Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards

Citation for published version:
Djeundje Biatat, VA & Crook, J 2018, 'Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards', *European Journal of Operational Research*, vol. 271, no. 2, pp. 697-709. https://doi.org/10.1016/j.ejor.2018.05.040

Digital Object Identifier (DOI):
10.1016/j.ejor.2018.05.040

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
European Journal of Operational Research

Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards.

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, j.crook@ed.ac.uk}

*Corresponding author

Abstract

Multistate delinquency models model the probability that an credit account transits from one state of delinquency to another between any two points in the life of the account. Using a large sample of credit card accounts we parametrise such models with flexible baselines defined in terms of splines, and investigate whether predictive accuracy is enhanced by the incorporation of account specific random effects as well as the incorporation of macroeconomic variables. We conclude that macroeconomic variables are statistically significant in such models, that the inclusion of random effects renders some fixed effects less statistically significant but does not enhance predictive accuracy.

Keywords

OR in banking; credit scoring; multi-state models; intensity models; credit cards.

Introduction

Financial lenders use credit scoring models to help to assess whether to lend to a new applicant (application scoring) and to predict the probability of default by a borrower who already has a credit product (behavioural scoring). Traditional cross section credit scoring models have a number of limitations that are addressed by survival models. For example survival model give more information than cross sectional models such as the probability that an event will occur in the next time period conditional on it not having happened before whereas cross sectional models give a prediction that an event will occur any time within a predefined time window. But a single event survival model predicts the probability of only that event occurring. A lender may wish to have even more information, such as predictions of the probability that an account will move from one specific state of delinquency to another state (either towards further delinquency or towards being more up to date - cure) between any two time periods in the life of the account. This would enable more accurate assessments of risk and so more accurate assessments of the appropriate interest rate to charge. It would enable the lender to predict expected cash flow in each month during a loan more accurately and so gain a more accurate estimate of the expected profit from a loan. It would also enable a lender to predict when different collections policies may be beneficially implemented. At the level of a portfolio, one could simulate values of macroeconomic variables or use specified scenarios to predict the number of accounts expected to transit between states and so the liquidity and funding requirements for the portfolio. It will also enable the lender to predict the amount of provisions more accurately. Such models are known as multistate intensity (sometimes just intensity) models.

In this paper we make three contributions. First unlike previous literature we show the parameterisations of such multistate intensity models for consumer loans including macroeconomic covariates. Second, we show the results of including highly flexible functional forms for the baseline intensities, specifically we model them using B-spline functions. The use of highly flexible functional forms is important because the time dependent probabilities are largely, but not exclusively, driven by the baseline intensities. Third, our paper is the first to account for unobserved heterogeneity between accounts (account level random effects) in such models. We find that many macroeconomic factors significantly affect predicted transition probabilities and that the baseline intensities differ noticeably between the types of transitions an account may experience. We also find that our models give reasonably high levels of predictive accuracy but that the inclusion of account specific random effects does not enhance the accuracy of the predictions.

There is quite a large literature on the parameterisation of survival distributions for consumer loan defaults (for example Banasik et al. 1999; Stepanova and Thomas 2002, 2001; Bellotti and Crook 2009, 2012, 2013). There is relatively little literature on multistate intensity models for any type of loans and most has concentrated on modelling ratings grade transitions for corporate debt and bonds. Two methodologies can be observed. First, the estimation of survival models for time to transit for each possible combination of states, the subsequent estimation of a generator matrix of integrated intensities and finally the estimation of the probability of a transition between any two states between any two time periods for any case using the product integral (see Andersen et al 1993 and Aalen et al 2008). A second method is to estimate ordered polytomous models with each state being observed in each time period or point in time; see Gagliardini and Gourieroux (2005). Examples of use of the first method include Jarrow et al. (1997) who estimated a transitions matrix between corporate bond ratings without covariates. Lando and Skødeberg (2002) estimate time non-homogeneous transition probabilities in continuous time in terms of a time varying covariates representing whether the last transition was an upgrade. Figlewski et al. (2012) estimate three ratings transitions between investment grade, speculative grade and default for corporate bonds using bond specific time varying covariates and macroeconomic variables. None of these papers attempts to make predictions and they omit any unobserved heterogeneity either over time or between observations; they include only observed covariates. Koopman et al. (2008) and Koopman et al. (2009) use the same methodology but do include time varying random effects in the first paper and indicate predictions but with no observables. In the second paper they include observables but without making predictions. The second method was employed by Gagliardini and Gourieroux (2005) and Creal et al. (2014). Gagliardini and Gourieroux also included unobserved heterogeneity and modelled corporate transitions using an ordered probit model with three unobserved factors. They indicated predictions though the accuracy of the predictions was not assessed. Creal et al used ordered logits with frailty to predict corporate ratings transitions, but these were not functions of duration time and predicitive accuracy was not assessed.

The only published multistate model parameterisations for retail loans is for credit cards by Leow and Crook (2014). They use the first methodology. This work however omits unobserved heterogeneity between borrowers and also omits macroeconomic variables yet in survival models there is ample evidence that for corporates transition probabilities depend on such variables (Figlewski et al., 2012; Lando and Skødeberg, 2002; Koopman et al., 2009) and papers using survival models for consumer loans also have the same finding (Bellotti and Crook 2009, 2012, 2013). The inclusion of random effects is important because if there are omitted variables that affect the hazard function, the estimated parameters of that function may be biased (see Cameron and Trevedi 2005).

The paper is organised as follows. Section 1 describes the modelling framework. Section 2 explores the output from the model based on a large dataset of individual card accounts from a major UK bank. Some discussion and concluding remark follow in Section 3.

1 Methodology

There exists a substantial literature on the incorporation of random effects into survival models and general competing risk models. See for example Andersen et al. (1993, Chap 9) or Parner (1997) on how to extend intensity models with random effects based on the theory of counting process. One characteristic of most credit risk datasets is that they are discrete in time (accounts are observed monthly), and this allows the possibility to model transitions between states using multinomial-type regressions (Enberg et al., 1990; Steele et al., 1996, 2004; Goldstein et al., 2004).

However, these models are often problematic, (especially in complex scenarios involving repeated episodes within individuals where there are multiple types of events which may vary across states over time which is the case here). A major obstacle lies in the implementation due to the intractability of the likelihood function. In practice, various approximations are used, but the Bayesian Markov Chain Monte Carlo (MCMC) approach has become a prominent method for implementing these models especially in the presence of random effects and recurrent events; see for example Gasbarra and Karia (2000), Steele et al. (2004), Kyung et al. (2010), Sen et al. (2010) among others.

One challenge of the MCMC method is its computational cost, especially for complex models involving a substantial number of parameters in the presence of a large training dataset. For example, the dataset that motivated this work gives rise to more that three millions months-exposure. An early investigation of fitting some competing risk models with random effects to this dataset using a MCMC method turned out to be very time-consuming to run, making such an approach impractical. This computational challenge worsens when one tries to allow for more flexibility by incorporating spline bases into the model. In this paper, we use a pragmatic approach that permits flexibility and allows one to account for heterogeneity. Our method is based on the marginal Bernoulli processes associated with the transition types.

Consider a portfolio of n credit card accounts. A number of states are defined and transition from a given state h at time point t to state j at time $(t+1)$ are driven by the characteristics of each individual account. We will denote by $\mathcal{S}$ the set of all permissible pairs (h, j).

To these transitions, let us associate the individual random processes, Y_{ihj}, $i \in \{1, ..., n\}$, $(h, j) \in \mathcal{S}$, $h \neq j$, defined by

$$Y_{ihj}(t) = \begin{cases} 1 & \text{if account } i \text{ is in state } j \text{ at time } t, \text{ given that it was at } h \text{ at } (t-1) \\ 0 & \text{if account } i \text{ is in state } h \text{ at time } t, \text{ given that it was at } h \text{ at } (t-1) \end{cases} \tag{1}$$

That is, the random variables $Y_{ihj}(t)$ take value 1 if account i moves from state h at time $(t-1)$ to state j at time t, and 0 if account i remains at state h at time t. Note that if there are no directional constraints, account i in state h at time $(t-1)$ can move into state $j' \neq j$. In this case $Y_{ihj}(t)$ is undefined; that is, when computing the marginal likelihood associated with the process Y_{ihj}, account i is interval-censored from time $(t-1)$ to t; we assume that censoring is non-informative.

These individual random processes are associated with individual transition probabilities which we denote by q_{ihj} such that

$$\begin{cases} Pr\{Y_{ihj}(t) = 1\} & = & q_{ihj}(t) \\ Pr\{Y_{ihj}(t) = 0\} & = & 1 - q_{ihj}(t), \end{cases} \tag{2}$$

with $i \in \{1, ..., n\}$, $(h, j) \in \mathcal{S}$, $h \neq j$.

In other words, $q_{ihj}(t)$ represents the probability that account i in state h at time $(t-1)$ will move into state j at time t, assuming that only transition to state j can be undertaken by this account at this time. Thus, these probabilities assumes that each transition type operates in isolation and therefore ignore the competing impact of other transition types. We shall describe how to derive the competing transition probabilities in Section 1.3.

The magnitude of these probabilities varies from account to account depending on the characteristics of each account holder as well as the state of the economy. We will denote by $\boldsymbol{X}_{ihj}(t)$ the vector of covariates on subjects i at time t; the type and number of covariates can differ between transition types. This comprises time-independent covariates (ie application variables) as well as those that change with time such as behavioural variables and macroeconomic variables. Credit risk models are designed with prediction in mind and therefore, the time-dependant covariates are often lagged.

1.1 Model specification

A common way to express the dependence of the transition probabilities on the covariates can be formulated as follows

$$q_{ihj}(t) = \mathcal{F}_{hj}(\, \alpha_{hj}(t) + \boldsymbol{\beta}_{hj}^{T} \boldsymbol{X}_{ihj}(t)\,) \tag{3}$$

where α_{hj} is a baseline function corresponding to transitions from stage h to stage j, $\boldsymbol{\beta}_{hj}$ are unknown vectors of coefficients, $\mathcal{F}_{hj}$ are one-to-one link functions. When fitting the models, we use logit links, ie $\mathcal{F}_{hj}(x) = 1/(1 + e^{-x})$.

However, formulation (3) assumes that two accounts with the same values of the covariates would have identical transition probabilities. This is a strong assumption because accounts' holders differ in so many ways that no set of measured covariates can fully capture all the variations among them (Collett, 1993; Allison, 2010). Additionally, it is very likely that some factors influencing transition intensities cannot be measured. Ignoring the impact of such unobserved factors can attenuate the estimates of the observed covariate effects; see for example Therneau and Grambsch (2000). Furthermore, in the present paper where each account can experience more than one transition of the same type, assuming model (3) ignores some dependence among the observations and can lead to biased estimates of standard errors and hypothesis tests.

To circumvent the impact of hidden variations and dependence on the transition probabilities, Vaupel et al. (1979) introduced the so-called frailty or random effects into standard survival models. Thus, we extend model (3) to

$$q_{ihj}(t) = \mathcal{F}_{hj}(\, \alpha_{hj}(t) + \boldsymbol{\beta}_{hj}^T \boldsymbol{X}_{ihj}(t) + u_{ihj}\,) \tag{4}$$

where u_{ihj} represents the random effect associated with accounts i during transitions from state h to state j. These u_{ihj} allow one to account for dependence between jumps undertaken by the same account and help to attenuate the impact of unobserved covariates.

Our formulation in (4) is flexible in the sense that it allows different random effects between and within transition types (although it can be extended further by allowing the random effects to be a function of time). For identifiability reasons however, some constraints must be placed on these random effects (Hoem, 1990). Hence, setting $\boldsymbol{u}_i = (u_{ihj})$, $(h, j) \in \mathcal{S}$, we assume that the $\boldsymbol{u}_i$ are i.i.d. according to the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0},\, \boldsymbol{\Phi}) \tag{5}$$

This choice is consistent with what is widely used in the literature; see for example Ripatti and Palmgren (2000) or Hougaard (2000) among others. For computational reasons we consider a diagonal covariance matrix in our application; some benefits of this simplification are discussed in Section 1.2.

In models (3) and (4), the baseline functions $\alpha_{hj}(t)$ are yet to be specified. One possibility is to assume some parametric form; see for example Bellotti and Crook (2013) or Leow and Crook (2015). However, parametric functions are usually not flexible enough to capture unanticipated or hidden patterns in the data (Ruppert et al., 2009; Djeundje, 2016). A flexible alternative is to model the baselines using spline functions.

Many types of splines are available in the literature including truncated polynomial, radial basis, B-splines, etc. In this work, we use B-splines, one reason being that they have compact supports and these yield better numerical properties compared to other spline bases. Additional benefits arising from using B-splines can be found in Eilers and Marx (2010) or Djeundje (2011). Thus, we express the baselines as

$$\alpha_{hj}(t) = \sum_{r=1}^{c} B_r(t)\, a_{hj,r} \tag{6}$$

where $B_l(t)$ are B-spline basis functions at points t, and $\boldsymbol{a}_{hj} = (a_{hj,1}, ..., a_{hj,c})$ is a vector of unknown coefficients to be estimated.

1.2 Parameter estimation

We now turn to the estimation. We want to estimate the regression parameters $\boldsymbol{\beta}_{hj}$, the baseline spline coefficients $\boldsymbol{a}_{hj} = (a_{hj,1}, ..., a_{hj,c})$, and the covariance matrix $\boldsymbol{\Phi}$. A standard way to perform this estimation is marginal likelihoods (Pinheiro and Bates, 1995; Searle et al., 2006).

Let us consider the multivariate Bernoulli process $\boldsymbol{Y} = \{Y_{ihj},\ (h,j) \in \mathcal{S},\ i = 1, ..., n\}$, and the joint vector of random effects $\boldsymbol{u} = \{u_{i,hj},\ \ (h,j) \in \mathcal{S},\ i = 1, ..., n\}$. Also, denote by $\boldsymbol{\beta}$ the joint vector of parameters $\boldsymbol{\beta}_{hj}$, and by $\boldsymbol{a}$ the joint vector of spline coefficients $\boldsymbol{a}_{hj}$. The joint likelihood of $(\boldsymbol{Y}, \boldsymbol{u})$ which we denote by $L_{(\boldsymbol{Y},\boldsymbol{u})}$, can be expressed as

$$L_{(\boldsymbol{Y},\boldsymbol{u})}\ (\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\Phi}) = L_{\boldsymbol{Y}|\boldsymbol{u}}\ (\boldsymbol{\beta}, \boldsymbol{a}) \times g_{\boldsymbol{u}}(\boldsymbol{\Phi}) \tag{7}$$

where $g_{\boldsymbol{u}}$ denotes the multivariate normal density given by

$$g_{\boldsymbol{u}}(\boldsymbol{\Phi}) \propto |\boldsymbol{\Phi}|^{-0.5n} \exp\left(-\frac{1}{2} \sum_i \boldsymbol{u}_i'\, \boldsymbol{\Phi}^{-1}\, \boldsymbol{u}_i\right), \tag{8}$$

and $L_{\boldsymbol{Y}|\boldsymbol{u}}$ represents the likelihood of $\boldsymbol{Y}$ conditional on the random effects $\boldsymbol{u}$:

$$L_{\boldsymbol{Y}|\boldsymbol{u}}\ (\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\Phi}) = \prod_{(h,j)\in\mathcal{S}} \prod_{t} \prod_{i\in\mathcal{R}_{hj}(t)} [q_{ihj}(t)]^{y_{ihj}(t)} \times [1 - q_{ihj}(t)]^{1-y_{ihj}(t)}. \tag{9}$$

In this representation, $\mathcal{R}_{hj}(t)$ represents the risk set for transitions from state h to state j at time t. At each time point t, accounts that transit from state h to states $k \neq j$ are censored and therefore excluded from the risk set $\mathcal{R}_{hj}(t)$.

The marginal likelihood, $L_{\boldsymbol{Y}}$, is obtained by averaging out the random effects from the joint likelihood (7):

$$L_{\boldsymbol{Y}}(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\Phi}) = \int L_{(\boldsymbol{Y},\boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{\Phi})\, d\boldsymbol{u}\cdot \tag{10}$$

The estimates of the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{a}$, as well as the covariance matrix $\boldsymbol{\Phi}$, are found by maximising the marginal likelihood (10). However it is important to point out that the integral defining this marginal likelihood is usually not available in a closed form. In practice, some restrictions are often placed on the structure of the covariance matrix $\boldsymbol{\Phi}$, and the integral is then approximated using a numerical method. One of the best known approximation methods in this context being the so-called Adaptive Gaussian Quadrature described by Pinheiro and Bates (1995). This method is available in many Statistical packages including R, Matlab and SAS.

The model presented in this paper has been fitted by maximising the marginal likelihood (10). In particular, when $\boldsymbol{\Phi}$ is a diagonal matrix, this marginal likelihood factors into a product of marginal likelihoods, one for each transition type. In such a case, the parameters are estimated separately for each transition type thereby by maximising the relevant component of marginal likelihood.

1.3 Deriving competing transition probabilities

The transition probabilities q_{ihj} in equation (2) and Section 1.1 are non-competing probabilities in the sense that they represent the probability that account i in state h at time $(t-1)$ will move into state j at time t assuming that only transition to state j can be undertaken by that account at this time. In other words, by assuming that each transition type operates in isolation, these probabilities ignore the competing aspect of other transition types.

Let us now denote by $\tilde{q}_{ihj}$ the competing probabilities; that is $\tilde{q}_{ihj}(t)$ represents probability that account i in state h at time $(t-1)$ will move into state j at time t, in the presence of all other transition types (i.e. while competing with other transition types). One way to obtain the competing transition probabilities $\tilde{q}_{ihj}(t)$ is to first estimate the underlying transition intensities, to form a generator matrix, and then compute the competing transition probabilities via the product integral (Leow and Crook, 2014; Lando and Skødeberg, 2002).

Alternatively, the competing transition probabilities $\tilde{q}_{ihj}$ can be derived directly from the transition probabilities q_{ihj}. Such a derivation is common in Actuarial Mathematics for life contingent risks. Specifically, if we assume that the non-competitive transitions occur uniformly over each month, it can be shown (see for example Luptakova and Bilikova (2014), Promislow (2006) and Dickson et al. (2009) among others) that the relationship between competing and non-competing transition probabilities is as follows:

$$\begin{aligned}\tilde{q}_{ihj}(t) = q_{ihj}(t) \times \Bigg(1 \; & - \; \frac{1}{2} \sum_{\substack{k \neq j; \\ where \\ (h,k) \in \mathcal{S}}} q_{ihk}(t) \\ & + \; \frac{1}{3} \sum_{\substack{k \neq j \neq r \\ where \\ (h,k) \in \mathcal{S} \\ (h,r) \in \mathcal{S}}} q_{ihk}(t)\, q_{ihr}(t) \\ & - \; \frac{1}{4} \sum_{\substack{k \neq r \neq s \neq j \\ where \\ (h,k) \in \mathcal{S} \\ (h,r) \in \mathcal{S} \\ (h,s) \in \mathcal{S}}} q_{ihk}(t)\, q_{ihr}(t)\, q_{ihs}(t) \\ & + \; \cdots\cdots\cdots\cdots \Bigg)\end{aligned} \tag{11}$$

In particular for the credit data and states defined in Section 2.1 below, formula (11) implies that the competing transition probabilities between states from time points $(t-1)$ to the next time point t are given by

$$\begin{cases} \tilde{q}_{i01}(t) &= q_{i01}(t) \\ \tilde{q}_{i10}(t) &= q_{i10}(t)\left(1-\frac{1}{2}q_{i12}(t)\right) \\ \tilde{q}_{i12}(t) &= q_{i12}(t)\left(1-\frac{1}{2}q_{i10}(t)\right) \\ \tilde{q}_{i20}(t) &= q_{i20}(t)\left(1-\frac{1}{2}(q_{i21}(t)+q_{i23}(t))+\frac{1}{3}q_{i21}(t)\,q_{i23}(t)\right) \\ \tilde{q}_{i21}(t) &= q_{i21}(t)\left(1-\frac{1}{2}(q_{i20}(t)+q_{i23}(t))+\frac{1}{3}q_{i20}(t)\,q_{i23}(t)\right) \\ \tilde{q}_{i23}(t) &= q_{i23}(t)\left(1-\frac{1}{2}(q_{i20}(t)+q_{i21}(t))+\frac{1}{3}q_{i20}(t)\,q_{i21}(t)\right) \end{cases} \tag{12}$$

Thus, the predicted transition probability matrices, $\tilde{\boldsymbol{P}}_i(t)$, are constructed as follows:

$$\tilde{\boldsymbol{P}}_i(t) = \begin{bmatrix} (1-\tilde{q}_{i01}(t)) & \tilde{q}_{i01}(t) & 0 & 0 \\ \tilde{q}_{i10}(t) & (1-\tilde{q}_{i10}(t)-\tilde{q}_{i12}(t)) & \tilde{q}_{i12}(t) & 0 \\ \tilde{q}_{i20}(t) & \tilde{q}_{i21}(t) & (1-\tilde{q}_{i20}(t)-\tilde{q}_{i21}(t)-\tilde{q}_{i23}(t)) & \tilde{q}_{i23}(t) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

These probability matrices can be used to explore various scenarios. For instance, the probabilities that account i in a given state $\delta_i(t_1)$ at time point t_1, will find itself in state 0, 1, 2, or 3 (respectively) at a latter time t_2, are given by the elements of the vector $\boldsymbol{\mu}_i(t_2)$ defined by the following matrix product

$$\boldsymbol{\mu}_i(t_2) = [\mathbb{1}_{\{\delta_i(t_1)=0\}},\ \mathbb{1}_{\{\delta_i(t_1)=1\}},\ \mathbb{1}_{\{\delta_i(t_1)=2\}},\ \mathbb{1}_{\{\delta_i(t_1)=3\}}]\ \tilde{\boldsymbol{P}}_i(t_1,t_2) \tag{14}$$

where $\mathbb{1}$ denotes the standard indicator operator, and $\tilde{\boldsymbol{P}}_i(t_1,t_2)$ represents the cumulative transition probability matrix defined by

$$\tilde{\boldsymbol{P}}_i(t_1,t_2) = \prod_{t=t_1+1}^{t_2} \tilde{\boldsymbol{P}}_i(t) \tag{15}$$

2 Application

2.1 Data and states definition

The data used for illustration is from a portfolio of credit card loans supplied by a major UK bank. This dataset of more than 35000 individual accounts is a random sample of credit card accounts which were accepted onto the books between 2005 and 2010, and observed monthly up to the first quarter of 2011. Some of the data have already been used by Leow and Crook (2014).

The dataset comprises both application variables (e.g. length of time at address, income and employment code) as well as behavioural variables collected at monthly time points (credit limit, repayment amount). In addition, macroeconomic variables (e.g. unemployment rate, credit card interest rate) were appended to the dataset. The variables used in this paper are listed in Table 1.

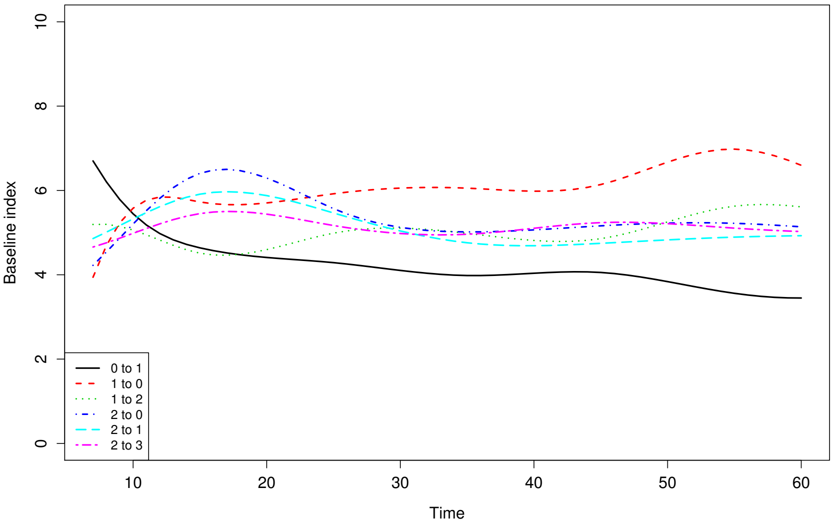
As in Leow and Crook (2014), we define 4 states: up-to-date (state 0), one month in arrears (state 1), two months in arrears (state 2) and default (state 3), where movements between the states depend on whether the borrower makes the minimum repayment for that month. The rules for transition between states remain as in Leow and Crook. These are as follows. All accounts start in state 0 that is up to date with repayments. If at any time during the observation period the repayment amount made is less than the minimum required the borrower advances to the next immediate state. A borrower who has missed a repayment before and is in states 1 or 2 but makes a repayment of some amount in the following month(s) will (a) remain in that state if the repayment made is greater than the minimum required but less than the sum of the amounts required in the current and previous month or (b) be moved to a one lower state if the repayment made exceeds the sum of the minimum required in the current and previous months but is less than the outstanding amount.

We fit the model using accounts that were opened before 2009; there are about 30000 such accounts. Accounts that were opened from January 2009 make up the independent subset with about 10000 unique accounts, and this subset would be used to explore predictions.

2.2 Parameter estimates

The baseline spline coefficients were jointly estimated together with the regression parameters as described in the Section 1.2. These estimated spline coefficients were then used to compute the baseline for each transition type via formula (6).

Figure 1: Fitted smooth baselines using B-splines.



In this section, we discuss the parameter estimates, baselines and random effects from the fitted model.

Baselines

The resulting baselines are displayed on Figure 1. The scale of the vertical axis is indexed for commercial confidentiality reasons. We note that the shapes of the baselines are quite versatile and vary from one transition type to another. The extraction of such flexible patterns has been made possible by the use of spline basis functions.

Focussing on the transitions from state 0 to state 1, its baseline indicates a higher chance of transition toward delinquency in the few early months after the account has been opened. However this chance decreases sharply and gradually tends to stabilise. Conversely, the graphic indicates that accounts in state 1 are more likely to recover than to move further toward delinquency except perhaps in the very early few months. It can also been noticed that accounts in state 2 are more likely to move toward recovery for the first part of the lifespan but become equally or more likely to default in latter years.

However, it is important to bear in mind that any isolated interpretation of the baselines must be undertaken with caution because such interpretation assumes that all covariates in the model are set to 0 (for continues covariates) and to the reference category (for categorical variables).

Regression coefficients

We now look at the relevance of the covariates. For comparison purposes, the same set of covariates was fed into each of the six sub-models. The fitted regression coefficients together with their relative significance are displayed in Table 1 for models including random effects and Table 2 for models without random effects. Starting with the first, most of the variables have the expected signs. For example older applicants have a lower probability of transiting from state 0 to state 1 whilst those with a higher credit limit or a higher proportion of credit limit drawn have a higher probability, and in most cases the opposite sign is observed for the reverse transition from state 1 to state 0. Older borrowers, those who had been with the bank longer, those with a higher credit limit and those with less history of improvement have a lower probability of transiting from one behind to two behind. On the other hand, those with a higher repayment amount and lower proportion of their limit drawn are more likely to recover from two behind to being up to date. The longer the applicant was at their address or with the bank the lower the chance of moving from two behind into default. Interestingly, the higher the credit limit and the proportion of the limit drawn, the lower the chance of moving into default.

Turning to the macroeconomic factors the higher the retail price index and the mortgage interest rate the higher the probability of transition from up to date to one payment behind and the lower the probability of recovery. The higher are house prices, the lower are retail prices and the lower are credit card interest rates the greater the chance of recovery from two behind to being up to date.

The estimates of the same parameter in the models excluding random effects are generally more significant than those in the models with random effects. Those covariates that are significant in the random effects models almost always have the same sign in the models without random effects.

Table 1: Parameter estimates together with their significance for the six sub-models (with random effects).

		0 −−− > 1		**1 −−− > 0**		**1 −−− > 2**		2 −−− > 0		2 −−− > 1		2 −−− > 3	
		Est.	p-val	Est.	p-val	Est.	p-val	Est.	p-val	Est.	p-val	Est.	p-val
Application variables	*Number of cards*	-0.002	0.741	-0.078	0.000	-0.004	0.810	-0.070	0.069	-0.057	0.052	0.005	0.835
	Indicator for presence of landline	-0.003	0.908	-0.178	0.001	-0.208	0.000	-0.133	0.396	-0.089	0.413	-0.134	0.142
	Time at address	-0.001	0.609	0.004	0.105	-0.002	0.445	-0.010	0.164	-0.021	0.000	-0.016	0.001
	Time with bank	0.000	0.001	0.001	0.000	-0.001	0.023	0.000	0.852	0.001	0.179	-0.001	0.013
	Indicator for missing time with bank	-0.016	0.567	0.153	0.011	-0.116	0.073	-0.209	0.254	0.107	0.401	-0.168	0.126
	Income	-0.024	0.082	0.548	0.000	0.320	0.000	0.079	0.479	0.043	0.595	0.013	0.853
	Indicator for missing or 0 income	-0.617	0.000	4.987	0.000	3.153	0.000	0.640	0.550	0.218	0.778	0.289	0.658
	Variable X, group B	0.213	0.000	0.047	0.435	0.168	0.010	-0.244	0.187	0.188	0.147	0.303	0.006
	Variable X, group C	0.207	0.000	-0.016	0.809	0.027	0.718	-0.687	0.002	-0.086	0.555	-0.039	0.753
	Variable X, group D	0.085	0.006	0.104	0.128	0.097	0.191	-0.111	0.582	0.284	0.049	0.086	0.493
	Variable X, group E	0.113	0.001	0.115	0.141	0.193	0.022	0.026	0.910	0.214	0.202	0.224	0.120
	Employment code, group B	0.070	0.004	0.116	0.024	0.182	0.001	-0.044	0.772	0.089	0.401	-0.115	0.219
	Employment code, group C	-0.063	0.146	0.358	0.001	0.419	0.002	-0.150	0.686	-0.244	0.360	0.003	0.991
	Employment code, group D	-0.254	0.000	0.034	0.733	0.157	0.139	0.360	0.241	-0.011	0.958	0.230	0.212
	Employment code, group E	0.013	0.744	-0.174	0.046	0.128	0.164	0.047	0.856	0.066	0.719	0.174	0.266
	Age at application, group 2	-0.047	0.117	0.103	0.128	-0.147	0.035	0.113	0.581	0.033	0.818	0.023	0.852
	Age at application, group 3	-0.096	0.006	0.072	0.346	-0.175	0.026	0.036	0.876	0.135	0.392	0.190	0.160
	Age at application, group 4	-0.143	0.000	-0.043	0.607	-0.163	0.059	0.194	0.432	0.009	0.958	0.077	0.599
	Age at application, group 5	-0.201	0.000	-0.134	0.125	-0.333	0.000	-0.090	0.729	-0.019	0.917	0.067	0.669
	Age at application, group 6	-0.186	0.000	-0.066	0.476	-0.196	0.044	-0.203	0.451	-0.132	0.487	0.007	0.967
	age at application, group 7	-0.265	0.000	-0.005	0.961	-0.385	0.000	-0.323	0.268	-0.177	0.384	-0.158	0.371
	Age at application, group 8	-0.304	0.000	-0.148	0.156	-0.608	0.000	-0.236	0.458	-0.274	0.232	-0.118	0.548
	Age at application, group 9	-0.398	0.000	-0.086	0.451	-0.698	0.000	-0.383	0.295	-0.168	0.515	-0.149	0.502
	Age at application, group 10	-0.551	0.000	0.388	0.003	-0.967	0.000	-0.222	0.609	-0.259	0.400	-0.160	0.547
Behavioural variables lagged 6 months	*Credit limit*	0.073	0.000	-0.355	0.000	-0.321	0.000	-0.142	0.068	-0.466	0.000	-0.298	0.000
	Repayment amount	0.069	0.000	0.057	0.000	-0.014	0.084	0.113	0.000	0.027	0.105	-0.024	0.119
	Proportion of credit drawn	1.033	0.000	-1.205	0.000	-0.073	0.134	-1.763	0.000	-0.493	0.000	-0.531	0.000
	Rate of total jumps	0.943	0.000	-0.208	0.036	0.409	0.000	1.339	0.000	0.719	0.001	0.384	0.053
	Improvement in state	-0.048	0.010	-0.023	0.630	0.124	0.039	-0.237	0.199	-0.064	0.583	0.100	0.385
Macroeconomic variables lagged 6 months	*Retail Price Index*	0.015	0.000	-0.028	0.002	-0.012	0.272	-0.081	0.009	-0.016	0.426	0.021	0.279
	Average Wage earning	0.008	0.000	0.000	0.870	0.002	0.547	-0.003	0.772	-0.012	0.046	0.004	0.470
	FTSE index	0.000	0.000	0.000	0.245	0.000	0.376	0.001	0.153	0.000	0.359	0.001	0.029
	Unemployement rate	-0.017	0.581	-0.781	0.000	-0.351	0.016	0.004	0.993	-0.194	0.463	-0.430	0.084
	Index of production	-0.011	0.000	0.003	0.245	0.003	0.421	-0.017	0.095	-0.004	0.516	-0.013	0.033
	House price index	-0.002	0.038	0.014	0.000	0.004	0.313	0.033	0.007	0.011	0.179	0.002	0.791
	Consumer confidence	0.003	0.058	-0.039	0.000	-0.017	0.083	-0.060	0.020	-0.012	0.439	-0.013	0.425
	Card interest rate	-0.146	0.000	-0.052	0.373	-0.052	0.464	-0.411	0.044	-0.116	0.396	-0.334	0.008
	Mortgate interest rate	0.067	0.032	-0.792	0.000	0.098	0.606	-0.036	0.947	0.187	0.569	0.255	0.417
	Total credit outstanding	-0.347	0.000	0.697	0.000	0.219	0.088	0.960	0.008	0.505	0.027	0.365	0.090

Table 2: Parameter estimates together with their significance for the six sub-models (without random effects).

		0 −−− > 1		**1 −−− > 0**		**1 −−− > 2**		2 −−− > 0		2 −−− > 1		2 −−− > 3	
		Est.	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***
Application variables	***Number of cards***	-0.016	0.000	-0.052	0.000	-0.006	0.578	-0.067	0.007	-0.056	0.001	-0.023	0.119
	Indicator for presence of landline	-0.004	0.763	-0.158	0.000	-0.176	0.000	-0.043	0.675	-0.007	0.915	-0.084	0.174
	Time at address	0.001	0.026	0.004	0.004	0.000	0.849	-0.008	0.096	-0.013	0.000	-0.012	0.000
	Time with bank	0.000	0.001	0.001	0.000	0.000	0.507	0.000	0.795	0.001	0.067	-0.001	0.016
	Indicator for missing time with bank	0.017	0.252	0.161	0.000	0.010	0.828	-0.142	0.238	0.183	0.023	-0.059	0.423
	Income	-0.026	0.001	0.491	0.000	0.287	0.000	0.109	0.117	0.085	0.092	0.160	0.000
	Indicator for missing or 0 income	-0.457	0.000	4.536	0.000	2.817	0.000	0.652	0.331	0.288	0.552	1.169	0.006
	Variable X, group B	0.129	0.000	0.020	0.592	0.132	0.005	-0.239	0.049	0.117	0.154	0.182	0.013
	Variable X, group C	0.168	0.000	-0.121	0.004	-0.030	0.572	-0.523	0.000	-0.056	0.531	-0.005	0.951
	Variable X, group D	0.102	0.000	0.034	0.422	0.039	0.460	-0.087	0.502	0.227	0.013	0.128	0.125
	Variable X, group E	0.087	0.000	-0.012	0.810	0.114	0.060	-0.101	0.488	0.181	0.080	0.180	0.055
	Employment code, group B	0.051	0.000	0.164	0.000	0.198	0.000	-0.056	0.566	0.021	0.757	-0.043	0.481
	Employment code, group C	-0.009	0.736	0.300	0.000	0.251	0.008	-0.207	0.383	-0.351	0.028	0.094	0.489
	Employment code, group D	-0.148	0.000	-0.005	0.938	0.133	0.083	0.455	0.019	0.244	0.064	0.591	0.000
	Employment code, group E	0.015	0.498	-0.031	0.562	0.125	0.057	-0.011	0.947	0.000	0.999	0.150	0.145
	Age at application, group 2	-0.042	0.013	0.050	0.247	-0.138	0.006	0.213	0.106	0.090	0.319	0.089	0.286
	Age at application, group 3	-0.063	0.001	-0.026	0.585	-0.211	0.000	0.007	0.962	0.126	0.206	0.131	0.155
	Age at application, group 4	-0.087	0.000	-0.131	0.013	-0.191	0.002	0.026	0.869	0.005	0.966	-0.021	0.831
	Age at application, group 5	-0.111	0.000	-0.256	0.000	-0.390	0.000	-0.073	0.661	0.007	0.951	0.019	0.858
	Age at application, group 6	-0.111	0.000	-0.139	0.016	-0.287	0.000	-0.274	0.114	-0.192	0.104	-0.183	0.085
	age at application, group 7	-0.126	0.000	-0.144	0.018	-0.403	0.000	-0.255	0.184	-0.090	0.483	-0.119	0.309
	Age at application, group 8	-0.151	0.000	-0.341	0.000	-0.671	0.000	-0.373	0.069	-0.484	0.001	-0.260	0.039
	Age at application, group 9	-0.187	0.000	-0.239	0.001	-0.693	0.000	-0.367	0.118	-0.231	0.141	-0.295	0.038
	Age at application, group 10	-0.294	0.000	0.031	0.703	-0.923	0.000	-0.175	0.535	-0.237	0.212	-0.276	0.103
Behavioural variables lagged 6 months	***Credit limit***	0.037	0.000	-0.374	0.000	-0.268	0.000	-0.165	0.001	-0.380	0.000	-0.292	0.000
	Repayment amount	0.078	0.000	0.084	0.000	-0.006	0.421	0.096	0.000	0.033	0.009	-0.001	0.962
	Proportion of credit drawn	1.042	0.000	-1.517	0.000	-0.256	0.000	-1.506	0.000	-0.576	0.000	-0.773	0.000
	Rate of total jumps	2.664	0.000	0.419	0.000	0.446	0.000	1.176	0.000	0.980	0.000	0.326	0.016
	Improvement in state	-0.005	0.761	-0.034	0.398	0.104	0.040	-0.176	0.190	0.128	0.160	0.037	0.668
Macroeconomic variables lagged 6 months	***Retail Price Index***	0.010	0.000	-0.048	0.000	-0.020	0.032	-0.084	0.000	-0.038	0.011	-0.039	0.006
	Average Wage earning	0.007	0.000	0.003	0.099	0.003	0.294	-0.006	0.372	-0.012	0.009	0.002	0.655
	FTSE index	0.000	0.000	0.001	0.000	0.000	0.287	0.001	0.018	0.001	0.012	0.001	0.000
	Unemployement rate	-0.113	0.000	-1.034	0.000	-0.642	0.000	-0.183	0.548	-0.461	0.017	-1.144	0.000
	Index of production	-0.011	0.000	0.005	0.048	0.003	0.396	-0.012	0.127	0.000	0.930	-0.010	0.047
	House price index	0.000	0.824	0.013	0.000	0.007	0.045	0.029	0.001	0.013	0.032	0.005	0.383
	Consumer confidence	-0.002	0.084	-0.061	0.000	-0.028	0.000	-0.068	0.000	-0.037	0.003	-0.056	0.000
	Card interest rate	-0.141	0.000	-0.095	0.048	-0.070	0.239	-0.307	0.040	-0.086	0.408	-0.061	0.512
	Mortgate interest rate	0.086	0.004	-0.455	0.000	-0.055	0.724	0.082	0.830	0.309	0.210	0.140	0.548
	Total credit outstanding	-0.256	0.000	0.885	0.000	0.381	0.000	0.961	0.000	0.600	0.000	0.915	0.000

Random effects

The random effects allow us to account for the correlation between different spells as well as the unobserved variations. When fitting the model, the covariance matrix $\mathbf{\Phi}$ was assumed to be diagonal. Table 3 displays the estimates of the variance of the random effects for each of the six transition types, together with their relative significance. The result in this Table indicates that the random effects are strongly significant.

Table 3: Variance of the random effects

	0 − − − > 1		**1 − − − > 0**		**1 − − − > 2**		**2 − − − > 0**		**2 − − − > 1**		**2 − − − > 3**	
	Est.	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***	***Est.***	***p-val***
σ^2_{hj}	1.14203	0.00000	1.59006	0.00000	0.97514	0.00000	2.35814	0.00000	1.77609	0.00000	1.93967	0.00000

2.3 Goodness of fit

A standard means of checking a model's fit is to look at the residuals, i.e. the standardised discrepancy between the actual data and what the model predicts. There are several types of residuals in the literature. In this work, we can take advantage of the discrete nature of the data and compute aggregate deviance residuals monthly.

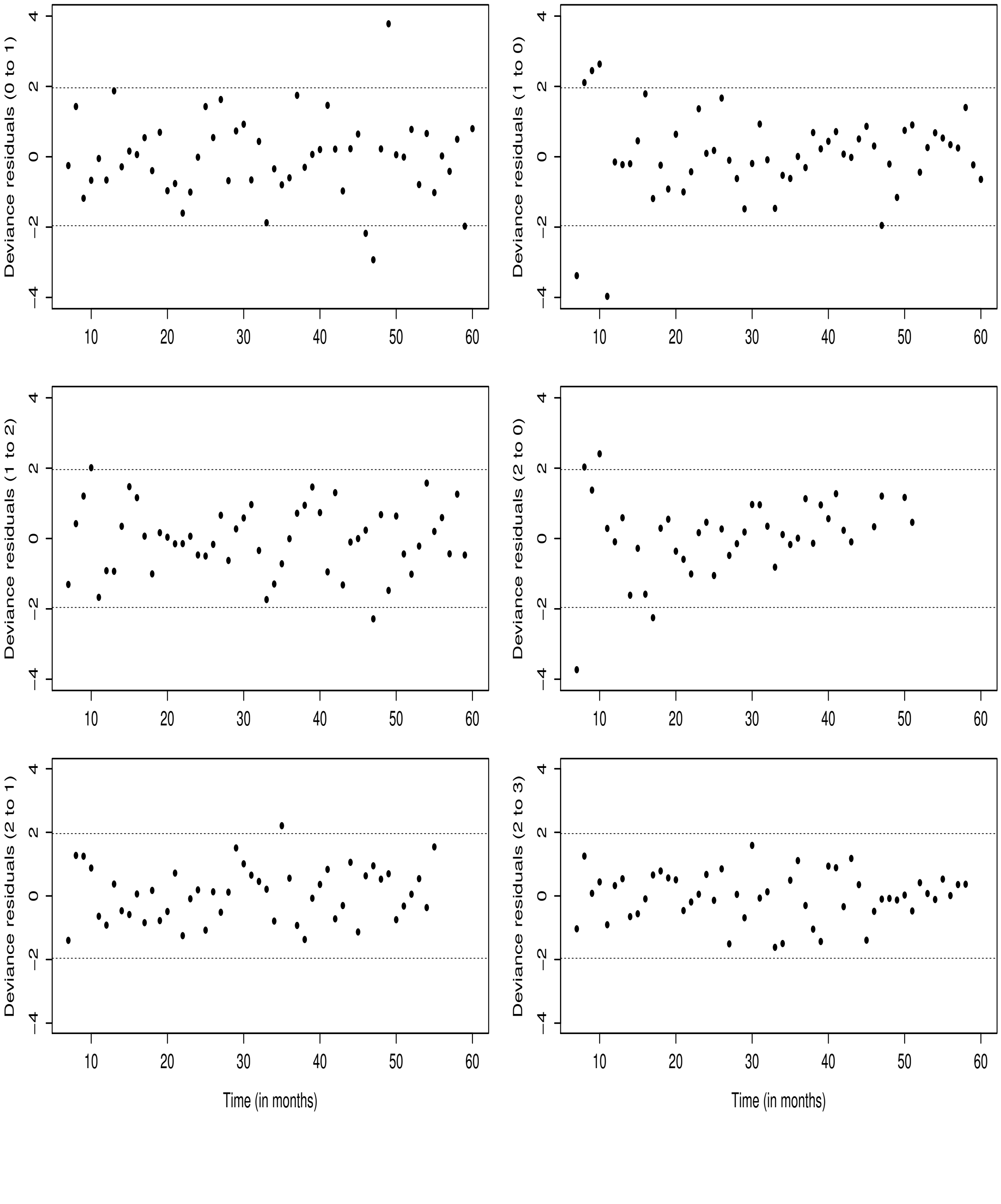
For a given transition type $(h, j) \in \mathcal{S}$, we compute the monthly deviance residuals arising from model (1)-(2) as follows:

$$D_{hj}(t) = \pm 2 \left[O_{hj}(t) \times \log \left(\frac{O_{hj}(t)}{E_{hj}(t)} \right) + (N_{hj}(t) - O_{hj}(t)) \times \log \left(\frac{N_{hj}(t) - O_{hj}(t)}{N_{hj}(t) - E_{hj}(t)} \right) \right] \cdot \qquad (16)$$

In this expression, $N_{hj}(t)$ is the number of accounts in the risk set $\mathcal{R}_{hj}(t)$; $O_{hj}(t)$ represents the total number of transitions from state h at time $(t-1)$ to state j at time t; $E_{hj}(t)$ denotes the predicted number of jumps from state h into state j, i.e. $E_{hj}(t) = \sum_{i \in \mathcal{R}_{hj}(t)} \hat{q}_{ihj}$ where hat ($\hat{}$) refers to the estimate.

A graphical illustration of these residuals is displayed in Figure 2. This shows that the residuals from each sub-model are broadly centred, with no discernible pattern (except perhaps for transitions from state 2 to state 0); in addition, more

Figure 2: Aggregate deviance residuals.



than 95% of the points lie between -2 and 2. These indicate that the sub-models fit the actual data well.

2.4 Predictions

The model described in Section 1 can be used to predict time-dependent transition probabilities for each account in the test set. These predicted probabilities encapsulate the baselines via the estimated spline coefficients and spline functions, as well as the predicted effects of each covariate based on the regression coefficients together with the values of the covariates in the test set.

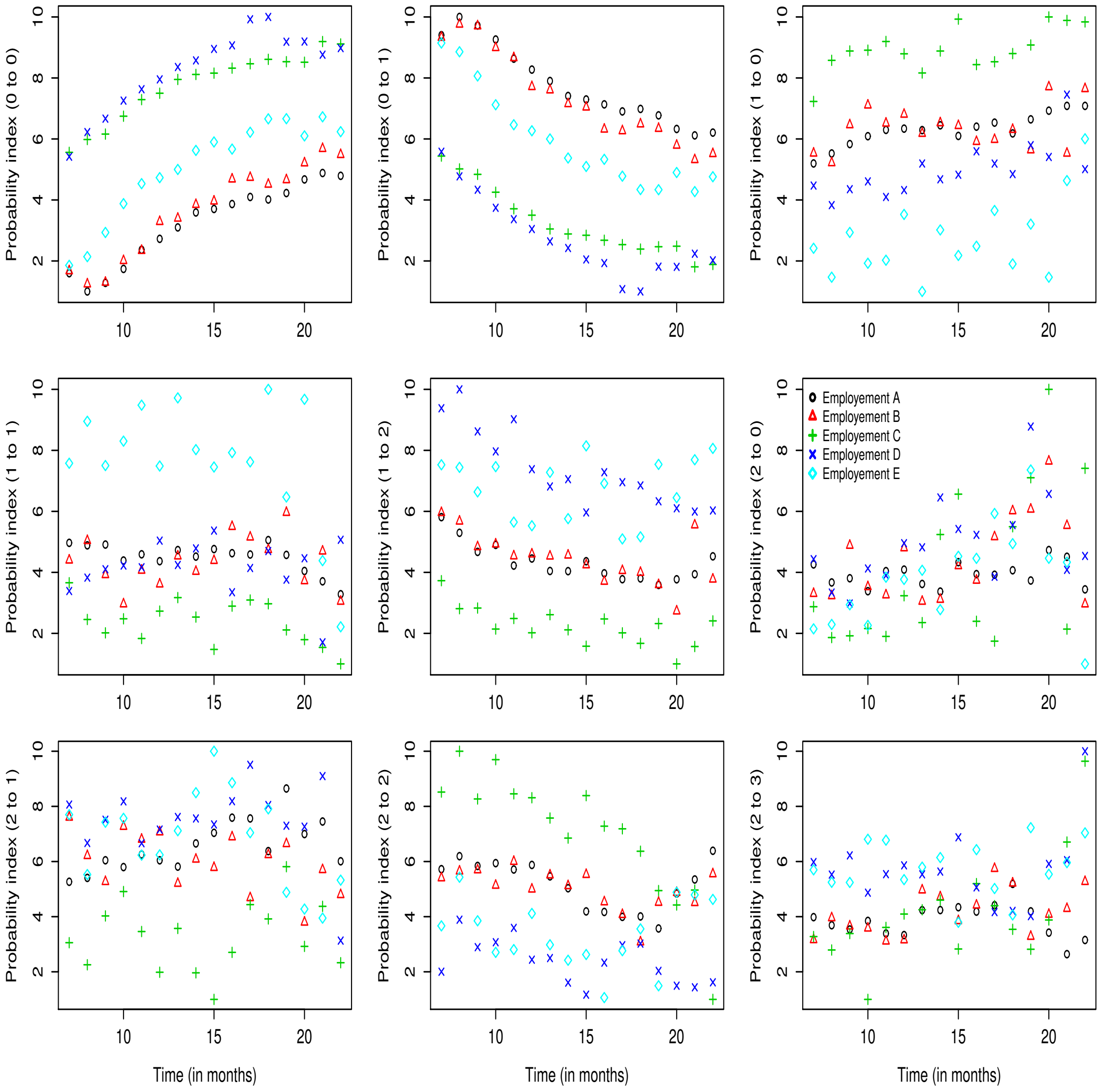
Insight from aggregated one-step prediction

In practice, rather than calculating probabilities for each individual account, there are situations where one often wants to explore predictions corresponding to specific values of a given covariate. This can be done in different ways.

One way is to set the unobservable random effects $\boldsymbol{u}_{ihj}$ to their expected value (i.e. to 0) for each account in the test set, and then average the individual predictions for accounts at each level of the targeted covariate at each time point. An illustration of such aggregated predictions by employment type in time is displayed in Figure 3. A number of conclusions can be drawn from these graphics.

First, there is a high and increasing chance for accounts to remain at state 0 for all employment types, with employment types C and D having the highest chance to remain. However, once an account has transited into state 1, there is a lower chance to recover if that account is from employment type D compared to types B, C and E. But overall, as the time the card is held advances there is a slightly increasing chance to recover from state 1, a decreasing chance to move from state 1 further into delinquency, and a broadly constant risk to remain at state 1 (with those in employment type E having the highest chance to remain in state 1). Also, there is an increasing chance of direct recovery from state 2, and a decreasing chance to remain in state 2. However, this chance of direct recovery is low compared to the chance of remaining in state 2, as well as that of defaulting or that of moving into state 1.

Figure 3: Aggregated predicted transition probabilities by employment type.



A similar graphic for aggregated prediction by age bands in shown in Figure 4. The overall patterns are broadly similar to those found in Figure 3. The increasing volatility as one moves from the right to the left of each panel is due to the decreasing number of accounts at risk in the test set.

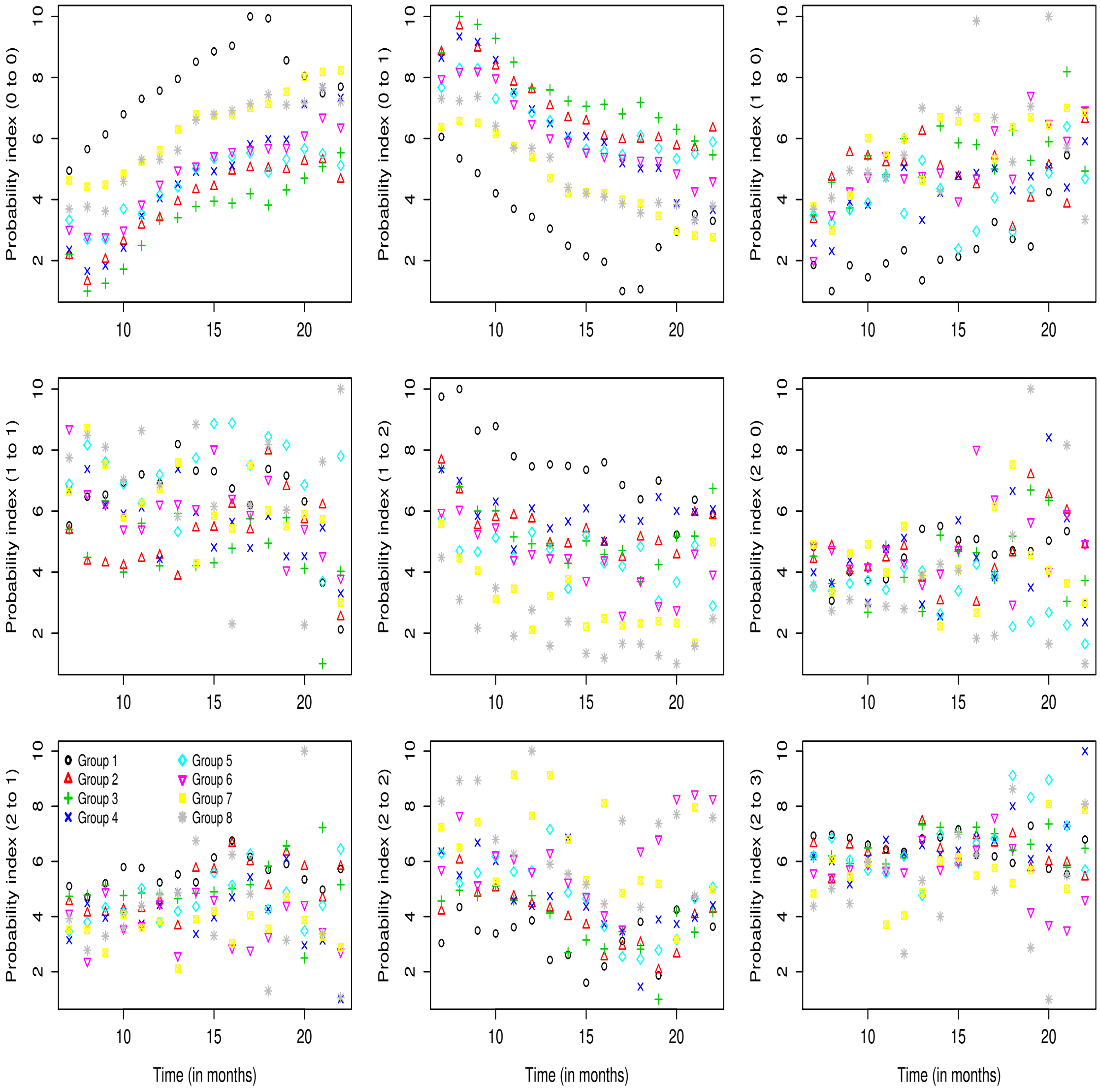
The second (and more realistic) approach to estimate aggregated predictions is to incorporate a reasonable amount of randomness into the process. For each account i and each transition type $(h, j) \in \mathcal{S}$, we generate a random deviate u_{ihj} from $\mathcal{N}(0, \hat{\sigma}^2_{hj})$ where the estimated variances $\hat{\sigma}^2_{hj}$ are those displayed in Table 3. We then add these simulated random effects to the linear predictor of each account. We next compute the competing transition probability at each time point, and then average these individual probabilities in each level of the targeted covariate at each time point. The overall conclusion drawn from this second procedure was broadly similar to that drawn from Figures 3 and 4, although the aggregated transition probabilities were slightly more spread than those seen in these Figures (due to the incorporation of random effects).

Insight from cumulative transition probabilities for typical accounts.

In the previous subsection, we looked at transition probabilities over a one month horizon. However, as described in Section 1.3, we can equally explore the likelihood of being in a given state at time t_2 given the state occupied by the account at an earlier time t_1. For illustration, we create a typical account for each employment type based on the test set as follows.

Each time-independent variable is set to the average (for continuous variables) or modes (for categorical variables) over the accounts in each employment category. Each behavioural variable at each time point is set to the mean or mode over the accounts at risk at that time. For macroeconomic variables, we consider two scenarios. Scenario 1 assumes that the account was open in January 2009, whereas Scenario 2 set the open date to January 2010. The macroeconomy was more bouyant in the latter period than the first with the index of production, the FTSE and average wage earnings all higher and the mortage rate and credit card rate lower.

Figure 4: Aggregated predicted transition probabilities by age group.



Tables 4 and 5 show the probabilities of transiting or staying in a given state in month 12, given the state occupied by the account at time 6. The difference between the two Tables is an indication of the impact of the change in the macroeconomic conditions under the two scenarios. Comparing the corresponding probabilities we can see that the effect of the change in the economy was relatively small across all states and employment types. In general, the probability of transiting from state 1 to 0 increases for all employment types, and that from state 1 to state 2 decreases as does that from state 2 to state 3. The probability of transiting from state 2 to state 1 increases and from state 2 to 3 decreases.These are all as expected as the economy improves. But those from state 0 to 1 or 2 or 3 all increase and those from state 2 to state 0 decrease, which are all contrary to expectations.

Accuracy of Predictions

In this section, we assess the ability of the model to predict future states. Since the outputs from the model are not the predicted states themselves, we will first describe how to derive predicted states from the predicted transition probabilities.

We propose to compute the predicted states based on the distance between the predicted probabilities and some pre-specified cut points. Let us denote by c_{k0}, c_{k1}, c_{k2} and c_{k3} the values of some pre-specified cut points corresponding to transitions from state k at time t_1 to states $j \in \{0, 1, 2, 3\}$ at a latter time t_2.

Consider a test account i, and let us denote by $\hat{p}_{ik0}, \hat{p}_{ik1}, \hat{p}_{ik2}$ and $\hat{p}_{i3}$ the predicted competing probabilities that the account will be in state $0, 1, 2$ and 3 at time t_2, given that the account was in state k at time t_1. These probabilities correspond to the k^{th} row of the cumulative probability matrix (15).

At time t_2 we predict that account i will be in state j such that

$$\hat{p}_{kj} - c_{kj} = \max\left\{\hat{p}_{k0} - c_{k0},\ \hat{p}_{k1} - c_{k1},\ \hat{p}_{k2} - c_{k2},\ \hat{p}_{k3} - c_{k3}\right\} \qquad (17)$$

In other words, we predict that account i will find itself in the state corresponding to the largest discrepancy between the transition probability and the corresponding cut point.

Table 4: Cumulative transition probability matrix, $\tilde{\boldsymbol{P}}(6, 12)$, by employment type for typical account opened in January 2009.

			To state			
			0	1	2	3
Employment A	From state	0	0.9020	0.0634	0.0167	0.0179
		1	0.7703	0.0591	0.0242	0.1463
		2	0.3361	0.0333	0.0256	0.6051
		3	0	0	0	1
Employment B	From state	0	0.9040	0.0595	0.0198	0.0167
		1	0.7569	0.0572	0.0357	0.1501
		2	0.3207	0.0365	0.0470	0.5958
		3	0	0	0	1
Employment C	From state	0	0.8960	0.0561	0.0207	0.0272
		1	0.7080	0.0497	0.0290	0.2133
		2	0.2489	0.0242	0.0260	0.7009
		3	0	0	0	1
Employment D	From state	0	0.9698	0.0235	0.0034	0.0032
		1	0.8983	0.0246	0.0123	0.0648
		2	0.3529	0.0213	0.0424	0.5834
		3	0	0	0	1
Employment E	From state	0	0.9351	0.0287	0.0110	0.0252
		1	0.7021	0.0241	0.0113	0.2625
		2	0.3099	0.0123	0.0071	0.6708
		3	0	0	0	1

Table 5: Cumulative transition probability matrices, $\tilde{\boldsymbol{P}}(6, 12)$, by employment type for typical account opened in January 2010.

			To state			
			0	1	2	3
Employment A	From state	0	0.8868	0.0730	0.0174	0.0227
		1	0.7722	0.0683	0.0233	0.1362
		2	0.2818	0.0340	0.0243	0.6599
		3	0	0	0	1
Employment B	From state	0	0.8891	0.0687	0.0206	0.0216
		1	0.7606	0.0655	0.0327	0.1412
		2	0.2727	0.0378	0.0441	0.6455
		3	0	0	0	1
Employment C	From state	0	0.8799	0.0645	0.0212	0.0343
		1	0.7145	0.0572	0.0268	0.2015
		2	0.2123	0.0241	0.0223	0.7413
		3	0	0	0	1
Employment D	From state	0	0.9650	0.0276	0.0035	0.0039
		1	0.9096	0.0284	0.0097	0.0523
		2	0.2887	0.0231	0.0424	0.6458
		3	0	0	0	1
Employment E	From state	0	0.9240	0.0332	0.0117	0.0310
		1	0.7157	0.0277	0.0111	0.2455
		2	0.2686	0.0121	0.0061	0.7133
		3	0	0	0	1

The cut point vectors $(c_{k0}, c_{k1}, c_{k2}, c_{k3})$ are estimated (based on the accounts in the training set) as the multi-dimensional maximisers of the objective functions f_k defined by

$$f_k(a_0, a_1, a_2, a_3) = \frac{1}{N_k(t_1)} \sum_{\substack{i, \\ with \\ \delta_i(t_1)=k}} \mathbb{1}_{\{\hat{\delta}_i(t_2|a_0,a_1,a_2,a_3)=\delta_i(t_2)\}} \tag{18}$$

In this expression, $N_k(t_1)$ represents the number of accounts in state k at time t_1, $\delta_i(t)$ denotes the true state occupied by account i at time t, and $\hat{\delta}_i(t|a_0, a_1, a_2, a_3)$ denotes the predicted state corresponding to the generic cut point vector (a_0, a_1, a_2, a_3). Thus, $f_k(a_0, a_1, a_2, a_3)$ is the proportion of accurate predictions corresponding to the cut point vector (a_0, a_1, a_2, a_3).

In some extreme scenarios, the discrepancy measure (17) above might tend to favour jumps toward transition types with larger predicted probabilities; this can lead to classification bias. One way to avoid this is to incorporate suitable scale factors. Thus, we consider two additional measures: the standardised discrepancy and the relative discrepancy.

Under the standardised discrepancy, the predicted state at time t_2 is the state j such that

$$\frac{\hat{p}_{kj} - c_{kj}}{\hat{s}_{kj}} = \max\left\{\frac{\hat{p}_{k0} - c_{k0}}{\hat{s}_{k0}}, \frac{\hat{p}_{k1} - c_{k1}}{\hat{s}_{k1}}, \frac{\hat{p}_{k2} - c_{k2}}{\hat{s}_{k2}}, \frac{\hat{p}_{k3} - c_{k3}}{\hat{s}_{k3}}\right\} \tag{19}$$

where the $\hat{s}_{kj}$ denote the empirical standard deviations of the predicted probabilities.

Under the relative discrepancy measure, the predicted state at time t_2 is the state j such that

$$\frac{\hat{p}_{kj} - c_{kj}}{c_{kj}} = \max\left\{\frac{\hat{p}_{k0} - c_{k0}}{c_{k0}}, \frac{\hat{p}_{k1} - c_{k1}}{c_{k1}}, \frac{\hat{p}_{k2} - c_{k2}}{c_{k2}}, \frac{\hat{p}_{k3} - c_{k3}}{c_{k3}}\right\} \tag{20}$$

We note that this classification framework is different to that found elsewhere. For example the cut points used by Leow and Crook (2014) were computed such that the proportion of accounts predicted to undergo transition is equal to the proportion that underwent transition in the training set, irrespective of whether the predicted states were correct or not. In addition, their classification algorithm

Table 6: Prediction performance at time 12, given state at time 6.

	Model without random effects			Model with random effects		
State at time 6	Discrepancy (17)	Discrepacy (19)	Discrepancy (20)	Discrepancy (17)	Discrepacy (19)	Discrepancy (20)
0	89%	90%	89%	90%	90%	90%
1	72%	73%	72%	72%	72%	72%
2	63%	62%	62%	63%	63%	63%

discarded the competing spirit of multi-state models. These concerns have been addressed in the framework described above.

A comparative illustration of the predictive performance at time $t_2 = 12$, given the state occupied at time $t_1 = 6$, is shown in Table 6 under our three discrepancy measures (17) (19) and (20). We notice first that the predictive accuracy is the same ragardless of the measure used. Second the predictive accuracy decreases at higher initial delinquency states. Third the model with random effects has almost identical predictive accuracy as the model without random effects, but is never less than 63%.

3 Conclusion

We have parameterised multistate models that predict the probability that a credit card account will transit between two delinquency states in the next time period and have used the estimated parameters to predict competing risk probabilities that an account will transit between states between two, not necessarily adjacent, time periods. For each possible transition, we have compared these probabilities to cut points to derive predicted jumps for an account and compared the predicted number of jumps with the observed number for the portfolio as a whole. We have made three contributions to the literature. We have included random effects in multistate models to account for unobserved heterogeneity and have observed the change in predictive accuracy this affords and the significance of macroeconomic variables that have been included in the models. We conclude first that the use of B-splines allows the detection of noticeably different baseline hazards between the jump processes. Second the inclusion of the random effects is supported by the highly significant variances of these effects for all models. Third the inclusion of random effects generally reduces the significance of the covariates. Fourth when a very flexible baseline function is used the inclusion of random effects does not enhance predictive accuracy.

Bibliography

Aalen,O. O., Borgan, O. and Gjessing, H. K. (2008) *Survival and Event History Analysis*. New York: Springer.

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

Andersen P. K. and Borgan Ø. and Gill R. D. and Keiding N. (1993) *Statistical Models Based on Counting Processes*. Springer.

Banasik J. and Crook J. and Thomas L. C. (1999) Not if but when borrowers default. *Journal of the 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. (2012) Loss given default models incorporating macroeconomic variables for credit cards. *International journal of Forecasting*, **28**, 171-182.

Bellotti T. and Crook J. (2009) Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, **60(12)**, 1699-1707.

Cameron A. C. and Trevedi P. K. (2005) *Microeconometrics*. Cambridge University Press.

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

Creal D. and Schwaab B. and Koopman S.J. and Lucas A.(2014) Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, **96(5)**, 898-915.

Crook J. and Bellotti T. (2010) Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society series A*, **173**, 283-305.

Dickson D. C. M. and Hardy M. and Hardy M.R. and Waters H.R. (2009) *Actuarial Mathematics for Life Contingent Risks.* Cambridge University Press.

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

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

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

Enberg J. and Gottschalk P. and Wolf D. (1990) A random-effects logit model of work–welfare transitions *Journal of Econometrics*, **43**, 63–75.

Figlewski S. and Frydman H. and Liang W. (2012) Modeling the effect of macroeconomic factors on corporate default and credit rating transitions *International Review of Economics and Finance*, **21**, 87-105.

Gagliardini and P. and Gourieroux and C. (2005) Stochastic migration models with application to corporate risk *Journal of Financial Econometrics*, **3(2)**, 188-226.

Gasbarra D. and Karia S. R. (2000) Analysis of Competing Risks by Using Bayesian Smoothing *Scandinavian Journal of Statistics*, **27**, 605-617.

Goldstein H. and Pan H. and Bynner J. (2004) A flexible procedure for analysing longitudinal event histories using a multilevel model. *Understanding Statistics*, **3**, 85–99.

Hoem J. M. (1990) Identifiability in hazards models with unobserved heterogeneity: The compatibility of two apparently contradictory results *Theoretical Population Biology*, **37**, 124-128.

Hougaard P. (2000) *Analysis of Multivariate Survival Data.* Springer.

Jarrow R. A. and Lando D. and Turnbull S.(1997) A Markov model for the term structure of credit risk spreads *Review of Financial Studies*, **20(2)**, 481-523.

Koopman S. J. and Kraussl R. and Lucas A. and Monteiro A. B. (2009) Credit cycles and macro fundamentals *Journal of Empirical Finance*, **16**, 42-54.

Koopman S. J. and Lucas A. and Monteiro A. (2008) The multi-state latent factor intensity model for credit rating transitions *Journal of Econometrics*, **142**, 399-424.

Kyung M. and Gill J. and Casella C. (2010) Estimation in Dirichlet random effects models. *Annals of Statistics*, **38**, 979–1009.

Lando D. and Skødeberg T. M. (2002) Analysing rating transitions and rating drift with continuous observations. *Journal of Banking and and Finance*, **26**, 423-444.

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 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.

Luptakova I.D.L. and Bilikova M. (2014) Actuarial Modeling of Life Insurance Using Decrement Models. *Journal of Applied Mathematics, Statistics and Informatics*, **10**, 81-91.

Parner E. (1997) *Inference in semiparametric frailty models.* University of Aarhus, Denmark.

Pinheiro J. C. and Bates D. (1995) Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, **4**, 12-35.

Promislow S. D. (2006) *Fundamentals of Actuarial Mathematics.* Willey.

Ripatti S. and Palmgren J. (2000) Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood. *Biometrics*, **56**, 1016-1022.

Ruppert D. and Wand M. P. and Carroll R. J. (2009) Semiparametric regression during 2003-2007. *Electron. J. Statist.*, **3**, 1193-1256.

Searle S. R. and Casella G. and McCulloch C. E. (2006) *Variance Components.* Willey.

Sen A. and Banerjee B. and Lib Y. and Noone A. M. (2010) A Bayesian approach to competing risks analysis with masked cause of death. *Statistics in Medicine*, **29**, 1681–1695.

Steele F. and Goldstein H. and Browne W. (2004) A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, **4**, 145–159.

Steele F. and Diamond I. and Wang D. (1996) The Determinants of the Duration of Contraceptive Use in China: A Multi-level Multinomial Discrete-Hazards Modeling Approach. *Demography*, **33**, 12-23.

Stepanova M. and Thomas L. C. (2002) Survival analysis methods for personal loan data. *Operations Research*, **50**, 277-289.

Stepanova M. and Thomas L. C. (2001) PHAB scores: Proportional hazards analysis behavioural scores. *Journal of the Operational Research Society*, **52**, 1007-1016.

Therneau T. M. and Grambsch P. M. (2000) *Modeling Survival Data: Extending the Cox Model.* Willey.

Vaupel J. W. and Manton K. G. and Stallard E. (1979) The impact of heterogeneity in individual frailty on the dynamic of Mortality. *Demography*, **16**, 439-454.