.15	3.93	4.52
$\tau_{U_{11}}$	25.00	25.58	1.60	22.63	28.91	22.55	0.99	20.61	24.49	23.62	0.81	22.06	25.24
$\tau_{U_{02}}$	4.00	3.81	0.24	3.35	4.30	4.01	0.18	3.65	4.35	3.79	0.13	3.54	4.05
$\tau_{U_{12}}$	25.00	26.09	1.61	23.06	29.40	25.26	1.10	23.07	27.41	23.70	0.81	22.15	25.34
ρ_{12}	−0.30	−0.20	0.04	−0.28	−0.11	−0.29	0.03	−0.35	−0.24	−0.25	0.02	−0.30	−0.21
ρ_{13}	0.30	0.36	0.04	0.28	0.44	0.32	0.03	0.26	0.37	0.30	0.02	0.26	0.34
ρ_{14}	0.30	0.24	0.04	0.15	0.32	0.30	0.03	0.24	0.35	0.28	0.02	0.23	0.32
ρ_{23}	0.30	0.32	0.04	0.24	0.40	0.33	0.03	0.27	0.38	0.34	0.02	0.29	0.38
ρ_{24}	0.30	0.32	0.04	0.23	0.39	0.29	0.03	0.24	0.34	0.30	0.02	0.26	0.34
ρ_{34}	−0.30	−0.29	0.04	−0.37	−0.21	−0.26	0.03	−0.32	−0.20	−0.33	0.02	−0.38	−0.29
$\lambda^{(1)}$	0.50	0.57	0.04	0.49	0.65	0.55	0.03	0.49	0.60	0.52	0.02	0.48	0.56
$\lambda^{(2)}$	−0.50	−0.55	0.04	−0.64	−0.47	−0.50	0.03	−0.56	−0.45	−0.48	0.02	−0.52	−0.44

a borrower who pays more than what is scheduled (slope below from the diagonal in Fig. 4) and whose internal score is high, would have a higher probability of paying the loan in full. This is further confirmed in the results shown in Section 4.3.

We mentioned that the joint model approach offers two advantages over traditional survival models. First, it can handle endogenous TVCs reducing possible estimation bias and, second, it presents a prediction framework that incorporates the mutual evolution of both the survival and the longitudinal processes. We cannot analyse the bias because we do not know the true data generation process. However, we can compare the predictions of each model.[5]

To validate and compare the models described in Section 4.2, we perform a 10-fold cross-validation analysis. For each validation set (out-of-sample), we assess the performance in terms of the discrimination and calibration metrics described in Section 2.4. Moreover, to assess the robustness of the results, we use the out-of-time dataset mentioned above. This dataset corresponds to 2516 borrowers with a total of 65,928 observations and there is no overlapping times with the data used in the cross-validation.

4.2. Models

In the analysis of a full-prepayment dataset, a bank is particularly interested in understanding how accurate is the model to correctly predict full prepayment. To investigate the predictive power of the joint model framework we propose here, we estimate four models. The first one is a discrete survival model where both longitudinal outcomes are included as standard TVCs (observed value), so no joint model framework is used. We denote this model as *Cox* model.[6] This model relates the event at month t with the last available observation of the longitudinal outcomes at month $t-1$. The limitation of this model is that when we are interested in predicting the probability of the event, for example, at $t+12$, we assume the longitudinal outcomes remained constant from the last available observations in t until $t+11$. The second model *Cox_Lag* is also a discrete survival model. The difference with the *Cox* model is that the event at month t is now related with the observations of the longitudinal outcomes lagged in 12 months[7], so when we predict the probability of the event at $t+12$ the model is already estimated to consider the observed values at t. These two models are the traditional survival approaches in credit literature when TVCs are present and, thus, are seen as our natural benchmarks (see, for example, Bellotti & Crook, 2013; Calabrese & Crook, 2020; Gross & Souleles, 2002; Wang et al., 2020).

The third and fourth models (*JM1* and *JM2*, respectively) are both multivariate joint models for discrete survival data. The only difference between them is in the assumed correlations for the random effects. The *JM1* model assumes correlation between the random effects belonging to each of the longitudinal outcomes, but no correlation between the random effects of different longitudinal outcomes. The *JM2* model, however, assumes correlation within and between the random effects of both longitudinal outcomes (fully correlated). This last setting aims to investigate if substantial improvements are gained when a more complex relationship between the longitudinal outcomes is used.

Following the notation introduced in Section 2, we define $Y^{(1)}_{i,t-1}$ and $Y^{(2)}_{i,t-1}$ as the cumulative ratio between the balances and the logarithm of the internal score, respectively, at time $t-1$ for borrower i. Moreover, we denote X_{it} as the binary variable that equals 1 if the borrower i fully prepays the loan at time t and 0 otherwise. $\mathbf{z}_i$ is the vector of time-fixed covariates for borrower i (for more details about these covariates, see Appendix C) and a^0_t is the baseline hazard. The four models' specifications of the event process follow $X_{it}|X_{i,t-1}=0, \eta^S_{it} \sim \text{Bernoulli}(\text{logit}^{-1}(\eta^S_{it}))$, the differences come in the assumed predictor η^S_{it} (Eq. (1)). Moreover, both longitudinal processes assume $Y^{(m)}_{i,t-1}|\eta^{(m)}_{i,t-1}, \tau^{(m)} \sim \mathcal{N}(\eta^{(m)}_{i,t-1}, 1/\tau^{(m)})$ for $m=1,2$, therefore, we only need to describe the corresponding predictors to fully specify the models. These are the following

Cox. Discrete survival model with TVCs. The event predictor is described as $\eta^S_{it} = a^0_t + \mathbf{z}_i^\intercal \boldsymbol{\gamma} + \lambda^{(1)} y^{(1)}_{i,t-1} + \lambda^{(2)} y^{(2)}_{i,t-1}$. Cox_Lag. Discrete survival model with lagged TVCs. The event predictor is described as $\eta^S_{it} = a^0_t + \mathbf{z}_i^\intercal \boldsymbol{\gamma} + \lambda^{(1)} y^{(1)}_{i,t-12} + \lambda^{(2)} y^{(2)}_{i,t-12}$. JM1. Joint model not fully correlated. The event predictor is described as $\eta^S_{it} = a^0_t + \mathbf{z}_i^\intercal \boldsymbol{\gamma} + \lambda^{(1)}(U^{(1)}_{0i} + U^{(1)}_{1i}(t-1)) + \lambda^{(2)}(U^{(2)}_{0i} + U^{(2)}_{1i}(t-1))$ and the corresponding longitudinal processes as

$$\eta^{(1)}_{i,t-1} = \beta^{(1)}_0 + \beta^{(1)}_1(t-1) + U^{(1)}_{0i} + U^{(1)}_{1i}(t-1)$$
$$\eta^{(2)}_{i,t-1} = \beta^{(2)}_0 + U^{(2)}_{0i} + U^{(2)}_{1i}(t-1)$$
$$(U^{(1)}_{0i}, U^{(1)}_{1i})^\intercal \sim N_2(\mathbf{0}, Q^{-1}_{\mathbf{U}_1})$$
$$(U^{(2)}_{0i}, U^{(2)}_{1i})^\intercal \sim N_2(\mathbf{0}, Q^{-1}_{\mathbf{U}_2}).$$

JM2. Joint model fully correlated. The event predictor has the same structure as the *JM1* model. However, the assumption over the random effects in the longitudinal outcomes is

$$\eta^{(1)}_{i,t-1} = \beta^{(1)}_0 + \beta^{(1)}_1(t-1) + U^{(1)}_{0i} + U^{(1)}_{1i}(t-1)$$
$$\eta^{(2)}_{i,t-1} = \beta^{(2)}_0 + U^{(2)}_{0i} + U^{(2)}_{1i}(t-1)$$
$$(U^{(1)}_{0i}, U^{(1)}_{1i}, U^{(2)}_{0i}, U^{(2)}_{1i})^\intercal \sim N_4(\mathbf{0}, Q^{-1}_{\mathbf{U}}).$$

We observe that the cumulative ratio has a linear trend (see Fig. 4) which explains the extra fixed effect term $\beta^{(1)}_1(t-1)$ in comparison with the internal score. Moreover, as we mention in Section 2, there is flexibility in how we link the event and the longitudinal processes. We find that linking them only through the random effects provides a good performance but this, by no means, is a restriction of this approach and further exploration can be pursed.

4.3. Results

4.3.1. Cross-validation

We show in Section 2.4 that the performance metrics depend on the pair of evaluation times we choose (c and $c+\Delta c$ denoted above). To make the comparison less arbitrary, we evaluate the full range of available starting points ($c=12,\dots,28$) with a fixed time window of 12 months ($\Delta c=12$), commonly used in the industry. Note that the starting point could have been $c=1$ but the *Cox_Lag* model limits the comparison due to the lagged observations.

In each validation fold we calculate the $\widehat{AUC}$ and $\widehat{PE}$ metrics presented in Section 2.4 for all the pairs of evaluation times $\{(c, c+12) : c=12,\dots,28\}$. Then, we summarise the metrics for the different pairs of times by calculating $C^{\Delta c=12}_{AUC}$ and $C^{\Delta c=12}_{PE}$ (Eqs. (7) and (8), respectively). Table 2 shows these metrics where each row represents one of the ten folds, the same for all models, so the results are comparable. The best performance out of the four models is marked in bold. The last row is the average among the 10 folds (Avg). First, we observe that in general terms the joint models outperform survival models in both discrimination and calibration metrics. Second, whenever one of the survival models predicts

[5] Few works (Dafni & Tsiatis, 1998; Sweeting & Thompson, 2011; Tsiatis & Davidian, 2001; Wu & Carroll, 1988) show under simulation studies, that ignoring the joint evolution of the processes produce biased results. For example, Wu & Carroll (1988) illustrates the bias produced when the dependence between the two processes is not controlled for and is the first work to argue for the use of shared random effects as a mechanism to control for this dependence.

[6] The model follows the discrete survival approach proposed in Cox (1972).

[7] This lag responds strictly to the time window of interest in the predictions and is a limitation in terms of flexibility for other time horizons.

Table 2
Comparison of the discrimination (C^{12}_{AUC}) and calibration (C^{12}_{PE}) metrics between the four models for a prediction window of 12 months. Each fold number represents the validation fold in the cross-validation analysis. The last row is the average (Avg) among the 10 folds and the bold number is the best performance metric within each validation fold.

	Cox		Cox_Lag		JM1		JM2	
Fold	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}
1	**0.6081**	0.5111	0.5874	0.3139	0.5853	**0.3119**	0.5869	**0.3119**
2	**0.6352**	0.4850	0.6240	0.4422	0.6125	**0.4047**	0.6135	0.4048
3	0.5504	0.4766	0.5499	0.3215	**0.6053**	**0.3155**	0.6051	0.3156
4	0.6156	0.4982	**0.6301**	0.3887	0.6228	**0.3672**	0.6191	0.3675
5	**0.5523**	0.5066	0.5170	0.3973	0.5416	**0.3689**	0.5408	0.3691
6	0.6136	0.7152	0.6741	**0.3044**	0.6996	0.3089	**0.7009**	0.3084
7	0.6459	0.4635	0.5850	**0.2713**	0.6716	0.2825	**0.6764**	0.2835
8	0.5944	0.5168	0.5976	0.3105	0.5995	0.3094	**0.6016**	**0.3092**
9	0.7076	0.4832	0.6890	0.3169	0.7702	0.3047	**0.7703**	**0.3046**
10	0.5721	0.5494	0.5653	0.3410	0.6178	0.3268	**0.6195**	**0.3263**
Avg	0.6095	0.5206	0.6019	0.3408	0.6326	**0.3301**	**0.6334**	**0.3301**

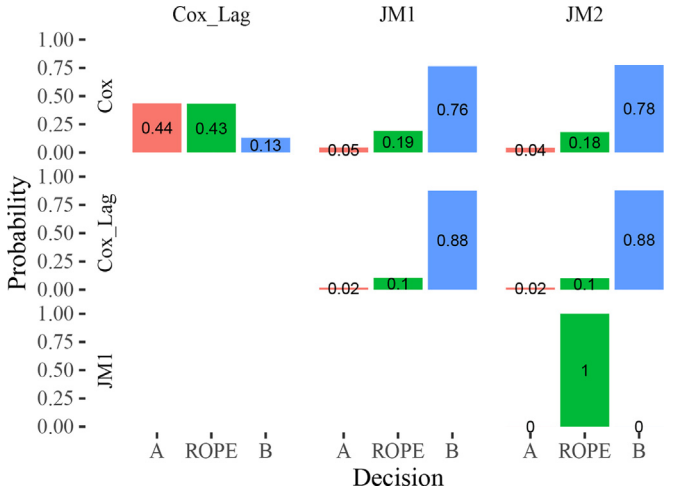

Fig. 6. Bayesian correlated *t*-test for the discrimination metric (C^{12}_{AUC}). It shows a 3 by 3 matrix of bar plots, where each plot is a comparison between the reference model named in the row (A) and the model we are comparing to in the column (B). The bars represent the posterior probabilities of the three possible decisions: A better than B (left bar in red), A practically equivalent to B (centre bar in green) and B better than A (right bar in blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

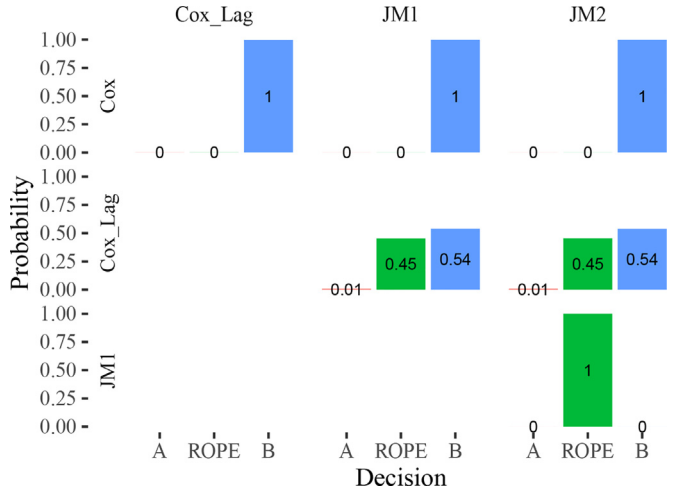

Fig. 7. Bayesian correlated *t*-test for the calibration metric (C^{12}_{PE}). It shows a 3 by 3 matrix of bar plots, where each plot is a comparison between the reference model named in the row (A) and the model we are comparing to in the column (B). The bars represent the posterior probabilities of the three possible decisions: A better than B (left bar in red), A practically equivalent to B (centre bar in green) and B better than A (right bar in blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

more accurately than the joint models, the difference in the metrics does not seem to be as significant as when we have the opposite[8].

We perform a Bayesian correlated *t*-test (Benavoli, Corani, Demšar, & Zaffalon, 2017) for both metrics (C^{12}_{AUC} and C^{12}_{PE}) in order to test the statistical validity of the differences shown in Table 2. The test is correlated because the metrics in each fold are not independent from each other since we have overlapping training sets (Nadeau & Bengio, 2000).

Following the recent work Gunnarsson, Vanden Broucke, Baesens, Óskarsdóttir, & Lemahieu (2021), we also consider two classifiers as practically equivalent when the mean difference of the metric is less than 0.01 and define the *Region of Practical Equivalence* (ROPE) as the interval $[-0.01, 0.01]$. In Figs. 6 and 7 we show the results for all combinations of model pairs for discrimination and calibration, respectively. On the left side of the figures are the reference models (A) and on top, the models we are comparing to (B). For instance, we estimate that the *Cox* model has a probability of 0.44 of being better in terms of discrimination than the *Cox_Lag* model, a probability of 0.43 of being equivalent and only 0.13 of being worse. We also observe that the joint models in comparison to both survival models are superior and there is no difference between the two joint models (ROPE-probability 1).

In terms of calibration (Fig. 7), we see that the *Cox* model performs poorly with respect to the three other models. When we compare the *Cox_Lag* model against the joint models, we observe that the probability that these models have the same calibration metrics is 0.45 and a probability of 0.54 in favour of the joint models. Moreover, since we strictly estimate the *Cox_Lag* model for predicting in a 12-month window (unlike the joint models), we expect it to do well in terms of calibration, but the evidence suggests that this model is not better than the joint models (probability of 0.01). Ultimately, we again see no difference between the two joint models (ROPE-probability 1).

So far, we have seen that the benchmarks cannot outperform the joint models for a 12-month forecast horizon ($\Delta c = 12$) for different starting months (c), but it is also of interest to study how these models behave when we vary the forecast horizon. For a

[8] We also compare the multivariate joint model with the univariate versions (see Appendix D). The main results are, first, the univariate joint models are superior to both Cox models (similar results are shown in Hu & Zhou (2019)). Second, the multivariate version can improve the performance in this data when compared to the univariate versions.

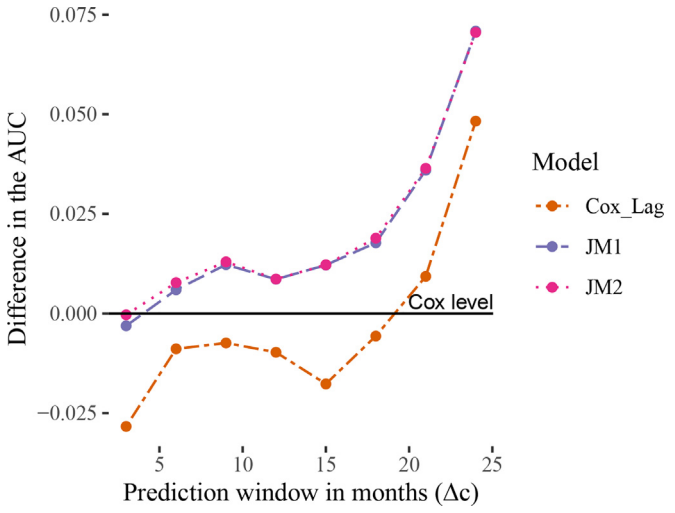

Fig. 8. Average difference in the $\widehat{AUC}$ with respect to the *Cox* model, for fixed $c = 12$ and variable Δc.

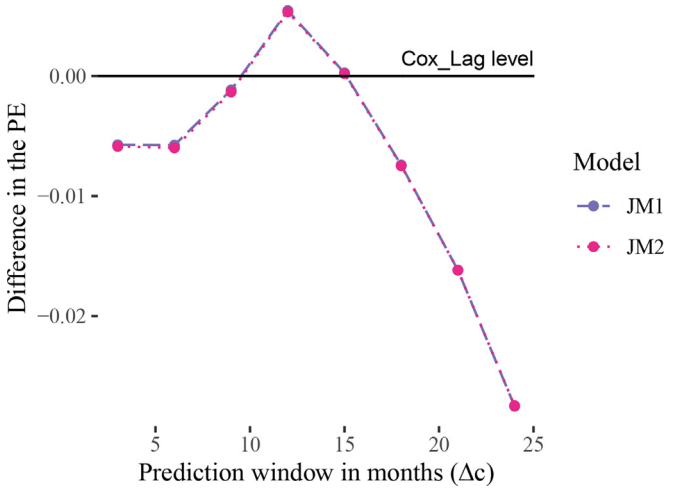

Fig. 9. Average difference in the $\widehat{PE}$ with respect to the *Cox_Lag* model, for fixed $c = 12$ and variable Δc.

fixed c we obtain the prediction for $[c + 1, c + \Delta c]$ for different values of Δc. Figure 8 shows the average difference in the $\widehat{AUC}$ with respect to the *Cox* model for Δc ranging from 3 to 24 months and $c = 12$, which is the first period the *Cox_Lag* model can predict. In general, we observe that both joint models have better discrimination for all time windows, a difference that is even more pronounced for longer horizons.

Analogously to Figs. 8, in 9 we show the average difference in the $\widehat{PE}$ with respect to the *Cox_Lag* model.[9] We observe that both joint models have practically the same calibration metric (overlapping lines) and that for almost all horizons this value is below (better) than the *Cox_Lag* level, especially for longer horizons. It is not surprising that for $\Delta c = 12$ the calibration of the reference model is better in comparison with the joint models since it was estimated for this Δc, but this only includes $c = 12$ and we have shown above that this result does not generalise when we consider all c.

To study the robustness of these results, in the next section we perform an out-of-time data analysis.

[9] We discard the *Cox* model from this plot because its performance is considerably inferior concerning the others.

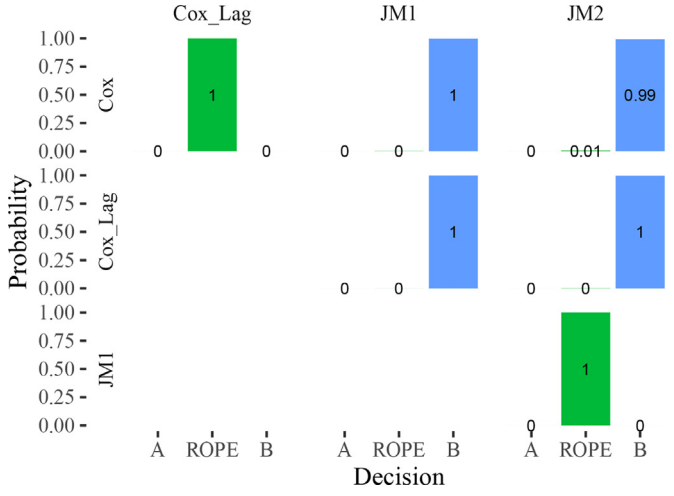

Fig. 10. Bayesian correlated t-test for the discrimination metric (C^{12}_{AUC}) shown as in Fig. 6 and applied to the out-of-time dataset.

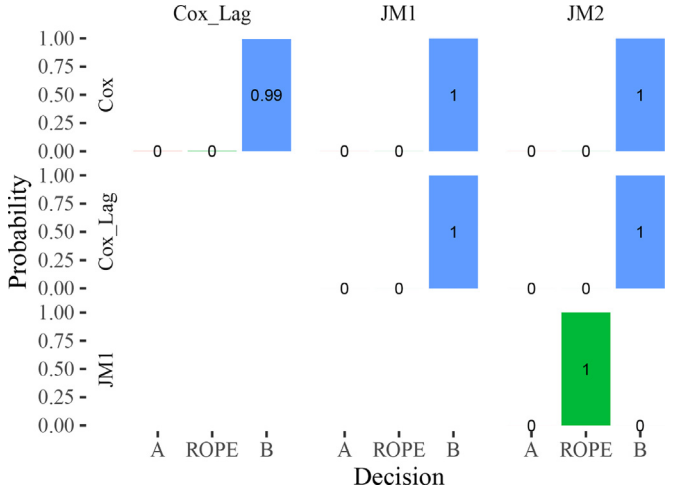

Fig. 11. Bayesian correlated t-test for the calibration metric (C^{12}_{PE}) shown as in Fig. 7 and applied to the out-of-time dataset.

4.3.2. Out-of-time validation

In practice, models are applied to new data that comes at later periods than the ones used in the construction stage. This for example happens when a bank is interested in classifying new customers. We study and compare how these models perform in an out-of-time scenario following a similar analysis to the previous one. We could estimate a model per specification using all the training data but since we have already estimated 10 models per specification, we use each of them to calculate the out-of-time performance.

In Table 3 we show the results for each model. We note that none of the traditional survival approaches can outperform the joint models, a result that is further supported by the Bayesian correlated t-test shown in Figs. 10 and 11 for discrimination and calibration, respectively. From these figures we note the following: the survival models are practically equivalent in terms of discrimination (ROPE-probability 1) but the *Cox_Lag* model outperforms the *Cox* in terms of calibration. Moreover, both joint models have essentially a probability of 1 of being better than the survival models, for both metrics, and there is not much difference between them (ROPE-probability 1).

The discrimination and calibration performances for different time windows Δc are shown in Figs. 12 and 13, respectively. We observe that both joint models have better $\widehat{AUC}$ than the survival

Table 3
Comparison of the discrimination (C^{12}_{AUC}) and calibration (C^{12}_{PE}) metrics between the four models for a prediction window of 12 months. Each fold number represents the hold-out fold when training the model. The predictions are done in the out-of-time dataset. The last row is the average (Avg) among columns and the bold number is the best performance metric per row.

	Cox		Cox_Lag		JM1		JM2	
Fold	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}
1	0.5531	0.5030	0.5520	0.4001	**0.5668**	**0.3755**	0.5664	0.3761
2	0.5531	0.4573	0.5523	0.4025	**0.5647**	**0.3761**	**0.5647**	0.3766
3	0.5533	0.4639	0.5525	0.4118	**0.5640**	**0.3749**	0.5629	0.3755
4	0.5532	0.4756	0.5521	0.4079	0.5676	0.3763	**0.5689**	**0.3755**
5	0.5549	0.4917	0.5523	0.4082	0.5707	**0.3749**	**0.5710**	0.3751
6	0.5509	0.6729	0.5520	0.4005	**0.5704**	0.3770	0.5691	**0.3767**
7	0.5534	0.4816	0.5513	0.4016	**0.5662**	0.3759	0.5629	**0.3754**
8	0.5519	0.5097	0.5520	0.4069	0.5692	0.3766	**0.5694**	**0.3761**
9	0.5506	0.4970	0.5527	0.4020	**0.5665**	**0.3752**	0.5658	0.3754
10	0.5532	0.5274	0.5522	0.4293	**0.5701**	**0.3762**	0.5687	0.3765
Avg	0.5528	0.5080	0.5521	0.4071	**0.5676**	**0.3759**	0.5670	**0.3759**

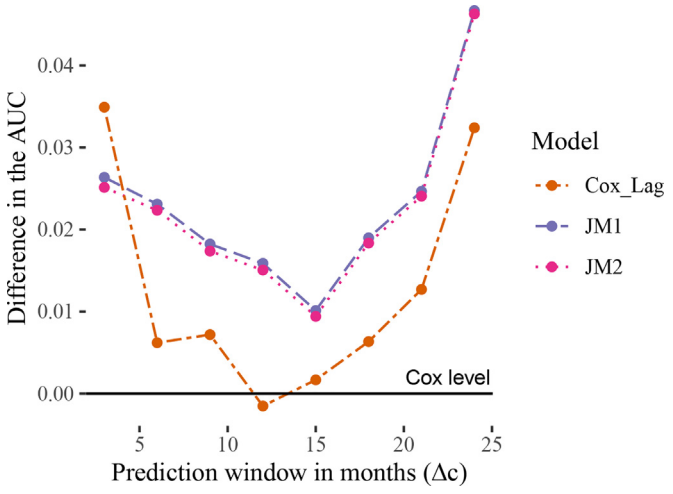

Fig. 12. Average difference in the $\widehat{AUC}$ with respect to the *Cox* model, for fixed $c = 12$ and variable Δc. Results from the out-of-time analysis.

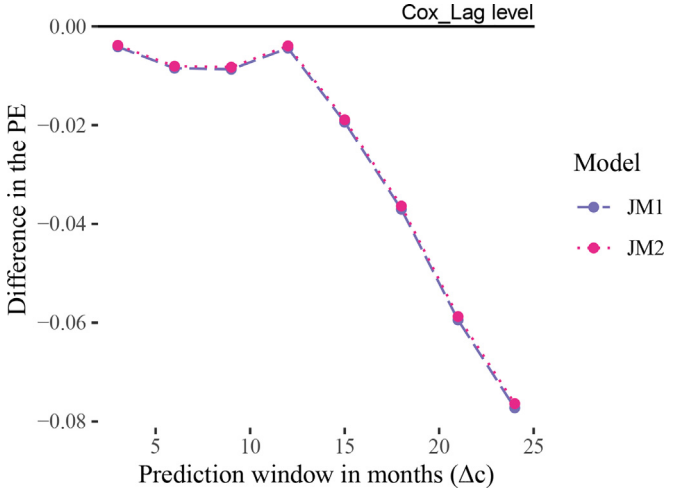

Fig. 13. Average difference in the $\widehat{PE}$ with respect to the *Cox_Lag* model, for fixed $c = 12$ and variable Δc. Results from the out-of-time analysis.

models for basically all the horizons. In terms of calibration we now see that for all the Δc, the $\widehat{PE}$ for each joint model is lower than for the *Cox_Lag* model. Also, the minimum difference is again obtained at $\Delta c = 12$ and it increases for longer horizons.

5. Conclusions

Survival models have been widely used in the credit risk literature since they can handle censored cases and answer when and which borrower is likely to experience the event. The inclusion of time-varying covariates (TVCs) into these models is very common. However, when the TVCs are endogenous to the borrower, the inclusion procedure has been limited to practices such as lagging these variables or simply considering their observed value as if they were exogenous. That leads to possible estimation biases and a lack of prediction framework that incorporates the mutual evolution of the survival process and the endogenous TVCs.

Joint models of longitudinal and survival data present a novel approach in this context, which allows one to not only incorporate these endogenous TVCs in sound statistical practice but also permits one to build a dynamic prediction framework that properly updates the probabilities once new information is collected. However, when maximum likelihood or MCMC schemes are used to estimate joint models, these approaches are computationally expensive. Cost that is further increased for large datasets and if more than one endogenous TVC is included (multivariate). This is commonly the case of credit risk applications.

In this paper, we make two methodological and two empirical contributions. First, we propose a fast and accurate joint model of bivariate longitudinal outcomes and discrete survival data based on the INLA framework. We study this model via simulation analysis. Second, we introduce a methodology for individual survival predictions using the Laplace method that leads to more accurate approximations than comparable approaches. From the empirical level, first, we introduce a multivariate joint model in the credit risk literature, specifically, for predicting the probability of full-prepayment in a consumer loan portfolio. Second, we show that for this particular application the multivariate joint models outperform standard survival approaches in both out-of-sample and out-of-time analyses.

As a new approach to credit risk modelling, many open questions remain that we believe could further enhance its use. For example, relaxing the assumption of linearity in the longitudinal process that could allow modelling challenging variables such as income (discrete changes in distant periods, missing or incorrect observations, etc.). We could also study new link structures between the event and longitudinal processes where the effect among them changes depending on the stage of the credit. The joint models approach offers the flexibility to explore these and other interesting topics.

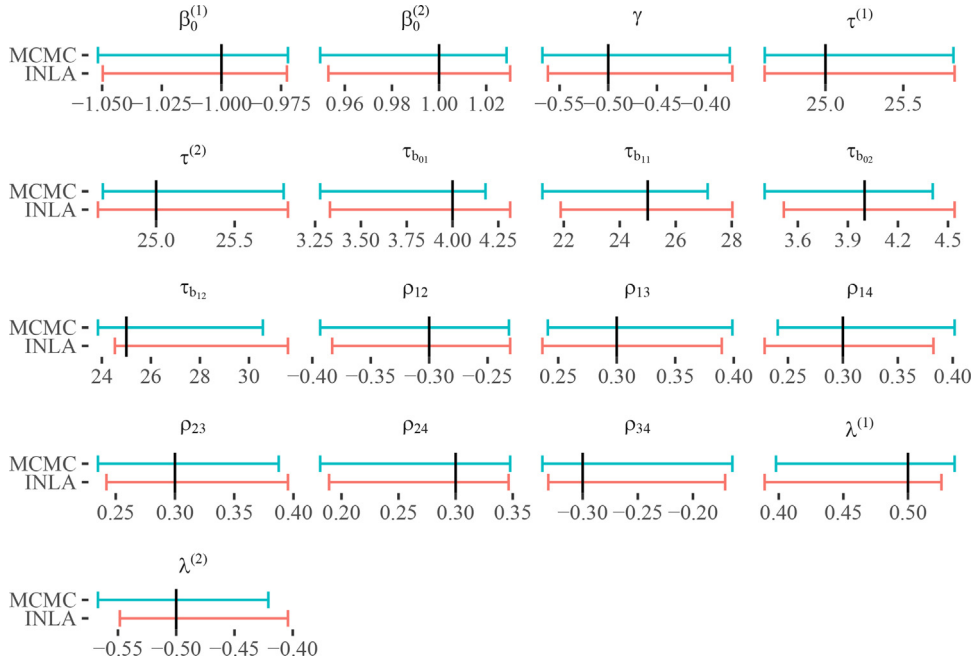

Fig. B.1. Credible intervals (2.5% – 97.5%) obtained by the MCMC and INLA implementations, for each parameter in the simulation analysis. The solid vertical line corresponds to the true parameter value.

Acknowledgements

Raffaella Calabrese acknowledges the support of the ESRC Project code ES/W010259/1.

Appendix A. Individual survival prediction under Laplace approximation

We aim to detail Eq. (6). The first approximation considers that we have already estimated the posterior distribution of the complete set of parameters Θ and we can rely on a point estimate denoted by $\hat{\Theta}$. Then, note that $P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta})$ can be marginalised by the random effects $\mathbf{U}$ as follows

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) = \int P(T_k > c + \Delta c | T_k > c, \mathbf{U}, \hat{\Theta}) P(\mathbf{U} | T_k > c, \mathcal{Y}_k, \hat{\Theta}) d\mathbf{U}$$

where we have used that T_k and $\mathcal{Y}_k$ are conditional independent given $\mathbf{U}$. Furthermore, $P(\mathbf{U}|T_k > c, \mathcal{Y}_k, \hat{\Theta})$ can be expressed as $P(T_k > c, \mathcal{Y}_k, \mathbf{U}|\hat{\Theta}) / P(T_k > c, \mathcal{Y}_k|\hat{\Theta})$ and the term $P(T_k > c, \mathcal{Y}_k|\hat{\Theta})$ can be marginalised as $\int P(T_k > c, \mathcal{Y}_k, \mathbf{U}|\hat{\Theta}) d\mathbf{U}$. This lets us write the following expression

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) = \frac{\int P(T_k > c + \Delta c | T_k > c, \mathbf{U}, \hat{\Theta}) P(T_k > c, \mathcal{Y}_k, \mathbf{U}|\hat{\Theta}) d\mathbf{U}}{\int P(T_k > c, \mathcal{Y}_k, \mathbf{U}|\hat{\Theta}) d\mathbf{U}}.$$

Define $-c \cdot h_k(\mathbf{U}) = \log\{P(T_k > c, \mathcal{Y}_k, \mathbf{U}|\hat{\Theta})\}$ and $g(\mathbf{U}) = P(T_k > c + \Delta c | T_k > c, \mathbf{U}, \hat{\Theta})$. This brings us to the following

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) = \frac{\int g(\mathbf{U}) \exp\{-c \cdot h_k(\mathbf{U})\} d\mathbf{U}}{\int \exp\{-c \cdot h_k(\mathbf{U})\} d\mathbf{U}}. \quad \text{(A.1)}$$

Since $g(\mathbf{U}) > 0$, this last expression can be further approximated using the Laplace method described by Tierney & Kadane (1986), as follows

$$P(T_k > c + \Delta c | T_k > c, \mathcal{Y}_k, \hat{\Theta}) = \frac{|\Sigma^*|^{1/2} \exp\{-c \cdot h_k^*(\mathbf{U}^*)\}}{|\tilde{\Sigma}|^{1/2} \exp\{-c \cdot h_k(\hat{\mathbf{U}})\}} + O\left(\frac{1}{c^2}\right)$$

where $-c \cdot h_k^*(\mathbf{U}) = -c \cdot h_k(\mathbf{U}) + \log g(\mathbf{U})$. The vectors $\mathbf{U}^*$ and $\hat{\mathbf{U}}$ are the arguments of the maxima of $-h_k^*(\cdot)$ and $-h_k(\cdot)$, respectively. Σ^* and $\tilde{\Sigma}$ are the inverse of the Hessians for h_k^* and h_k, respectively, evaluated at $\mathbf{U}^*$ and $\hat{\mathbf{U}}$.

Note also that we can recover the approach of Rizopoulos (2012) described in Section 2.3 as the first order approximation of Eq. (A.1) by applying the Laplace method separately in the numerator and denominator of Eq. (A.1). Explicitly,

$$\begin{aligned}
&\frac{\int g(\boldsymbol{U}) \exp\{-c \cdot h_k(\boldsymbol{U})\} d\boldsymbol{U}}{\int \exp\{-c \cdot h_k(\boldsymbol{U})\} d\boldsymbol{U}} \\
&= \frac{g(\hat{\boldsymbol{U}}_k)(2\pi/c)^{p/2} |\tilde{\Sigma}|^{1/2} \exp\{-c \cdot h_k(\hat{\boldsymbol{U}}_k)\}[1 + O_1(1/c)]}{(2\pi/c)^{p/2} |\tilde{\Sigma}|^{1/2} \exp\{-c \cdot h_k(\hat{\boldsymbol{U}}_k)\}[1 + O_2(1/c)]} \\
&= g(\hat{\boldsymbol{U}}_k)\left[1 + \frac{O_1(1/c) - O_2(1/c)}{1 + O_2(1/c)}\right] \\
&= g(\hat{\boldsymbol{U}}_k) + O\left(\frac{1}{c}\right) \\
&= \frac{P(T_k > c + \Delta c | \hat{\boldsymbol{U}}_k, \hat{\Theta})}{P(T_k > c | \hat{\boldsymbol{U}}_k, \hat{\Theta})} + O\left(\frac{1}{c}\right)
\end{aligned}$$

Appendix B. Comparison between MCMC and INLA estimations

The purpose of this section is to illustrate how fast and accurate is the INLA methodology in comparison with an MCMC sampling scheme for the multivariate joint model presented in this manuscript. To this extend, we implement the multivariate joint model using the platform for statistical modelling Stan with the No-U-Turn Sampler (NUTS Hoffman & Gelman, 2014), which is regarded as a faster extension to Hamiltonian Monte Carlo algorithm.[10] To assess convergence of the NUTS sampler, we performed the sampling from 3 independent chains with overdis-

[10] As additional experiments, we also implemented the model in Stata, using the `merlin` package (Crowther, 2020). However, we faced convergence problems and longer computation periods than the MCMC approach.

Table B.1
Time required, in minutes, for model estimation through MCMC and INLA schemes, as a function of the number of loans (N_{id}).

N_{id}	T_{MCMC}(min)	T_{INLA}(min)
250	106.03	1.07
300	127.79	1.31
350	159.52	1.50
400	187.70	1.99
450	207.00	2.00
500	256.12	2.42

persed starting points and, following the general diagnosis detailed in Betancourt (2017), no convergence problems were detected.

For the simulation setting described in Section 3, we estimate the model via MCMC and INLA using exactly the same computational resources (6 CPU cores, each with 4 GB of memory). We measure the times each procedure takes for different numbers of simulated loans (N_{id}), ranging from 250 to 500. The times, in minutes, are shown in the Table B.1. We can observe that considering, for example, a sample with 300 simulated loans, which is fairly small, the time required by the MCMC estimation is more than 2 hours, whereas for the INLA version is less than 2 minutes.

Moreover, for the biggest sample size simulated in this analysis ($N_{id} = 500$), we save the marginal posterior distributions for each parameter in the simulation setting (see Section 3), and estimate their credible intervals obtained by both implementations. The comparison of the 2.5% – 97.5% credible intervals are shown in Figure B.1. First, we notice that both implementations estimate intervals that include the true parameter value for all the parameters in the simulation setting and, second, both interval estimations are fairly similar which evidences and supports the quality of the Bayesian inference approximation performed by INLA for our model specification.

In the original paper (Rue et al., 2009), it is demonstrated how accurate and fast INLA is compared to MCMC schemes. However, INLA is not considered a replacement for MCMC. As we mentioned in Section 2.2, INLA is used on the class of latent Gaussian models (LGMs) and, although this is a broad class, if the purpose is to estimate a model outside the realm of LGMs, we would need to use other approaches. Furthermore, the computational cost is exponential with respect to the number of hyperparameters. Thus, the main computational advantage is obtained when the number of hyperparameters is moderate as in our case (15). Further comparisons in this regard can be found in Rue et al. (2009), Held, Schrödle, & Rue (2010), Carroll et al. (2015).

Appendix C. Time-fixed covariates distributions

Figure C.1 shows the distribution for the six fixed covariates that were provided: 4 continuous and 2 categorical. These covariates are in addition to the TVCs mentioned in Section 4.1. Due to data confidentiality agreements, not all covariates can be named and the x-axis of the plots are omitted. We also show next to the covariate name, in parentheses, the parameter estimate sign. The signs are consistent among all the estimated models. Note that the effect of age and debt-to-income is negative for the probability of prepayment, however, the effect is positive for the interest rate given at origination and for the amount of the loan.

Similar consumer loan datasets can be found, for example, in Tong, Mues, & Thomas (2012), Luong & Scheule (2021).

Appendix D. Univariate versus multivariate joint models

To complement our empirical analysis, in this section we estimate two additional joint models. Both are univariate versions of the proposed model described in Section 2, where each incorporates one of the two longitudinal processes. Namely, *Uni_1* and *Uni_2* include the first and the second longitudinal outcome, respectively.

The aim of this analysis is to illustrate what is the advantage of using a joint model that includes both longitudinal processes compared to the univariate version. Table D.1 shows the results of the metrics used in Section 4. In this respect, we can mention two things. First, if we compare the results obtained above (Table 2), we observe that the univariate joint models applied to this data are superior to both Cox models. These results are consistent with those shown in Hu & Zhou (2019). Second, we see that when using the joint models with both longitudinal processes, on average, the

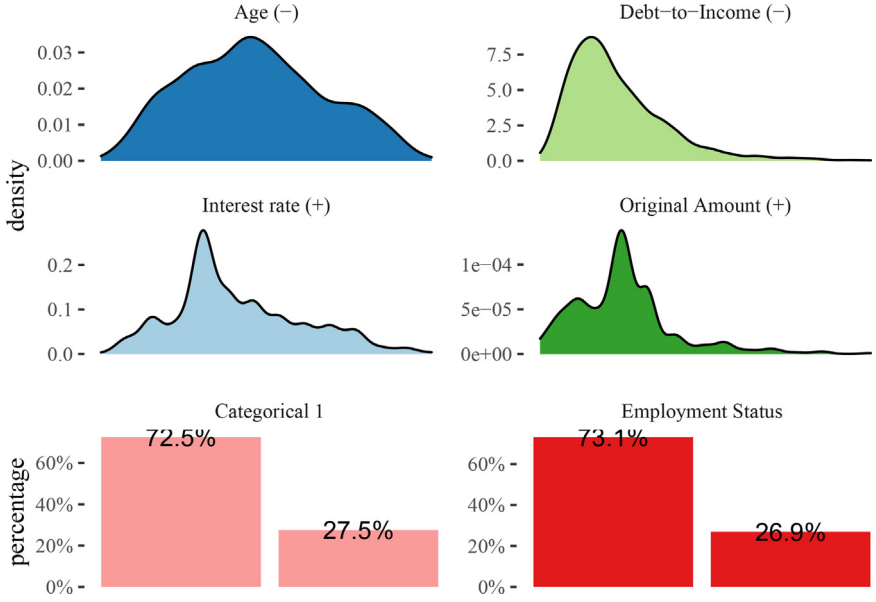

Fig. C.1. Distribution of the time-fixed covariates included in the survival model. For the bank privacy concerns, some information is omitted. The sign in parentheses is the sign of the parameters estimates.

Table D.1
Comparison of the discrimination (C^{12}_{AUC}) and calibration (C^{12}_{PE}) metrics between the uni- and multivariate versions of the joint models for a prediction window of 12 months. Each fold number represents the validation fold in the cross-validation analysis. The last row is the average (Avg) among the 10 folds and the bold number is the best performance metric within each validation fold.

	Uni_1		Uni_2		JM1		JM2	
Fold	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}	C^{12}_{AUC}	C^{12}_{PE}
1	0.5691	0.3193	0.5806	0.3152	0.5853	**0.3119**	**0.5869**	**0.3119**
2	0.6012	0.4090	0.6071	0.4075	0.6125	**0.4047**	**0.6135**	0.4048
3	**0.6098**	0.3169	0.5981	0.3201	0.6053	**0.3155**	0.6051	0.3156
4	0.6132	0.3715	0.6154	0.3706	**0.6228**	**0.3672**	0.6191	0.3675
5	0.5208	0.3738	0.5391	0.3727	**0.5416**	**0.3689**	0.5408	0.3691
6	0.6985	0.3105	0.6970	0.3096	0.6996	0.3089	**0.7009**	**0.3084**
7	0.6639	0.2849	0.6706	0.2865	0.6716	**0.2825**	**0.6764**	0.2835
8	**0.6344**	0.3106	0.5921	0.3137	0.5995	0.3094	0.6016	**0.3092**
9	**0.7752**	0.3082	0.7664	0.3075	0.7702	0.3047	0.7703	**0.3046**
10	0.5971	0.3298	0.6181	0.3285	0.6178	0.3268	**0.6195**	**0.3263**
Avg	0.6283	0.3334	0.6285	0.3332	0.6326	**0.3301**	**0.6334**	**0.3301**

models' performance increases in comparison with their univariate versions.

Appendix E. Robustness checks

Despite using weakly-informative priors in this work (see Section 2.2), we want to study how robust the results are under different prior specifications. We re-estimate the most complex joint model (*JM2*) using laxer priors over the parameters. To be precise, for all fixed effects, we use zero-mean Gaussian priors with a precision parameter equal to 0.001, instead of 0.01. For the precisions of the longitudinal outcomes and the baseline terms, we preserve the already weakly priors and for the precision matrix $Q_{\mathbf{U}}$ we assumed a less informative Wishart distribution than before with $\mathcal{W}_4(\mathbf{I}, 6)$.

Table E.1 shows the difference for the mean and the standard deviation between the estimations obtained with the different priors. Due to confidentiality concerns, we cannot disclose parameter estimates, hence the difference is expressed in percentage relative to the estimations obtained with the priors used in this work. The differences suggest that the results obtained with the priors described in Section 2.2 are consistent with the results obtained by using less informative ones.

Table E.1
The difference in parameters and standard deviations between the estimations done with the chosen priors and laxer ones. Due to confidentiality issues, the difference is expressed in percentage relative to the estimations obtained with the priors used in this work.

	Mean (%)	SD (%)
$\beta_0^{(1)}$	0.01	0.03
$\beta_1^{(1)}$	0.00	0.02
$\beta_0^{(2)}$	0.00	0.02
Interest rate	0.04	0.03
Age	0.21	0.01
Debt-to-Income	0.23	0.03
Categorical 1	0.23	0.01
Original Amount	0.37	0.00
Employment Status	0.01	0.01
$\tau^{(1)}$	0.00	0.20
$\tau^{(2)}$	0.00	0.01
τ_0	0.63	0.85
$\tau_{U_{01}}$	0.03	0.38
$\tau_{U_{11}}$	0.02	0.22
$\tau_{U_{02}}$	0.02	0.01
$\tau_{U_{12}}$	0.03	0.07
ρ_{12}	0.02	0.33
ρ_{13}	0.20	0.03
ρ_{14}	0.03	0.03
ρ_{23}	0.39	0.05
ρ_{24}	0.00	0.03
ρ_{34}	0.02	0.14
$\lambda^{(1)}$	0.11	0.15
$\lambda^{(2)}$	0.30	0.01

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2023.03.012.

References

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology, 13*, 61–98.
Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J., & Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine, 33*(18), 3167–3178.
Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society, 50*(12), 1185–1190.
Bartus, T. (2017). Multilevel multiprocess modeling with gsem. *The Stata Journal, 17*(2), 442–461.
BCBS (2019). The basel framework. *Technical Report*. Bank for International Settlements. https://www.bis.org/basel_framework/
Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society, 60*(12), 1699–1707.
Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting, 29*(4), 563–574.
Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *The Journal of Machine Learning Research, 18*(1), 2653–2688.
Betancourt, M. (2017). A conceptual introduction to hamiltonian Monte Carlo. arXiv: 1701.02434.
Bhattacharya, A., Wilson, S. P., & Soyer, R. (2019). A Bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research, 274*(3), 1112–1124.
Brown, E. R., Ibrahim, J. G., & DeGruttola, V. (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics, 61*(1), 64–73.
Calabrese, R., & Crook, J. (2020). Spatial contagion in mortgage defaults: A spatial dynamic survival model with time and space varying coefficients. *European Journal of Operational Research, 287*(2), 749–761.
Campbell, J. Y., & Cocco, J. F. (2015). A model of mortgage default. *The Journal of Finance, 70*(4), 1495–1554.
Carroll, R., Lawson, A., Faes, C., Kirby, R. S., Aregay, M., & Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical poisson modeling in disease mapping. *Spatial and Spatio-Temporal Epidemiology, 14*, 45–54.
Chi, Y.-Y., & Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics, 62*(2), 432–445.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187–220.
Crook, J., & Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(2), 283–305.
Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447–1465.

Crowther, M. J. (2020). merlin–A unified modeling framework for data analysis and methods development in stata. *The Stata Journal, 20*(4), 763–784.
Dafni, U. G., & Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 1445–1462.
De Andrade, F. W. M., & Thomas, L. (2007). Structural models in consumer credit. *European Journal of Operational Research, 183*(3), 1569–1581.
Djeundje, V. B., & Crook, J. (2019). Identifying hidden patterns in credit risk survival data using generalised additive models. *European Journal of Operational Research, 277*(1), 366–376.
Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research, 249*(2), 427–439.
Foote, C. L., Gerardi, K., & Willen, P. S. (2008). Negative equity and foreclosure: Theory and evidence. *Journal of Urban Economics, 64*(2), 234–245.
Furgal, A. K., Sen, A., & Taylor, J. M. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review, 87*(2), 393–418.
Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological), 14*(1), 107–114. http://www.jstor.org/stable/2984087
Gross, D. B., & Souleles, N. S. (2002). An empirical analysis of personal bankruptcy and delinquency. *The Review of Financial Studies, 15*(1), 319–347.
Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research, 295*(1), 292–305.
Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning, 77*(1), 103–123.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.
Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association, 247*(18), 2543–2546.
Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical modelling and regression structures* (pp. 91–110). Springer.
Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics, 1*(4), 465–480.
Henderson, R., Diggle, P., & Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics, 3*(1), 33–50.
Hickey, G. L., Philipson, P., Jorgensen, A., & Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology, 16*(1), 117.
Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593–1623.
Hu, W., & Zhou, J. (2019). Joint modeling: An application in behavioural scoring. *Journal of the Operational Research Society, 70*(7), 1129–1139.
Kalbfleisch, J. D., & Prentice, R. L. (2002). The statistical analysis of failure time data. *Wiley Series in Probability and Statistics* (2nd ed.). John Wiley & Sons.
Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963–974.
Leow, M., & Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research, 249*(2), 457–464.
Lindgren, F., & Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics, 35*(4), 691–700.
Luong, T. M., & Scheule, H. (2021). Benchmarking forecast approaches for mortgage credit risk for forward periods. *European Journal of Operational Research*.
Medina-Olivares, V., Calabrese, R., Crook, J., & Lindgren, F. (2022). Joint models for longitudinal and discrete survival data in credit scoring. *European Journal of Operational Research*. Accepted/in press 2022
Nadeau, C., & Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems* (pp. 307–313).
Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics, 67*(3), 819–829.
Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press.
Rizopoulos, D., & Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine, 30*(12), 1366–1380.
Rizopoulos, D., Molenberghs, G., & Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal, 59*(6), 1261–1276.
Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 71*(2), 319–392.
Shi, Y., Peng, Y., Xu, W., & Tang, X. (2002). Data mining via multiple criteria linear programming: Applications in credit card portfolio management. *International Journal of Information Technology & Decision Making, 1*(01), 131–151.
Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 1–28.
Sweeting, M. J., & Thompson, S. G. (2011). Joint modelling of longitudinal and time–to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal, 53*(5), 750–763.
Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*: vol. 2. Society for Industrial and Applied Mathematics.
Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81*(393), 82–86.
Tong, E. N., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218*(1), 132–139.
Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika, 88*(2), 447–458.
Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time–to-event data: An overview. *Statistica Sinica, 14*(3), 809–834.
Tutz, G., Schmid, M., et al., (2016). *Modeling discrete time-to-event data*. Springer.
Van Niekerk, J., Bakka, H., & Rue, H. (2019). Joint models as latent Gaussian models - not reinventing the wheel. arXiv:1901.09365
Wang, Z., Crook, J., & Andreeva, G. (2020). Reducing estimation risk using a Bayesian posterior distribution approach: Application to stress testing mortgage loan default. *European Journal of Operational Research, 287*(2), 725–738.
Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association, 64*(327), 1073–1078.
Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 175–188.
Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics, 53*(1), 330–339.