Dynamic survival models for credit risks.

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

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, for example they allow the incorporation of time-varying factors which may be specific to an account or represent systemic factors. The literature shows that the parameters of such models changed from those before the financial crisis of 2008 to different values after the crisis.

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, and to do so in a way that is predictable. 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

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). A 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 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 developments can found in Fan and Zhang (2008), Ferguson et al. (2007), 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 an 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 has been previously assumed in discrete survival functions 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 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 outline 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 the applications and benefits of these models, and 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 (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 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 (Allison, 2010; Efrom, 1977). 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 Parametrisation of survival models for credit risk

In practice, credit risk data are usually discrete in time. If we 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, it can be shown (Allison, 2010) that model (3) simplifies to

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

where g is the complementary log function defined by $g(x) = \log[\log(x)]$, and $h_{0,\tau}$ is the transformed baseline. This discrete representation facilitates the implementation of survival models for credit risk data. In particular, if we replace the complementary log link 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 refereed to as application variables), time-dependent and account-dependent factors (often refereed 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

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

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 (2017). 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 at the beginning of month τ, 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 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 time-varying coefficient models.

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

$$g(q_{i,\tau}) = h_{0,\tau} + \boldsymbol{X}_{i,\tau}\, \boldsymbol{\beta}(\tau), \tag{8}$$

where the components of the joint parameter vector $\boldsymbol{\beta}(\tau)$ are allowed to vary over 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.

In general, the the effects of a covariate can vary not only over time but also according to some attributes or risk factor. In this case, the vector of coefficients $\boldsymbol{\beta}^{(1)}(\tau)$ takes the form $\boldsymbol{\beta}^{(1)}(C_{i,\tau})$, 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

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

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 $\boldsymbol{\beta}^{(1)}(\tau)$?

2.1 Parametric specification

A simple way to model the components of the varying coefficients $\boldsymbol{\beta}^{(1)}(\tau)$ is to consider a simple parametric shape. For example, assuming that they can be described by straight lines, they take 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 $X_{i,\tau,j}^{(1)}\,\beta_j^{(1)}(\tau) = a_j\,X_{i,\tau,j}^{(1)} \;+\; b_j\,\tau\,X_{i,\tau,j}^{(1)}$. In other words, 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. In general, each component of $\beta_j^{(1)}(\tau)$ can be expressed as a combination of simple standard functions. In this work, we consider the following family

$$\left\{1,\,\tau,\,\tau^2,\sqrt{\tau},\,\frac{1}{\tau},\,\log(\tau),\,[\log(\tau)]^2\right\}, \tag{11}$$

in which case $\beta_j^{(1)}(\tau)$ takes the form

$$\beta_j^{(1)}(\tau) = a_j \;+\; b_j\,\tau \;+\; c_j\,\tau^2 \;+\; d_j\,\sqrt{\tau} \;+\; \frac{e_j}{\tau} \;+\; f_j\,\log(\tau) \;+\; g_j\,[\log(\tau)]^2, \tag{12}$$

where $(a_j,\,b_j,\,c_j,\,d_j,\,e_j,\,f_j,\,g_j)$ are unknown parameters to be estimated. Similarly the baseline $h_{0,\tau}$ can be expressed as a linear combination of the family of basis functions in (11).

With this in place, all the regression coefficients can 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.

2.2 Flexible splines specification

Although the family (11) 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 (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).

An alternative approach is to express the baselines and varying coefficients using a bases of splines. Such bases have been used extensively in the literature to model complex variabilities; this includes radial basis, backward and forward truncated lines basis (Ruppert et al., 2003; Djeundje, 2016), as well as B-splines (Eilers and Marx, 1996; Brown et al., 2005). The latter boasts better numerical properties compared to other spline bases as discussed by Eilers and Marx (2010).

In terms of B-splines, the varying coefficients $\beta_j^{(1)}(\tau)$ are expressed as

$$\beta_j^{(1)}(\tau) = \sum_r \mathcal{B}_j(\tau)\phi_{j,r} \tag{13}$$

where $\mathcal{B}_j(\tau)$ are cubic B-spline functions at time point τ, and $\phi_{j,r}$ are unknown splines coefficients to be estimated. The baseline is expressed in a similar form (but with different coefficients) and all the parameters (including the splines coefficients) are jointly estimated by maximising the likelihood defined in (7).

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 2008 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)
	Employement type	Categorical (5 groups)
	Age at application	Categorical (10 groups)
Behavioural variables	Repayment amount	Continuous
	Prop one month delinquency	Continuous
	Prop one month delinquency	Continuous
Macroeconamic variables	Index of production	Continuous
	Consumer confidence	Continuous
	FTSE index	Continuous
	Unemployement 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 (2017).

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. Thus, several models were implemented, starting from the ones 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 (12) and next under the spline specification (13), 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 coefficients 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 on application variables
M2	model with varying coefficients for behavioral variables
M3	model with varying coefficients 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 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 (12) and (13) for the baseline only.

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 areas increases (see coefficients for *Prop one month delinquency* and *Prop one month delinquency*).

3.2.1 Exploring baselines

Figure 1 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 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

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 (11); the baseline in the second model is expressed in terms of B-splines.

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.

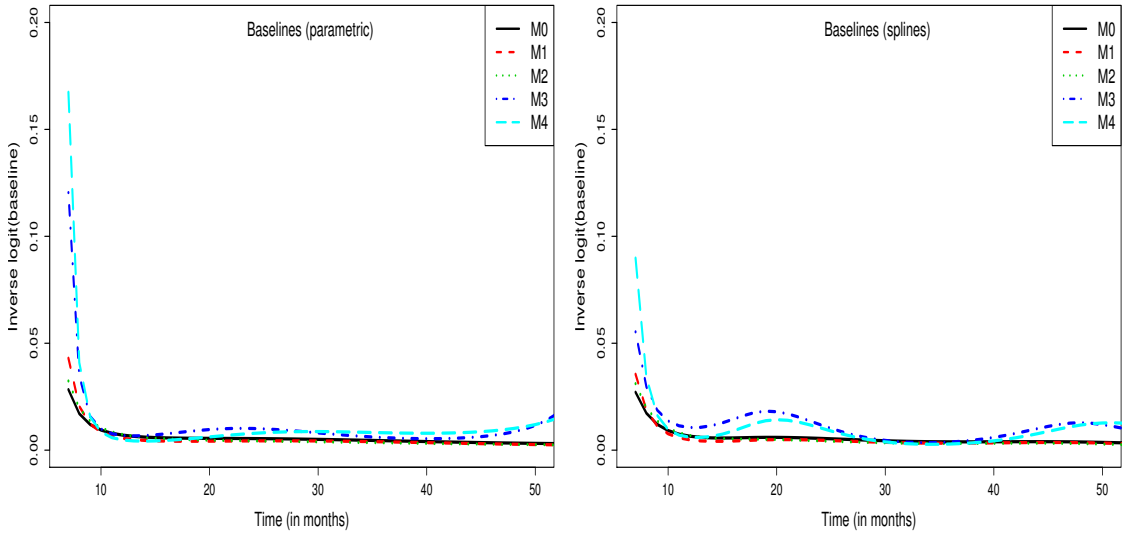
3.2.2 Exploring the relative effects of the application variables

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 describe 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 2 shows the fitted coefficients 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 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 (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.

Figure 1: Fitted baselines.



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

Similar comments apply to the other application variables. For example, Figure 3 reveals a quick change in the sign and magnitude of the varying coefficients associated with categories E and D of employment types. See also Figure 4 for the coefficients associated with *Variable X* as well as Figure 5 for Age. The latter seems to indicate that an assumption of constant coefficients for most ages bands looks reasonable. This can be tested formally as indicated in Section 2.1.

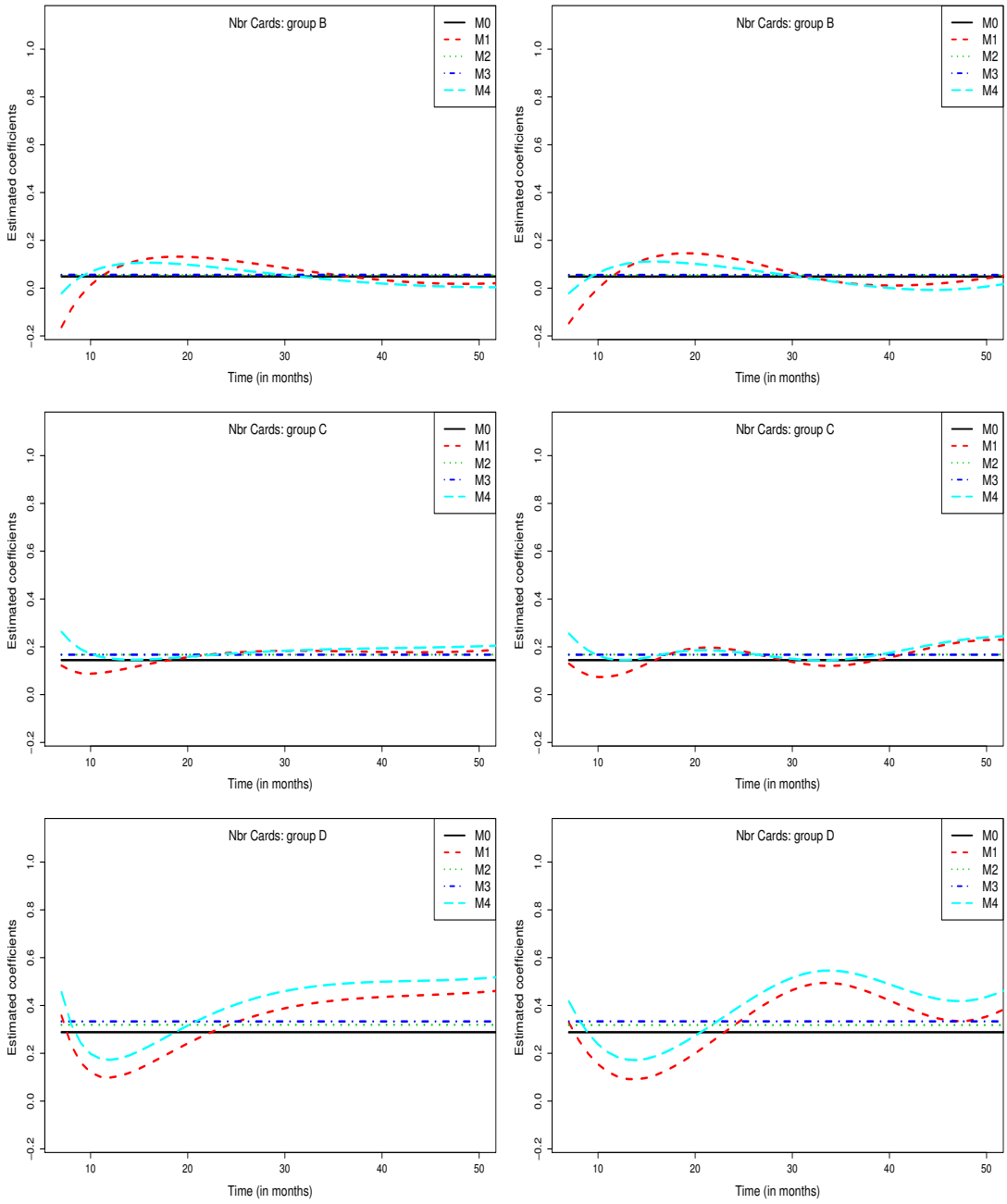
3.2.3 Exploring the relative effects of the behavioural variables

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 6. The coefficients associated with the variable *Repayment amount* are essentially the same (two top panels) regardless of the model considered. A similar remark applies to the variable called *Proportion of credit drawn*. However the four lower panels reveal that the varying coefficients associated with the other two behavioural variables increase monotonically with time.

3.2.4 Exploring the relative effects of the macroeconomic variables

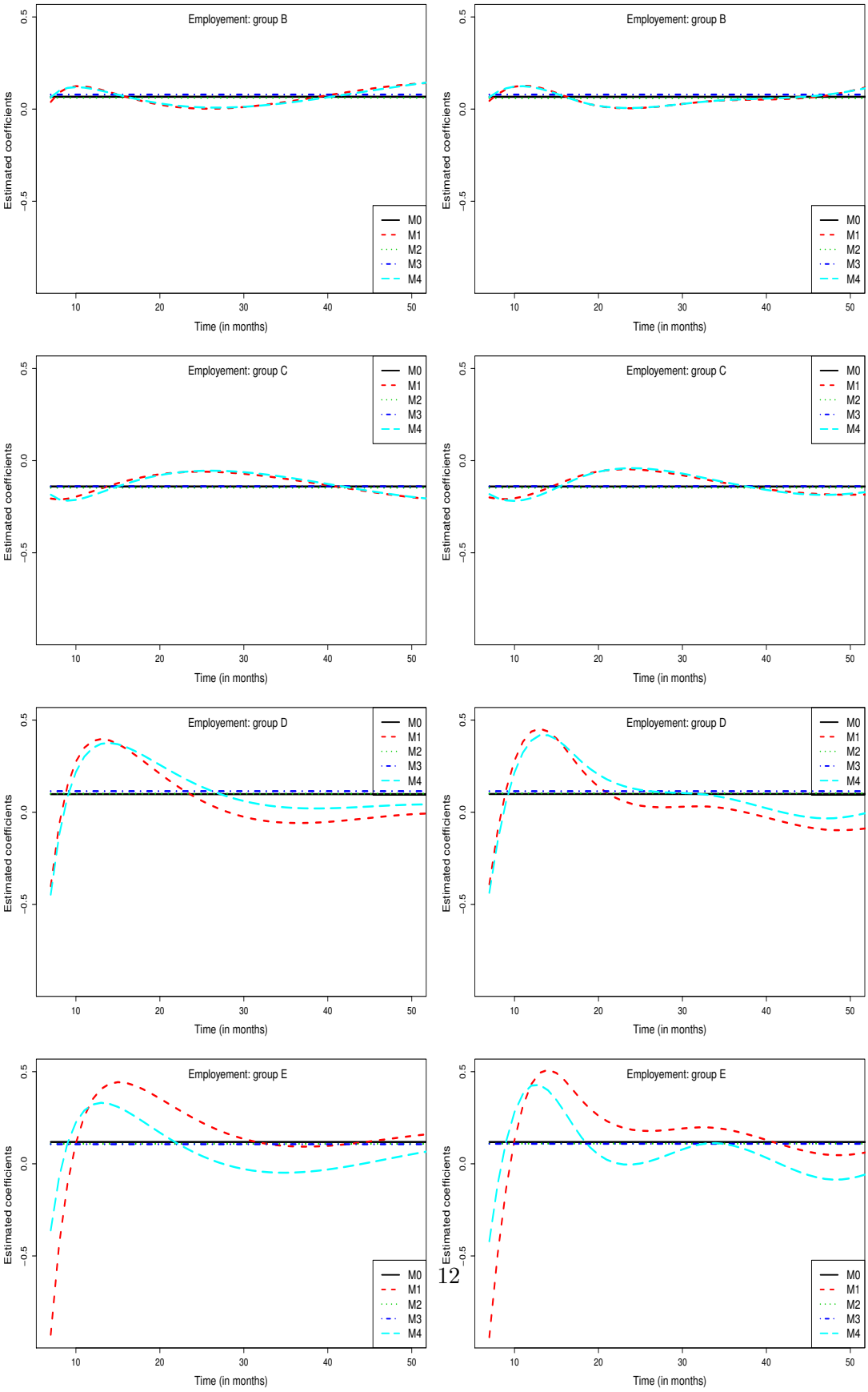
Finally we consider the effects of macroeconomic variables; the result is shown on Figure 7. Unlike the coefficients associated with the first three macroeconomic variables, we observe greater variability in the varying coefficient associated with *Unempleyement rate*. In particular, the effect of unemployment rate decreases steeply until month 15,

Figure 2: Fitted coefficients for number of cards.



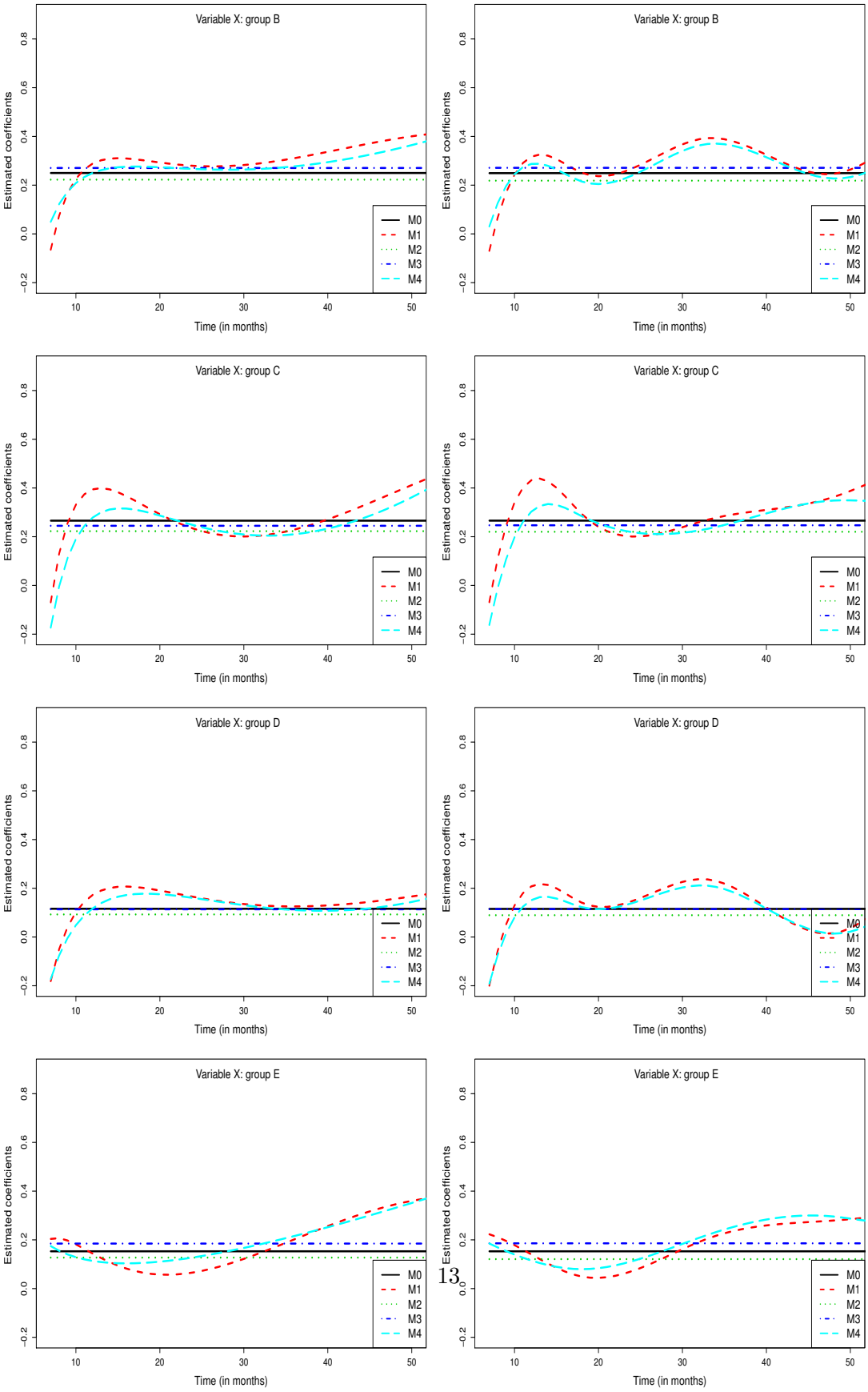
Left: baselines and time-varying coefficients are parametric based. Right: baselines and time-varying coefficients are splines based.

Figure 3: Fitted coefficients for employment type.



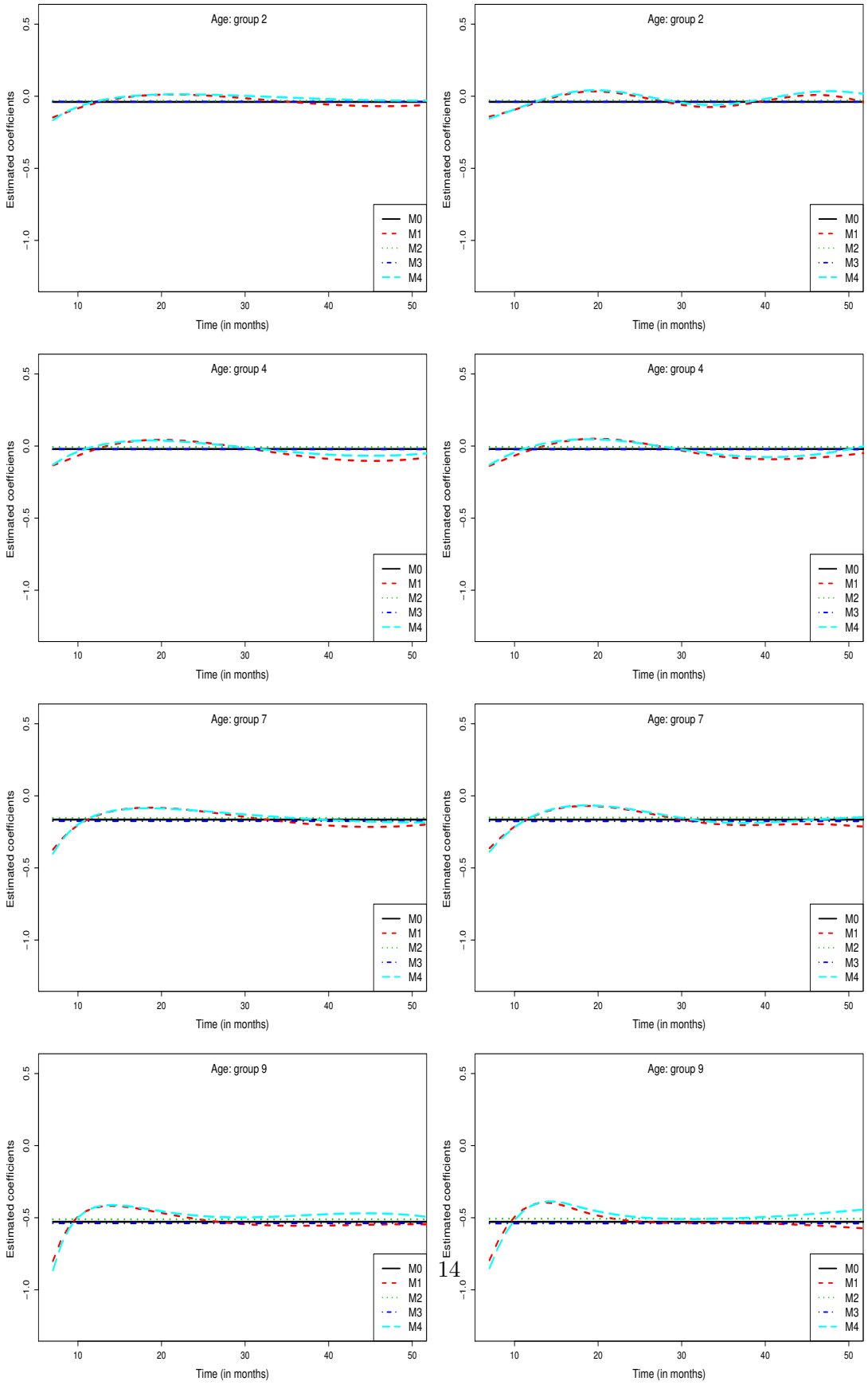
Left: baselines and time-varying coefficients are parametric based. Right: baselines and time-varying coefficients are splines based.

Figure 4: Fitted coefficients for variable X.



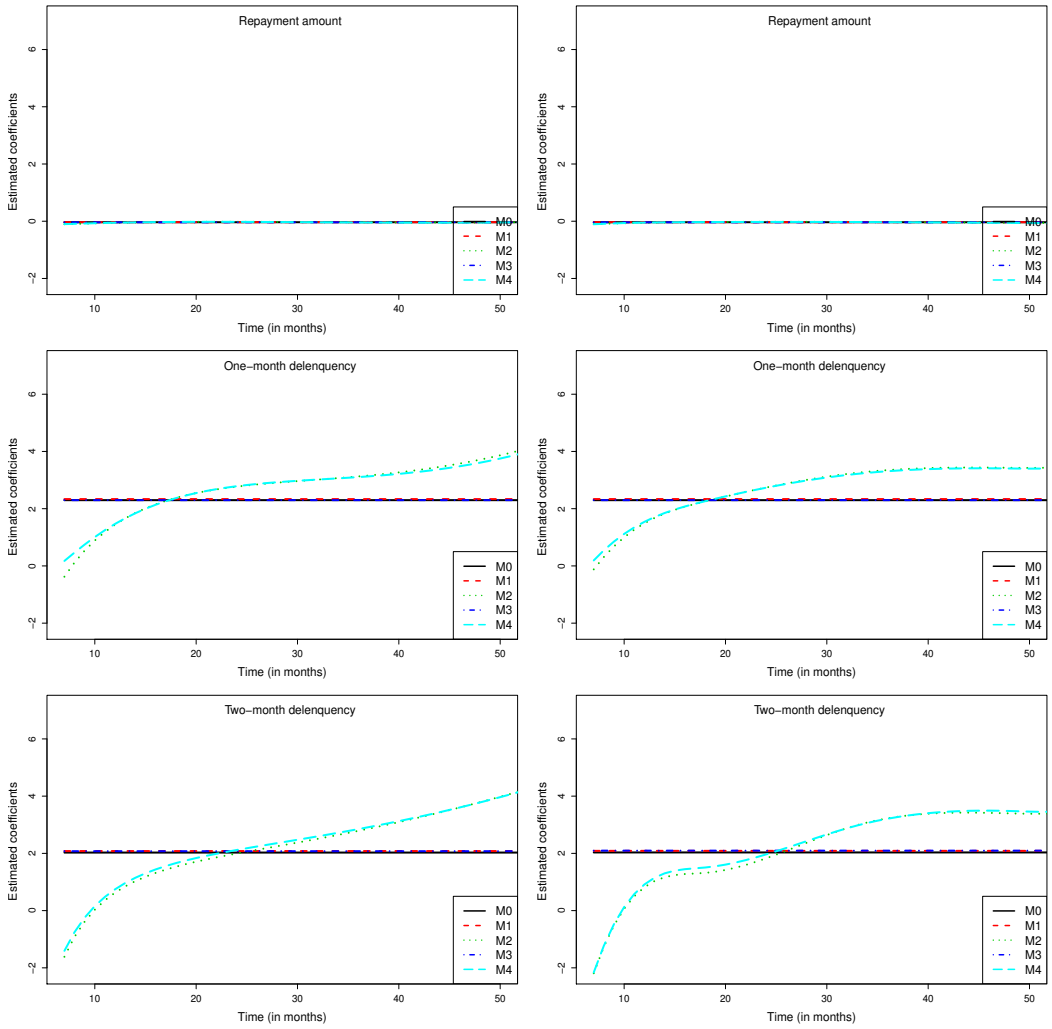
Left: baselines and varying coefficients are parametric based. Right: baselines and varying coefficients are splines based.

Figure 5: Fitted coefficients for Age.



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

Figure 6: Fitted coefficients for behavioural variables.



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

remains constant and then declines in both specifications.

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:

$$D_\tau = \pm 2 \left[O_\tau \times \log\left(\frac{O_\tau}{E_\tau}\right) + (N_\tau - O_\tau) \times \log\left(\frac{N_\tau - O_\tau}{N_\tau - E_\tau}\right) \right] \tag{14}$$

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 8. 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

$$AIC = -2\hat{\ell} + 2p \tag{15}$$

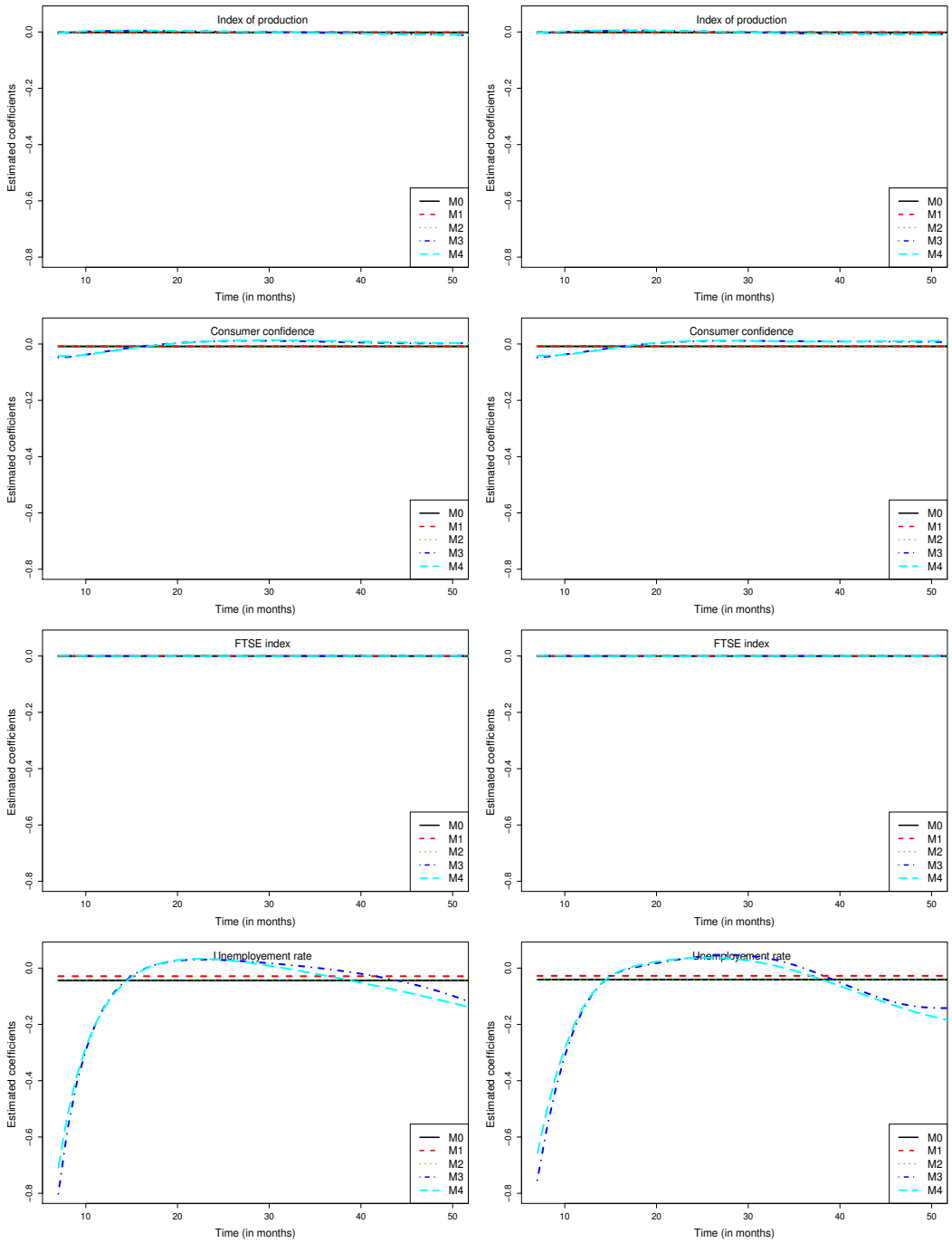
where p represents the number of parameters in the model, and $\hat{\ell}$ is the maximised value of the log-likelihood function. In general models with lower AIC would be preferred.

The AIC's 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 covariate 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.

4.3 Prediction performance

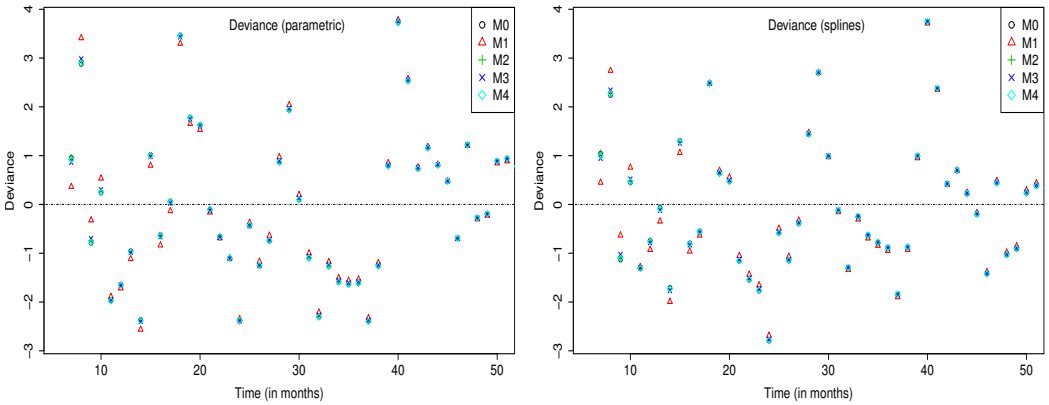
A simple way to assess and compare the prediction performance of classifiers is to look at the proportion of correctly classified cases. For typical credit risk data however, such an approach can be misleading due to the higher imbalance between good and bad

Figure 7: Fitted coefficients for macroeconomic variables.



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

Figure 8: Deviance residuals.



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

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

cases (Brown and Mues, 2012; Chawla et al., 2002). In the dataset used in this analysis for instance, the class of defaulters has far fewer cases in comparison to the class of non-defaulters.

In addition 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. A similar approach was used by Bellotti and Crook (2009) when assessing the importance of macroeconomic variables in dynamic models for credit risk.

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 status 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 5: 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 5 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 time-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 of interest to investigate the usefulness of varying coefficients in such a general multistate setting.

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.

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

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.

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.

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 multi-level longitudinal data. *Statistical Methodology*, **32**, 203-217.

Djeundje V. A. B. and Crook J. (2016) Accounting for heterogeneity and macroeconomic condition when estimating transition probabilities in a portfolio of credit card loans. *To appear*.

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. (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.

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.

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.