Dynamic survival models with time varying coefficients for credit risks

Citation for published version:
Djeundje , VB & Crook, J 2019, 'Dynamic survival models with time varying coefficients for credit risks', *European Journal of Operational Research*, vol. 275, no. 1, pp. 319-333.
https://doi.org/10.1016/j.ejor.2018.11.029

Digital Object Identifier (DOI):
10.1016/j.ejor.2018.11.029

Published In:
European Journal of Operational Research

Dynamic survival models with varying coefficients for credit risks.

Viani Biatat Djeundje and Jonathan Crook*
Credit Research Centre, University of Edinburgh

Credit Research Centre, University of Edinburgh Business School,
29 Bucceleuch Place, Edinburgh EH8 9JS, UK
{viani.djeundje@ed.ac.uk, j.crook@ed.ac.uk}

*Corresponding author

Abstract

Single event survival models predict the probability that an event will occur in the next period of time, given that the event has not happened before. In the context of credit risk, where one may wish to predict the probability of default on a loan account, such models have advantages over cross sectional models. The literature shows that the parameters of such models changed after compared with before the financial crisis of 2008. But there is also the possibility that the sensitivity of the probability of default, to say behavioural variables, changes over the life of an account.

In this paper we make two contributions. First, we parameterise discrete time survival models of credit card default using B-splines to represent the baseline relationship. These allow a far more flexible specification of the baseline hazard than has been adopted in the literature to date. This baseline relationship is crucial in discrete time survival models and typically has to be specified ex-ante. Second, we allow the estimates of the parameters of the hazard function to themselves be a function of duration time. This allows the relationship between covariates and the hazard to change over time. Using a large sample of credit card accounts we find that these specifications enhance the predictive accuracy of hazard models over specifications which adopt the type of baseline specification in the current literature and which assume constant parameters.

Keywords:

OR in Banking; Risk Analysis; Risk Management; Multivariate Statistics; Splines

Introduction

Credit scoring models are extensively used by financial institutions to evaluate the risk associated with a loan. At its core, a credit scoring model involves predicting the probability that an account will default over a future time period based on a number of observed variables, or attributes, that characterise account holders or applicants. Traditional scoring methods were based essentially on the attributes of the applicants measured at the time of application. Yet, many characteristics of the applicants change with time. Survival analysis techniques provide an attractive platform to address the limitations of traditional methods.

Survival models are not new. They have been used widely in many fields over the past 50 years, especially in medicine (Altman et al., 1995; Collett, 1993; Hougaard, 2012). In the credit risk context, the applications of these models have grown rapidly over the past decade and became an area of intensive investigations (Banasik et al., 1999; Ciochetti et al., 2002; Andreeva, 2006; Bellotti and Crook, 2009). An important advantage of these models is that they facilitate the incorporation of different types of time-varying risk factors (including behavioural variables and macroeconomic conditions) into the scoring process as suggested by Banasik et al. (1999) and tested by many authors (Stepanova and Thomas, 2001, 2002; Bellotti and Crook, 2009). In addition, survival models provide a dynamic framework for the prediction and assessment of different types of credit events (Leow et al., 2011; Bellotti and Crook, 2014, 2013). These models are being increasingly used in a variety of contexts by banks, for example in profit prediction, accept-reject decisions for mortgages and for provision calculations (IFRS9).

Most applications of survival models encountered in the literature assume that the impact of each risk factor on the probability of default remains constant over the business cycle. While this assumption is appropriate in some cases, it is questionable in general especially when the modelling period is not short. In this work, we investigate the validity of such an assumption in the context of retail banking. Specifically we consider a class of flexible models in which the marginal impacts of the risk factors are free to vary. We then propose a parametric formulation and a spline specification to capture the dynamic patterns of the impacts of the risk factors. Finally, we show that the varying coefficients approach consistently improves the overall model quality and yields more accurate predictions than the traditional constant coefficient approach.

Varying coefficients models have been used elsewhere to explore patterns. An overview of some methodological and theoretical development can be found in Hastie and Tibshirani (1993), Fan and Zhang (2008), Ferguson et al. (2007), and Park et al. (2015) among others. These models have been applied successfully to predict corporate defaults. For example, Kauermann et al. (2005) applied varying coefficient models on time-homogeneous factors to analyse the survival of newly founded firms in Germany. Hwang (2012) used varying coefficient models to illustrate how the effects of firm-specific covariates depend on the dynamics of macroeconomic factors. However, the investigation of varying covariates models in retail banking has received very little attention. The only exception is Leow and Crook (2015) but they compared the parameters in only two periods, before and after the financial crisis. We fill the gap with this paper, using a large portfolio of credit card loans comprising several time-homogeneous and time-dependent risk factors. In addition, we present a simple parametric and a flexible spline formulation of the varying effects that can be implemented using standard statistical packages.

We make three contributions to the literature. First, we present a flexible method to estimate the parameters of a survival model where the parameters themselves vary with duration time, thus allowing for the effect of each covariate to vary over time, which is highly likely to be a more realistic assumption than that of assuming the coefficients are constant. Second, we show the effects of assuming a more flexible baseline specification than what has been previously assumed in the context of discrete survival models for credit risk. Third, we illustrate, using a large sample of credit card accounts, the extend to which time-varying parameters boost predictive accuracy compared with the standard constant coefficients model. From a practical point of view these contributions are very important because when survival models are used they are used to make predictions several periods into the future so the robustness of the marginal relationships between the duration time and covariate over time is crucial to the accuracy of the predictions.

The paper is organised as follows. Section 1 introduces some notation and outlines the formulation of survival models in continuous and discrete settings. Section 2 presents varying coefficients survival models in the credit risk context and describes the parametric and splines specifications. Section 3 presents examples of parameter estimates and Section 4 compares the predictive performance of parameteric and spline function specifications against the standard constant coefficient model. We close with some concluding remarks in Section 5.

1 Survival analysis

1.1 Standard survival model

Survival analysis is the term used to describe the study of time between entry to a study and a subsequent event (such as death or default). Thus, the modelling of defaults in credit risk lends itself naturally to the survival analysis framework. The most commonly used survival model is the so-called Cox model (Cox, 1972). Let us denote by $\lambda_i(t)$ the hazard function for account i at duration time t; that is:

$$\lambda_i(t) = \lim_{\Delta t \to 0} \frac{Pr\{i \text{ will default before time } t + \Delta t, \text{ given that } i \text{ was still active at time } t\}}{\Delta t} \tag{1}$$

In its simplest form, the Cox model specifies the hazard function as

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta}), \tag{2}$$

where $\lambda_0(t)$ is an unspecified and non-negative function of time, $\boldsymbol{X}_i$ is the $(1 \times p)$ vector of covariates for account i, and $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of coefficients. The function λ_0 can be interpreted as the hazard function for an account whose covariates all have the value of 0. Thus it is usually referred to as the baseline hazard. An important feature of formulation (2) is that the ratio of the hazard of two individuals is independent of time. In order words, the value of the hazard of any individual is a fixed proportion of the hazard for any other individual. Thus, this model is generally refereed to as the proportional hazard model (Collett, 1993; Allison, 2010).

In typical credit portfolios however, many potential risk factors change over time. In general, let us denote by $\boldsymbol{X}_i(t)$ the $(p \times 1)$ joint vector of all covariates for account i at time t; this includes the time-dependent covariates as well as the time-homogeneous ones. The basic Cox model (2) is extended to

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{X}_i(t) \boldsymbol{\beta}) \tag{3}$$

As remarked by Cox, the likelihood function arising from these models factors into two components: a first component that depends on both $\lambda_0(t)$ and $\boldsymbol{\beta}$, and a second component that depends on $\boldsymbol{\beta}$ alone. This second component is usually referred to as the partial likelihood function and in most applications of the Cox model, inference about the relative significance of the risk factors is based on this partial likelihood. An attractive feature of this approach is that it does not require any constraint on the shape of the baseline function $\lambda_0(t)$. However, as pointed out by many authors, such a truncation of the likelihood function yields estimates that are not fully efficient although in practice the lost of efficiency is generally small (Efron, 1977; Allison, 2010). An alternative estimation approach that guarantees full efficiency is to maximise the full likelihood upon some restriction on the form of the baseline.

1.2 Parameterisation of survival models for credit risk

In practice, credit risk data are usually discrete in time. Let us denote by $q_{i,\tau}$ the conditional probability that account i will default in month τ given that it is still active at the beginning of the month. That is, $q_{i,\tau}$ is a discrete-time approximation to the hazard function. Assuming constant values for each covariate within months, Model (3) can be approximated (Cameron and Trivedi, 2005) as follows

$$g(q_{i,\tau}) = h_{0,\tau} + \boldsymbol{X}_{i,\tau} \boldsymbol{\beta}, \tag{4}$$

where g is the complementary *log-log* function defined by $g(x) = \log(-\log(1-x))$ and $h_{0,\tau}$ is a transformed baseline. This discrete representation facilitates the implementation of survival models for credit risk data. In particular, if we replace the complementary *log-log* function by the *logit* function, we obtain the standard logistic regression model.

Risk factors used in credit risk models fall into three classes: time-homogeneous but account-dependent factors (often referred to as application variables), time-dependent and account-dependent factors (often referred to as behavioural variables) and time-dependent but account-independent factors (e.g. the macroeconomic variables). To allow prediction to take place, the time-dependent covariates are usually lagged. Thus, model (4) takes the following expanded form

$$g(q_{i,\tau}) \;=\; h_{0,\tau} \;+\; \boldsymbol{U}_i\boldsymbol{\alpha} \;+\; \boldsymbol{V}_{i,\tau-\tau_o}\,\boldsymbol{\delta} \;+\; \boldsymbol{Z}_{i,\tau-\tau_o}\,\boldsymbol{\gamma}, \tag{5}$$

where $\boldsymbol{U}_i$ represents the row-wise vector of application variables, $\boldsymbol{V}_{i,\tau-\tau_o}$ denotes the behavioural variables, $\boldsymbol{Z}_{i,\tau-\tau_o}$ represents the macroeconomic conditions, and τ_o is the lag. Since the macroeconomic conditions are the same for all accounts observed at the same calendar time, the dependence of $\boldsymbol{Z}_{i,\tau-\tau_o}$ on i is only due to the fact that accounts are opened at different points in calendar time. Correspondingly to model (4), we have:

$$\boldsymbol{X}_{i,\tau} = [\boldsymbol{U}_i,\, \boldsymbol{V}_{i,\tau-\tau_o},\, \boldsymbol{Z}_{i,\tau-\tau_o}] \text{ and } \boldsymbol{\beta} = [\boldsymbol{\alpha}^T,\, \boldsymbol{\delta}^T,\, \boldsymbol{\gamma}^T]^T. \tag{6}$$

A central objective of credit risk models is to quantify not only the relative importance of the risk factors, but also the full probability of default. This requires estimation of the baseline and the regression parameters. Thus, the model specification is completed by imposing some reasonable structure on the baseline. Both rigid parametric structures and flexible splines specification are possible; see for example Crook and Bellotti (2010), Luo et al. (2016), Djeundje and Crook (2018). With this in place, if we denote by $\boldsymbol{\theta}$ the joint vector of all parameters in the model (including the parameters that define the baseline), an efficient estimate for $\boldsymbol{\theta}$ can be found by maximising the likelihood function, L, given by

$$L(\boldsymbol{\theta}) \;\propto\; \prod_{\tau} \prod_{i\in\mathcal{R}(\tau)} (q_{i,\tau})^{y_{i,\tau}} \times (1-q_{i,\tau})^{1-y_{i,\tau}} \tag{7}$$

where $\mathcal{R}(\tau)$ represents the set of accounts that are active and so at risk of default at the beginning of month τ, $q_{i,\tau}$ is the conditional default probability defined in Equation (4), and $y_{i,\tau}$ is the indicator function taking value 1 if account i has defaulted during month τ and 0 otherwise. †

2 Modelling varying effects in credit risk

The models described in the previous section assume that the magnitude of the impact of the risk factors (ie $\boldsymbol{\beta}$) remains constant over time. This is a strong and

†The likelihood (7) arises by rearranging the product of individual contributions, L_i, given by

$$L_i(\boldsymbol{\theta}) \;=\; \begin{cases} \prod_{\tau=1}^{\tau_i-1} (1-q_{i,\tau})\, q_{i,\tau_i} & \text{if account } i \text{ has defaulted} \\ \prod_{\tau=1}^{\tau_i} (1-q_{i,\tau}) & \text{if account } i \text{ is censored} \end{cases}$$

where τ_i denotes the last observed month associated with account i.

questionable assumption especially when the modelling period is not short. For instance, using a dataset on credit card loans, Leow and Crook (2015) built two survival models (based on accounts that were opened before and after the 2008 crisis) and show that the magnitudes of the impact of the risk factors from the two models were statistically different from each other. The aim of this section is to describe how to allow for changes in the magnitude of the impacts of the risk factors, when building a dynamic model for credit risk. In a later section we will illustrate how this improves the quality of the model. The modelling framework that we adopt for this is that of varying coefficient models.

In the varying coefficients approach, model (4) is generalised to

$$g(q_{i,\tau}) = h(\tau) + \boldsymbol{X}_{i,\tau}\,\boldsymbol{\beta}(\tau), \tag{8}$$

where the components of the joint parameter vector $\boldsymbol{\beta}(\tau)$ are allowed to vary over duration time.[‡] That is, some (or all) regression parameters are now functions of time. An investigation of these functions can be used to validate or to reject the assumption of constant parameters commonly used in dynamic models for credit risk.

Notice that we model how the parameters change as an account ages, rather than how calendar time per se affects parameters. We argue that the sensitivity of future hazard probabilities to information gained at the time of application, and the sensitivity of future hazards to behavioural factors, all vary over the life of an account. For example we expect the affect of application variables to decline over time. It is also possible that the sensitivity of the hazard to macroeconomic variables also changes over an account's life.

In general, the effects of a covariate can vary not only over time but also according to some attributes or risk factors. In this case, the vector of coefficients $\boldsymbol{\beta}(\tau)$ takes the form $\boldsymbol{\beta}(C_{i,\tau})$, where C represents the conditions or attributes driving the magnitude of the impacts of the risk factors $\boldsymbol{X}_{i,\tau}$. However, formulation (8) is general enough to illustrate the importance of varying coefficients when modelling portfolios of credit loans, as we shall see in Section 3.

For the sake of clarity, we express equation (8) as

$$g(q_{i,\tau}) = h(\tau) + \boldsymbol{X}^{(0)}_{i,\tau}\,\boldsymbol{\beta}^{(0)} + \boldsymbol{X}^{(1)}_{i,\tau}\,\boldsymbol{\beta}^{(1)}(\tau), \tag{9}$$

where $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(1)}$ represent the time-independent and time-varying components of $\boldsymbol{\beta}$ respectively. A natural question that emerges is how should we specify and estimate the components of $\boldsymbol{\beta}^{(1)}(\tau)$?

[‡]In equation (8), $h(\tau)$ is still the baseline; ie $h(\tau) = h_{0,\tau}$. Its specification under two parameterisations is provided in the Sections 2.1 & 2.2 below.

2.1 Parametric specification

A naive way to model the varying coefficients $\boldsymbol{\beta}^{(1)}(\tau)$ in equation (9) is to consider a simple parametric shape. For example, assuming that they can be described by straight lines, the jth component $\beta_j^{(1)}$ of the coefficients vector $\boldsymbol{\beta}^{(1)}$ takes the form

$$\beta_j^{(1)}(\tau) = a_j + b_j\,\tau, \tag{10}$$

where the intercepts a_j and slopes b_j are parameters to be estimated. In this case, the elements $X_{i,\tau,j}^{(1)}\,\beta_j^{(1)}(\tau)$ of $\boldsymbol{X}_{i,\tau}^{(1)}\,\boldsymbol{\beta}^{(1)}(\tau)$ expand into

$$X_{i,\tau,j}^{(1)}\,\beta_j^{(1)}(\tau) = a_j\,X_{i,\tau,j}^{(1)} \;+\; b_j\,\tau\,X_{i,\tau,j}^{(1)}. \tag{11}$$

That is, the varying coefficient model (9) falls into the family of time-dependent covariate model (4), but with extra pseudo covariates given by the $\tau \times X_{i,\tau,j}^{(1)}$.

Clearly the straight line assumption would not hold in many cases. One possibility is to capture non-linear effects through appropriate basis functions. For example, Bellotti and Crook (2013) investigated the effectiveness of using a family of four standard functions to capture the shape of the baseline hazard in the credit risk context; the same approach was adopted by Leow and Crook (2015). In this work, as a benchmark starting point, we consider the following family (comprising the functions used by Bellotti and Crook 2013)

$$\left\{1,\, \tau,\, \tau^2, \sqrt{\tau},\, \frac{1}{\tau},\, \log(\tau),\, [\log(\tau)]^2\right\}. \tag{12}$$

Hence, the components $\beta_j^{(1)}$ of $\boldsymbol{\beta}^{(1)}$ in equation (9) take the form

$$\beta_j^{(1)}(\tau) = a_j \;+\; b_j\,\tau \;+\; c_j\,\tau^2 \;+\; d_j\,\sqrt{\tau} \;+\; \frac{e_j}{\tau} \;+\; f_j\,\log(\tau) \;+\; g_j\,[\log(\tau)]^2 \tag{13}$$

where $(a_j,\, b_j,\, c_j,\, d_j,\, e_j,\, f_j,\, g_j)$ are unknown parameters to be estimated. The baseline $h(\tau)$ can be expressed in a similar form.[§] In the rest of the paper we shall use the term *parametric specification* whenever the baseline and varying coefficients are modelled using the family (12).

With this specification, all the regression coefficients can then be jointly estimated by maximising the likelihood function (7), and the standard error estimates of the parameters are used to carry out statistical tests. In particular, tests on the shape of the varying coefficients are based on the ratios of the estimates $(\hat{a}_j,\, \hat{b}_j,\, \hat{c}_j,\, \hat{d}_j,\, \hat{e}_j,\, \hat{f}_j,\, \hat{g}_j)$ to their standard errors.

This can be implemented in a standard statistical package for generalised linear models upon computation of extra pseudo covariates similar to (11) but with $\beta_j^{(1)}(\tau)$ as in equation (13).

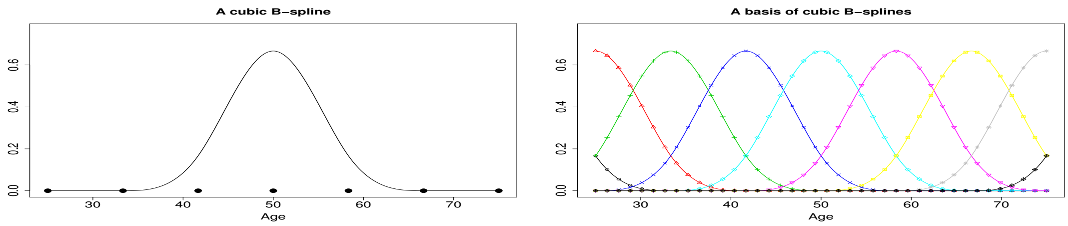
[§]The baseline $h(\tau)$ can also be expressed as a linear combination of the family of basis functions in (12). That is, $h(\tau) = a_0 + b_0\,\tau + c_0\,\tau^2 + d_0\,\sqrt{\tau} + \frac{e_0}{\tau} + f_0\,\log(\tau) + g_0\,[\log(\tau)]^2$, where $(a_0, b_0, c_0, d_0, e_0, f_0, g_0)$ are parameters to be estimated.

2.2 Flexible B-splines specification

Although the family (12) is broad enough to handle some complex forms of baseline and varying coefficients, it can suffer from the global dependence of these functions on local properties of the data (De Boor, 1978). In other words, a given month can exert an unexpected influence on remote parts of the fitted $\hat{\beta}_j^{(1)}(\tau)$, and such behaviour can potentially lead to unstable predictions with poor interpolation properties, as illustrated in Djeundje (2011).

A more attractive approach is to express the baselines and varying coefficients using a basis of splines. Such bases have been used extensively in the literature to model complex variabilities; this includes radial basis, backward and forward truncated lines (Ruppert et al., 2003; Djeundje, 2016), as well as B-splines (Eilers and Marx, 1996; Brown et al., 2005).

Figure 1: Illustration of B-splines.



Left: One cubic B-spline; the dots represent the knots. Right: Basis of cubic B-splines.

A B-spline can be described as a combination of truncated polynomials. An illustration of B-splines is shown on Figure 1. Each B-spline has a compact support and this makes them numerically advantageous over other spline bases. For a complete description of B-splines, we refer the reader to De Boor (1978) or Eilers and Marx (1996). We use cubic B-spline basis in this paper; some motivations of this preference are discussed by Green and Silverman (1995).

In terms of B-splines, the jth component $\beta_j^{(1)}$ of the coefficients vector $\boldsymbol{\beta}^{(1)}$ in equation (9) takes the form

$$\beta_j^{(1)}(\tau) = \sum_r \mathcal{B}_{j,r}(\tau)\, \phi_{j,r} \tag{14}$$

where $\mathcal{B}_{j,r}(\tau)$ are cubic B-spline functions at time point τ, and $\phi_{j,r}$ are unknown splines coefficients to be estimated; The baseline can be expressed in a similar form, yet with different coefficients.¶

¶ $h(\tau) = \sum_r \mathcal{B}_{0,r}(\tau)\, \phi_{0,r}$

In the rest of the paper, we use the term *splines specification* whenever the baseline and varying coefficients are expressed in terms of B-splines as in equation (14). Under this specification, extra pseudo covariates can be computed as in Section 2.1, but with the $\beta_j^{(1)}(\tau)$ given by (14), and all the parameters (including the splines coefficients) can then be jointly estimated by maximising the likelihood defined in (7) using standard packages for regression models.

When fitting models using B-splines however, an important point to address is the number and positions of the knots. Indeed, this can have a detrimental impact on the values and shapes of the fitted varying coefficients. For example, at one extreme, using too many splines can lead to over-fitting; at the other end, using an insufficient number of splines or poor knot locations can yield a model that does not fit the data well. This can negatively affect the predictive performance of the model. In the literature, there are two major approaches to avoid this problem.

One approach is to cover the data range with a sufficiently large number of B-splines and then penalise the roughness in adjacent spline coefficients to achieve smoothness (Eilers and Marx, 1996; Wood, 2006). With this approach, smoothing parameters are introduced (one for each varying coefficient) and used to tune the amount of smoothing. Optimal values of these smoothing parameters must be chosen carefully because large (small) values can lead to under (over) fitting. In practice, these parameters can be selected via information criteria (or via MCMC simulations especially in the context of a large number of smoothing parameters).

An alternative approach is to carefully and parsimoniously select the number of splines and knot positions; see Friedman and Silverman (1989). In this work for example, we considered various scenarios separately for each varying coefficient (including equi-spaced and irregular knot spacing) and selected scenarios corresponding to lower values of the Akaike Information Criteria; see (16). [††]

All models presented in this paper were implemented using `SAS` software. But there are functions in other standard statistical software that can be used to estimate these models as well. For instance, varying coeffieints models expressed in terms of B-splines as in Equation (14) can be seen as extension of GAMs (Hastie and Tibshirani, 1990; Eilers and Marx, 2002); thus, packages developped for GAMs such as `mgcv` (Wood, 2016, 2006) or `R-INLA` (Rue et al., 2009) in `R` can be adapted to fit some of the varying coefficients models described in this paper.

[††]The models presented in this paper use up to 10 internal knots for each varying coefficient. Using a larger set of knots on our dataset generally yielded less attractive models in terms of Akaike Information Criteria and prediction performance.

3 Application

3.1 Data and risk factors

For illustration we consider a dataset of credit card accounts supplied by a major UK bank. This consists of more than 200,000 individual accounts opened from 2002 to 2011 on different books. The dataset contains several variables collected at the time of application as well as behavioural variables collected monthly. In addition, some macroeconomic variables were appended to the dataset. The complete list of variables used in this investigation is shown in Table 1.

Table 1: Risk factors used in this investigation.

Application variables	Number of cards	Categorical (4 groups)
	Variable X	Categorical (5 groups)
	Employment type	Categorical (5 groups)
	Age at application	Categorical (10 groups)
Behavioural variables	Repayment amount	Continuous
	Prop one-month delinquency	Continuous
	Prop two-month delinquency	Continuous
Macroeconomic variables	Index of production	Continuous
	Consumer confidence	Continuous
	FTSE index	Continuous
	Unemployment rate	Continuous

The dataset was split into three parts: a training set, a retrospective test set and a prospective test set. The training data set consists of a random sample of 80% of all the accounts which were opened from January 2002 to December 2008. The retrospective test set consists of the 20% out of sample of accounts opened from January 2002 to December 2008. The prospective test consists of accounts opened from January 2009 onwards. Thus, relative to the training set, the retrospective test set is out of sample but in time, whereas the prospective test set is out of sample and out of time.

Models were fitted using the training dataset whereas prediction performance of different models was assessed and compared using the retrospective and prospective test sets. An account was defined as being in default if it has missed three payments. Note that these missed payments need not to be in consecutive months. This definition is consistent with that used in Leow and Crook (2014) and Djeundje and Crook (2018).

3.2 Models and outputs

The main purpose of this work is to investigate the improvements arising from the incorporation of time-varying coefficients into survival models in a credit risk context. We also wish to see time varying parameters if spline, rather than parametric parameterisations, enhance predictive accuracy more for application models than for behavioural models. Thus, several models were implemented, starting from the static models without time-varying coefficients through to models with time-varying coefficients on several risk factors simultaneously. The list of models discussed in this paper is shown in Table 2.

Each model in Table 2 was first fitted under the parametric assumption (13) and next under the spline specification (14), giving rise to 10 models comprising two without time-varying coefficients and eight with time-varying coefficients. In each case, all the covariates in Table 1 were included, with a time-varying coefficient specification on the relevant covariates according to the description in Table 2.

Table 2: List of models.

Model code	Description
M0	model without varying coefficients
M1	model with varying coefficients only on application variables
M2	model with varying coefficients only for behavioral variables
M3	model with varying coefficients only for macroeconomic variables
M4	model with varying coefficients for application, behavioral and macroeconomic variables

Each model listed in this table was implemented under both parametric and splines specifications, giving rise to 10 models in total. In addition to these, models with varying coefficients on single variables were also investigated. Model M0 was implemented with the assumptions (13) and (14) for the baseline only.
All models contain the application, behavioural and macroeconomic variables listed in Table 1.

The fitted coefficients from the two models without varying coefficients are shown in Table 3. Several conclusions can be drawn from this table. For example, we note that the estimated coefficients from both models are very similar and highly significant. Also, this table indicates that holding more credit cards increases the risk of default. Furthermore, it shows that the risk of default increases as the proportion of time spent with one or two payments in arrears increases (see coefficients for *Prop one-month delinquency* and *Prop two-month delinquency*).

Exploring baselines

Figure 2 shows the fitted baselines from the 10 models described in Table 2. For each model, there is some difference between the fitted baseline from parametric and splines specifications; in particular, the panel on the right reveals the potential

Table 3: Parameter estimates from the two models without time-varying coefficients.

		M0 with parametric baseline		M0 with splines baseline	
		Est.	*p-val*	*Est.*	*p-val*
Application Variables	*Number of cards, group B*	0.02435	0.00959	0.02409	0.01037
	Number of cards, group C	0.10269	0.00000	0.10261	0.00000
	Number of cards, group D	0.21284	0.00000	0.21285	0.00000
	Variable X, group B	0.41530	0.00000	0.41518	0.00000
	Variable X, group C	0.49518	0.00000	0.49539	0.00000
	Variable X, group D	0.22397	0.00000	0.22380	0.00000
	Variable X, group E	0.36725	0.00000	0.36740	0.00000
	Age at application, group 2	-0.11765	0.00000	-0.11750	0.00000
	Age at application, group 3	-0.15831	0.00000	-0.15825	0.00000
	Age at application, group 4	-0.14723	0.00000	-0.14722	0.00000
	Age at application, group 5	-0.16270	0.00000	-0.16272	0.00000
	Age at application, group 6	-0.22887	0.00000	-0.22894	0.00000
	age at application, group 7	-0.34847	0.00000	-0.34856	0.00000
	Age at application, group 8	-0.52694	0.00000	-0.52712	0.00000
	Age at application, group 9	-0.78460	0.00000	-0.78477	0.00000
	Age at application, group 10	-1.10357	0.00000	-1.10374	0.00000
	Employment code, group B	0.12299	0.00000	0.12288	0.00000
	Employment code, group C	-0.10384	0.00013	-0.10397	0.00013
	Employment code, group D	0.03055	0.02933	0.03119	0.02613
	Employment code, group E	0.14443	0.00000	0.14405	0.00000
Behavioural variables lagged 6 months	*Prop one-month delinquency*	3.74603	0.00000	3.74746	0.00000
	Prop two-month delinquency	3.04395	0.00000	3.04566	0.00000
	Repayment amount	0.05972	0.00000	0.05974	0.00000
Macroeconomic variables lagged 6 months	*Index of production*	-0.00103	0.20647	-0.00099	0.22424
	Consumer confidence	-0.00590	0.00000	-0.00584	0.00000
	FTSE index	-0.00009	0.00000	-0.00008	0.00000
	Unemployement rate	-0.02444	0.00010	-0.02332	0.00021

The same covariates are used in both models; the only difference is the structure of their baselines. In the first model, the baseline is expressed using the parametric family in (12); the baseline in the second model is expressed in terms of B-splines.

of the splines approach in terms of its ability to capture hidden patterns from the data (provided care is taken to avoid under(over)-smoothing).

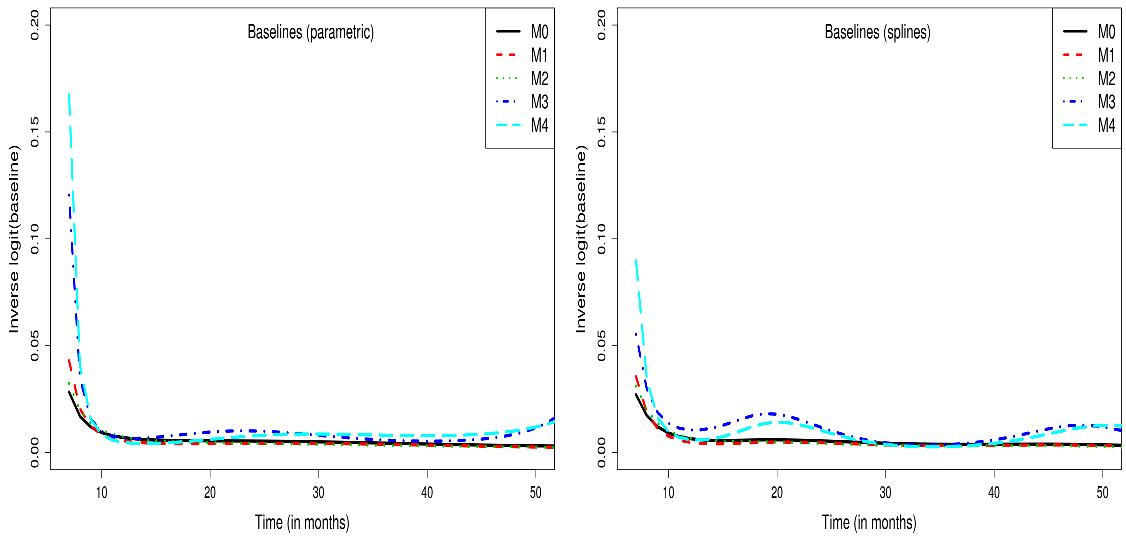
The two panels show that the risk of default is higher around the 6th month following the opening date of the account, but decreases sharply during the second semester and tends to stabilise thereafter. However, one should bear in mind that any interpretation of these baselines should be done with caution because these baselines are not representative of all the covariate patterns in the training data.

Exploring the relative effects of the application variables

In this section, we explore the fitted coefficients associated with the application variables. These variables are categorical giving rise to several indicator variables, one for each category. As described in Table 2, varying coefficients were first allowed on the application variables alone (see model M1) and then simultaneously on all variables (see M4).

Figure 3 shows the fitted coefficients at each month that are associated with the variable *number of cards* for each model listed in Table 2. Each panel refers to the coefficient of the indicator for the relevant category. The left panels are based on the parametric formulation of the baselines and varying coefficients, whereas

Figure 2: Fitted baselines.



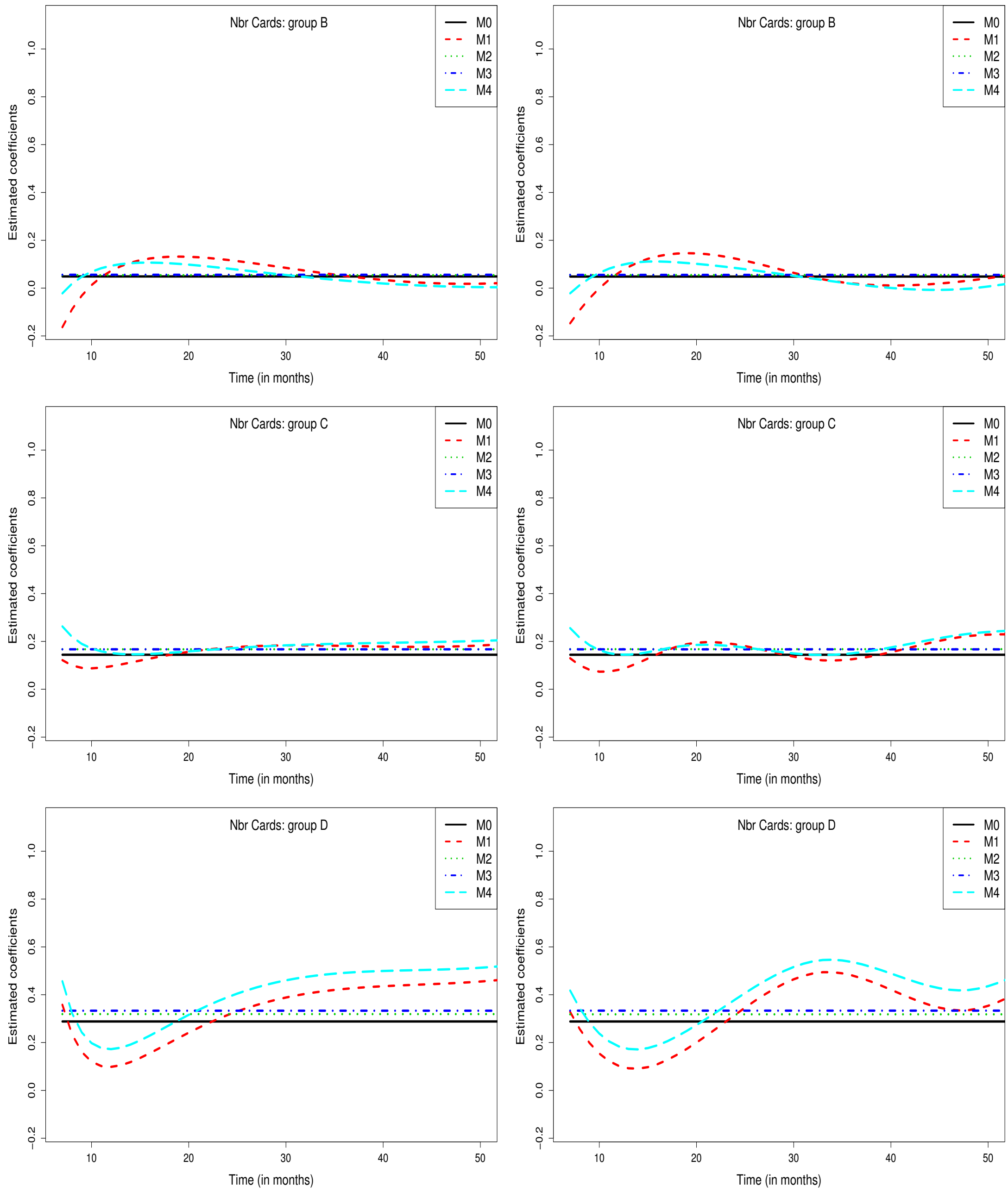
Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines-based.

the right panels are splines based.

A number of conclusions can be drawn from these graphics. For example, the coefficients of the accounts in group C of *number of cards* is broadly the same for the models without and with varying coefficients M0 and M1 to M4, respectively (see the two middle panels). But in general, the shape of the coefficients varies from one class of *number of cards* to another. The spline formulation is able to detect a more granular change in the magnitude of the coefficients compared to its parametric counterpart. Under each formulation, the difference between models M1 and M4 is mainly due to the interaction between varying coefficients over time. For example consider group B which corresponds to few cards. Increased card age may increase the hazard. This may be because the outstanding balance on the card may increase and with few cards cannot be spread over many other cards, though this effect dwindles. For a larger number of cards at time of application (group D), taking out another card has an increasing effect on the hazard over time but by declining amounts. Between 15 and 33 months the effect increases and this may be due to a greater volume of debt that may be built up over a larger number of other cards, rather than no cards, which is the reference category.

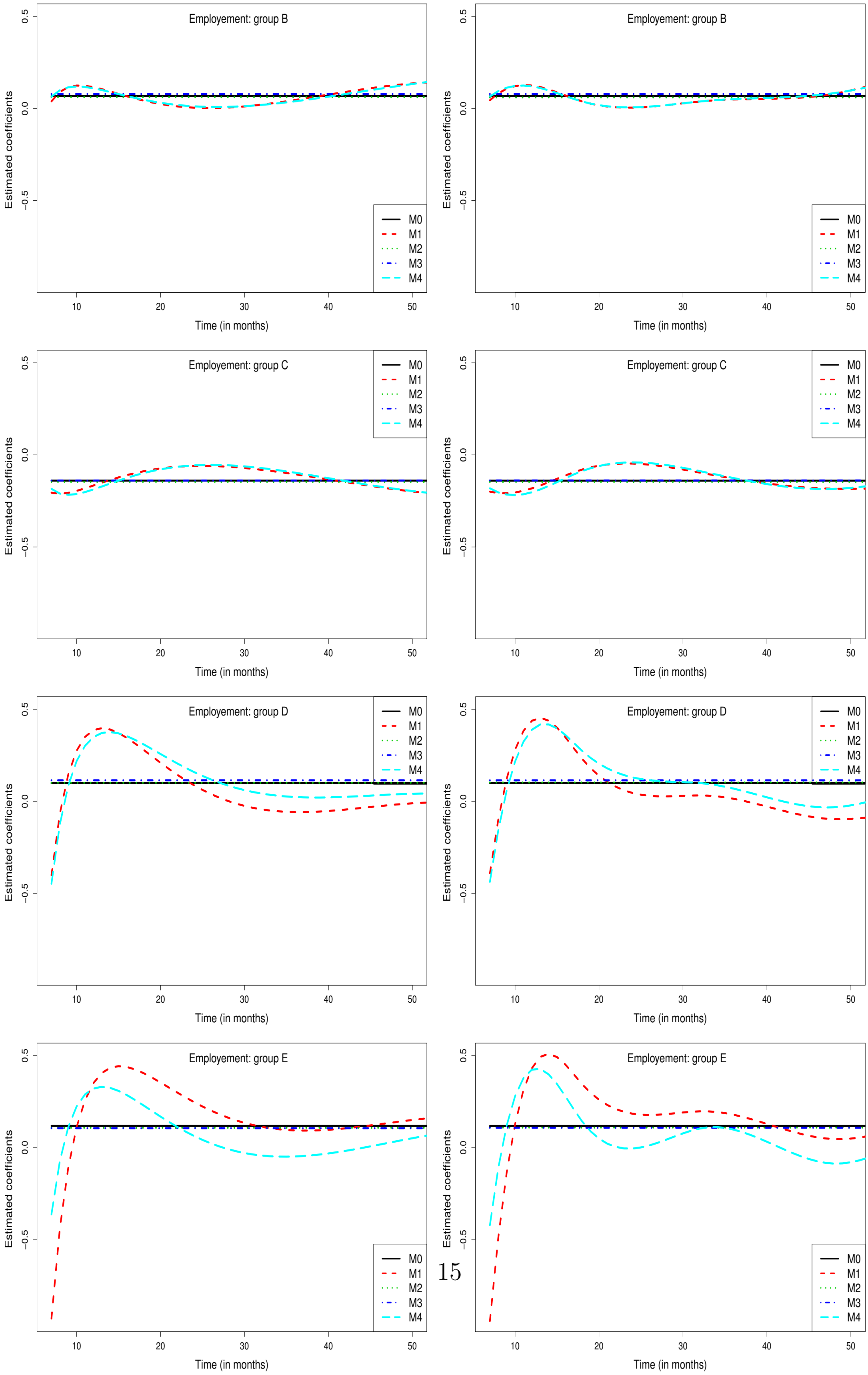
Differences in the time pattern of coefficients also occurs for other application variables. For example, Figure 4 reveals a quick change in the sign and magnitude of the varying coefficients associated with categories E and D of employment types. This can be tested formally as indicated in Section 2.1. Employment categories

Figure 3: Fitted coefficients for each group of variable *number of cards*.



Left: baselines and time-varying coefficients are parametric based. Right: baselines and time-varying coefficients are splines based.

Figure 4: Fitted coefficients for each *employment type*.



Left: baselines and time-varying coefficients are parametric based. Right: baselines and time-varying coefficients are splines based.

D and E (students and no information) show the same time pattern of the coefficients: when an account is opened this group has a lower hazard than the reference category but as the account ages the effect increases, reaches a maximum around 12 months and then decreases. Intuitively this may be because the balance may build up over time but then card holders get used to managing their repayments and they may gain employment and this may account for the effects declining to zero. For both the number of cards and employment group, both at the time of application, the marginal effect would be expected to decline as the account ages because this information becomes increasingly less indicative of the card holders ability to repay.

The marginal effects of other application variables vary over time. For example Figure A1 for the coefficients associated with *Variable X* as well as Figure A2 for *Age*, both in the Appendix. For each age group the time pattern of coefficients is broadly similar. One intuitive explanation might be that the card we are modelling provides additional debt capacity but as it is used and the outstanding balance may increase so does the hazard. But the cardholder may become more used to making payments as his/her income rises so the hazard assymptotes.The latter seems to indicate that an assumption of constant coefficients for most ages bands looks reasonable.

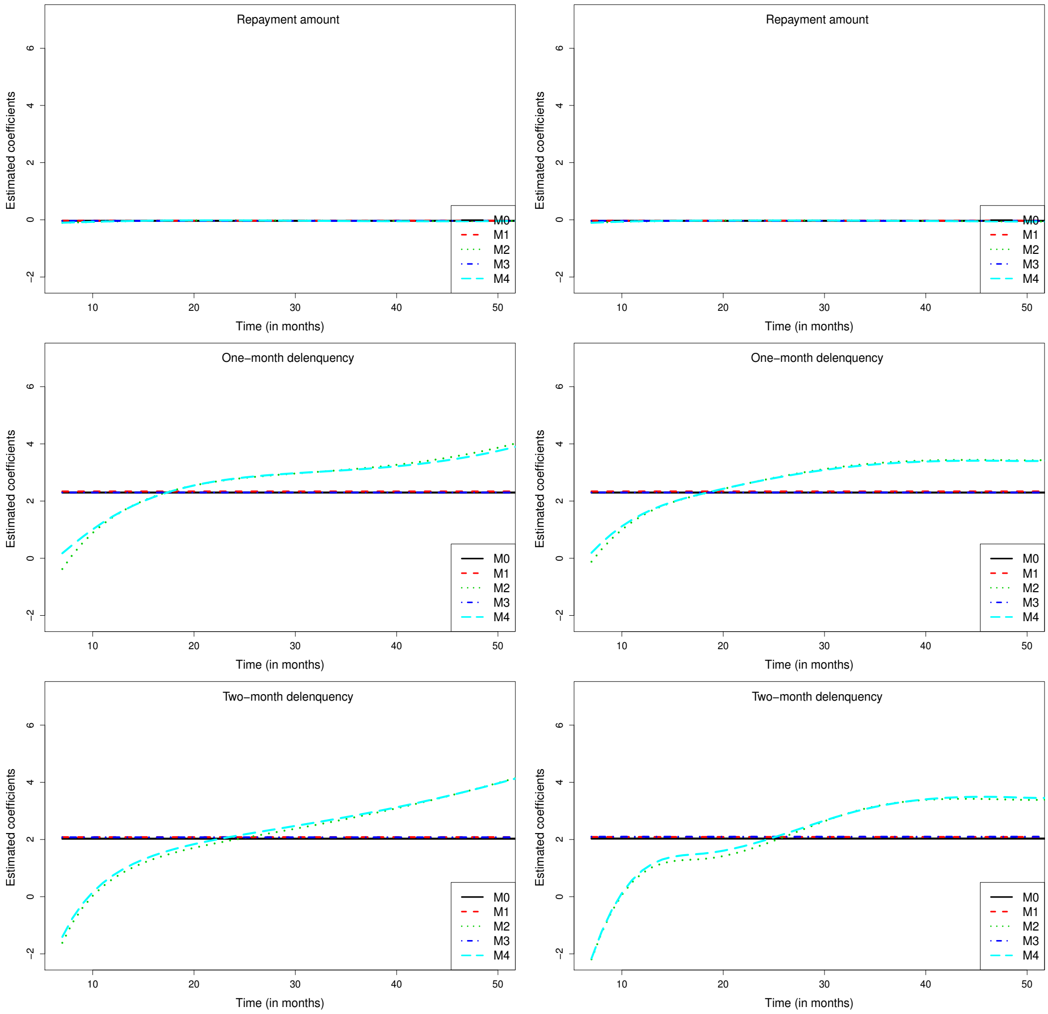
Exploring the relative effects of the behavioural variables

We now turn to the behavioural variables; there are three of them all continuous and time-dependent. The fitted coefficients associated with these variables are shown in Figure 5. The coefficients associated with the variable *Repayment amount* are essentially the same (two top panels) regardless of the model considered. However the four lower panels reveal that the longer the card is held the larger the marginal effect for both the proportion of survival time with one outstanding payment and for proportion of survival time with two such payments. Intuitively there may be an interaction: the longer the card is held the greater the values of these covariates are likely to be. But, given the proportions, the effect is also larger possibly because over time other loans are also acquired which a given disposable income has to repay.

Exploring the relative effects of the macroeconomic variables

Finally we consider the effects of macroeconomic variables; the result is shown on Figure 6. Unlike the coefficients associated with the first three macroeconomic variables, we observe greater variability in the varying coefficient associated with *Unemployment rate*. In particular, the effect of unemployment rate decreases (in absolute value) steeply until month 15, remains constant and then declines in both

Figure 5: Fitted coefficients for behavioural variables.



Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines based.

specifications. An intuitive explanation may be that just after opening the account the card holder has more credit availability which can be used in unemployed spells but over time this may reduce if debt is built up as might be expected over time. After around 30 months only the good payers remain and these may be more financially robust to changes in the macroeconomy. In the case of consumer confidence where the variability in marginal effects is small, the longer a card is held the more confident a card holder may be about being able to repay so that if confidence generally increases those who have held a card longest may be especially confident and overextend themselves in terms of ability to repay.

4 Assessment and comparison

4.1 Model checking

The analysis of residuals is a crucial step for checking model assumptions in regression type models. The monthly deviance residuals for each model were calculated as follows:

$$D_\tau = \pm\, 2 \left[O_\tau \times \log\left(\frac{O_\tau}{E_\tau}\right) + (N_\tau - O_\tau) \times \log\left(\frac{N_\tau - O_\tau}{N_\tau - E_\tau}\right)\right] \tag{15}$$

where N_τ represents the number of accounts at risk at the beginning of month τ; O_τ is the total number of defaults that occurred during month τ, and E_τ denotes the corresponding expected number of defaults.

A graphical illustration of these residuals from the 10 models in Table 2 is displayed in Figure 7. These residuals are broadly similar across the 10 models and all show more variations during the first months. But overall the residuals from each model are centred with no discenible patterns.

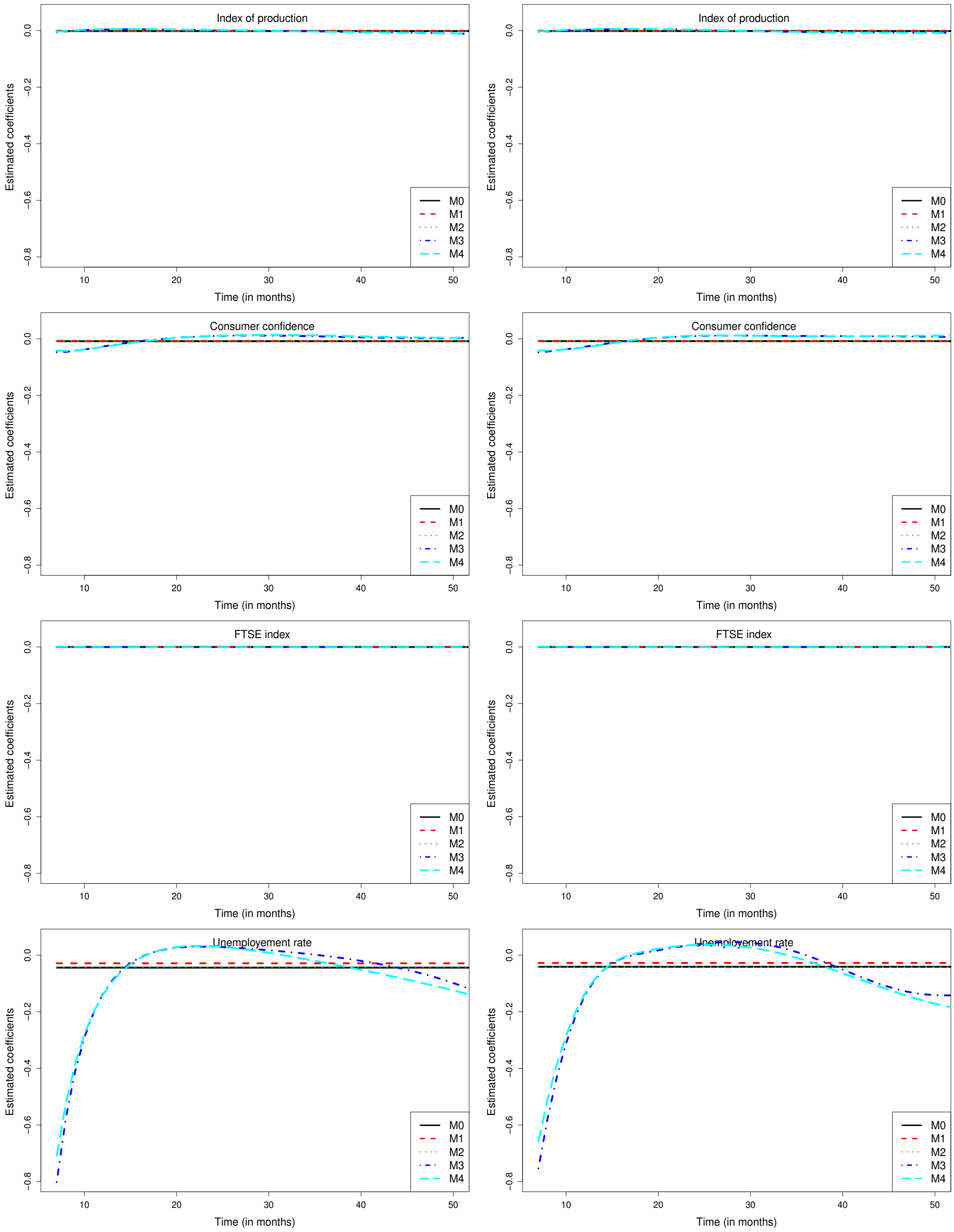
4.2 Overall model quality

In general it is always possible to improve model fit by adding in a new variable; but doing so can lead to overfitting and poor predictive power. A penalty against model complexity allows one to avoid this problem. The Akaike Information Criteria (AIC) measures the relative goodness of fit of a statistical model with a suitable penalty term for complexity. It is defined by

$$AIC = -2\hat{\ell} + 2p \tag{16}$$

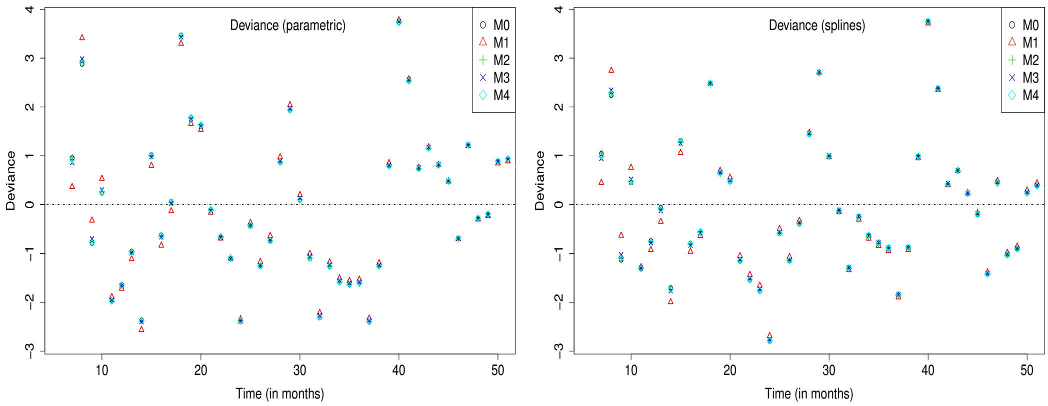
where p represents the number of parameters in the model, and $\hat{\ell}$ is the maximised value of the log-likelihood function. In general models with lower AIC would be preferred.

Figure 6: Fitted coefficients for macroeconomic variables.



Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines based.

Figure 7: Deviance residuals.



Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines based.

The AICs from the models described in Table 2 are shown in Table 4. Several conclusions can be drawn. First, under parametric or spline specifications, the models with varying coefficients outperform the standard survival model M0. Second, allowing for varying coefficients on several covariates simultaneously (see model M4) yields a further improvement compared to models with varying coefficients on a restricted set of covariates. Third, the varying coefficients models with spline specification tend to be better than their parametric counterparts.

Table 4: Comparative AIC from different models.

	Parametric-based		**Splines-based**		**Parametric vs Splines**
	AIC	*Drop in AIC relative to M0*	*AIC*	*Drop in AIC relative to M0*	*From parametric to spline: Drop in AIC*
M0	804860	0	804779	0	82
M1	802897	1964	802748	2031	149
M2	802700	2160	802638	2141	62
M3	801902	2959	801778	3001	124
M4	798728	6133	798489	6289	238

4.3 Prediction performance

We use two metrics to compare the predictive accuracy of the models. First, comparisons of Receiver Operating Characteristics (ROC) curves and second, the costs of misclassification. For the ROC the estimated parameters of the models were used to predict the probabilities of default for each account in the test sets, and these probabilities were used in turn to construct ROC curves, separately for each model. An illustration of these curves for the prospective test set over a twelve-month window is shown on Figure 8. The estimated area under curves are also shown in Table 5. Overall, the varying coefficients models (parametric-based and splines-based) perform better than the standard model with constant coefficients.

Figure 8: ROC curves.

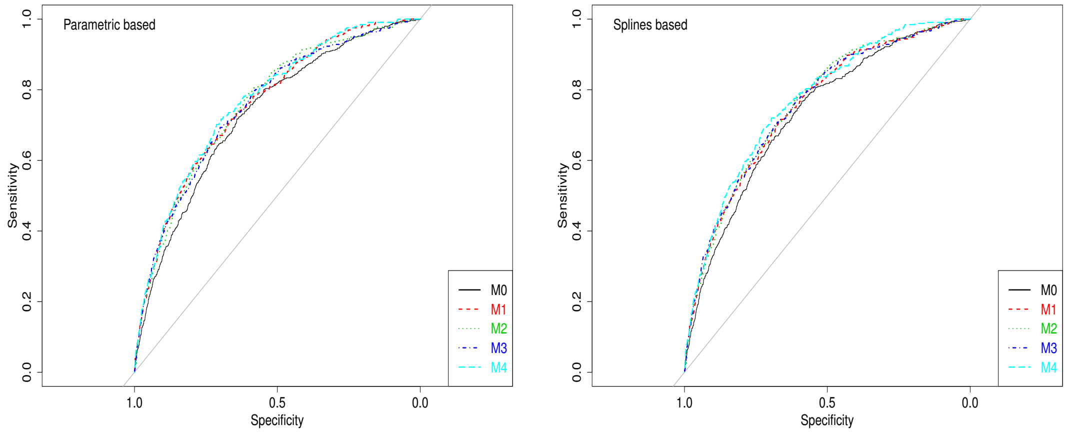


Table 5: Areas under the ROC curves.

Model code	M0	M1	M2	M3	M4
Parametric based	0.731	0.759	0.758	0.756	0.769
Splines based	0.731	0.753	0.758	0.756	0.770

Second, we consider the percentage of cases misclassified, weighted by the relative costs of each type of error. Lenders view good and bad cases very differently because the cost associated with the misclassification of a bad case is generally larger that that associated with the misclassification of a good case. To reflect this reality, we assign a cost of £0 to accounts that are correctly classified, £1 to a good account misclassified as bad, and a higher cost (for example £5 or £10) to a bad case misclassified as good. We use two alternative figures because there is very little published evidence on the true figures. A similar approach was used by Bellotti and Crook (2009) when assessing the importance of macroeconomic variables in dynamic models for credit risk. We then compare the total misclassification of each model.

For illustration, we focus on a 6-month horizon (although all the models implemented in this work allow one to compute the probability of default at any given time point). We consider the accounts still active at the end of the first year and compare the models in terms of their ability to predict the status of the account over the next six months horizon.

Since the outputs from the models are not the predicted statuses themselves, we first score our datasets with monthly probability; we then derive the six month survival and default probabilities, and then predict status according to some cut points. The cut points were estimated as the minimizer of the total cost based on the training set, separately for each model.

Table 6: Predicted mean cost for prospective and retrospective test sets.

		Cost=5		Cost=10	
		Parametric-based	Splines-based	Parametric-based	Splines-based
Prospective	M0	0.2917	0.2916	0.5081	0.5081
	M1	0.2901	0.2898	0.5047	0.5043
	M2	0.2912	0.2913	0.5082	0.5081
	M3	0.2892	0.2891	0.5052	0.5047
	M4	0.2876	0.2874	0.5015	0.5015
Retrospective	M0	0.2940	0.2877	0.5088	0.5080
	M1	0.2872	0.2870	0.5015	0.4996
	M2	0.2893	0.2883	0.5067	0.5069
	M3	0.2868	0.2874	0.5022	0.5009
	M4	0.2857	0.2851	0.4966	0.4970

Table 6 shows the predicted mean cost for each model on our two test sets when the cost of misclassifying a bad case is £5 and £10 respectively. A number of conclusions can be drawn from this table. For example, on both test sets and under the two cost scenarios, the models with varying coefficients consistently outperform models M0 (note that this good prediction performance of varying coefficients models can be improved further by dropping the weakest pseudo-variables from these models). Overall there is no clear winner between the parametric and splines formulations of varying coefficients regarding these predicted costs; but the spline formulation gives slightly higher predictive accuracy in most instances.

5 Concluding remarks

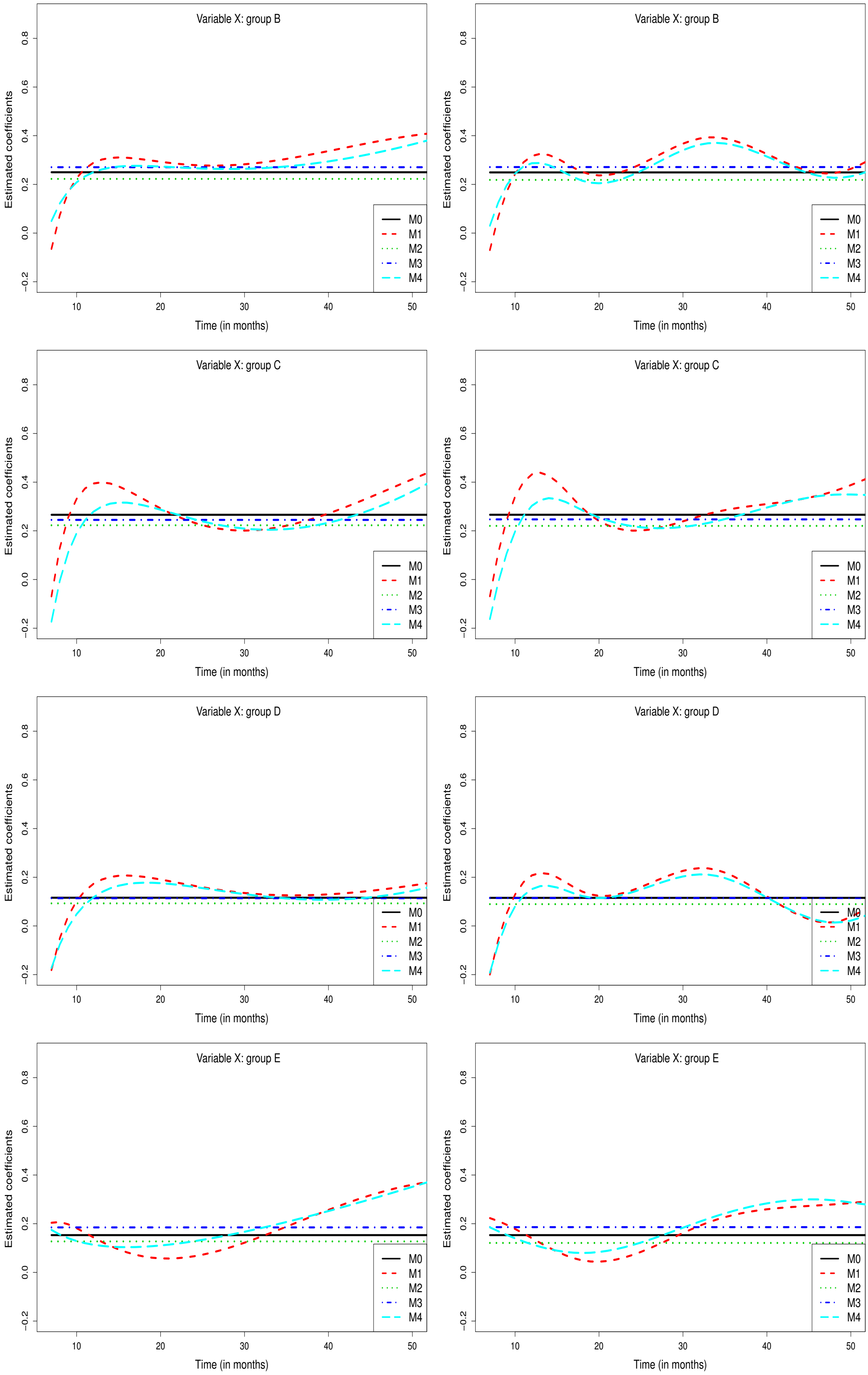
The main aim of this work was to investigate if and how patterns change in the effects of risk factors on the probability of default in retail banking. This has been achieved using time-varying coefficients survival models with application to a large portfolio of credit card loans from a major UK bank. We started by describing the framework of varying coefficients survival models with simple parametric specifications and a more flexible specification in terms of B-splines. We then fitted several models under each specification.

We found that (*i*) in terms of overall model quality and prediction accuracy, the varying coefficients models outperform standard survival models with constant coefficients. (*ii*) Using varying coefficients simultaneously on several risk factors can help to boost the overall goodness of fit and prediction accuracy. However this requires some care because of the risk of overfiting. Also, this does not translate systematically into better predictions when the model is used to score an independent dataset. (*iii*) In terms of overall model quality, the spline formulation of varying coefficients is to be preferred over their parametric counterpart.

In this work we have focussed on the importance of varying coefficients models for a single event, namely default. In practice however, the lender may want to predict the probability that an account would move from one stage of delinquency to another before eventually defaulting. It would be of interest to investigate the usefulness of varying coefficients in such a general multistate setting.

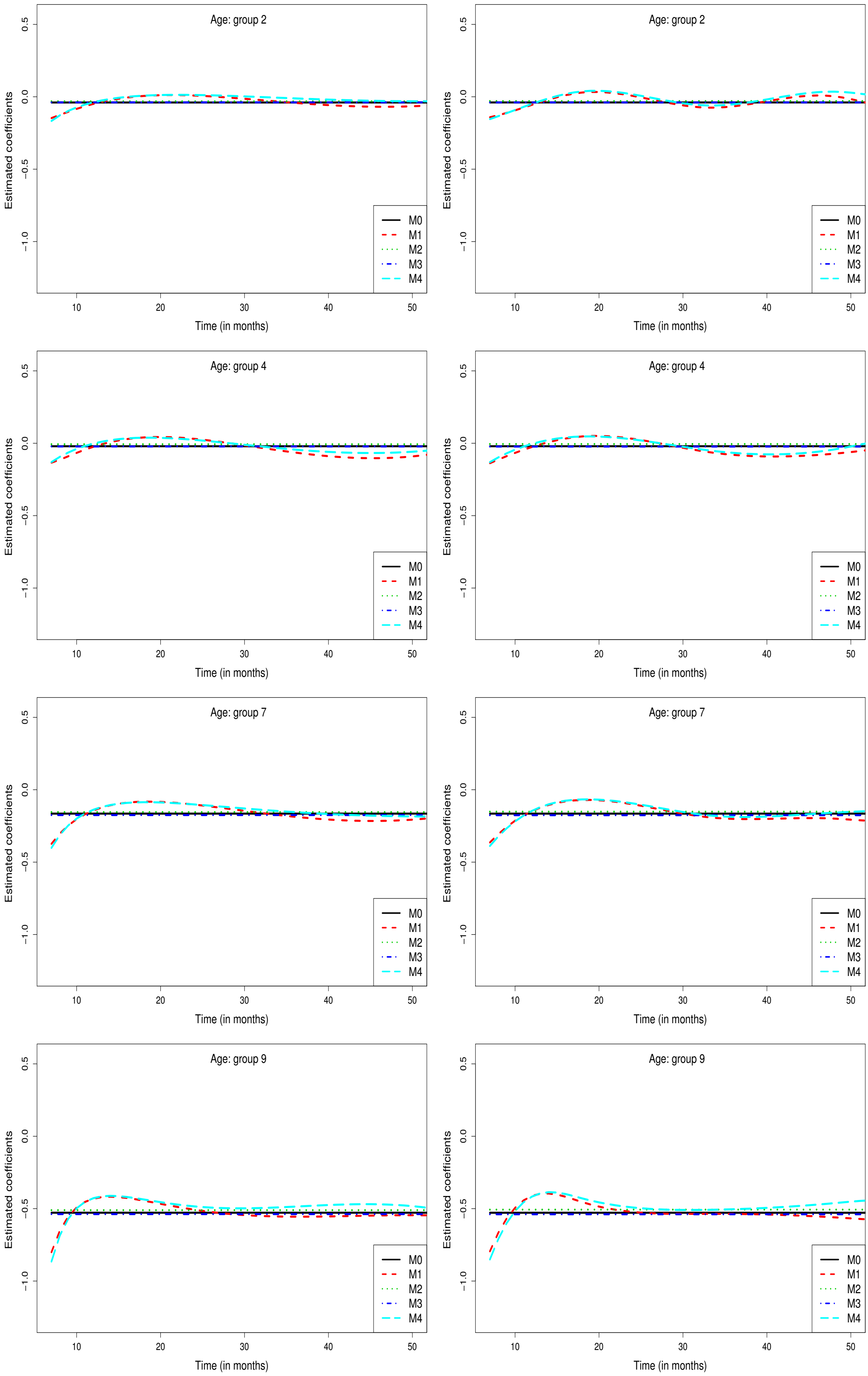
Appendix

Figure A1: Fitted coefficients for variable X.



Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines based.

Figure A2: Fitted coefficients for Age.



Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines based.

Bibliography

Allison P.D. (2010) *Survival analysis using SAS: A Practical Guide, Second Edition*. Cary, NC: SAS Institute Inc.

Altman D. G. and Stavola B. L. D. and Love S. B. and Stepniewska K. A. (1995) Review of survival analyses published in cancer journals. *British Journal of Cancer*, **72**, 511-518.

Andreeva G. (2006) European generic scoring models using survival analysis. *The Journal of Operational Research Society*, **57**, 1180-1187.

Banasik J. and Crook J. N. and Thomas L. C. (1999) Not if but when will borrowers default. *The Journal of Operational Research Society*, **50**, 1185-1190.

Bellotti T. and Crook J. (2013) Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, **29**, 563-574.

Bellotti T. and Crook J. (2014) Retail credit stress using discrete hazard models with macroeconomic factors. *The Journal of Operational Research Society*, **65**, 340-350.

Bellotti T. and Crook J. (2009) Credit Scoring with Macroeconomic Variables Using Survival Models. *The Journal of Operational Research Society*, **60**, 1699-1707.

Brown E. R. and Ibrahim J. G. and DeGruttola V. (2005) A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival. *Biometrics*, **61**, 64-73.

Brown I. and Mues C. (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, **39**, 3446–3453.

Cameron A. C. and Trivedi P. K. (2005) *Microeconometrics*. Oxford University Press.

Chawla N. V. and Bowyer K. W. and Hall L.O. and Kegelmeyer W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

Ciochetti, D. Deng, Y. and Gao, B.(2002) The termination of commercial mortgage contracts through pre-payment and default: a proportional hazards approach with competing risks. *Real Estate Economics*, **30(4)**, 595-633.

Collett D. (1993) *Modelling Survival Data in Medical Research.* Chapman and Hall.

Cox D. R. (1972) Regression models and life-tables (with discussion). *Journal of Royal Statistic Society, Series B*, **74**, 187-220.

Crook J. and Bellottia T. (2010) Time varying and dynamic models for default risk in consumer loans. *R. Statist. Soc. A*, **173**, 283-305.

De Boor C. (1978) *A practical guide to splines.* Springer.

Djeundje V. A. B. (2011) *Hierarchical and multidimensional smoothing with applications to longitudinal and mortality data.* Heriot-Watt University, United Kingdom.

Djeundje V. A. B. (2016) Systematic deviation in smooth mixed models for multilevel longitudinal data. *Statistical Methodology*, **32**, 203-217.

Djeundje V. A. B. and Crook J. (2018) Accounting for heterogeneity and macroeconomic condition when estimating transition probabilities in a portfolio of credit card loans. *European Journal of Operational Research*, **271**, 697-709.

Efron B. (1977) The efficiency of Cox's likelihood function for censored data *Journal of the American Statistical Association*, **76**, 312-319.

Eilers P. H. C. and Marx B. D. (1996) Flexible smoothing with B-splines and penalties *Statistical Science*, **11**, 89-121.

Eilers P. H. C. and Marx B. D. (2002) Generalized Linear Additive Smooth Structures *Journal of Computational and Graphical Statistics*, **11**, 758-783.

Eilers P. H. C. and Marx B. D. (2010) Splines, knots, and penalties *Computational Statistics*, **2**, 637-653.

Fan J. and Zhang W. (2008) Statistics and Its Interface *Statistics and Its Interface*, **1**, 179–195.

Ferguson C. A. and Bowman A. W. and Scott E. M. (2007) Model comparison for a complex ecological system *Journal of Royal Statistical Society - A*, **170**, 691–711.

Friedman J. H. and Silverman B. W. (1989) Flexible parsimonious smoothing and additive modelling *Technometrics*, **31**, 3-21.

Green P. J. and Silverman B. W. (1995) *Nonparametric regression and generalized linear models*. Chapman and Hall.

Hand D. J. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve *Machine Learning*, **77**, 103-123.

Hastie T. J. and Tibshirani R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hastie T. and Tibshirani R. (1993) Varying-coefficient models *Statistical Methodology - B*, **55**, 757–796.

Hougaard P. (2012) *Analysis of Multivariate Survival Data*. Springer.

Hwang R. C. (2012) A varying-coefficient default model *International Journal of Forecasting*, **28**, 675–688.

Kauermann G. and Tutz G. and Brüderl J.(2005) The survival of newly founded firms: a case-study into varying-coefficient models *Journal of Royal Statistical Society - A*, **168**, 145–158.

Leow M. and Crook J. (2014) Intensity models and transition probabilities for credit card loan. *European Journal of Operational Research*, **236**, 685-694.

Leow M. and Mues C. and Thomas L. (2011) Competing risk survival model for mortgage loans with simulated distributions. *Credit Scoring and Credit Control III, Edinburgh.*

Leow M. and Crook J. (2015) 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**, 457-64.

Luo S. and Kong X. and Nie T. (2016) Spline based survival model for credit risk modelling. *European Journal of Operational Research*, **253**, 869-879.

Park B. U. and Mammen M. and Lee Y. K. and Lee E. R. (2015) Varying Coefficient Regression Models: A Review and New Developments. *International Statistical Review*, **83**, 36–64.

Rue H. and Martino S. and Chopin N. (2009) Approximate bayesian inference for latent gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319-392.

Ruppert D. and Carroll R. J. and Wand M. P. (2003) Semiparametric Regression. *Cambridge University Press.*.

Stepanova M. and Thomas L. C. (2001) PHAB Scores: Proportional Hazards Analysis Behavioural Scores. *The Journal of the Operational Research Society*, **52**, 1007-1016.

Stepanova M. and Thomas L. C. (2002) Survival analysis for personal loan data. *The Journal of the Operational Research Society*, **50**, 277-289.

Wood S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC.

Wood S. N. (2016). *https://cran.r-project.org/web/packages/mgcv/mgcv.pdf*. Package mgcv.