Joint models for longitudinal and discrete survival data in credit scoring

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 for longitudinal and discrete survival data in credit scoring

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

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

ARTICLE INFO

Article history:
Received 22 December 2020
Accepted 12 October 2022
Available online xxx

Keywords:
OR in banking
Bayesian joint models
Discrete time
Autoregressive process

ABSTRACT

The inclusion of time-varying covariates into survival analysis has led to better predictions of the time to default in behavioural credit scoring models. However, when these time-varying covariates are endogenous, there are two major problems: estimation bias of the survival model and lack of a prediction framework for future values of both the event and the endogenous time-varying covariates. Joint models for longitudinal and survival data is an appropriate framework to model the mutual evolution of the survival time and the endogenous time-varying covariates. To the best of our knowledge, this paper explores for the first time the application of discrete-time joint models to credit scoring. Moreover, we propose a novel extension to the joint model literature by including autoregressive terms in modelling the endogenous time-varying covariates. We present the method via simulations and by applying it to US mortgage loans. The empirical analysis shows, first, that discrete joint models can increase the discrimination performance compared to survival models. Second, when an autoregressive term is included, this performance can be further improved.

1. Introduction

Payment default is a specific event in credit risk analysis and refers to the inability of a borrower to pay its debts in a timely way. The Basel capital framework defines the event as the moment at which the borrower is past due more than 90 days in any credit obligation (BSBS, 2004) and is the standard definition used among practitioners[1] Credit scoring models aim to estimate the probability of default (PD) for each borrower or potential applicant based on past payment behaviour. These models allow banks to quickly assess the creditworthiness of new applicants (application scoring), monitoring the default risk of ongoing borrowers (behavioural scoring) and help to calculate provisions and capital levels for both expected and unexpected losses (BSBS, 2017). We are interested in predicting when and who is going to default in the presence of endogenous time-varying covariates for fixed-rate US mortgages.

Survival approaches, widely used in credit risk modelling (Bellotti & Crook, 2014; Calabrese & Crook, 2020; Djeundje & Crook, 2019; Leow & Crook, 2014; Stepanova & Thomas, 2002), are flexible enough to model when and who is likely to experience the event without the need to pre-define a performance period as required by classification approaches (Thomas, Crook, & Edelman, 2017). The ubiquitous survival model assumes a proportional hazard for continuous-time, or proportional odds model for discrete time, both easily extended to handle time-varying covariates (TVCs) only if these are exogenous (Allison, 1982; Cox, 1972). TVCs are generally included in credit survival models to either increase the accuracy of the predictions (Calabrese & Crook, 2020; Crook & Bellotti, 2010; Stepanova & Thomas, 2002) or enhance the understanding of why borrowers default (Dirick, Bellotti, Claeskens, & Baesens, 2019; Djeundje & Crook, 2018). However, in this context there is little work that addresses, first, the distinction among the types of the TVCs included (exogenous or not) and, second, how to model the time to default when we have endogenous TVCs (see Section 2 for further details of exogenous versus endogenous TVCs).

A rapidly evolving field of statistical methodology, known as "joint models for longitudinal and time-to-event data" (joint models, henceforth) (Henderson, Diggle, & Dobson, 2000; Rizopoulos, 2012; Tsiatis, Degruttola, & Wulfsohn, 1995), addresses the problem of endogeneity by modelling both the time to event and the

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

[1] The Basel definition of the default has also a qualitative component that relies on the institution's criteria. We only consider here the quantitative definition.

https://doi.org/10.1016/j.ejor.2022.10.022

endogenous TVCs[2], simultaneously. This approach, in addition to avoiding estimation biases by considering the mutual evolution of both processes (Tsiatis et al., 1995; Wulfsohn & Tsiatis, 1997), allows us to develop a dynamic risk prediction model that can make greater use of the newly collected data. Other research fields also show that joint models increase the model accuracy (Rizopoulos, Hatfield, Carlin, & Takkenberg, 2014).

We make two contributions to the literature. First, to the best of our knowledge, this is the first time joint models for discrete survival data are investigated in credit risk prediction. Although a recent work, Hu & Zhou (2019), applies joint models for predicting early repayment events on mortgage loans and default events on a peer-to-peer data set, the authors consider survival time as continuous, which is the usual assumption in the joint model literature (Hickey, Philipson, Jorgensen, & Kolamunnage-Dona, 2016; Lawrence Gould et al., 2015) and in the variety of software available to estimate them (see Furgal, Sen, & Taylor (2019) for a comparison among different computational approaches). In credit risk analysis, however, there are three reasons why the discrete time approach should be preferred over the continuous one. First, account records are observed monthly so events are intrinsically discrete. Therefore, the continuous approach would end up being an approximation (Tutz & Schmid, 2016). Second, many events naturally happen in the same month. The continuous approach implies that there will be no tied events, however, with the discrete approach, ties are perfectly fine. Finally, with the discrete approach, the probability predictions require simple summations over time-points, rather than integrations which are complex when TVCs are included (Bellotti & Crook, 2013; Leow & Crook, 2016; Rizopoulos, 2012). This makes the model computationally efficient and capable of scaling to sample sizes like the one presented here with more than 285K observations.

Second, in relation to the endogenous TVCs, it is standard in the joint model literature to address subjects' heterogeneity via random effects (see details Section 4). In our case, and making use of the fact that the observations are equally-spaced and indexed by a discrete variable (time), we propose to also include autoregressive terms to model the dependency among borrower's observations. This longitudinal approach belongs to the category known as linear mixed-effects models (LME) with serial correlation (Hedeker & Gibbons, 2006). It is motivated by, first, the serial correlation found in our application and, second, the possible implications for the predictions that this correction might have. The empirical autocorrelation functions (ACF, see Pinheiro & Bates, 2006) for the longitudinal outcome used in this work showed autocorrelated residuals for two common LME specifications in the joint model literature, namely *random intercept* and *random intercept and slope* (see Appendix A). Although these are not joint models, they provide relevant bases for the existence of serial correlation in the longitudinal outcome and therefore to support our proposal.

We implement six models in total, all of them coded in the platform for statistical modelling *Stan* and available in the supplementary material. One, discrete survival model, provides our benchmark, and five other discrete joint models which differ in the specification of random effects and the inclusion of the autoregressive term. The simulation analysis shows good convergence diagnostics and recovery of the true parameter values. To estimate a scoring model for mortgage loans, we use the Single Family Loan-Level Dataset from Freddie Mac that is publicly available. We perform a cross-validation analysis that shows, first, that the discrete joint models approaches can increase the discrimination performance compared to the discrete survival model. Second, when an autoregressive term is included, this performance can be further improved. These results are enhanced when more historical data are included in the prediction.

The paper is organised as follows. In Section 2, we describe the differences among exogenous and endogenous TVCs in the discrete survival context. Section 3 shows the relevant literature on joint models. Section 4 details the methodology for the discrete-time setting with the extension of additional autoregressive terms, how the inference and the individual survival predictions are performed and assessed. Section 5 presents a simulation study of the discrete joint model with autoregressive terms and Section 6 illustrates an application to US mortgages. The concluding remarks follow in Section 7.

2. Exogenous versus endogenous TVCs

TVCs can be broadly categorised into two general classes: exogenous and endogenous (Kalbfleisch & Prentice, 2002; Rizopoulos, 2012). Exogenous TVCs are variables whose future paths are not affected by the occurrence of the event (in our case default) or that are not correlated with variables that are omitted from the model but also affect the outcome. In contrast, endogenous TVCs are variables whose path is influenced by the survival status (default) of the individual and therefore carries direct information on the time to the default or that are correlated with omitted variables that also affect the occurrence of default. More specifically, assume $y_1, \ldots, y_t$ is the sequence of observations until time t of a generic TVC represented by $\{Y_s\}_{s\leq t}$. Denote the survival time as T and represent it by an indicator variable X_t such that $(x_1, \ldots, x_{t^*}) = (0, \ldots, 0, 1)$ if $T = t^*$. Hence, the joint probability of the stochastic process $\{X_s, Y_s\}_{s\leq t}$ can be written as

$$P(\{X_s, Y_s\}_{s\leq t}) = P(X_t, Y_t|\{X_s, Y_s\}_{s<t}) \\ P(X_{t-1}, Y_{t-1}|\{X_s, Y_s\}_{s<t-1}) \cdot \ldots \cdot P(X_1, Y_1),$$

where any term on the right hand side follows

$$P(X_t, Y_t|\{X_s, Y_s\}_{s<t}) = P(X_t|Y_t, \{X_s, Y_s\}_{s<t})P(Y_t|\{X_s, Y_s\}_{s<t}).$$

The main difference between exogenous and endogenous TVCs is in the assumption of $P(Y_t|\{X_s, Y_s\}_{s<t})$. For the former, it is assumed that Y_t is independent of the survival status X_s and thus the term $P(Y_t|\{X_s, Y_s\}_{s<t})$ does not affect the parameters estimation in the hazard $P(X_t|Y_t, \{X_s, Y_s\}_{s<t})$. For the endogenous case though, that does not hold.

Examples of exogenous TVCs in the credit modelling context are the macroeconomic variables such as the inflation rate, GDP and unemployment rate (Bellotti & Crook, 2009), where their paths may influence the rate of default over time but their future values are not affected by a loan's default. Some examples of endogenous case are the spending and repayment amounts, outstanding balance and arrears in instalments, among others.

The consequence of including a TVC into a Cox model is that, for any individual, the probability of surviving longer than time t given that we have measured the TVC until t is a survival function when the TVC is exogenous, meaning that the usual relationship between the hazard and survival functions holds and the estimation is obtained by simply maximising the Cox's partial likelihood (Cox, 1975). However, when the TVC is endogenous, the probability of surviving longer that t given that we have measured the TVC until t is equal to 1 (and is no longer a survival function) since we know that the individual is still "alive" at t and will, for sure, survive longer than t (see Kalbfleisch & Prentice (2002), chap. 6 for a detailed discussion). This mutual evolution between the survival data and the endogenous TVC has direct implications for how the prediction and estimation are done and we can no longer rely on the standard Cox procedure, requiring, consequently, methodological alternatives. The joint model approach addresses this problem

[2] To unify the jargons between the literature of joint models and credit scoring, the endogenous TVCs will be also termed here as longitudinal outcomes.

basically by assuming conditional independence between X_t and Y_t given an underlying random effect U (see Section 4).

3. Literature review

The inclusion of TVCs in the prediction framework changes the traditional approach of estimating the probability of an event occurring t^* months after origination into a dynamic one in which the borrower is event-free after t months and historical information is available. Mathematically, this is $P(T > t^*|T > t, \{Y_s\}_{s\leq t})$. The vast literature on credit scoring includes TVCs for prediction, either by keeping the last observed values or by estimating hazard models with lagged TVC values (Bellotti & Crook, 2009; 2013; 2014; Crook & Bellotti, 2010; Divino & Rocha, 2013; Malik & Thomas, 2010; Thackham & Ma, 2020; Wang, Crook, & Andreeva, 2020). However, these papers do not control for potential endogeneity in the TVCs.

Other approaches to include TVCs in dynamic prediction have been made: multistate intensity models (Crook & Bellotti, 2010; Djeundje & Crook, 2018; Leow & Crook, 2014), Markov chain models (Crook & Bellotti, 2010; Thomas, Ho, & Scherer, 2001), Markov for discrimination (Volkov, Benoit, & Van den Poel, 2017), lifecycle and forward models (Luong & Scheule, 2021), survival models with flexible link functions (Calabrese & Crook, 2020) and boosting algorithms (Xia, He, Li, Fu, & Xu, 2021). Yet all of these approaches have omitted the simultaneous estimation of the time-to-event parameters and those relating to the TVC processes.

In the literature on dynamic prediction not connected to credit-related applications, there are mainly four approaches that address the relationship between endogenous TVCs and the survival process: backward modelling, latent class models, landmarking and forward modelling (van Houwelingen & Putter, 2011). The backward approach estimates the conditional distribution $\{Y_s\}_{s\leq t}|T = t^*$ and the marginal distribution of $T = t^*$. Hence, Bayes' theorem gives the dynamic prediction (Fieuws, Verbeke, Maes, & Vanrenterghem, 2008). This approach, however, requires imputing censored observations, which might be problematic for some survival distributions.

In the latent class models, the assumption is that Y_t and T are conditionally independent given an unobserved latent class whose predictive role is similar to the one in a survival model with frailties (Henderson et al., 2000; Proust-Lima & Taylor, 2009). For some applications, however, the assumption of different classes of subjects might not be suitable. Landmarking, on the other hand, estimates the probability at t^* by building a Cox model only with the subjects at risk at t (Van Houwelingen, 2007). However, since no joint modelling for T and Y_t is performed, it does not offer a major understanding of the underlying link between them. Finally, the forward approach, which we follow in this work, estimates the conditional distribution $T|\{Y_s\}_{s\leq t}$ directly through the hazard function and the prediction is based on the posterior predictive distribution (Rizopoulos, 2012; Tsiatis et al., 1995; Wulfsohn & Tsiatis, 1997). Rizopoulos, Molenberghs, & Lesaffre (2017) show predictive benefits of the forward modelling approach over landmarking due mainly to its flexibility when modelling the longitudinal outcome.

Most of the literature on joint models comes from medical research where the interest lies in the association between the repeated measurements of a biomarker for a patient and her survival time (Tsiatis et al., 1995), but the approach can be applied in any area where the link between both processes is of interest. The standard joint model is formed by two sub-models, one for the survival data and the other for the longitudinal outcome, both assumed to be conditionally independent given a latent structure. The survival process is commonly modelled by assuming a Cox model and a linear mixed-effects model for the longitudinal part (Rizopoulos, 2012). Both sub-models are associated through a functional form that could adopt many different structures (see Hickey et al. (2016)). A thorough review of this topic can be found in Tsiatis & Davidian (2004) who clarifies the main assumptions employed in the likelihood function. The textbooks Rizopoulos (2012) and Fitzmaurice, Davidian, Verbeke, & Molenberghs (2008, Ch.13–16) provide a comprehensive explanation of the technique, its inference and possible extensions. Moreover, Alsefri, Sudell, García-Fiñana, & Kolamunnage-Dona (2020) gives a summary of the recent developments and issues.

Although most of the literature assumes survival time as continuous, there are works that study the discrete case. Albert & Shih (2010b) propose a two-stage approximation method for estimation in which the discrete hazard is modelled on the probit scale, which was extended later in Albert & Shih (2010a) to handle multiple longitudinal outcomes. Jaffa, Woolson, & Lipsitz (2011) are more interested in the longitudinal process rather than the survival. They introduce a joint model with bivariate longitudinal outcomes adjusted by informative right censoring using a discrete survival approach, then extended in Jaffa, Gebregziabher, & Jaffa (2014) for a high dimensional multivariate case. More in line with our work, Barrett, Diggle, Henderson, & Taylor-Robinson (2015) propose an exact likelihood inference when the discrete hazard adopts a probit model by using distributional properties of the skew normal family. They also include an unobserved stationary Gaussian process in the longitudinal model to bring more flexibility when the follow-up period is relatively long. Moreover, Bacci, Bartolucci, & Pandolfi (2018) assume a logit model for the discrete process, as in this work, and consider random intercepts in the longitudinal model to change over time according to an autoregressive process of order 1. In the present work, we follow a similar approach to Bacci et al. (2018), but we propose instead to consider autoregressive terms explicitly in the longitudinal process and not restricted to order 1. That allows us to make prediction more straightforward and interpretable since we directly estimate the influence of past longitudinal observations on the forecast. Accordingly, the proposed model determines a correlation structure that assumes both the subject-specific correlation, through random effects, and that due to the natural evolution of the longitudinal outcome, through autoregressive terms. The goal is twofold. First, we control for the serial correlation found in our application (see Appendix Appendix A) and that the aforementioned approaches do not consider. Second, we expect that by assuming a more flexible correlative structure, the predictive performance can be improved, as later confirmed in Section 6.

In the credit modelling context and as far as we know, there is only one published paper, Hu & Zhou (2019), that applies joint modelling for behavioural scoring and supports the superiority over the Cox model with TVCs in discrimination, for prediction time windows of 2 and 3 months, and in calibration, for prediction time windows of 3 and 6 months. The authors point out that, due to the complexity and lack of software, the joint models approach has not been widely used until the recent decade with the appearance of some statistical packages. They used the R package `JM` (Rizopoulos, 2010) which does not allow the inclusion of autoregressive terms and only handles time as continuous. The potential of this approach in credit-related applications is also investigated in a Working Paper by Medina-Olivares, Lindgren, Calabrese, & Crook (2022). They estimate a discrete joint model but unlike this paper whilst it has more than one longitudinal outcome it omits autoregressive terms. They also suggest new ways to evaluate individual survival predictions.

Recent works in machine learning show interesting approaches related to survival analysis. Luck, Sylvain, Cardinal, Lodi, & Bengio (2017) and Katzman et al. (2018) use deep neural networks to exploit the ability to learn complex interactions of the covariates and show better performance than traditional survival analysis. How-

ever, these approaches are limited to time-fixed covariates. Alaa & van der Schaar (2017) and Bellot & Schaar (2018), also using deep architectures, develop respectively, a non-parametric Bayesian model and a tree-based Bayesian mixture model that can capture subject-specific representations similar to joint models but the survival prediction is restricted to the use of the last available measurement. Lee, Yoon, & Van Der Schaar (2019), aware of this limitation, introduce a deep network architecture that learns high-level relationships between the longitudinal outcome and the survival prediction and shows better discrimination performance than the traditional joint models approach. The adoption of deep architectures in credit scoring though, has been restricted largely due to the low interpretability of the predictions which is required if the model is used to make decisions since those who were rejected must be given a reason for their rejection. The search of interpretability mechanisms is a current research topic (Dastile, Celik, & Potsane, 2020).

4. Methodology

4.1. Framework

Assume we wish to model the time to default $T_i \in \mathbb{Z}_+$ for subject i $(i = 1, \ldots, N)$ in terms of time-invariant covariates $\mathbf{z}_i$ and a longitudinal outcome Y_{is} that is observed at times s with $s \in \{0, 1, 2, \ldots, t_i - 1\}$ where t_i is the time where either the event or the end of the follow-up happens. In theory, the number of observed values for the longitudinal outcome can differ from the survival times, but in our case we have equally spaced times and no missing observations before t_i, so we can unify the notation to $Y_{i,s-1}$ with $s = 1, \ldots, t_i$. As in Albert & Shih (2010a), we use $s-1$ because we relate the survival time with the immediate previous observed value of the longitudinal outcome[3] Analogously to the notation introduced in Section 2, we represent T_i as a sequence of binary indicators X_{is} which is 1 if the event happens at time s and 0 otherwise. The key assumption in the joint modelling approach is that X_{is} and $Y_{i,s-1}$ are conditional independent given the random effects $\mathbf{U}_i$, i.e. $P(\{X_{is}, Y_{i,s-1}\}_{s\le t_i}) = \int P(\{X_{is}\}_{s\le t_i}|\mathbf{U}_i)P(\{Y_{i,s-1}\}_{s\le t_i}|\mathbf{U}_i)p(\mathbf{U}_i)d\mathbf{U}_i$ and interest is now turned on how to model each of the three elements of the integrand.

For the survival part $P(\{X_{is}\}_{s\le t_i}|\mathbf{U}_i)$ and following Allison (1982), the probability that the event occurs at t_i is given by

$$P(\{X_{is}\}_{s\le t_i}|\mathbf{U}_i) = \prod_{s=1}^{t_i}[p_{is}]^{X_{is}}[1-p_{is}]^{1-X_{is}}, \tag{1}$$

where $p_{is} = P(X_{is} = 1|X_{is-1} = 0, \mathbf{U}_i)$. Assuming a logit link function, we can add the covariates as follows

$$p_{is} = \text{logit}^{-1}(a_s^0 + \mathbf{z}_i^\intercal \boldsymbol{\gamma} + \lambda_f f(\mathbf{U}_i, s)), \tag{2}$$

where a_s^0 represents the baseline event time distribution. Following Djeundje & Crook (2018), we specify the baseline with cubic B-spline functions, i.e. $a_s^0 = \mathbf{B}(s)^\intercal \boldsymbol{\gamma_0}$ with $\mathbf{B}$ the vector of B-spline functions and $\boldsymbol{\gamma_0}$ the corresponding vector of regression coefficients. Moreover, $\boldsymbol{\gamma}$ is the vector of coefficients for the covariates $\mathbf{z}_i$ and λ_f is known as the association coefficient between the survival and longitudinal processes. The function f relates both processes through the random effects $\mathbf{U}_i$ and, eventually, the time s. As mentioned before, f can adopt different structures (Hickey et al., 2016). In this work we study a set of combinations detailed in Section 6.2.

For the longitudinal part, assume that $Y_{i,s-1}$ can be described by an underlying signal $m_{i,s-1}$ and mutually independent noise terms $\epsilon_{i,s-1}$ as $Y_{i,s-1} = m_{i,s-1} + \epsilon_{i,s-1}$. Further, denote as $\mathbf{w}_i$ a vector of time-invariant covariates and as $\mathbf{q}_{i,s-1}$ a vector of time-varying exogenous covariates measured at time $s-1$. Then, assume that $m_{i,s-1}$ can be decomposed into fixed effects, $\mathbf{w}_i^\intercal\boldsymbol{\alpha} + \mathbf{q}_{is-1}^\intercal\boldsymbol{\beta}$, and random effects, $\mathbf{d}_{is-1}^\intercal\mathbf{U}_i$, where $\mathbf{d}_{i,s-1}$ is the design vector at time $s-1$. This leads to the following mixed-effect model (Laird & Ware, 1982)

$$Y_{i,s-1} = \underbrace{\mathbf{w}_i^\intercal\boldsymbol{\alpha} + \mathbf{q}_{i,s-1}^\intercal\boldsymbol{\beta} + \mathbf{d}_{i,s-1}^\intercal\mathbf{U}_i}_{m_{i,s-1}} + \epsilon_{i,s-1}, \quad s = 1, \ldots, t_i, \tag{3}$$

where the subject-level $\mathbf{U}_i$ are assumed as mutually independent and coming from a zero-mean multivariate Gaussian distribution of dimension d, $\mathbf{U}_i \sim \mathcal{N}_d(0, \Sigma)$. The error terms are assumed normally distributed $\epsilon_{i,s-1} \sim \mathcal{N}(0, \sigma^2)$, mutually independent and independent from the subject-level random effects $\mathbf{U}_i$.

Suppose now that the longitudinal process described above is also explained by an additional autoregressive structure of order p (Hedeker & Gibbons, 2006), then Eq. 3 can be modified as

$$Y_{i,s-1} = \underbrace{\mathbf{w}_i^\intercal\boldsymbol{\alpha} + \mathbf{q}_{i,s-1}^\intercal\boldsymbol{\beta} + \mathbf{d}_{i,s-1}^\intercal\mathbf{U}_i + \sum_{r=1}^{p}\phi_r Y_{i,s-1-r}}_{m_{i,s-1}} + \epsilon_{i,s-1}, \quad s = p+1, \ldots, t_i, \tag{4}$$

where ϕ_r $(r = 1, \ldots, p)$ represents the coefficient for the rth autoregressive term. Note that the endogenous variable is now correlated with both its own history and the subject-level random effects, but the conditional dependence structure for $Y_{i,s-1}$ given U_i follows a simple autoregression structure with conditional expectation $m_{i,s-1}$ and conditional variance σ^2 (see Eq. 7). To make the model well-specified, it is now assumed that none of the events occurred in $s \le p$ and Eqs. 1 and 2 are modified correspondingly and detailed below.

4.2. Estimation of the joint model with autoregressive terms

Denote the observed survival data for subject i $(i = 1, \ldots, N)$ as $\mathcal{X}_i = \{X_{is} : s = p+1, \ldots, t_i\}$ and its longitudinal measurements as $\mathcal{Y}_i = \{Y_{i,s-1} : s = 1, \ldots, t_i\}$, and represent by $\mathcal{D}_N = \{\mathcal{Y}_i, \mathcal{X}_i : i = 1, \ldots, N\}$ the complete observed data[4] The parameters to estimate are the B-spline coefficients $\boldsymbol{\gamma_0}$, the covariate coefficients $\boldsymbol{\gamma}$, the association parameter λ_f, the coefficients of the fixed effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the set of autoregressive coefficients $\{\phi\} = \{\phi_r : r = 1, \ldots, p\}$, the covariance matrix of the random effects Σ and the variance of the error terms σ^2. Denote the set of all these parameters as $\Theta = \{\boldsymbol{\gamma_0}, \boldsymbol{\gamma}, \lambda_f, \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\phi\}, \Sigma, \sigma^2\}$, thus the likelihood of the joint model with autoregressive terms $\mathcal{L}(\Theta|\mathcal{D}_N)$ is written as

$$\begin{aligned}\mathcal{L}(\Theta|\mathcal{D}_N) &= \prod_{i=1}^{N}\int P(\mathcal{X}_i, \mathcal{Y}_i|\mathbf{U}_i, \Theta)P(\mathbf{U}_i|\Theta)\mathrm{d}\mathbf{U}_i \\ &= \prod_{i=1}^{N}\int P(\mathcal{X}_i|\mathcal{Y}_i, \mathbf{U}_i, \Theta)P(\mathcal{Y}_i|\mathbf{U}_i, \Theta)P(\mathbf{U}_i|\Theta)\mathrm{d}\mathbf{U}_i.\end{aligned} \tag{5}$$

Following the Gaussian assumption on $\mathbf{U}_i$, the last term of the integrand in Eq. 5 is

$$\begin{aligned}P(\mathbf{U}_i|\Theta) &= P(\mathbf{U}_i|\Sigma) \\ &= (2\pi)^{-d/2}\det(\Sigma)^{-1/2}\exp\left(-\mathbf{U}_i^\intercal\Sigma^{-1}\mathbf{U}_i/2\right).\end{aligned} \tag{6}$$

Moreover, we apply the chain rule and the assumption that the error terms are zero-mean Gaussian distributed, hence the second term of the integrand in Eq. 5 is

[3] We assume that no subjects experience a default at $s = 0$.

[4] Note that all the covariates previously mentioned are also observed but we intentionally omit them to avoid excess of notation.

$$
\begin{aligned}
P(\mathcal{Y}_i|\mathbf{U}_i,\Theta) &= P(Y_{i,t_i-1},.,Y_{i0}|\mathbf{U}_i,\Theta)\\
&= P(Y_{i,t_i-1}|Y_{i,t_i-2},\ldots,Y_{i0},\mathbf{U}_i,\Theta)P(Y_{i,t_i-2},\ldots,Y_{i0}|\mathbf{U}_i,\Theta)\\
&= \prod_{s=1}^{t_i-p} P(Y_{i,t_i-s}|Y_{i,t_i-s-1},\ldots,Y_{i0},\mathbf{U}_i,\Theta)P(Y_{i,p-1},\ldots,Y_{i0}|\mathbf{U}_i,\Theta)\\
&\propto \prod_{s=1}^{t_i-p} P(Y_{i,t_i-s}|Y_{i,t_i-s-1},\ldots,Y_{i,t_i-s-p},\mathbf{U}_i,\Theta)\\
&= \prod_{s=1}^{t_i-p}(2\pi\sigma^2)^{-1/2}\exp\left(-(Y_{i,t_i-s}-m_{i,t_i-s})^2/2\sigma^2\right), \qquad (7)
\end{aligned}
$$

where we know that Y_{i,t_i-s} $(s=1,\ldots,t_i-p)$ only depends on the previous p lags and the terms $Y_{i,p-1},\ldots,Y_{i0}$ are not informative to the parameters. m_{i,t_i-s} is defined as in Eq. 4.

Note that for the first term of the integrand in Eq. 5, $P(\mathcal{X}_i|\mathcal{Y}_i,\mathbf{U}_i,\Theta)$, the conditional dependency on $\mathcal{Y}_i$ will hold subject to the chosen structure for the link function $f(\mathbf{U}_i,s)$ (see Eq. 2). For example, if we consider the simple case where $f(\mathbf{U}_i,s)=\mathbf{d}_{i,s-1}^{\intercal}\mathbf{U}_i$, then $P(\mathcal{X}_i|\mathcal{Y}_i,\mathbf{U}_i,\Theta)=P(\mathcal{X}_i|\mathbf{U}_i,\Theta)$, which is the common assumption used in joint modelling (Rizopoulos, 2012). However, if $f(\mathbf{U}_i,s)=m_{i,s-1}$ with $m_{i,s-1}$ following Eq. 4, then we know that $m_{i,s-1}$ not only depends on the random effects but also on the previous p lag values of the longitudinal outcome. Assume this last case since generalises the other. Hence, following Eq. 1 we write

$$
P(\mathcal{X}_i|\mathcal{Y}_i,\mathbf{U}_i,\Theta)=\prod_{s=p+1}^{t_i}[p_{is}]^{X_{is}}[1-p_{is}]^{1-X_{is}}, \qquad (8)
$$

where p_{is} (see Eq. 2)

$$
p_{is}=\text{logit}^{-1}\left(\mathbf{B}(s)^{\intercal}\boldsymbol{\gamma_0}+\boldsymbol{z}_i^{\intercal}\boldsymbol{\gamma}+\lambda_f m_{i,s-1}\right). \qquad (9)
$$

Eqs. 6 –9 completely specify the likelihood in Eq. 5. Conceptually, this model could be estimated by maximising the log-likelihood. The standard algorithms such as EM, Newton's method or modifications of them with asymptotic approximations have been used in the literature for other joint models (Rizopoulos, 2012). However, the Bayesian approach has some advantages in this context such as that approximations are not required and the computational implementation is easier and more flexible (Ibrahim, Chen, & Sinha, 2014).

To complete the Bayesian model specification we define the prior distributions on the parameters $P(\Theta)$. For $\boldsymbol{\gamma}$, λ_f, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\{\phi\}$ and σ, we use noninformative uniform priors across each parameter's domain. For the B-spline coefficients, $\boldsymbol{\gamma_0}$, we assume a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0},\upsilon^2 I)$, where υ is a hyperparameter with a *half-Cauchy* prior with a scale of 25 (this is large enough to be "noninformative" (Gelman et al., 2013)). For the prior of the covariance matrix, Σ, we work on its decomposition between a vector of variances and a correlation matrix. For the variances we set noninformative uniform priors in the positive domain and for the correlation matrix the LKJ distribution with a regularisation parameter[5] of 2 (Lewandowski, Kurowicka, & Joe, 2009).

We implement this and the other models specified in Section 4 in the platform for statistical modelling Stan with the No-U-Turn Sampler (Hoffman & Gelman, 2014) which is a faster extension to Hamiltonian Monte Carlo algorithm (HMC). The code for this work can be found in the supplementary material.

4.3. Individual survival predictions

In this section we describe the methodology to predict how likely is the default for a new subject k not originally included in the observed data $\mathcal{D}_N$. Assume that this new subject has not defaulted yet at least until time c and we collect the longitudinal outcome up to time $c-1$. Denote this set of measurements by $\mathcal{Y}_k=\{Y_{k,s-1}:s=1,\ldots,c\}$. We are now focus on the conditional probability of surviving time $c+\Delta c>c$ $(\Delta c\in\mathbb{Z}_+)$ given that it has survived up to c, i.e. $P(T_k>c+\Delta c|T_k>c,\mathcal{Y}_k,\mathcal{D}_N)$. For the purpose of readability, denote this last term as $\pi_k(c+\Delta c|c)$.

However, since subject k is new to $\mathcal{D}_N$, we have no estimation of its random effects $\mathbf{U}_k$. One procedure to get the prediction is to include subject k into the training sample and rerun the estimation as described in Section 4.2, but this would be computationally expensive and not feasible if we apply it in a real-time fashion. A fast alternative is to adopt a first order approximation by using empirical Bayes estimates for the random effects as explained below (see Rizopoulos (2012) for a detailed description of the continuous time-setting). Formally, the conditional probability can be marginalised as

$$
\pi_k(c+\Delta c|c)=\int P(T_k>c+\Delta c|T_k>c,\mathcal{Y}_k,\Theta)P(\Theta|\mathcal{D}_N)\mathrm{d}\Theta, \qquad (10)
$$

where $P(\Theta|\mathcal{D}_N)$ is the posterior distribution of the parameters given the sample $\mathcal{D}_N$. The first term of the integrand can be written as

$$
\begin{aligned}
&P(T_k>c+\Delta c|T_k>c,\mathcal{Y}_k,\Theta)\\
&=\int P(T_k>c+\Delta c|T_k>c,\mathcal{Y}_k,\mathbf{U}_k,\Theta)\\
&\quad\times P(\mathbf{U}_k|T_k>c,\mathcal{Y}_k,\Theta)\mathrm{d}\mathbf{U}_k. \qquad (11)
\end{aligned}
$$

Joining together Eqs. 10 and 11, the first order approximation is given by $\pi_k(c+\Delta c|c)\approx P(T_k>c+\Delta c|T_k>c,\mathcal{Y}_k,\hat{\mathbf{U}}_k,\hat{\Theta})$. $\hat{\Theta}$ denotes the posterior point-estimate, the random effects estimates solve $\hat{\mathbf{U}}_k=\text{argmax}_{\mathbf{U}}\{\log P(T_k>c,\mathcal{Y}_k,\mathbf{U}|\hat{\Theta})\}$ and the prediction is performed as

$$
\begin{aligned}
\hat{\pi}_k(c+\Delta c|c) &= \frac{P(T_k>c+\Delta c|\mathcal{Y}_k,\hat{\mathbf{U}}_k,\hat{\Theta})}{P(T_k>c|\mathcal{Y}_k,\hat{\mathbf{U}}_k,\hat{\Theta})}\\
&= \frac{\prod_{s=p+1}^{c+\Delta c}(1-\hat{p}_{ks})}{\prod_{s=p+1}^{c}(1-\hat{p}_{ks})}\\
&= \prod_{s=c+1}^{c+\Delta c}(1-\hat{p}_{ks}), \qquad (12)
\end{aligned}
$$

where $\hat{p}_{ks}$ follows Eq. 9. The standard error of the above expression can be estimated through Monte Carlo simulation schemes as proposed in Rizopoulos (2011) and Proust-Lima & Taylor (2009).

4.4. Performance measures

We are interested in assessing the models by discrimination and calibration performance in the presence of right-censoring and given that we know that the loans have not yet defaulted up to a time point c. The metrics come from the literature mentioned below and we adapt the notation to the discrete case as follows.

For discrimination, a common measure is the Area Under the ROC curve (AUC) (Fawcett, 2006) which is the area enclosed by the curve formed by the proportion of correctly predicted events versus the proportion of incorrectly classified events overall threshold values. An AUC of 1 represents a perfect classifier and 0.5 a random one. An alternative interpretation of the AUC between evaluation times c and $c+\Delta c$ reads that for any random pair of subjects $\{i,j\}$ the AUC can be formulated as (Hanley & McNeil, 1982)

$$
AUC_c^{\Delta c}=P(\pi_i(c+\Delta c|c)<\pi_j(c+\Delta c|c)|\{T_i\in(c,c+\Delta c]\}\cap\{T_j>c+\Delta c\}),
$$

where $\pi_i(c+\Delta c|c)$ follows Eq. 10. For correcting by censored cases, we follow Rizopoulos et al. (2017) who propose to use

[5] A regularisation parameter of 1 represents a jointly uniform distribution over all possible correlation matrices. For values larger than 1, the mode of the LKJ distribution is the identity matrix and as larger the value, the more sharply peaked at the mode.

model-based estimators of the censoring distribution by counting the concordant pairs of subjects as

$$\widehat{AUC}_c^{\Delta c} = \widehat{AUC}_1(c, \Delta c) + \widehat{AUC}_2(c, \Delta c) + \widehat{AUC}_3(c, \Delta c) + \widehat{AUC}_4(c, \Delta c), \quad (13)$$

where each of the AUC components is estimated over the four sets of combinations of concordant pairs, $\Omega_{ij}^{(l)}$, $l = 1, \ldots, 4$, defined as follows

1. $\Omega_{ij}^{(1)}$: Subject i suffers the event between $c+1$ and $c+\Delta c$, and subject j survives longer than $c+\Delta c$,
2. $\Omega_{ij}^{(2)}$: Subject i is censored between $c+1$ and $c+\Delta c$, and subject j survives longer than $c+\Delta c$,
3. $\Omega_{ij}^{(3)}$: Subject i suffers the event between $c+1$ and $c+\Delta c$, and subject j is censored between $c+1$ and $c+\Delta c$,
4. $\Omega_{ij}^{(4)}$: Both subjects, i and j, are censored between $c+1$ and $c+\Delta c$.

We can now specify the estimates of each component of Eq. 13 as

$$\widehat{AUC}_l(c, \Delta c) = \frac{\sum_i^N \sum_{j\neq i}^N I(\hat{\pi}_i(c+\Delta c|c) < \hat{\pi}_j(c+\Delta c|c)) \cdot I(\Omega_{ij}^{(l)}) \cdot \hat{v}_{ij}^{(l)}}{\sum_i^N \sum_{j\neq i}^N I(\Omega_{ij}^{(l)}) \cdot \hat{v}_{ij}^{(l)}},$$
$$l = 1, 2, 3, 4$$

where $I(\cdot)$ is the indicator function, $\hat{\pi}_i$ follows Eq. 12 and the terms $\hat{v}_{ij}^{(l)}$ account for the probability that the pairs are comparable. Thus, $\hat{v}_{ij}^{(1)} = 1$, $\hat{v}_{ij}^{(2)} = 1 - \hat{\pi}_i(c+\Delta c|T_i)$, $\hat{v}_{ij}^{(3)} = \hat{\pi}_j(c+\Delta c|T_j)$ and $\hat{v}_{ij}^{(4)} = (1 - \hat{\pi}_i(c+\Delta c|T_i))\hat{\pi}_j(c+\Delta c|T_j)$.

The calibration, which measures how accurate are the predictions, is commonly assessed in survival models by the expected error of predicting future events (Rizopoulos et al., 2017). The expected prediction error is written as

$$EPE(c+\Delta c|c) = \mathbb{E}(L\{N_i(c+\Delta c), \pi_i(c+\Delta c|c)\})$$

where $L(\cdot,\cdot)$ is the loss function (Brier score, absolute error, ignorance score, among others). $N_i(c+\Delta c) = I(T_i > c+\Delta c)$ is the true event status at time $c+\Delta c$ and the expectation is taken with respect to the distribution of the event times. As an error measurement, the lower is the value, the better the calibration.

To account for censored cases, we follow Henderson, Diggle, & Dobson (2002) who propose an estimate of $EPE(c+\Delta c|c)$ that reads

$$\widehat{EPE}(c+\Delta c|c) = n(c)^{-1} \sum_{i:T_i>c} \{S_i(c+\Delta c|c) + E_i(c+\Delta c|c) + C_i(c+\Delta c|c)\} \quad (14)$$

where $n(c)$ is the number of subjects at risk at time c and the terms inside the sum are

$$\begin{aligned} S_i(c+\Delta c|c) &= I(T_i > c+\Delta c) L\{1, \hat{\pi}_i(c+\Delta c|c)\} \\ E_i(c+\Delta c|c) &= \delta_i I(T_i \leq c+\Delta c) L\{0, \hat{\pi}_i(c+\Delta c|c)\} \\ C_i(c+\Delta c|c) &= (1-\delta_i) I(T_i \leq c+\Delta c) \big[\hat{\pi}_i(c+\Delta c|T_i) L\{1, \hat{\pi}_i(c+\Delta c|c)\} + \\ &\quad + (1 - \hat{\pi}_i(c+\Delta c|T_i)) L\{0, \hat{\pi}_i(c+\Delta c|c)\}\big] \end{aligned}$$

where δ_i is the censor index that equals 1 if the subject experiences the event and 0 otherwise.

The loss function L we use in this work corresponds to the popular Brier score (Brier, 1950) defined as mean squared error for the probabilistic forecasts. Hence, $\widehat{EPE}(c+\Delta c|c)$ (Eq. 14), measures the mean square deviation at a time $c+\Delta c$ with historical data collected until c. Henderson et al. (2002) also propose to measure the expected predicted error as an average over the interval $[c+1, c+\Delta c]$, following

$$\widehat{PE}_c^{\Delta c} = \frac{\sum_{i:c<T_i\leq c+\Delta c} \delta_i w(c, T_i) \widehat{EPE}(T_i|c)}{\sum_{i:c<T_i\leq c+\Delta c} \delta_i w(c, T_i)}, \quad (15)$$

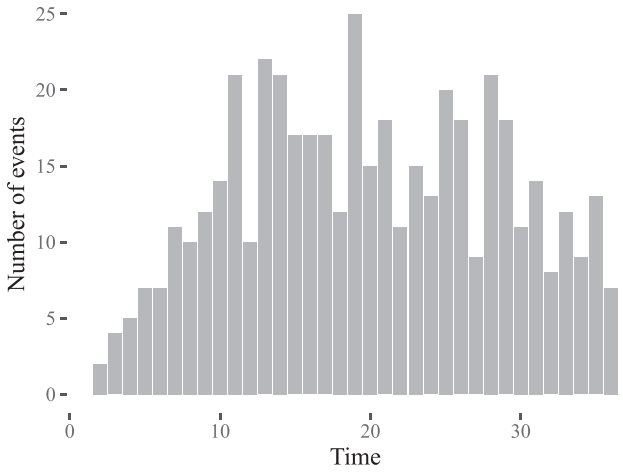

Fig. 1. Distribution of the events in time for simulated data with 10,000 subjects.

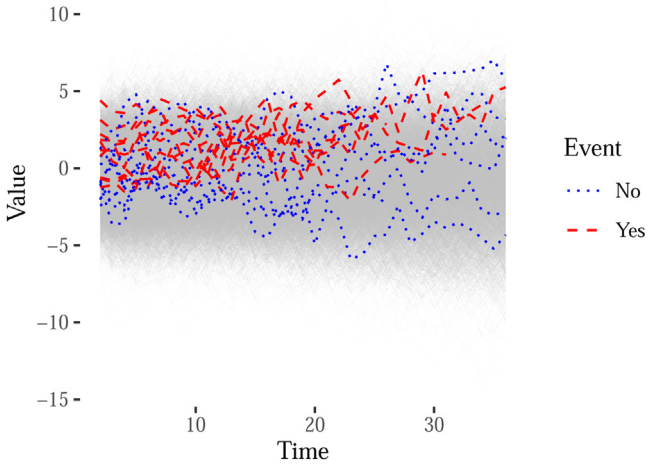

Fig. 2. Simulated longitudinal outcome versus time. Highlighted are 10 subjects that experience the event (dashed line) and 10 that are censored (dotted line).

where $w(c, T_i) = \widehat{KM}(c+1)/\widehat{KM}(T_i)$ are weights to compensate for the loss of censored cases and $\widehat{KM}(\cdot)$ is the Kaplan-Meier estimator. One way of measuring useful statistics of Eq. 15 is to perform a Monte Carlo approach for the estimation of the conditional posterior probabilities $\pi_i(c+\Delta c|c)$, as described in Rizopoulos (2011), but this is beyond the scope of this work.

5. Simulation

We study the discrete joint model with autoregressive terms introduced in Section 4 via simulation. To explore the MCMC sampling behaviour of our implementation, we fit the model for three sample sizes with 1,000, 5000 and 10,000 subjects, respectively, over a maximum of 36 periods (3 years), representing 24,424, 124,184 and 245,789 observations respectively. Fig. 1 illustrates the distribution of the events over time for the largest sample and Fig. 2 shows the evolution of the simulated longitudinal outcome where we have highlighted 10 subjects that experience the event (dashed line) and 10 that do not (dotted line). The setup for the simulation is motivated by the models applied in Section 6. Although in the application we explore the discrete joint model approach by fitting different link functions f (see Eq. 2), here we study the case where $f(\mathbf{U}_i, s) = m_{i,s-1}$ with $m_{i,s-1}$ following Eq. 4 that corresponds to the most general one.

Table 1
Estimations of joint model with autoregressive term over the different simulated samples.

		$N = 1,000$				$N = 5,000$				$N = 10,000$			
	True	Mean	SD	5%	95%	Mean	SD	5%	95%	Mean	SD	5%	95%
γ_1	2.00	2.100	0.239	1.724	2.502	2.033	0.097	1.873	2.190	1.999	0.070	1.886	2.114
γ_2	1.00	1.056	0.198	0.733	1.389	0.984	0.076	0.859	1.110	0.941	0.053	0.854	1.028
λ_f	1.00	1.027	0.103	0.862	1.204	1.002	0.043	0.932	1.075	0.997	0.031	0.944	1.048
α_0	-0.30	-0.320	0.040	-0.387	-0.255	-0.292	0.017	-0.318	-0.264	-0.289	0.012	-0.308	-0.270
ϕ	0.40	0.407	0.007	0.396	0.418	0.412	0.003	0.407	0.417	0.414	0.002	0.411	0.418
σ	1.00	1.001	0.005	0.993	1.009	1.000	0.002	0.996	1.003	1.003	0.002	1.000	1.005
$\sigma_{U_{0i}}$	1.20	1.196	0.034	1.141	1.252	1.170	0.015	1.146	1.194	1.154	0.010	1.138	1.171
$\sigma_{U_{1i}}$	0.05	0.049	0.002	0.046	0.052	0.048	0.001	0.047	0.049	0.049	0.001	0.048	0.050
$\rho_{U_{01}}$	-0.20	-0.182	0.040	-0.246	-0.117	-0.179	0.018	-0.209	-0.150	-0.184	0.013	-0.205	-0.163

Table 2
Specification of the models. **Id** is the identifier of the model, **Type** is either survival or joint model, **R-E** specifies the random effects used (only intercept or intercept and slope), **AR1** if the model has autoregressive term of order 1. $f(\mathbf{U}_i, s)$ is the link function (for the survival is the TVC) and $m_{i,s-1}$ the longitudinal predictor.

Id	Type	R-E	AR1	$f(\mathbf{U}_i, s)$	$m_{i,s-1}$
M_0	Survival	-	-	$Y_{i,s-1}$	–
M_1	Joint	Int	No	U_{0i}	$\alpha_0 + U_{0i}$
M_2	Joint	Int	Yes	U_{0i}	$\alpha_0 + U_{0i} + \phi Y_{i,s-2}$
M_3	Joint	Int-slope	No	$\alpha_0 + U_{0i} + U_{1i}(s-1)$	$\alpha_0 + U_{0i} + U_{1i}(s-1)$
M_4	Joint	Int-slope	Yes	$\alpha_0 + U_{0i} + U_{1i}(s-1)$	$\alpha_0 + U_{0i} + U_{1i}(s-1) + \phi Y_{i,s-2}$
M_5	Joint	Int-slope	Yes	$\alpha_0 + U_{0i} + U_{1i}(s-1) + \phi Y_{i,s-2}$	$\alpha_0 + U_{0i} + U_{1i}(s-1) + \phi Y_{i,s-2}$

Table 3
Summary of the posterior distributions of each model's parameters with fold 1 kept out.

Parameter	M_0			M_1			M_2		
	Mean	5%	95%	Mean	5%	95%	Mean	5%	95%
fico	-0.701	-0.819	-0.584	-0.698	-0.813	-0.583	-0.697	-0.816	-0.576
cltv	0.515	0.334	0.705	0.544	0.361	0.732	0.542	0.367	0.728
orig_upb	-0.151	-0.294	-0.011	-0.183	-0.328	-0.037	-0.182	-0.323	-0.044
dti	0.152	0.025	0.284	0.165	0.033	0.292	0.166	0.041	0.294
n_borr	-0.260	-0.513	-0.007	-0.268	-0.521	-0.019	-0.264	-0.510	-0.020
loan_purpose	-0.977	-1.246	-0.697	-0.992	-1.268	-0.717	-0.987	-1.265	-0.713
λ_f	1.456	1.135	1.778	0.345	0.150	0.550	1.053	0.480	1.703
α_0				-0.472	-0.495	-0.450	-0.206	-0.216	-0.195
$\sigma_{U_{0i}}$				1.209	1.193	1.226	0.504	0.496	0.512
σ				0.935	0.932	0.937	0.787	0.785	0.789
ϕ							0.587	0.584	0.591
Parameter		M_3			M_4			M_5	
	Mean	5%	95%	Mean	5%	95%	Mean	5%	95%
fico	-0.700	-0.820	-0.581	-0.716	-0.842	-0.589	-0.701	-0.821	-0.581
cltv	0.517	0.337	0.705	0.477	0.290	0.667	0.516	0.333	0.703
orig_upb	-0.156	-0.297	-0.015	-0.088	-0.233	0.060	-0.155	-0.300	-0.014
dti	0.151	0.018	0.282	0.150	0.016	0.282	0.152	0.021	0.283
n_borr	-0.276	-0.536	-0.014	-0.288	-0.548	-0.029	-0.270	-0.527	-0.018
loan_purpose	-0.969	-1.239	-0.695	-0.969	-1.248	-0.686	-0.971	-1.246	-0.696
λ_f	1.165	0.781	1.572	3.555	2.587	4.541	1.317	0.895	1.771
α_0	-0.444	-0.465	-0.422	-0.278	-0.292	-0.264	-0.280	-0.294	-0.266
$\sigma_{U_{0i}}$	1.860	1.836	1.885	1.236	1.218	1.254	1.237	1.219	1.255
σ	0.734	0.732	0.736	0.706	0.704	0.708	0.706	0.704	0.708
ϕ				0.357	0.354	0.361	0.357	0.353	0.360
$\sigma_{U_{1i}}$	0.086	0.085	0.088	0.053	0.052	0.054	0.053	0.052	0.054
ρ_U	-0.782	-0.790	-0.775	-0.810	-0.817	-0.803	-0.811	-0.818	-0.804

The longitudinal outcome $Y_{i,s-1}$ is described with one fixed effect (intercept) and two random effects (intercept and slope) in addition to an autoregressive process of order one ($p = 1$), specifically

$$Y_{i,s-1} = \underbrace{\alpha_0 + U_{0i} + U_{1i}(s-1) + \phi Y_{i,s-2}}_{m_{i,s-1}} + \epsilon_{i,s-1}$$

where $(U_{0i}, U_{1i})^\intercal \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$ and $\epsilon_{i,s-1} \sim \mathcal{N}(0, \sigma^2)$. We define the event process to depend on two time-invariant covariates $z_i^{(1)}$ and $z_i^{(2)}$

$$p_{is} = \text{logit}^{-1}(a_s^0 + \gamma_1 z_i^{(1)} + \gamma_2 z_i^{(2)} + \lambda_f m_{i,s-1}).$$

The term a_s^0 is generated from a cubic polynomial function so that the default rate is of the order of magnitude of the one observed in mortgage loans[6]

To assess convergence of the HMC inference in each setup, we sampled from 3 independent chains with overdispersed starting points per setup, each with 4000 and 2000 iterations for the warm-up and sampling periods, respectively. In regard to the general diagnosis of the HMC inference, none of the chains suffered from transitions that hit the maximum treedepth or were divergent. The energy Bayesian fraction of missing information (E-BFMI) was satisfactory for all transitions. Moreover, all the parameters

[6] Approximately 4.5% of the subjects experienced the event over the 36 periods.

Table 4
Mean difference of $\widehat{AUC}_c^{\Delta c}$ (Eq. 13) with respect to model M_0 (Cox model) and prediction window of 12 months ($\Delta c = 12$). The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis (corrected by the overlapping training sets). The largest increment of the corresponding row is marked in bold.

Time(c)	$\widehat{AUC}_c^{12}$ M_0	$\Delta\widehat{AUC}_c^{12}$					
		M_1	M_2	M_3	M_4	M_5	JM_3
6	0.732	**0.068** (0.046)	0.067 (0.045)	0.021 (0.028)	-0.010 (0.020)	0.021 (0.029)	0.017 (0.028)
7	0.750	0.050 (0.052)	**0.050** (0.050)	0.014 (0.028)	-0.024 (0.024)	0.013 (0.036)	0.007 (0.028)
8	0.796	0.025 (0.017)	**0.025** (0.017)	-0.003 (0.008)	-0.059 (0.010)	-0.013 (0.005)	-0.013 (0.008)
9	0.792	**0.010** (0.017)	0.010 (0.016)	-0.008 (0.020)	-0.034 (0.011)	-0.005 (0.006)	-0.014 (0.017)
10	0.791	0.004 (0.009)	0.004 (0.009)	-0.012 (0.027)	-0.012 (0.017)	**0.007** (0.008)	-0.020 (0.027)
11	0.799	0.001 (0.013)	0.001 (0.013)	-0.008 (0.033)	-0.013 (0.029)	**0.011** (0.018)	-0.027 (0.034)
12	0.790	**0.009** (0.015)	0.008 (0.015)	-0.011 (0.030)	-0.032 (0.033)	-0.003 (0.023)	-0.039 (0.032)
13	0.794	0.005 (0.018)	0.004 (0.017)	**0.006** (0.023)	-0.009 (0.023)	-0.011 (0.023)	-0.026 (0.023)
14	0.778	0.002 (0.015)	0.002 (0.014)	0.041 (0.043)	**0.042** (0.044)	0.000 (0.026)	0.002 (0.035)
15	0.785	-0.002 (0.013)	-0.003 (0.013)	0.046 (0.040)	**0.060** (0.040)	0.004 (0.030)	0.006 (0.038)
16	0.783	-0.006 (0.014)	-0.007 (0.013)	0.050 (0.026)	**0.075** (0.026)	0.006 (0.022)	0.004 (0.029)
17	0.779	-0.009 (0.012)	-0.010 (0.011)	0.057 (0.036)	**0.089** (0.033)	0.018 (0.033)	0.009 (0.035)
18	0.768	-0.008 (0.013)	-0.008 (0.012)	0.066 (0.035)	**0.105** (0.031)	0.029 (0.034)	0.013 (0.035)
19	0.767	-0.006 (0.011)	-0.007 (0.011)	0.061 (0.031)	**0.105** (0.028)	0.032 (0.031)	0.018 (0.035)
20	0.762	-0.009 (0.011)	-0.009 (0.011)	0.060 (0.031)	**0.111** (0.027)	0.035 (0.034)	0.015 (0.038)
21	0.774	-0.006 (0.011)	-0.007 (0.010)	0.054 (0.038)	**0.104** (0.033)	0.037 (0.039)	0.020 (0.038)
22	0.761	-0.006 (0.012)	-0.006 (0.012)	0.069 (0.051)	**0.123** (0.046)	0.055 (0.052)	0.035 (0.050)
23	0.750	-0.007 (0.010)	-0.008 (0.009)	0.073 (0.049)	**0.132** (0.045)	0.062 (0.052)	0.035 (0.046)
24	0.757	-0.016 (0.005)	-0.016 (0.004)	0.062 (0.044)	**0.124** (0.044)	0.053 (0.047)	0.109 (0.050)

Table 5
Mean difference of $\widehat{PE}_c^{\Delta c}$ (Eq. 15) with respect to model M_0 (Cox model) and prediction window of 12 months ($\Delta c = 12$). The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis (corrected by the overlapping training sets). The largest reduction of the corresponding row is marked in bold.

Time(c)	$\widehat{PE}_c^{12}$ M_0	$\Delta\widehat{PE}_c^{12}$					
		M_1	M_2	M_3	M_4	M_5	JM_3
6	0.367	-0.021 (0.009)	**-0.022** (0.009)	-0.020 (0.008)	-0.017 (0.008)	-0.018 (0.008)	-0.009 (0.009)
7	0.397	-0.019 (0.007)	**-0.020** (0.007)	-0.018 (0.007)	-0.017 (0.007)	-0.017 (0.006)	-0.004 (0.007)
8	0.428	-0.014 (0.008)	-0.014 (0.009)	-0.015 (0.009)	**-0.017** (0.009)	-0.015 (0.009)	0.002 (0.009)
9	0.467	-0.010 (0.006)	-0.010 (0.006)	-0.009 (0.006)	**-0.014** (0.006)	-0.011 (0.006)	0.007 (0.006)
10	0.487	-0.007 (0.004)	-0.008 (0.004)	0.002 (0.007)	**-0.012** (0.004)	-0.009 (0.004)	0.014 (0.004)
11	0.530	-0.006 (0.003)	-0.006 (0.003)	0.032 (0.022)	**-0.008** (0.003)	-0.007 (0.003)	0.039 (0.016)
12	0.590	-0.005 (0.002)	**-0.005** (0.002)	0.089 (0.022)	0.010 (0.007)	-0.004 (0.002)	0.094 (0.021)
13	0.617	-0.003 (0.001)	**-0.004** (0.001)	0.168 (0.059)	0.071 (0.040)	0.002 (0.003)	0.177 (0.067)
14	0.680	-0.003 (0.001)	**-0.004** (0.001)	0.373 (0.135)	0.337 (0.131)	0.022 (0.013)	0.431 (0.176)
15	0.744	-0.002 (0.002)	**-0.002** (0.002)	0.516 (0.250)	0.671 (0.322)	0.054 (0.038)	0.624 (0.337)
16	0.805	-0.001 (0.003)	**-0.001** (0.002)	0.601 (0.315)	1.022 (0.471)	0.103 (0.073)	0.748 (0.434)
17	0.806	0.000 (0.003)	**-0.001** (0.003)	0.668 (0.357)	1.389 (0.605)	0.161 (0.108)	0.859 (0.497)
18	0.851	0.000 (0.003)	**-0.001** (0.003)	0.730 (0.368)	1.789 (0.710)	0.230 (0.141)	0.964 (0.510)
19	0.911	0.000 (0.003)	0.000 (0.003)	0.691 (0.364)	1.968 (0.807)	0.257 (0.162)	0.916 (0.488)
20	0.918	0.000 (0.004)	**-0.001** (0.003)	0.646 (0.342)	2.068 (0.885)	0.291 (0.173)	0.852 (0.459)
21	0.951	0.000 (0.004)	0.000 (0.004)	0.591 (0.258)	2.135 (0.762)	0.305 (0.150)	0.762 (0.354)
22	0.916	0.000 (0.004)	0.000 (0.004)	0.468 (0.245)	1.889 (0.810)	0.264 (0.155)	0.593 (0.328)
23	0.919	0.002 (0.002)	0.002 (0.002)	0.386 (0.138)	1.783 (0.610)	0.247 (0.103)	0.487 (0.196)
24	0.908	0.004 (0.002)	0.003 (0.003)	0.373 (0.127)	1.720 (0.553)	0.269 (0.116)	0.535 (0.179)

*For ease of visualisation, all values are multiplied by 100.

Table 6
Estimations of M_0 (Cox) for the largest simulated sample.

		M_0			
	True	Mean	SD	5%	95%
γ_1	2.00	1.780	0.061	1.681	1.878
γ_2	1.00	0.836	0.049	0.756	0.918
λ_f	1.00	0.752	0.024	0.713	0.792

had satisfactory effective sample sizes $\hat{n}_{\text{eff}}$, which plays a similar role as the number of independent draws in the standard central limit theorem. Also, they all showed satisfactory potential scale reduction factors $\hat{R}$ that measures the consistency between chains by quantifying the between-chain over the within-chain variability. Hence, no problems were detected. Further details on HMC diagnosis and the problems associated to these metrics are presented in Betancourt (2017).

The final 6000 sampling iterations per setup (2,000 per chain) are summarised in Table 1 by their means, standard deviations and 5%-95% posterior credible intervals in addition to the true generating parameter values.

The baseline hazard is generated from a cubic polynomial but modelled through cubic B-spline functions. We use three equally spaced internal knots placed at the 25th, 50th and 75th percentiles of the distribution of the event times. That implies 7 spline

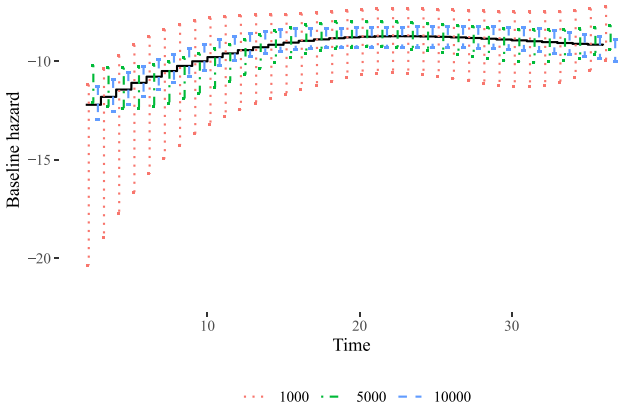

Fig. 3. True baseline hazard a_s^0 (solid line) and the corresponding estimations for the three sample size settings with their 5–95% posterior credible intervals.

coefficients to estimate (degree of 3, 3 knots and 1 intercept)[7] Fig. 3 shows the true baseline hazard a_s^0 (solid line) and the estimated 5–95% posterior credible intervals for the 3 settings[8] It can be seen that using this spline configuration the intervals of the three settings cover the true value and when more data is added, the intervals are narrower, as expected.

Since we know the data generation process, we can quantify the bias in the parameter estimates when the discrete survival model is used, as we did in Appendix B. The results from this analysis show that the true parameter values are outside the 95% credible intervals. However, in the empirical analysis shown below, we cannot analyse the bias because we do not know the true data generation process, but rather we can compare the predictions of each model.

6. Application: credit scoring for US mortgages

6.1. Data

We analyse the Single Family Loan-Level Dataset publicly available from Freddie Mac[9]. The dataset contains loan-level origination and monthly performance for fixed-rate US mortgages and it is periodically updated. Freddie Mac provides a randomly selected sample dataset of 50,000 loans for each vintage year, starting from 1999 onwards. Due to computational limitations, we use the loans originated from October to December of the year 1999 from the sample dataset and follow their performance for the next 36 months. The final number of loans in our training sample is 10,399 that corresponds to 285,462 observations. 2.3% of these loans defaulted in the period of analysis and Fig. 4 shows how they are distributed in time.[10] Appendix D shows descriptive statistics of the data. To the best of our knowledge, this is the largest sample size used in the literature on joint models.

For the longitudinal outcome, we consider the difference between the implicit interest rate and the fixed interest rate granted at origination since it nicely balances the scheduled versus the observed information in the following way. Considering that the data contains the original amount of the loan (P_0), the fixed interest rate and the loan term, we can then calculate the original instalment amount (A) and the scheduled unpaid principal balance if neither

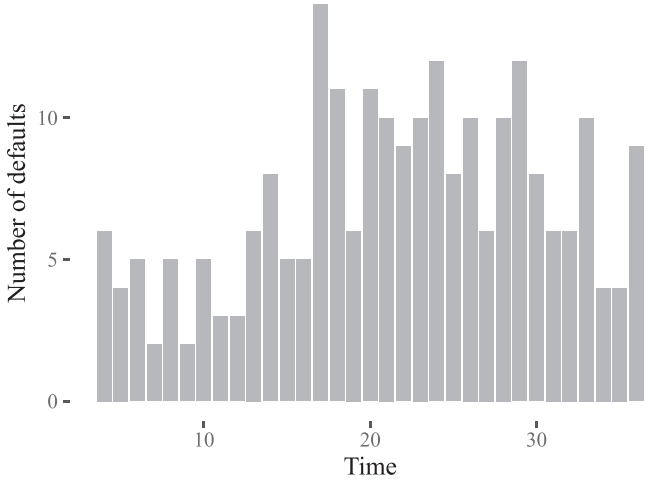

Fig. 4. Distribution of the defaults in time for the training sample.

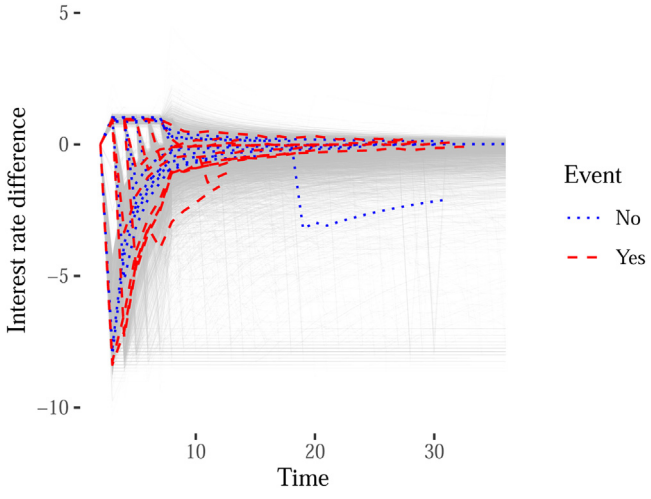

Fig. 5. The evolution in time of the difference between the implicit and granted interest rate. Highlighted are 10 borrowers that defaulted (dashed) and 10 who are censored (dotted).

an early nor delayed repayment is made. The implicit rate (i) is thus calculated as the one that corresponds to the observed unpaid principal balance (P_t) for the remaining period of the mortgage following Eq. 16. That means that if the payments are made as scheduled, the implicit interest rate will be the same as the original fixed interest rate. Otherwise, if there is any unscheduled principal changes it will increase or decrease the implicit interest rate.

$$P_t = P_0(1+i)^t - A\frac{(1+i)^t}{i} + \frac{A}{i} \tag{16}$$

Fig. 5 shows the evolution in time of the interest rate difference in which, for illustrative purposes, we have highlighted ten borrowers who do and do not experience default (dashed and dotted line, respectively). We observe that in the first six months the series either goes up or down. This happens because the data provider reports the current unpaid principal balance to the nearest $1000 for the first 6 months of each loan and it is also reflected in the implicit rate (if the rounded number is above or below the scheduled).

We use the following time-invariant covariates in the survival processes (Hu & Zhou, 2019; Wang et al., 2020)

- **fico** is a number summarising the borrower's creditworthiness (credit score) developed by FICO. Generally, the number dis-

[7] We explored configurations with different number of knots but no major improvements were obtained.

[8] All effective sample sizes $\hat{n}_{\text{eff}}$ of the HMC sampling are above 6000.

[9] http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

[10] We use the definition of default that corresponds to the time when the borrower is 90 days or more past due.

closed is the score known at the time of acquisition and is the score used to originate the mortgage.
- **cltv** is the loan-to-value ratio based on the original mortgage loan amount plus any other mortgage loan amount divided by the mortgaged purchase price of the property.
- **orig_upb** is the original unpaid principal balance of the mortgage on the note date.
- **dti** is the debt to income ratio based on the sum of the borrower's monthly debt payments divided by the total monthly income used to underwrite the loan.
- **n_borr** is the number of borrower(s) who are obligated to repay the mortgage. Either one borrower (= 0, 38% of the loans) or more than one (= 1, 62% of the loans).
- **loan_purpose** indicates whether the mortgage loan purpose is a refinance (= 0, 26% of the loans) or a purchase (= 1, 74% of the loans).

To measure how well each model performs in an out-of-sample scenario, we assess them by 5-fold cross-validation analysis. Since the default rate is low, the folds are created in such a way as to preserve the overall default rate (see Appendix D).

6.2. Models and results

We estimate six models, all of them using the same time-invariant covariates mentioned above. The differences come from the assumptions made about the link function f (Eq. 2) and longitudinal outcome structure (Eqs. 3 or 4), and are summarised in Table 2. M_0 is a discrete survival Cox model in which the interest rate difference is included as observed, i.e. as exogenous TVC. This model corresponds to the standard approach in credit risk survival modelling (Bellotti & Crook, 2013; Crook & Bellotti, 2010; Wang et al., 2020) and it serves as benchmark to the joint models. The other five models come from combining random intercept or random intercept and slope with or without autoregressive term as detailed in the table.

For each model specification, we perform a 5-fold cross-validation with 3 independent chains for the MCMC sampling procedure, i.e. 90 model estimations in total (6 specifications, 5 folds and 3 chains), each of them with a warm-up period of 4000 iterations and 2000 sampling draws. To make the computation more efficient, we implement each specification using the within chain parallelisation feature already available in the *CmdStan* interface (see https://mc-stan.org/users/interfaces/cmdstan.html) with 4 CPU cores of 16GB of memory[11] The computational resources were provided by the Edinburgh Compute and Data Facility (ECDF, http://www.ecdf.ed.ac.uk/) and the codes with explanatory comments are available with the supplementary material. With respect to the general diagnosis of the HMC inference, following the metrics mentioned in Section 5, no problems were detected. That is, none of the transitions were divergent or hit the maximum treedepth, all had satisfactory E-BFMI as well as the $\hat{n}_{\text{eff}}$ and $\hat{R}$ for all the parameters (see Appendix Appendix J for further details on convergence). Moreover, to substantiate that the results are not strongly dependent on the choice of prior distributions described in Section 4.2, we conducted a robustness study in Appendix Appendix I.

Table 3 summarises the parameter estimations for the six models for one of the five cross-validation results (keeping fold 1 out in this case)[12] First, we observe that for almost all parameters their 5–95% posterior credible intervals do not include 0 and, second, there is strong evidence that the autoregressive coefficient ϕ for models M_2, M_4 and M_5 is significant. Furthermore, the posterior mean of the parameters associated with the time-invariant covariates have fairly similar estimates among the six models with coherent signs. For example, greater *fico*, meaning better creditworthiness, has lower probability of default. Moreover, the greater the mortgage loan with respect to the purchase price (*cltv*), the greater the probability of defaulting, analogous results for the debt to income ratio *dti*. If there is more than one borrower responsible of paying the loan (*n_borr*), we observe that the probability of defaulting is also lower and the same is obtained when the purpose of the loan is to purchase the mortgage instead of refinance it. The exception comes from *orig_upb* estimated by model M_4 where its credible interval does include 0 and its estimated mean drops 50% in relation to the others models. In addition, the posterior of the association parameter λ_f shows differences among the models as expected, since the linking variables are not strictly comparable (for example, constant versus linear tendency) but all the intervals are far from 0. The signs are all positive which can be interpreted as if the level of the difference between the implicit and the original interest rate increases, then also the probability of default increases.

We measure the performance of the individual survival predictions under the discrimination and calibration metrics described in Section 4.4. Both the AUC and the PE depend on the evaluation times c and $c + \Delta c$. We study the predictions for the range of $c \in [6, 24]$ and $\Delta c = 12$ to analyse how the models behave when more information is collected in time. For instance, if $c = 6$ we use the collected data until the sixth month and predict the probability of default for months 7 to 18. Further, all the predictions are done for the unknown fold, so the new collected data is not used for estimating the parameters of the models but rather to estimate the random effects that serve to the individual predictions as described in Section 4.3.

To compare the models against the benchmark (M_0), we calculate the difference in the AUC for all the values of c within their respective fold. Table 4 shows the means and standard deviations[13] of the difference in the AUC considering the 5 folds. The number in the first column corresponds to c. We observe that for c between 6 and 9, models M_1 and M_2 outperform the benchmark in terms of discrimination for the forecast window of 12 months but for greater c, both remain practically the same to M_0. Moreover, for $c \leq 13$ there is not a great difference for models M_3, M_4 and M_5 with respect to M_0, however, for $c \geq 14$, the discrimination increases considerably, specially for M_4, with an average increase of more than 0.1 in the AUC.

Table 4 also shows under the name JM_3 the results of the continuous-time version of M_3. This model, unlike the joint models M_1 to M_5, has been estimated with the R package *JMbayes* (Rizopoulos, 2014). This additional analysis aims to compare the performance between discrete and continuous-time versions of the joint model without autoregressive terms. We note that, in general terms, the JM_3 model presents better discrimination than M_0, which is also shown in Hu & Zhou (2019) for a prepayment predictive model. Yet, for this data, the discrete-time version M_3 presents slightly better discrimination results (more details on this analysis in Appendix Appendix H).

Analogously, Table 5 shows the mean differences and standard deviations of PE with respect to M_0, for the range of c and the forecast window of 12 months. For ease of viewing, all the values are scaled by a factor of 100. We observe, for $c < 12$, that the calibration of the joint models are, in general, better than the benchmark,

[11] As a reference, one run of one chain of 6000 samples for model M_5 (the most complex one) took 12 hours to finish.

[12] We only disclose this first fold since the results obtained in the others are consistent with the results we have shown.

[13] The standard deviation includes the additional correlation term detailed in Nadeau & Bengio (2000) that accounts for the overlapping training sets in the cross-validation.

in particular, for models M_4 and M_5. For $c \geq 12$, however, models M_3 and M_4 start to increase the expected predictive error in comparison to M_0. Model M_5 also increases the predictive error but not as much as models M_3 and M_4 which can be seen as a good balance between improvement in discrimination without affecting too much the calibration. Furthermore, models M_1 and M_2 recover the same performance levels as the benchmark. Finally, the calibration performance of the continuous-time version of M_3, JM_3, follows a similar trend to its counterpart without presenting sustainable improvements (see Appendix Appendix H).

Models M_3, M_4 and M_5 show better discrimination than the benchmark when more historical information is collected but the same is not true in terms of calibration. This discrepancy largely comes from the fact that this data is highly unbalanced, i.e. the number of defaults is considerably lower than that of non defaults. Under these circumstances, it could happen that any model, for instance, that assigns a survival probability of 1 to all, still has a relatively good calibration, so it is important to take this metric with caution and understand where the major contributions come from.

Table 10 in Appendix E shows the 5–95% probability ranges estimated by each model and separated by non-defaulters (value 0) versus defaulters (value 1). We observe that, for $c > 12$, the joint models M_3, M_4 and M_5 start to have a broader range than the benchmark for both labels, which is also when the differences in the calibration metric appear. In other words, the joint models can identify better the defaulters versus the non-defaulters, since they have better discrimination performance, and assign lower probabilities of surviving to the defaulters than the benchmark. However, these models also assign lower probabilities to the non-defaulters and, because of the large number of these cases in the data, the calibration is made worse.

To investigate at what extend the models are sensitive to class imbalance, we re-estimate the models M_0 (benchmark) and M_5 (joint model with autoregressive term) by controlling the proportion of non-defaulters in the data. Appendix F shows the results for two scenarios. The first one, randomly reduces the number of non-defaulters in such a way that 75% of the loans are non-defaulters and, the second one, has equal number of defaulters and non-defaulters. From these results, we see that the calibration of the joint model shows major improvements to class imbalance compared with the Cox model, reducing the difference between them ($\Delta\widehat{PE}_c^{12}$ M_5) for $c \geq 15$, in more than 50% when comparing to the results shown in Table 5.

Finally, to complement the comparative analysis, in Appendix Appendix K we compare the models from an economic perspective. The results of this analysis reveal, first, that the joint models M_1 and M_3 show similar average costs as the Cox model. Second, there are cost reductions when we include autoregressive terms in both the longitudinal and link parts (M_5).

7. Concluding remarks

The inclusion of TVCs into survival credit scoring models is widely applied in the literature to either improve the predictions or enhance the understanding of why borrowers default (Bellotti & Crook, 2009; 2014; Calabrese & Crook, 2020; Dirick et al., 2019; Wang et al., 2020). However, there are few works that focus on distinguishing the type of variable included (Dirick et al., 2019; Hu & Zhou, 2019), thus treating endogenous and exogenous variables equally. This practice can lead to two main problems if the TVC is endogenous. First, from a statistical standpoint, we might encounter biased parameter estimations (Section 2). Second, from a forecast perspective, we lack a dynamic prediction framework that takes advantage of the mutual evolution between the TVC and the survival time, forcing the prediction to keep the last observed value fixed or estimating the model with lagged values of the TVC (Bellotti & Crook, 2013; Crook & Bellotti, 2010; Wang et al., 2020).

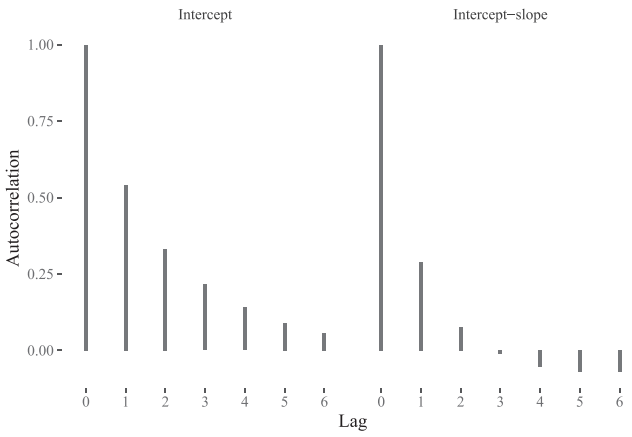

Fig. 6. Empirical autocorrelation functions for the longitudinal outcome. On the left, the linear mixed-effect model with random intercept and, on the right, the linear mixed-effect model with random intercept and slope.

To address the inclusion of endogenous TVCs into survival scoring models, we explore the joint modelling approach and adapt it to handle features typical of credit risk applications. First, to the best of our knowledge, this is the first work that uses joint models in discrete time for credit scoring models. Second, from a methodological angle and making use that observations are equally-spaced and indexed by a discrete variable (time), we propose an extended joint model that incorporates autoregressive terms into the longitudinal outcome. This extension is motivated by the autoregressive components seen in the data (see Fig. 6 in Appendix A) and how, by applying this extension, the accuracy on the predictions could be improved.

In total, we implement six models, a traditional survival model (M_0) that is our benchmark and five joint models (M_1, M_2, M_3, M_4, M_5), all of them following the Bayesian approach, coded in Stan, using `CmdStan` interface (Stan Development Team, 2018) with withing chain parallelisation feature and available with the supplementary material. We study the most general case of the implementations (M_5) via simulation analysis which shows satisfactory converging diagnosis for three independent sampling chains and true value recovery for different sample sizes.

Furthermore, we apply all the models to US mortgage loans data and compare them via cross-validation analysis. The results show that the joint models that assume the longitudinal outcome with only random intercepts, either with or without autoregressive term (M_1 and M_2, respectively), only improve the discrimination measure with respect to the benchmark when not much historical information of the new borrowers is known. However, the other three joint models (M_3, M_4 and M_5) show a higher improvement in terms of discrimination when more historical data is collected, specially the model M_4 that includes autoregressive correction in the longitudinal outcome.

In terms of calibration, we see that when using the historical data up to the first year (12 months), the joint models are, in general, better than the benchmark. Moreover, when more historical data are considered, models M_1 and M_2 preserve the same level of calibration as the benchmark. However, for models M_3, M_4 and M_5 the calibration error grows in comparative terms. This is mainly because these models estimate posterior probability distributions with higher variability than the benchmark for the non-defaulters when more historical data are considered and, given that these data are highly imbalanced, greater variability in the probabilities is more detrimental to the overall quality of the calibration in com-

parison to the benchmark. Nevertheless, when we control for the class imbalance, we see that this difference is considerably reduced (Appendix F).

In this paper, we include only one longitudinal outcome in the model with only one autoregressive term in the implementations. A potential extension of this work might be to consider a multivariate longitudinal case with different autoregressive orders, where more complex payment patterns can be recognised and included into the time to default prediction. For example, being able to measure and incorporate correlations between the use of the credit card and the implicit interest rate through a bivariate longitudinal model.

However, a major drawback of this methodology is the computational cost. Typically financial institutions estimate scoring models on big sample size, on the order of thousands or millions of data. In order to scale this model by including a multivariate longitudinal process and make it feasible for real life applications, some approximations in the estimation procedure can be applied. For instance, if the event and the multivariate longitudinal processes are assumed to have a linear Gaussian association structure, then it can be seen as a latent Gaussian model (LGM). Thus, the Bayesian inference can be approximated, for example, with the integrated nested Laplace approximation (INLA) (Rue, Martino, & Chopin, 2009).

To conclude, use of joint models is a promising approach to investigate in credit risk applications in which we usually have a variety of endogenous TVCs that could bring relevant predictive information. We believe our extension to include autoregressive terms can be further exploited to extract predictive behaviours and to better understand the dynamic nature of credit defaults.

Acknowledgements

Raffaella Calabrese was supported by ESRC funding (ESRC Award Number ES/W010259/1).

Appendix A. Empirical autocorrelation functions

Appendix B. Estimation of Cox model for joint model simulated data

We use the largest simulated data detailed in Section 5 (10,000 subjects) and estimate a Cox model where the longitudinal outcome is included as observed (see Table 2). We sampled from 3 independent chains with overdispersed starting points, each with 4000 and 2000 iterations for the warm-up and sampling periods, respectively. We follow the same general diagnosis procedure described in Section 5 with respect to the HMC inference and no problems were detected. Table 6 summarises the parameter estimations. For this model specification under this simulated data, we observed that the 5–95% credible intervals do not include the true parameter value. This evidences the estimation bias in the parameters when the data representation is well-described by a joint stochastic process between the survival and longitudinal variables, but the latter is considered deterministically, as the Cox approach does.

Appendix C. Comparing simulations with and without autoregressive term

We are interested in quantifying the relevance of adding at least one autoregressive term in the longitudinal outcome when compared to the case with no autoregressive terms. To do so, we perform two simulations analysis. The first one uses the same simulated data from Section 5 (10,000 subjects) and estimate a joint model without the autoregressive term ($\phi = 0$), which is analogous to the specification M_3 in the empirical analysis. We call this model $\widetilde{M}_3$. The second one simulates data as if it were generated by a joint model without an autoregressive term and estimate a joint model with an autoregressive term. We call this model $\widetilde{M}_5$. The results are shown in Table 7. We observed that for both models, despite that they are misspecified for the data, the 5%-95% credible intervals for the parameters related to the event process include the true parameters. The differences come from the longitudinal part. We observed that $\widetilde{M}_3$ tries to compensate for misspecification overestimating the fixed effect α_0, the variability of the random effects ($\sigma_{U_{0i}}$ and $\sigma_{U_{1i}}$) and the variability of the error terms (σ). However, when the data is generated by a joint model without an autoregressive term and we estimate a joint model with an autoregressive term ($\widetilde{M}_5$), we observe that the parameters related to the longitudinal outcome are closer to the true values.

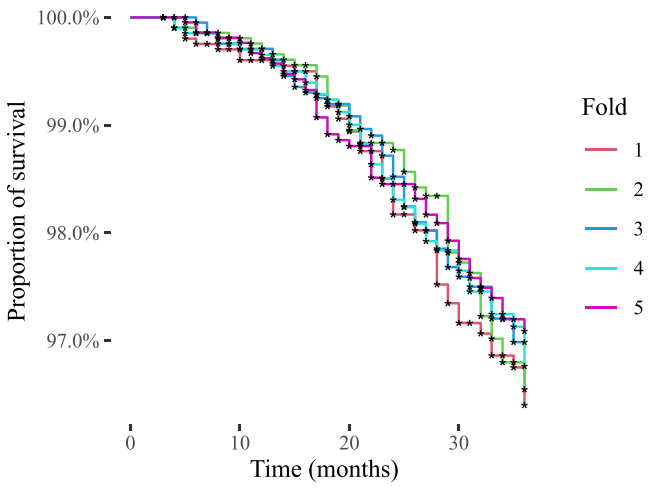

Fig. 7. Kaplan–Meier curves for each fold.

Appendix D. Descriptive statistics of the data

Table 8 provides descriptive statistics for the numeric covariates. To ease the MCMC sampling, we standardise these covariates to have a zero-mean and standard deviation of 1. Table 9 gives the number of loans and default rates for each fold and Fig. 7 shows the corresponding Kaplan-Meier curves.

Appendix E. Survival probability ranges

Table 10 shows the probability ranges (5-95%) for non-defaulters (value 0) and defaulters (value 1) for the 6 estimated models.

Appendix F. Calibration sensitivity analysis

Our interest is to investigate how sensitive is the calibration of the joint model M_5 to the class imbalance in comparison to the benchmark. To this end, we perform a 5-fold cross-validation analysis similar to the one described in Section 6 but we now randomly reduce the non-defaulters proportion in the training folds of the empirical data (down-sampling). We perform the analysis for two different non-defaulters proportions, one corresponding to 75% of the loans and the other to 50%. Table 11 shows the mean differences and standard deviations of PE with respect to M_0, for the range of c and the forecast window of 12 months for both class proportions. Although we observe that both models, M_0 and M_5, are sensible to class imbalance showing improvements in their calibration when compared to the results shown in Table 5, the joint

Table 7
Estimations of $\widetilde{M}_3$ (joint model without AR1) for data coming from $\widetilde{M}_5$ (left) and estimations of $\widetilde{M}_5$ (joint model with AR1) for data coming from $\widetilde{M}_3$ (right).

		$\widetilde{M}_3$ with data from $\widetilde{M}_5$					$\widetilde{M}_5$ with data from $\widetilde{M}_3$			
	True	Mean	SD	5%	95%	True	Mean	SD	5%	95%
γ_1	2.00	1.989	0.070	1.873	2.107	2.00	2.086	0.084	1.949	2.225
γ_2	1.00	0.937	0.054	0.849	1.026	1.00	1.114	0.070	1.000	1.230
λ_f	1.00	0.990	0.032	0.937	1.043	1.00	1.021	0.051	0.939	1.105
α_0	-0.30	-0.473	0.019	-0.505	-0.442	-0.30	-0.291	0.012	-0.311	-0.271
ϕ	0.40					0.00	0.008	0.002	0.004	0.011
σ	1.00	1.044	0.002	1.042	1.047	1.00	1.002	0.001	1.000	1.005
$\sigma_{U_{0i}}$	1.20	1.945	0.014	1.921	1.969	1.20	1.168	0.010	1.152	1.184
$\sigma_{U_{1i}}$	0.05	0.090	0.001	0.089	0.092	0.05	0.050	0.001	0.049	0.051
$\rho_{U_{01}}$	-0.20	-0.196	0.011	-0.215	-0.177	-0.20	-0.175	0.013	-0.196	-0.154

Table 8
Descriptive statistics for numeric covariates in the dataset.

Covariate	N	Mean	SD	$Q_{2.5\%}$	$Q_{25\%}$	$Q_{50\%}$	$Q_{75\%}$	$Q_{95\%}$
fico	10,399	710.70	52.50	619.00	672.00	716.00	753.00	786.00
cltv	10,399	78.05	15.42	46.00	72.00	80.00	90.00	95.00
orig_upb*	10,399	122.35	53.49	48.00	80.00	115.00	155.00	228.00
dti	10,399	33.79	10.50	16.00	27.00	34.00	41.00	50.00

*1,000 USD.

Table 9
Number of loans (N) and default rate (DFR) per fold.

Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
N	DFR (%)	N	DFR (%)	N	DFR (%)	N	DFR (%)	N	DFR (%)
2036	2.50	2093	2.25	2092	2.15	2037	2.26	2141	2.15

Table 10
Survival probability ranges (5-95%) for non-defaulters (value 0) and defaulters (value 1) (see $\hat{\pi}_k(c+12|c)$ in Eq. 12). The Time(c) column represents c, the known history of the subjects.

Time(c)	M_0		M_1		M_2		M_3		M_4		M_5	
	0	1	0	1	0	1	0	1	0	1	0	1
6	0.95-1.00	0.90-1.00	0.98-1.00	0.96-1.00	0.98-1.00	0.96-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.94-1.00
7	0.95-1.00	0.89-1.00	0.97-1.00	0.95-1.00	0.98-1.00	0.95-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.94-1.00
8	0.97-1.00	0.92-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.93-1.00
9	0.97-1.00	0.93-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.95-1.00	0.97-1.00	0.93-1.00	0.98-1.00	0.94-1.00	0.97-1.00	0.93-1.00
10	0.97-1.00	0.92-1.00	0.97-1.00	0.94-1.00	0.97-1.00	0.94-1.00	0.95-1.00	0.92-1.00	0.98-1.00	0.95-1.00	0.97-1.00	0.94-1.00
11	0.96-1.00	0.91-1.00	0.97-1.00	0.93-1.00	0.97-1.00	0.94-1.00	0.92-1.00	0.80-1.00	0.97-1.00	0.92-1.00	0.97-1.00	0.93-1.00
12	0.96-1.00	0.91-1.00	0.96-1.00	0.93-1.00	0.96-1.00	0.93-1.00	0.87-1.00	0.66-1.00	0.94-1.00	0.86-1.00	0.96-1.00	0.93-0.99
13	0.96-1.00	0.91-1.00	0.96-1.00	0.93-1.00	0.96-1.00	0.93-1.00	0.80-1.00	0.51-1.00	0.87-1.00	0.67-1.00	0.95-1.00	0.89-0.99
14	0.96-1.00	0.90-1.00	0.96-1.00	0.92-0.99	0.96-1.00	0.92-0.99	0.74-1.00	0.39-1.00	0.73-1.00	0.43-1.00	0.94-1.00	0.85-0.99
15	0.95-1.00	0.90-1.00	0.96-1.00	0.92-0.99	0.96-1.00	0.92-0.99	0.70-1.00	0.21-1.00	0.55-1.00	0.12-1.00	0.92-1.00	0.77-0.99
16	0.95-1.00	0.89-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.67-1.00	0.18-1.00	0.37-1.00	0.04-1.00	0.90-1.00	0.71-0.99
17	0.95-1.00	0.89-1.00	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.66-1.00	0.16-1.00	0.24-1.00	0.01-1.00	0.87-1.00	0.62-1.00
18	0.95-1.00	0.90-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.67-1.00	0.20-1.00	0.17-1.00	0.01-1.00	0.85-1.00	0.62-1.00
19	0.95-1.00	0.91-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.69-1.00	0.21-1.00	0.14-1.00	0.00–1.00	0.84-1.00	0.56-0.99
20	0.95-1.00	0.91-1.00	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.71-1.00	0.37-0.99	0.13-1.00	0.01-1.00	0.83-1.00	0.59-0.99
21	0.95-1.00	0.91-0.99	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.74-1.00	0.37-0.99	0.14-1.00	0.01-1.00	0.83-1.00	0.58-0.99
22	0.95-1.00	0.90-1.00	0.95-1.00	0.92-0.99	0.95-1.00	0.92-0.99	0.77-1.00	0.41-1.00	0.16-1.00	0.01-1.00	0.83-1.00	0.58-1.00
23	0.95-1.00	0.91-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.78-1.00	0.45-1.00	0.19-1.00	0.02-1.00	0.83-1.00	0.59-1.00
24	0.95-1.00	0.90-1.00	0.95-1.00	0.92-1.00	0.95-1.00	0.92-1.00	0.79-1.00	0.52-1.00	0.23-1.00	0.03-1.00	0.83-1.00	0.62-1.00

model has fairly decreased the difference in the PE ($\Delta\widehat{PE}_c^{12}$ M_5), especially for $c \geq 15$ where the largest differences were observed before.

Appendix G. Comparison within the sample

Table 12 shows the value of the log predictive density for the within-sample estimation. We observe that the joint models with two random effects (M_3, M_4 and M_5) obtain higher values compared to the joint models with only one random intercept (M_1 and M_2) and to the Cox model (M_0).

Appendix H. Comparison between continuous and discrete time

The purpose of this section is to measure the performance differences between the discrete-time joint model versus its continuous-time counterpart. To do so, we first note that the discrete joint model with autoregressive terms cannot be fully compared with its continuous version unless the autoregressive formulation is generalised to be handled in continuous time, which is beyond the scope of this paper. Therefore, instead, we work with the discrete joint model specification that does not include autore-

Table 11
Mean difference of $\widehat{PE}_c^{\Delta c}$ (Eq. 15) between models M_5 and M_0 for prediction window of 12 months ($\Delta c = 12$) considering two down-sampling settings; 75% and 50% of non-defaulters. The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis. The first two columns are the corresponding results from Table 5 and are copied here for ease of comparison.

Time(c)	No down-sampling		25%-75%		50%-50%	
	$\widehat{PE}_c^{12}$ M_0	$\Delta\widehat{PE}_c^{12}$ M_5	$\widehat{PE}_c^{12}$ M_0	$\Delta\widehat{PE}_c^{12}$ M_5	$\widehat{PE}_c^{12}$ M_0	$\Delta\widehat{PE}_c^{12}$ M_5
6	0.367	-0.018 (0.008)	0.354	**-0.010** (0.006)	0.348	**-0.005** (0.003)
7	0.397	-0.017 (0.006)	0.384	**-0.009** (0.006)	0.379	**-0.004** (0.003)
8	0.428	-0.015 (0.009)	0.418	**-0.007** (0.006)	0.414	**-0.004** (0.004)
9	0.467	-0.011 (0.006)	0.458	**-0.005** (0.004)	0.455	**-0.003** (0.002)
10	0.487	-0.009 (0.004)	0.481	**-0.004** (0.003)	0.478	**-0.002** (0.002)
11	0.530	-0.007 (0.003)	0.523	**-0.003** (0.002)	0.521	-0.001 (0.001)
12	0.590	-0.004 (0.002)	0.584	0.000 (0.002)	0.582	0.002 (0.000)
13	0.617	0.002 (0.003)	0.611	0.005 (0.004)	0.609	0.007 (0.003)
14	0.680	0.022 (0.013)	0.673	0.029 (0.030)	0.670	0.021 (0.010)
15	0.744	0.054 (0.038)	0.737	0.067 (0.083)	0.734	0.040 (0.030)
16	0.805	0.103 (0.073)	0.797	0.110 (0.152)	0.796	0.061 (0.053)
17	0.806	0.161 (0.108)	0.798	0.165 (0.212)	0.797	0.089 (0.074)
18	0.851	0.230 (0.141)	0.845	0.217 (0.263)	0.842	0.117 (0.092)
19	0.911	0.257 (0.162)	0.905	0.237 (0.284)	0.903	0.124 (0.099)
20	0.918	0.291 (0.173)	0.912	0.255 (0.289)	0.910	0.132 (0.103)
21	0.951	0.305 (0.150)	0.946	0.250 (0.243)	0.946	0.131 (0.081)
22	0.916	0.264 (0.155)	0.913	0.220 (0.239)	0.914	0.110 (0.085)
23	0.919	0.247 (0.103)	0.914	0.184 (0.143)	0.914	0.099 (0.043)
24	0.908	0.269 (0.116)	0.904	0.200 (0.153)	0.905	0.110 (0.055)

*For ease of visualisation, all values are multiplied by 100.

Table 12
Comparison of the log predictive density within the sample. The value in parentheses is the standard deviation.

	M_0	M_1	M_2	M_3	M_4	M_5
log_lik	-1364.52 (4.14)	-1386.12 (4.08)	-1384.18 (4.52)	-1343.86 (7.66)	-1222.42 (16.41)	-1349.10 (7.35)

Table 13
Mean difference of $\widehat{AUC}_c^{\Delta c}$ (Eq. 13) and $\widehat{PE}_c^{\Delta c}$ (Eq. 15) with respect to the version of model M_3 in continuous time and prediction window of 12 months ($\Delta c = 12$). The Time(c) column represents c, the known history when making the prediction. The number in parentheses is the standard deviation of the cross-validation analysis (corrected by the overlapping training sets). The best performance of the corresponding row is marked in bold.

Time(c)	$\widehat{AUC}_c^{12}$ JM_3	$\Delta\widehat{AUC}_c^{12}$ M_3	$\widehat{PE}_c^{12}$ JM_3	$\Delta\widehat{PE}_c^{12}$ M_3
6	0.749	**0.004** (0.002)	0.359	**-0.011** (0.002)
7	0.757	**0.007** (0.002)	0.393	**-0.014** (0.003)
8	0.782	**0.010** (0.002)	0.430	**-0.016** (0.002)
9	0.777	**0.006** (0.005)	0.473	**-0.016** (0.003)
10	0.771	**0.008** (0.005)	0.502	**-0.013** (0.004)
11	0.772	**0.019** (0.009)	0.568	**-0.007** (0.006)
12	0.751	**0.028** (0.012)	0.683	**-0.005** (0.003)
13	0.768	**0.032** (0.010)	0.794	**-0.009** (0.009)
14	0.780	**0.039** (0.012)	1.112	**-0.059** (0.043)
15	0.791	**0.040** (0.006)	1.369	**-0.109** (0.089)
16	0.786	**0.046** (0.007)	1.554	**-0.148** (0.119)
17	0.788	**0.048** (0.006)	1.665	**-0.190** (0.140)
18	0.781	**0.053** (0.007)	1.815	**-0.233** (0.142)
19	0.785	**0.043** (0.007)	1.827	**-0.225** (0.125)
20	0.777	**0.045** (0.008)	1.769	**-0.205** (0.118)
21	0.793	**0.034** (0.004)	1.713	**-0.171** (0.095)
22	0.796	**0.035** (0.006)	1.509	**-0.125** (0.083)
23	0.784	**0.038** (0.005)	1.406	**-0.101** (0.059)
24	0.866	-0.047 (0.012)	1.443	**-0.162** (0.058)

gressive terms (model M_3) and compare it with a similar model formulated in continuous time.

The continuous-time version of M_3 is estimated using the R package *JMbayes* (Rizopoulos, 2014). Performance comparison is done analogous to Section 6, i.e. using the discrimination and calibration metrics described in Section 4.4. The results are shown in Table 13 where JM_3 denotes the continuous joint model. First, we observe that the joint model in continuous time obtains, in general, better discrimination performance than model M_0 (see Table 4), which is also seen in the early repayment model presented in Hu & Zhou (2019). However, the calibration performance is only better for the first evaluation periods and afterwards, it follows a similar trend as the ones seen for the other joint models (see Table 5). Second, in comparison with model M_3, we note that the discrimination and calibration metrics are slightly better for the discrete-time model in practically all the evaluation periods.

In terms of computational costs, we have measured the estimation times of the discrete-time and continuous-time versions. The continuous version needed 3.73 h and its discrete counterpart, 3.69 h. However, we must warn that these results should be taken with caution since both MCMC implementations are different which makes the comparison problematic. The *JMbayes* package uses a tailored MCMC procedure for this type of model and our version does not, which opens new lines of future research.

Appendix I. Robustness checks

To study the robustness of the results shown in Table 3, we re-estimate the model that has the most complex structure, M_5, using different priors. We keep the noninformative uniform priors for $\boldsymbol{\gamma}$, λ_f, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and ϕ. Moreover, for the covariance matrix Σ, we set the scale parameter of the LKJ distribution to 1, which corresponds to the uniform density over correlation matrices. In addition, for both variability terms, σ and υ, instead of using a uniform and a half-Cauchy priors, respectively, we use for both the inverse Gamma with shape 1 and scale 0.001, as suggested by Ibrahim et al. (2014, Ch.7). Recall that σ is the standard deviation of the error terms (Eq. 4) and υ is the hyperparameter of the vector of B-spline coefficients $\boldsymbol{\gamma}_0$ (Eq. 9). That is to say, instead of assuming $\upsilon \sim$ half-Cauchy(25) for $\boldsymbol{\gamma}_0 \sim \mathcal{N}(\mathbf{0}, \upsilon^2 I)$, we assume $\upsilon \sim$ inverse-Gamma$(1, 0.001)$. To illustrate how different these two distributions are, Table 14 shows various percentiles for each of them.

Table 14
Comparison of percentiles between half-Cauchy with a scale of 25 and a inverse-Gamma with shape 1 and scale 0.001.

	10%	25%	50%	75%	90%
Half-Cauchy(25)	3.95961	10.35534	25.00000	60.35534	157.84379
Inverse-Gamma(1, 0.001)	0.00043	0.00073	0.00145	0.00351	0.00958

Table 15
Summary of parameter estimates of the model M_5 using different prior distributions and with fold 1 kept out. To ease comparison, the three columns below M_5 are copied from Table 3 and the three below $\tilde{M}_5$ are the new results.

Parameter	M_5			$\tilde{M}_5$		
	Mean	5%	95%	Mean	5%	95%
fico	-0.701	-0.821	-0.581	-0.699	-0.820	-0.579
cltv	0.516	0.333	0.703	0.514	0.339	0.703
orig_upb	-0.155	-0.300	-0.014	-0.154	-0.304	-0.012
dti	0.152	0.021	0.283	0.150	0.022	0.283
n_borr	-0.270	-0.527	-0.018	-0.277	-0.528	-0.029
loan_purpose	-0.971	-1.246	-0.696	-0.970	-1.243	-0.706
λ_f	1.317	0.895	1.771	1.316	0.901	1.750
α_0	-0.280	-0.294	-0.266	-0.280	-0.293	-0.268
$\sigma_{U_{0i}}$	1.237	1.219	1.255	1.237	1.218	1.256
σ	0.706	0.704	0.708	0.706	0.704	0.708
ϕ	0.357	0.353	0.360	0.357	0.354	0.360
$\sigma_{U_{1i}}$	0.053	0.052	0.054	0.053	0.052	0.054
ρ_U	-0.811	-0.818	-0.804	-0.811	-0.819	-0.803

Table 16
Parameter estimates associated with the vector of B-spline functions of the model M_5 using different prior distributions and with fold 1 kept out.

Parameter	M_5			$\tilde{M}_5$		
	Mean	5%	95%	Mean	5%	95%
υ	8.344	5.206	13.371	7.122	4.609	11.034
γ_{01}	-8.577	-9.605	-7.655	-8.534	-9.584	-7.638
γ_{02}	-8.047	-9.227	-6.880	-8.064	-9.193	-6.948
γ_{03}	-6.583	-7.475	-5.695	-6.550	-7.401	-5.713
γ_{04}	-6.245	-6.877	-5.617	-6.252	-6.870	-5.624
γ_{05}	-6.052	-6.893	-5.239	-6.051	-6.884	-5.244
γ_{06}	-6.209	-7.098	-5.350	-6.191	-7.053	-5.372
γ_{07}	-6.315	-7.051	-5.632	-6.318	-7.064	-5.657

In the Table 15, under the name M_5 and to facilitate comparison, we show again the results of the parameters associated with model M_5 from Table 3. Moreover, under the name of $\tilde{M}_5$ are the results obtained using these new priors. We observe that the parameter estimates remain consistent when using these different prior distributions.

Likewise, in Table 16 we show the results for the hyperparameter, υ, and the B-spline coefficients $\boldsymbol{\gamma}_0$. We can see that although there are differences between the estimates of the hyperparameter υ when using both priors, the results corresponding to the B-splines coefficients remain practically the same, which agrees with the results in Table 15.

Appendix J. Convergence analysis

One way to check the convergence of the MCMC sampling is to compare the behaviour of randomly initialised chains. This is the motivation of the potential scale reduction factor, known as $\hat{R}$ (Gelman et al., 2013, Ch.11). This factor measures the consistency between chains by quantifying the between-chain over the within-chain variability. A value of $\hat{R} = 1$ means that all the chains are at equilibrium and values greater than one indicates that the chains have not converged to a common distribution. Gelman et al. (2013, Ch.11) recommends stopping the sampling when $\hat{R}$ has reached a value lower than 1.1 for each parameter, and *Stan* uses 1.05 as a threshold. Table 17 shows the $\hat{R}$ obtained for the parameters of each model with fold 1 kept out. All values are below 1.05. The results for the other 4 folds are consistent.

Table 17
Potential scale reduction factor ($\hat{R}$) of the parameters of each model with fold 1 kept out.

Parameter	M_0	M_1	M_2	M_3	M_4	M_5
fico	1.000	1.000	1.000	1.000	1.000	1.000
cltv	1.001	1.000	1.000	1.000	1.000	1.000
orig_upb	1.000	1.000	1.000	1.000	1.000	1.000
dti	1.000	1.000	1.000	1.000	1.000	1.000
n_borr	1.000	1.000	1.000	1.000	1.000	1.000
loan_purpose	1.000	1.000	1.000	1.000	1.000	1.000
λ_f	1.000	1.000	1.001	1.000	1.001	1.000
α_0		1.016	1.001	1.027	1.000	1.015
$\sigma_{U_{0i}}$		1.017	1.003	1.004	1.002	1.001
σ		1.000	1.001	1.000	1.000	1.000
ϕ			1.000		1.000	1.000
$\sigma_{U_{1i}}$				1.017	1.004	1.009
ρ_U				1.012	1.008	1.004

Appendix K. Economic considerations

The model comparison in Section 6 was performed using the two most common dimensions, namely discrimination and calibration. In this section, our goal is to compare the models from an economic perspective. Following the work of Bellotti & Crook (2009), we determine the value of a prediction by assigning costs to misclassification. We know that the relative cost of categorising a good account as bad is lower than categorising a bad account as good. To estimate the total cost of the predictions, we use the cost function proposed in Bellotti & Crook (2009) that assigns: a cost of 0 to a correctly categorised account, a cost of 1 to a good account predicted as bad, and a cost of 20 to a bad account predicted as good.

For each model, we estimate the cut-off threshold that minimises the total prediction cost on the training data and use it to categorise the accounts in the test set. Table 18 shows the average cost on the test set (fold 1) for different evaluation periods. The results reveal, first, that the joint models M_1 and M_3 show the same average costs as the Cox model. Second, the joint models, M_3 and M_4, which do not have autoregressive terms in the link expression between both processes (Table 2), do not seem to experience economic benefits relative to the Cox model. Finally, the model M_5, which includes autoregressive terms in both the longitudinal and the link parts, exhibits lower costs than M_3 and M_4. Furthermore, for the evaluation periods 12 and 18, M_5 can also reduce the cost compared to the Cox model and this is largely due to the autoregressive term included in the survival predictor.

Table 18
Average cost of predictions on test set.

Time(c)	M_0	M_1	M_2	M_3	M_4	M_5
6	0.1190	0.1294	0.1215	0.1573	0.1424	0.1414
12	0.2235	0.2390	0.2499	0.2385	0.2106	0.1993
18	0.3285	0.3227	0.3227	0.3430	0.3718	0.3181
24	0.2734	0.2798	0.2741	0.3366	0.3743	0.3452

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2022.10.022.

References

Alaa, A. M., & van der Schaar, M. (2017). Deep multi-task Gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 2326–2334).
Albert, P. S., & Shih, J. H. (2010a). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics, 4*(3), 1517.
Albert, P. S., & Shih, J. H. (2010b). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics, 66*(3), 983–987.
Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology, 13*(1982), 61–98.
Alsefri, M., Sudell, M., García-Fiñana, M., & Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: A methodological review. *BMC Medical Research Methodology, 20*, 1–17.
Bacci, S., Bartolucci, F., & Pandolfi, S. (2018). A joint model for longitudinal and survival data based on an ar (1) latent process. *Statistical Methods in Medical Research, 27*(5), 1285–1311.
Barrett, J., Diggle, P., Henderson, R., & Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: Flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 77*(1), 131.
Bellot, A., & Schaar, M. (2018). Tree-based Bayesian mixture model for competing risks. In *International conference on artificial intelligence and statistics* (pp. 910–918). PMLR.
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.
Bellotti, T., & Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society, 65*(3), 340–350. https://doi.org/10.1057/jors.2013.91.
Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. arXiv:1701.02434,.
Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.
BSBS (2004). International convergence of capital measurement and capital standards: a revised framework. *Technical Report*. Bank for International Settlements.
BSBS (2017). Basel III: Finalising post-crisis reforms. *Technical Report*. Bank for International Settlements.
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.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2). 87–22
Cox, D. R. (1975). Partial likelihood. *Biometrika, 62*(2), 269–276.
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.
Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing, 91*, 106263.
Dirick, L., Bellotti, T., Claeskens, G., & Baesens, B. (2019). Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. *Journal of Business & Economic Statistics, 37*(1), 40–53.
Divino, J. A., & Rocha, L. C. S. (2013). Probability of default in collateralized credit operations. *The North American Journal of Economics and Finance*. https://doi.org/10.1016/J.NAJEF.2012.06.015.
Djeundje, V. B., & Crook, J. (2018). Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. *European Journal of Operational Research, 271*(2), 697–709. https://doi.org/10.1016/j.ejor.2018.05.040.
Djeundje, V. B., & Crook, J. (2019). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research, 275*(1), 319–333.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
Fieuws, S., Verbeke, G., Maes, B., & Vanrenterghem, Y. (2008). Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics, 9*(3), 419–431.
Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.
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.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.
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.
Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*: 451. John Wiley & Sons.
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.
van Houwelingen, H., & Putter, H. (2011). *Dynamic prediction in clinical survival analysis*. CRC Press.
Hu, W., & Zhou, J. (2019). Joint modeling: An application in behavioural scoring. *Journal of the Operational Research Society, 70*(7), 1129–1139.
Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2014). *Bayesian survival analysis*. Springer Science & Business Media.
Jaffa, M. A., Gebregziabher, M., & Jaffa, A. A. (2014). A joint modeling approach for right censored high dimensional multivariate longitudinal data. *Journal of Biometrics & Biostatistics, 5*(4).
Jaffa, M. A., Woolson, R. F., & Lipsitz, S. R. (2011). Slope estimation for bivariate longitudinal outcomes adjusting for informative right censoring by using a discrete survival model: application to the renal transplant cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 387–402.
Kalbfleisch, J. D., & Prentice, R. L. (2002). The statistical analysis of failure time data. *Wiley Series in Probability and Statistics* (2nd). John Wiley & Sons.
Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology, 18*(1), 1–12.
Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963–974.
Lawrence Gould, A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., & Bois, F. Y. (2015). Joint modeling of survival and longitudinal non–survival data: Current methods and issues. report of the DIA Bayesian joint modeling working group. *Statistics in Medicine, 34*(14), 2181–2195.
Lee, C., Yoon, J., & Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering, 67*(1), 122–133.
Leow, M., & Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research, 236*(2), 685–694. https://doi.org/10.1016/j.ejor.2013.12.026.
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.
Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001.
Luck, M., Sylvain, T., Cardinal, H., Lodi, A., & Bengio, Y. (2017). Deep learning for patient-specific kidney graft survival analysis. arXiv preprint arXiv:1705.10245,.
Luong, T. M., & Scheule, H. (2021). Benchmarking forecast approaches for mortgage credit risk for forward periods. *European Journal of Operational Research*. https://doi.org/10.1016/J.EJOR.2021.09.026.
Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *The Journal of the Operational Research Society, 61*(3), 411–420. https://doi.org/10.1057/jors.2009.123.
Medina-Olivares, V., Lindgren, F., Calabrese, R., & Crook, J. (2022). Joint models of multivariate longitudinal outcomes and discrete survival data with INLA: An application to credit repayment behaviour. *Working Paper*. The University of Edinburgh.
Nadeau, C., & Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems* (pp. 307–313).
Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
Proust-Lima, C., & Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment psa: A joint modeling approach. *Biostatistics, 10*(3), 535–549.
Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online), 35*(9), 1–33.
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*. Chapman and Hall/CRC.
Rizopoulos, D. (2014). The R package jmbayes for fitting joint models for longitudinal and time-to-event data using MCMC. arXiv preprint arXiv:1404.7625,.
Rizopoulos, D., Hatfield, L. A., Carlin, B. P., & Takkenberg, J. J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using bayesian model averaging. *Journal of the American Statistical Association, 109*(508), 1385–1397. https://doi.org/10.1080/01621459.2014.931236.
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.
Stan Development Team (2018). Cmdstan: The command-line interface to stan. http://mc-stan.org.

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research, 50*(2), 277–289. https://doi.org/10.1287/opre.50.2.277.426.

Thackham, M., & Ma, J. (2020). On maximum likelihood estimation of competing risks using the cause-specific semi-parametric cox model with time-varying covariates – an application to credit risk. *Journal of the Operational Research Society*. https://doi.org/10.1080/01605682.2020.1800418.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*: 2. Society for Industrial and Applied Mathematics.

Thomas, L. C., Ho, J., & Scherer, W. (2001). Time will tell: Behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*. https://doi.org/10.1093/IMAMAN/12.1.89.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time–to-event data: An overview. *Statistica Sinica, 14*(3), 809–834.

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association, 90*(429), 27–37. https://doi.org/10.1080/01621459.1995.10476485.

Tutz, G., & Schmid, M. (2016). *Modeling discrete time-to-event data*. New York: Springer.

Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics, 34*(1), 70–85.

Volkov, A., Benoit, D. F., & Van den Poel, D. (2017). Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination. *Decision support systems, 98*, 59–68. https://doi.org/10.1016/j.dss.2017.04.008.

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.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics, 53*(1), 330–339.

Xia, Y., He, L., Li, Y., Fu, Y., & Xu, Y. (2021). A dynamic credit scoring model based on survival gradient boosting decision tree approach. *Technological and Economic Development of Economy, 27*(1), 96–119.