Identifying hidden patterns in credit risk survival data using Generalised Additive Models

Citation for published version:
Djeundje, VAB & Crook, J 2019, 'Identifying hidden patterns in credit risk survival data using Generalised Additive Models', *European Journal of Operational Research*, vol. 277, no. 1, pp. 366-376. https://doi.org/10.1016/j.ejor.2019.02.006

Digital Object Identifier (DOI):
10.1016/j.ejor.2019.02.006

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
European Journal of Operational Research

Identifying hidden patterns in credit risk survival data using Generalised Additive Models.

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

viani.djeundje@ed.ac.uk, j.crook@ed.ac.uk

Credit Research Centre, University of Edinburgh Business School,
29 Bucceleuch Place, Edinburgh EH8 9JS, UK

Abstract

Modelling patterns in credit risk using survival analysis techniques have received considerable and increasing attention over the past decade. In these models, the predictor of the hazard of default is often expressed as a simple linear combination of the risk factors. In this work, we discuss how these models can be enhanced using Generalised Additive Models (GAMs). In the GAMs framework, the predictor is formulated as a combination of flexible univariate functions of the risk factors. In this paper, we parametrise GAMs for credit risk data in terms of penalised splines, outline the implementation via frequentist and Bayesian MCMC methods, apply them to a large portfolio of credit card accounts, and show how GAMs can be used to improve not only the application, behavioural and macro-economic components of survival models for credit risk data at individual account level, but also the accuracy of predictions. From a practitioner point of view, this work highlights that some accounts may actually become more (less) attractive to the lender if flexible smooth functions are used whereas the same applicant may be denied (accepted) a loan if the linearity assumption is forced.

Key words

OR in Banking; Risk Analysis; Smoothing; Multivariate Statistics; Predictions Accuracy.

Introduction

Modelling and predicting credit behaviour patterns is a topic of crucial importance to lenders of credit card loans. In this context, survival models have attracted impressive attention over the recent past, and are increasingly being used both in academia and industry. Survival models offer several advantages over traditional statistical methods. Some of these benefits are documented in Allison (2010), Andreeva (2006), Bellotti and Crook (2013), Stepanova and Thomas (2002), Djeundje and Crook (2018) among others.

At its core, modelling credit risk data using survival analysis involves expressing a major component of the expected hazard rate as a linear combination of the risk factors. These risk factors comprise categorical variables (employment type, etc) as well as scale variables (macro-economic variables, age, etc). While this simple linear assumption may hold in some cases, it is often not flexible enough for some scale variables and as a result, standard survival models are unable to detect some important hidden patterns in the data. This assumption of linearity can be relaxed through Generalised Additive Models (GAMs).

GAMs is a simple yet attractive technique for extracting patterns from data. Unlike standard survival models used in credit risk, GAMs involve a combination of flexible smooth functions of the covariates. Early methodological work can be found in Hastie and Tibshirani (1986, 1990), Friedman (1991) and Wood (2000, 2008) among others. GAMs have also attracted strong attention over the recent past. Recent developments including fast implementation algorithms for large datasets are detailed in Wood et al. (2015, 2016). GAM techniques have been implemented successfully in various application areas of statistics, including medicine, demography, environment, economics, etc (Sapra, 2013; Drexler and Ainsworth, 2013; Djeundje, 2016). In the the credit risk context, GAMs have been used to enhance predictive accuracy. For example, Berg (2006) applied GAMs on firm-specific variables to enhance bankruptcy predictions, and this was extended by Dakovica et al. (2010) with firm-specific time-varying covariates at yearly time intervals. In the retail context, some investigations of GAM techniques in simple cross-sectional logistic regressions have been reported; see for example Liu et al. (2009).

Although the application of survival models in credit risk data is growing rapidly, the integration of GAM techniques into these models has received very little attention in retail banking. The contribution of this paper is show how GAMs via penalised splines can be used to improve not only the application, behavioural and macroeconomic components of survival models for credit risk data at individual account level, but also the accuracy of predictions. Simultaneously, we also show that it may not be appropriate to apply GAMs blindly to all covariates; we demonstrate this by comparing the increased predictive accuracy when a GAM specification is applied, on the one hand to behavioural and macroeconomic variables and on the other hand to application variables. This is the first time GAMs techniques have been applied to credit cards data for survival models.

In the paper we present two implementation methods of these models: the frequentist approach and the Bayesian approach. Both methods can be implemented using standard statistical software including R, SAS and STATA. The estimation algorithm arising from the frequentist approach is generally faster. But conversely, the Bayesian approach via MCMC provides the opportunity to explore the full posterior distributions of the parameters of interest. We apply both methods to a large dataset of credit card accounts and show how they allow one to extract hidden patterns in the data and yield improved predictions.

The paper is organised as follows. Section 1 outlines standard survival models as applied to credit risk data and sets some notations for the rest of the paper. Section 2 presents GAMs for discrete time survival data with penalised splines. Section 3 describes the implementation methods from frequentist and Bayesian points of view. Section 4 introduces the data that motivated this work and presents some applications of GAMs. A simulation exercise is undertaken in Section 5 and we close with some concluding remarks in Section 6.

1 Survival models for credit risk data

We consider a portfolio of n credit card accounts. The objective is to model the time to default. Let T_i denote the true survival time for account i, $1 \leq i \leq n$. In our applications, time is measured in months, from the opening date of the accounts. Some accounts may not experience default by the end of the study, in which case their survival times would be right censored. We assume that censoring is non-informative. The discrete hazard function of default for account i at duration time t is defined by

$$q_{i,t} = Prob\{T_i = t \mid T_i \geq t\} \tag{1}$$

That is, $q_{i,t}$ represents the default rate associated with account i at time t, conditional on the account still being active just before time t. The values taken by the hazard function are driven by various factors. Some of these factors are observable but others are not. This includes the application variables (i.e. variables obtained from the application process), the behavioural variables (i.e. time-dependant and account-dependant variables), and the macroeconomic conditions.

For a given account i, we will denote by $\boldsymbol{U}_i$ the $1 \times a$ vector of application variables, and by $\boldsymbol{V}_{i,t-t_o}$ and $\boldsymbol{Z}_{i,t-t_o}$ the vector of behavioral variables and macroeconomic variables respectively, lagged (t_o) at time t. Since the macroeconomic conditions are the same for all accounts observed at the same calendar time, the dependence of $\boldsymbol{Z}_{i,t-t_o}$ on i is due to the fact that accounts are opened at different points in calendar time. To quantify the effects of these variables on the risk of default, a standard option is to use a link function (Allison, 2010; Therneau and Grambsch, 2000):

$$\begin{aligned} g(q_{i,t}) &= h_{0,t} + \boldsymbol{U}_i\boldsymbol{\alpha} + \boldsymbol{V}_{i,t-t_o}\boldsymbol{\beta} + \boldsymbol{Z}_{i,t-t_o}\boldsymbol{\delta} \\ &= h_{0,t} + \sum_{k=1}^{a}\alpha_k \times u_{i,k} + \sum_{k=1}^{b}\beta_k \times v_{i,t-t_o,k} + \sum_{k=1}^{m}\delta_k \times z_{i,t-t_o,k} \end{aligned} \tag{2}$$

In these expressions g represents the link function, $h_{0,t}$ is some baseline, and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_a)^T$ is the $a \times 1$ vector of unknown of regression coefficients associated with the application variables; its components quantify the effect of the application variables on the risk of default. Similarily, $\boldsymbol{\beta} = (\beta_1, ..., \beta_b)^T$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_m)^T$ represent the regression coefficients associated with the behavioural and the macroeconomic variables respectively. To complete the model specification, some restrictions are often placed on the shape of the baseline $h_{0,t}$.

Denoting by $\boldsymbol{\alpha}_o$ the parameters that define the shape of $h_{0,t}$, all the unknown parameters in the model can be estimated jointly by maximisation of the likelihood function given by

$$L(\boldsymbol{\alpha}_o, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n}\prod_{t=1}^{\tau_i}(q_{i,t})^{y_{i,t}} \times (1 - q_{i,t})^{1-y_{i,t}} \tag{3}$$

In this expression, $y_{i,t}$ denotes the indicator function taking value 1 if account i has defaulted in month t and 0 otherwise; and τ_i represents the age of the account at the time of default (if account i has defaulted) or at the end of the study (if account i is censored).

2 GAMs for discrete time survival data using penalised splines

Model 2 assumes that each covariate is linearly associated with the predictor. This is a strong assumption for scale variables. The GAMs approach relaxes this assumption. For example, if the first application variable is a scale variable, then each component $\alpha_1 \times u_{i,1}$ in Equation (2) is substituted by $\mathcal{S}(u_{i,1})$, where $\mathcal{S}(\cdot)$ is a flexible smooth function.

Without loss of generality, we suppose that the first a_1 application variables are dummies and the other $a - a_1$ application variables are scales. Also, let us assume similar repartition of the behavioural and macroeconomic variables. The full GAMs extension of Equation (2) is as follows

$$
\begin{aligned}
g(q_{i,t}) \quad = \quad & h_{0,t} \\
& + \sum_{k=1}^{a_1} \alpha_k \times u_{i,k} + \sum_{k=a_1+1}^{a} \mathcal{S}_{u_k}(u_{i,k}) \\
& \quad + \sum_{k=1}^{b_1} \beta_k \times v_{i,t-t_o,k} + \sum_{k=b_1+1}^{b} \mathcal{S}_{v_k}(v_{i,t-t_o,k}) \\
& \quad\quad + \sum_{k=1}^{m_1} \delta_k \times z_{i,t-t_o,k} + \sum_{k=m_1+1}^{m} \mathcal{S}_{z_k}(z_{i,t-t_o,k})
\end{aligned}
\tag{4}
$$

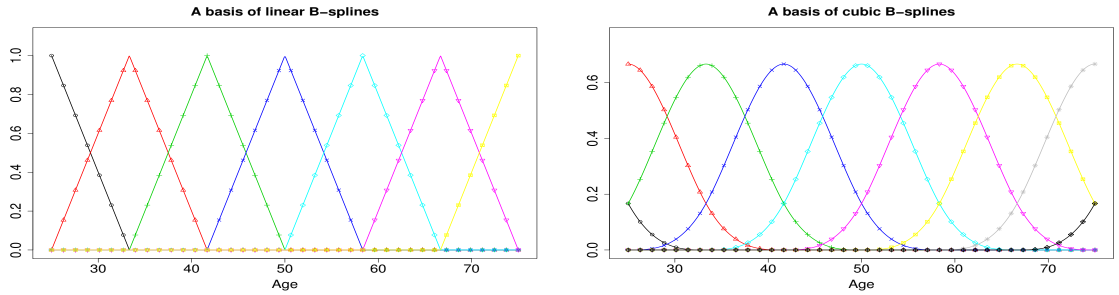
where the $\mathcal{S}_{u_k}(\cdot)$, $\mathcal{S}_{v_k}(\cdot)$ and $\mathcal{S}_{z_k}(\cdot)$ are unknown flexible smooth functions to be estimated. This extended expression assumes that the impact of each scale variable is modelled using a flexible smooth function. In practice the assumption of simple linearity may hold for some scale variables, in which case the smooth function would be applied only on an appropriate subset of the scale variables. In Section 4 for example, we shall investigate models involving smooth functions on application variables alone.

A natural question that emerges is how do we estimate the smooth function involved in (4)? Indeed, the method should be flexible enough to capture hidden trends in the data. A good candidate is the P-splines methodology of Eilers and Marx (1996). This method shares several features with standard regressions. In particular, it involves expressing a smooth function as a linear combination of a basis of B-splines:

$$
\mathcal{S}_w(x) = \sum_{r=1}^{c_w} \theta_{w,r} \times B_{w,r}(x) \tag{5}
$$

where $B_{w,r}(\cdot)$ are B-splines along the scale variable w, r is the index of the B-splines, c_w is the number of B-splines, and $\theta_{w,r}$ are unknown splines coefficients to be estimated. Details on the estimation is given in Section 3 below.

Figure 1: B-splines along age.



A B-spline can be described as a combination of truncated polynomials. An illustration of linear and cubic B-splines is shown on Figure 1. Each B-spline has a compact support and this makes them advantageous over other spline bases. For a complete description of B-splines, we refer the reader to De Boor (1978) or Eilers and Marx (2010). We use cubic B-splines in this paper; some motivations of this preference are discussed by Green and Silverman (1995).

3 Estimation of GAMs

We want the smooth functions $\mathcal{S}_w(\cdot)$ to be as flexible as possible in order to capture hidden patterns in the data. However it is imperative to ensure that we do not over-fit the data. An attractive way to achieve this is to penalise the differences in adjacent B-spline coefficients; this is know as the method of penalised splines or simply P-splines (Eilers and Marx, 1996; Wood, 2006). With this approach, B-splines with equi-spaced knots can be used, and for each covariates w in equation (5), the number of B-splines c_w is chosen large enough so that there are enough B-splines to capture the important features in the data while penalisation ensures smoothness.

In this paper, we implement GAM via penalised splines in two ways. First we follow the frequentist approach based on the optimisation of the penalised likelihood, and second, we use a Bayesian MCMC method. The former provides point estimates of the parameters of interest whereas the latter gives access to the full posterior distributions of the parameters. We can compare the estimates from the two methods.

3.1 Frequentist estimation

The penalised log-likelihood arising from (4) can be expressed as

$$\ell_P \;=\; \ell \;+ \sum_{k=a_1+1}^{a} \lambda_{u_k} \times \mathcal{P}(\boldsymbol{\theta}_{u_k}) + \sum_{k=b_1+1}^{b} \lambda_{v_k} \times \mathcal{P}(\boldsymbol{\theta}_{v_k}) + \sum_{k=m_1+1}^{m} \lambda_{m_k} \times \mathcal{P}(\boldsymbol{\theta}_{m_k}) \qquad (6)$$

where ℓ is the ordinary log-likelihood function arising from (2), λ_w is the smoothing parameter associated with the scale variable w, and $\mathcal{P}(\cdot)$ denotes the penalty function acting on the spline coefficients to ensure smoothness. We use the second order difference penalties in this paper, in which case $\mathcal{P}(\cdot)$ is given by

$$\mathcal{P}(\boldsymbol{\theta}_w) = (\theta_{w,3} - 2\theta_{w,2} + \theta_{w,1})^2 \;+\; \cdots \;+\; (\theta_{w,c_w} - 2\theta_{w,c_{w-1}} + \theta_{w,c_{w-2}})^2 \qquad (7)$$

For fixed values of the smoothing parameters, the value of the regression parameters and splines coefficients that maximise the penalised log-likelihood (6) can be computed via the penalized iteratively re-weighted least squares algorithm (Green and Silverman, 1995; Wood, 2008). For large datasets, more efficient algorithms can help to boost speed and convergence properties as described in Wood et al. (2015, 2016).

So far, we have overlooked a very important issue: the choice of the smoothing parameters. From the objective function (6), it can be seen that the λ_w's quantify the trade-off between fidelity to the data as measured by the log-likelihood, and the smoothness of the model as measured by the difference penalty terms. Hence, the smoothing parameters play a central role in the model specification and their choice falls in the bias-variance trade-off paradigm. In practice, optimal values of the smoothing parameters can be selected via an information metric such as the restricted maximum likelihood (Wood, 2011) or the Akaike Information Criteria (AIC) defined by

$$\text{AIC} = \text{Deviance} \; + \; 2p, \tag{8}$$

where p represents the effective dimension of the model.

The implementation of this procedure can be facilitated by using the `mgcv` package in `R` (Wood, 2016). In particular, two functions are available in this package: `gam`() and `bam`(). Both functions facilitate the estimations of a variety of flexible models (including GAMs) the latter being a more efficient implementation specifically developed for large datasets.

3.2 Estimation via Bayesian MCMC method

In the Bayesian paradigm, the unknown regression parameters and spline coefficients are treated as random variables and have to be supplemented with appropriate prior distributions. The prior distributions are often non-informative. But they can also be specified so as to incorporate some external or expert judgements about the parameters of interest or the default rates themselves.

For consistency with Section 3.1, we assume non-informative priors about the regression parameters. For the splines coefficients however we impose smoothness. Thus, following Lang and Brezger (2004), we replace the second order difference penalties in (7) by their stochastic analogues (i.e. second-order random walks) as follows

$$\theta_{w,j} \quad \sim \quad \mathcal{N}\left(2\theta_{w,j-1} - \theta_{w,j-2},\; \sigma_w^2\right), \tag{9}$$

with diffuse prior for $\theta_{w,1}$ and $\theta_{w,2}$.

In this case, the amount of smoothness is controlled by the variance parameters σ_w^2, which correspond to the inverse of the smoothing parameters in Section 3.1. These variance parameters are unknown themselves. Thus, they are also treated as random and hyper priors are assigned to them. A common choice of prior for such variance parameters is a non-informative prior specified using the inverse Gamma distribution; See for example Lang and Brezger (2004) and Crainiceanu et al. (2005).

With this in place, the regression parameters, variance parameters and spline coefficients can be investigated by Bayesian inference via MCMC simulations, and this entails updating full conditionals of single parameters or blocks of parameters. However, single-move steps, which update each parameter separately can suffer from problems with convergence and mixing (Fahrmeir and Lang, 2001) especially in models comprising a large number of unknown parameters. Thus, in this paper, MCMC samples were generated and updated in blocks based on the Metropolis-Hasting algorithm with iterative weighted least square proposals; see Gamerman (1997), Fahrmeir and Lang (2001), Brezger and Lang (2006). This procedure can be implemented using `BayesX`, a software package designed to fit structured additive regression models using MCMC (Brezger et al., 2005).

4 Applying GAMs to a credit risk dataset

The dataset that motivated this work is a large sample of credit card accounts from a major UK bank. It consists of more than 60,000 individual accounts opened from 2002 to 2011 on different books. The dataset contains several variables collected at the time of application as well as behavioural variables collected monthly. In addition, some macroeconomic variables were appended to the dataset.

In this analysis, an account is said to have 'defaulted' if and when it became three months in arrears. Note that the three missed payments need not to be in consecutive months. We computed a minimum payment using constant parameters for each account and so consistently throughout the period. Hence, whilst this definition is consistent with that used in Djeundje and Crook (2018), it differs from that used for example by the data provider.

These data lend themselves naturally into a survival analysis framework. In this framework, an important tool for aggregate data exploration is the overall survival func-

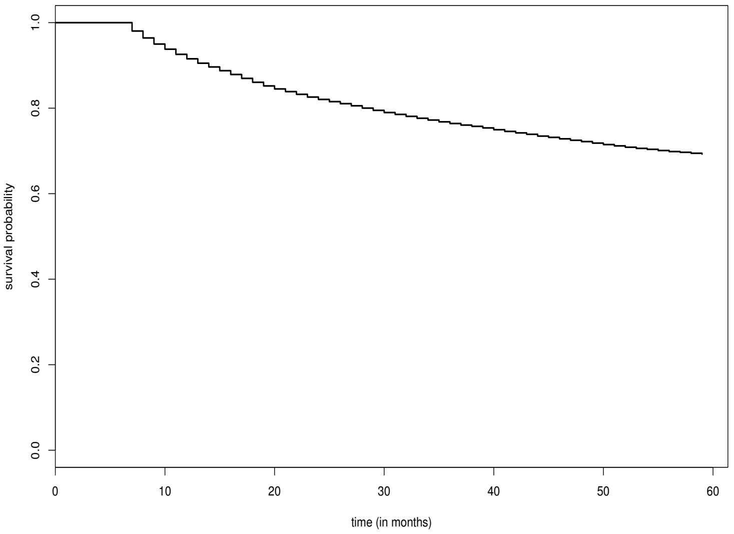


Figure 2: Survival curve.

tion (Kaplan and Meier, 1958). It allows one to visualise the fraction of accounts still active at certain times after the opening date. The survival function for our dataset is displayed in Figure 2. Several conclusions can be drawn. For example, it shows that the probability of surviving the first year in this portfolio is about 90%. That is, nearly 10% of the accounts have missed three payments or more during the first year. Similarly, about 82% were able to avoid 3 missing payments during the first two years, and more than $\frac{3}{4}$ of the accounts were still active after the first three years.

Separate survival curves can also be constructed for different blocks of business; for example by employment type within age groups. But this entails splitting the data into different sub-blocks. As such, the Kaplan-Meir survival function is limited in its ability to estimate the probability of default/survival adjusted for covariates. This can be achieved using appropriate survival models.

Following Djeundje and Crook (2018) we split this dataset into three separate sets: a training set, a retrospective test set and a prospective test set. The training set comprises a random sample of 80% of the accounts opened from 2002 to 2008. We use it to estimate the models. The retrospective test comprises the rest of the 20% of accounts opened from 2002 to 2008, whereas the prospective test set consists of all the accounts opened from 2009 onward. Thus, the prospective test set is out-of-sample and out-of-time relative to the training set, whereas the retrospective set is out-of-sample but in-time. Both test sets are used to assess and compare the predictive performance of the models.

The dataset contains a number of categorical and scale variables. The variables used in this paper are those shown in Table 1. The application and behavioral variables were calculated directly from the data supplied by the data provider, whereas the macroeconomic variables are those of the Office of National Statistics in the UK, subject to some scale and location adjustments.

Table 1: Risk factors used in this investigation.

Application variables	Age at application	Numeric
	Number of cards group	Categorical (4 groups)
	Employement type	Categorical (5 groups)
	Variable X	Categorical (5 groups)
Behavioural variables	Repayment amount	Numeric
	% Time with one outstanding payment	Numeric
	% Time with two outstanding payments	Numeric
Macroeconamic variables	Average wage earnings	Numeric
	Consumer confidence	Numeric
	Unemployement rate	Numeric

In order to identify and quantify the impact of GAMs for credit risk data, a number of models with various GAMs specifications were implemented. In this paper, we narrow the presentation to those shown in Table 2. Each model in this table was fitted in two ways. First, via maximisation of the penalised log-likelihood as described in Section 3.1; and second, by Bayesian MCMC method as discussed in Section 3.2.

4.1 Models output

This section presents some of the main output from the models described in Table 2. We start with `Model0`; that is the model without GAMs specification. The parameters were estimated by maximum penalised likelihood and by MCMC simulations. In both cases, the baseline was specified in terms of B-splines, and penalties were applied on the spline coefficients to achieve smoothness. An illustration of the MCMC samples from the posterior distribution of some of the baseline spline coefficients and regression

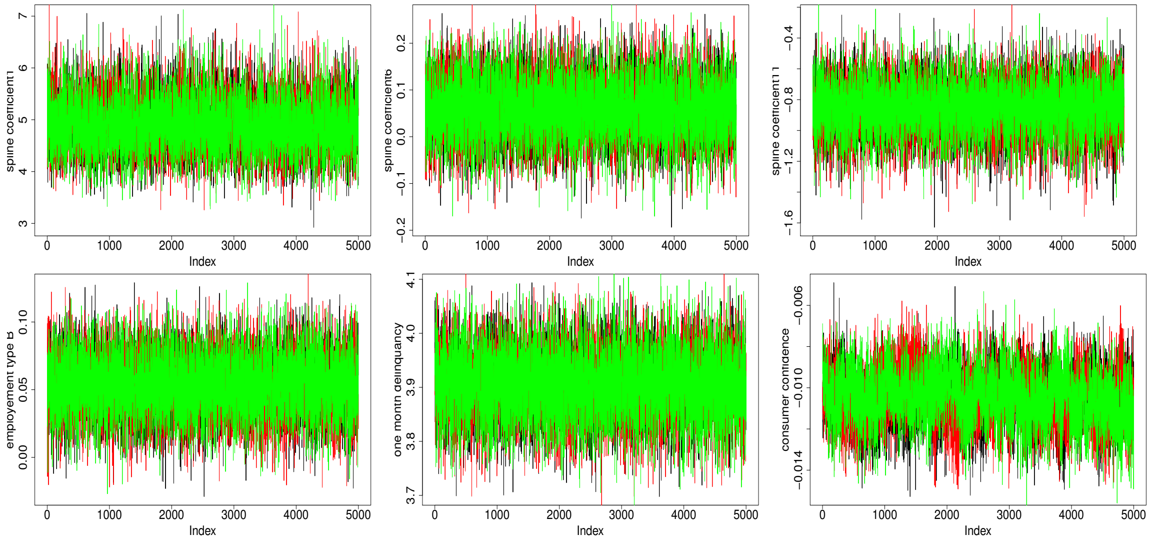
Table 2: List of models.

Model code	Description
`Model0`	model without GAMs specification
`Model1`	model with GAMs specification on application variables
`Model2`	model with GAMs specification on behavioral variables
`Model3`	model with GAMs specification on macroeconomic variables
`Model4`	model with GAMs specification on application, behavioral and macroeconomic variables

Each model listed in this table was implemented using the frequentist and Bayesian approaches. In addition to these, models with GAMs specification on single variables were also investigated.

parameters is shown in Figure 3.

Figure 3: MCMC samples from the posterior distribution of some of the baseline spline coefficients and regression parameters for `Model0`.



A comparative summary of the parameters estimates is given in Table 3. This shows that estimates from both methods are very similar and their signs are broadly as expected. For example, having a larger number of cards or repaying a larger amount are associated with increased risk of default whereas increased consumer confidence is associated with reduced risk; one possible explanation to this direction of the impact of the repayment amount is that when people default their repayment amount is larger because they are trying to pay off larger balances outstanding.

We now look at the ability of GAMs to capture patterns in the data. We start by the marginal effects of the variable *Age* from each of our five models. These effects are shown

Table 3: Estimated parameters from `Model0`, i.e. without GAMs effects.

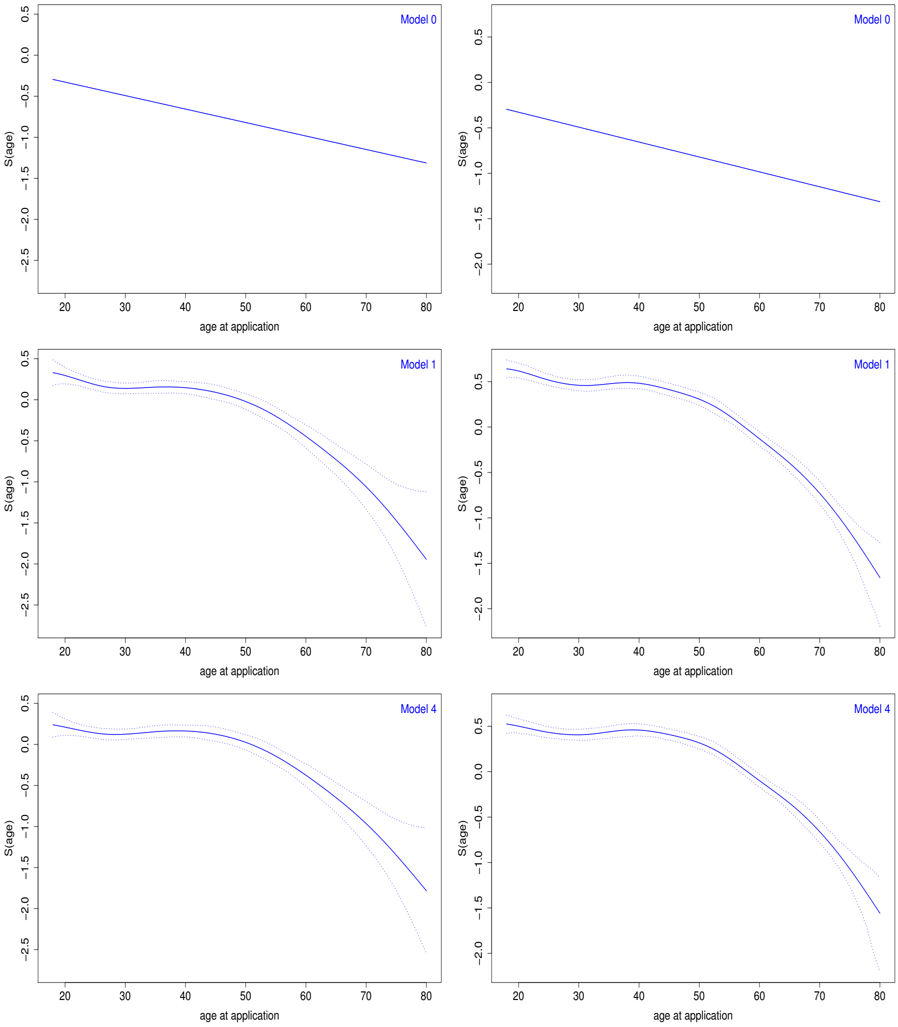
		Maximum Penalised Likelihood			**Bayesian MCMC**	
		Coefficient	*Std Error*	*p-val*	*Coefficient (mean)*	*Std Dev*
Application variables	*Age at application*	-0.0164	0.0009	0.0000	-0.0164	0.0010
	Number of cards, group B	0.0531	0.0224	0.0177	0.0528	0.0220
	Number of cards, group C	0.1569	0.0248	0.0000	0.1574	0.0248
	Number of cards, group D	0.1827	0.0909	0.0444	0.1805	0.0893
	Variable X, group B	0.4012	0.0285	0.0000	0.4014	0.0282
	Variable X, group C	0.4620	0.0336	0.0000	0.4623	0.0331
	Variable X, group D	0.2024	0.0316	0.0000	0.2024	0.0318
	Variable X, group E	0.3849	0.0314	0.0000	0.3850	0.0311
	Employment code, group B	0.0653	0.0299	0.0288	0.0657	0.0294
	Employment code, group C	-0.3301	0.0577	0.0000	-0.3304	0.0570
	Employment code, group D	-0.0556	0.0319	0.0814	-0.0558	0.0323
	Employment code, group E	0.1544	0.0266	0.0000	0.1544	0.0269
Behavioral variables	*%time with one oustanding payment*	3.9058	0.0620	0.0000	3.9058	0.0608
	%time with two oustanding paymens	3.0611	0.1485	0.0000	3.0725	0.1504
	Repayment amount	0.0647	0.0036	0.0000	0.0642	0.0037
Macroeconomic variables	*Consumer confidence*	-0.0106	0.0014	0.0000	-0.0106	0.0014
	Unemployement rate	-0.0516	0.0143	0.0003	-0.0518	0.0143
	Average wage earnings	0.0003	0.0011	0.7880	0.0003	0.0011

In both methods, the baseline were fitted using penalised splines via maximising the penalised likelihood and via Bayesian method. Each behavioural and macroeconomic variable was lagged 6 months.

in Figure 4 with approximative 95% confidence bands. The panels on the left hand side are based on maximising the penalised likelihood as described in Section 3.1 whereas those on the right hand side are calculated from the MCMC samples. We recall that `Model0`, `Model2` and `Model3` assume that the marginal effect of *Age* is linear; `Model1` and `Model4` relax this assumption through GAMs specifications. A number of conclusions can be drawn from these graphics. For example, the panels corresponding to `Model1` and `Model4` show that, indeed, the shape of the marginal effect of *Age* is not linear. Overall, as with `Model0`, the summary estimates of the marginal effects from the frequentist and Bayesian methods are very similar. From now on, we shall omit most of the graphics obtained from the Bayesian MCMC method.

Let us consider the marginal effects for *repayment amount*; these are shown on Figure 5. `Model0`, `Model1` and `Model3` assume a linear marginal effect for the *repayment amount*. However, `Model2` and `Model4` do not make this restrictive assumption, and their outputs demonstrate that the risk of default increases steadily only over the lower values of *repayment amounts* and then become almost flat for larger values of *repayment amounts*. Similar, yet distinct comments apply to the marginal effects of other variables in the models. See for example Figure 6 for the marginal effects of %*time with one outstanding payment*, and in the Appendix Figure A1 for *average wage earnings*, or

Figure 4: Marginal effects of *age*.



Left: maximum penalised likelihood estimates. Right: empirical estimates from MCMC samples.

The marginal effect of age from models 2 & 3 are essentially the same as in `Model0`.

Figure A2 for *consumer confidence*. In all cases the considerable deviation from linearity implies that, if used in practical applications, some applicants may actually become more attractive to the lender if a flexible spline-based function is used whereas the same applicants may be denied a loan if a linear function is used.

Figure 5: Marginal effects of *repayment amount*.

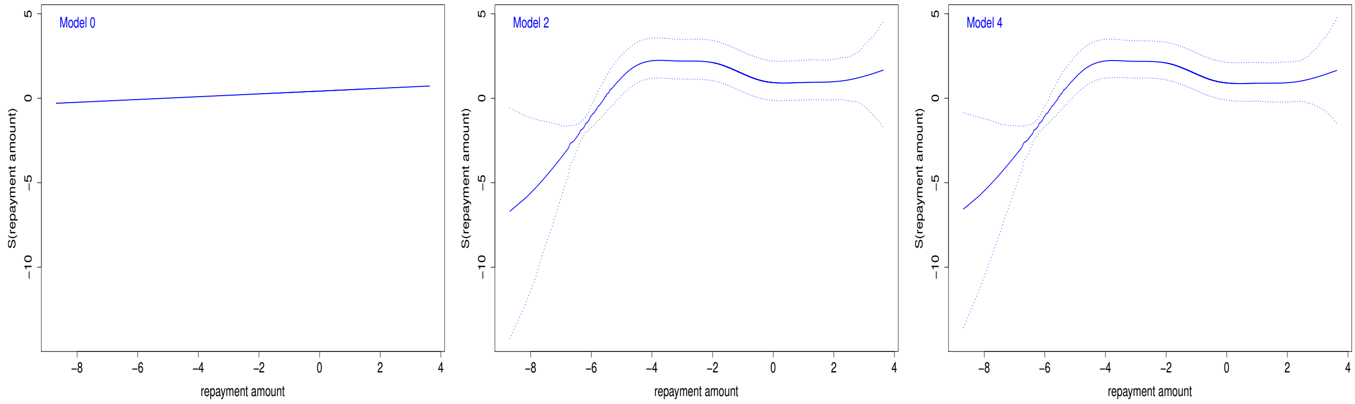
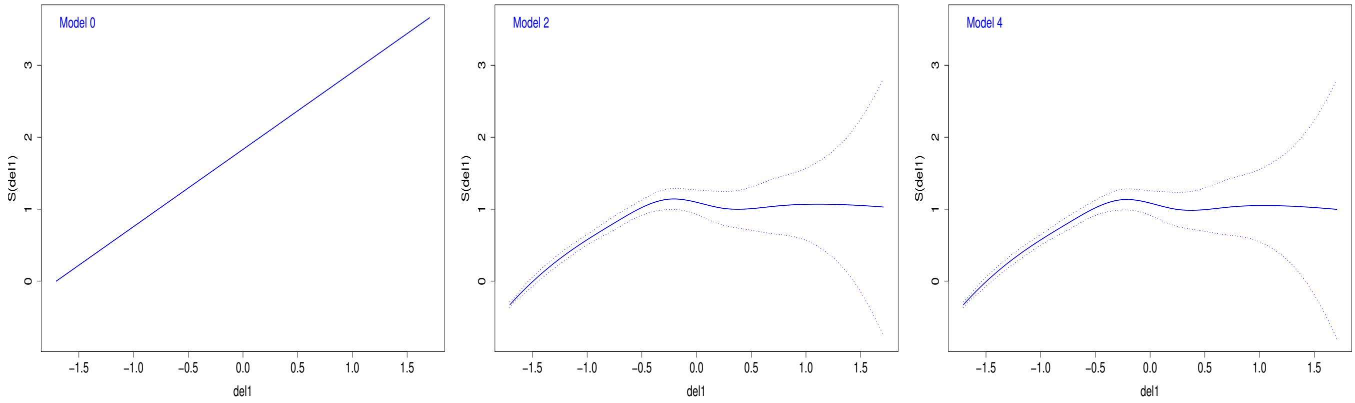


Figure 6: Marginal effects of %*time with one outstanding payment*.



4.2 Implications for hazard functions

We now consider some implications of using a GAMs specification on the shape of the hazard functions for a typical account. For illustration we consider accounts where the borrower has one of five employment types. At each time point within each employment type, we set the values of each covariate at its mean (for scale variables) and mode (for categorical variables), and then calculated the predicted probabilities of default using the estimated regression parameters and splines coefficients. An illustration of the outcome is shown on Figure 7. Different observations can be made.

First, applying GAMs as in `Model1`, `Model2`, `Model3` or `Model4` might cause the hazard relationship with time to differ between the models. In this paper that is what we observe. Second we observe that employing GAMs on time varying covariates in this case increases the probability of default at any duration time for any given employment category. This is because the GAMs specification results in marginal effects (i.e. the $\mathcal{S}_w$ in equation (4)) whose values for the time varying covariates are greater than those when a linear function is chosen, at the mean or mode of each covariate. Third, we also notice that employing GAMs on all scale variables results in greater variation over time in the hazards. This can clearly be seen from employment type D. In this case the use of GAMs on all variables (top line) results in the probability of default increasing noticeably after month 12 whereas if a linear function is used (`Model0`) the probability is almost constant. For employment type B the hazard declines more steeply over time when a GAM is used rather than when a linear form is assumed.

4.3 Model assessment and comparison

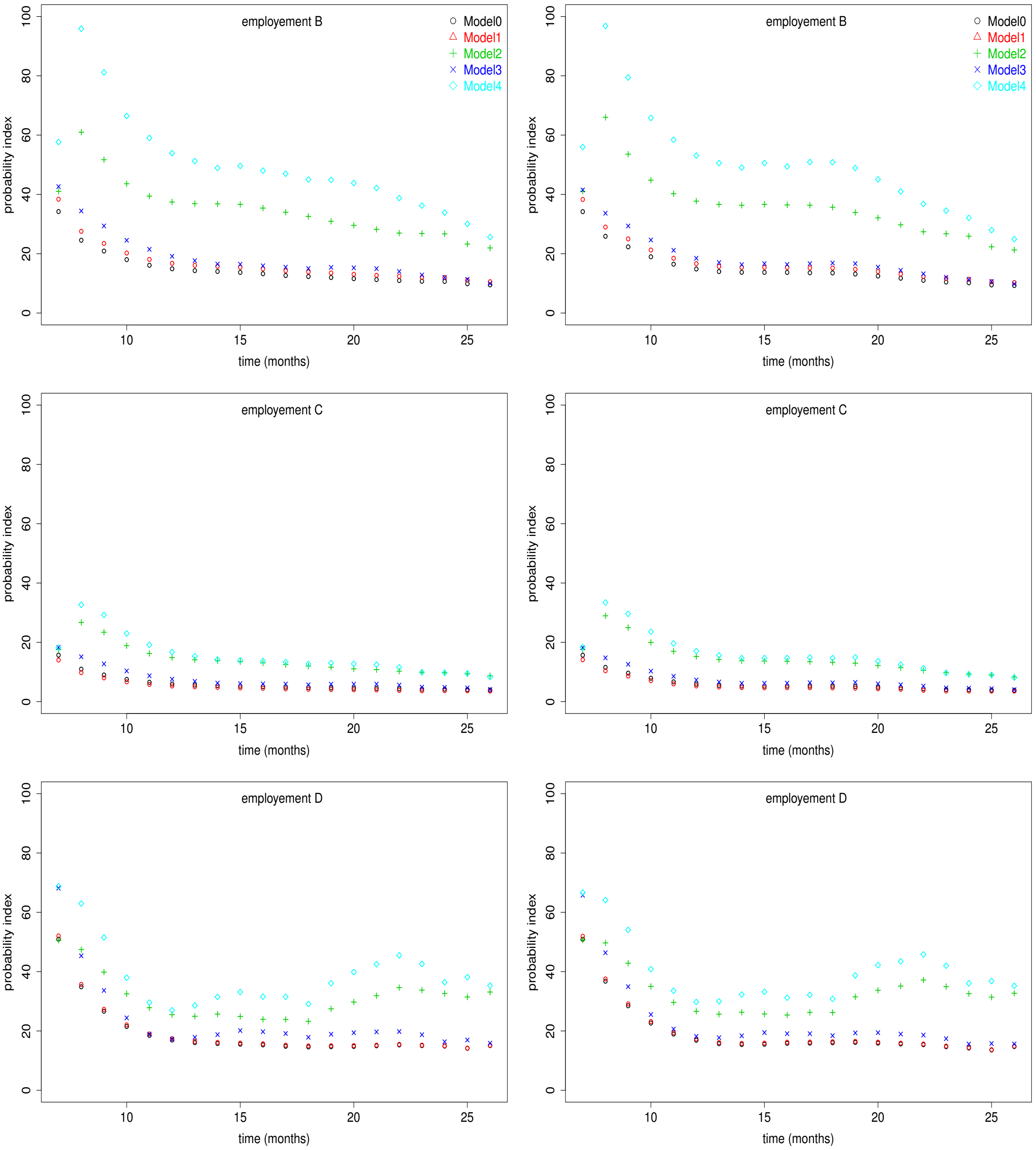
In the previous section, we illustrated the ability of GAMs to extract patterns from the data. In credit risk however, the focus is usually on predictions. In this section therefore we compare the models in terms of overall quality and predictive power.

4.3.1 Overall model quality

In practice it is always possible to improve model fit by adding a new variable or more splines into the model; but doing so can lead to over-fitting and poor predictions. A penalty against model complexity allows one to avoid this problem. In particular, AIC provides a measure of relative goodness of fit of a statistical model with a suitable penalty term for complexity as shown in (8). In general, models with lower AIC would be preferred.

Table 4 shows comparative AIC from our five models. A number of conclusions can be drawn. First, all the four models with GAMs specification on one or many variables outperform the standard model (i.e. `Model0`). The best model based on AIC statistics is `Model4`, i.e. the model with GAMs specification simultaneously on *Age*, the behavioral

Figure 7: Predicted probabilities of default for typical accounts based on the medians/modes of the covariates by employment type in the prospective test set.



Left: predictions based on the penalised log-likelihood estimates of the spline coefficients and regression parameters. Right: based on the MCMC estimates of the spline coefficients and regression parameters.

variables and the macroeconomic variables. However, the largest contribution to the drop in AIC is from the behavioral variables as revealed by AIC corresponding to `Model2`. Nonetheless, allowing GAMs specification for *Age* or the macroeconomic variables also improve the model significantly; see AICs from `Model1` and `Model3`.

Table 4: Comparative AIC.

	AIC	Drop in AIC
`Model0` (i.e. without GAMs)	139415	0
`Model1` (i.e. GAMs on application variables)	139256	158
`Model2` (i.e. GAMs on behavavioural variables)	135967	3448
`Model3` (i.e. GAMs on macroeconomic variables)	138999	416
`Model4` (i.e. GAMs on appl., behav. and macroec. variables)	135378	4037

4.3.2 Comparing predictive performance

A standard method to compare the predictive performance of binary-response models is to use the Receiver Operating Characteristic curve, also known as ROC curve. An attractive feature of the ROC is that, besides the graph of the ROC curves themselves, the accuracy of the models can be assessed by measuring the area under the curves.

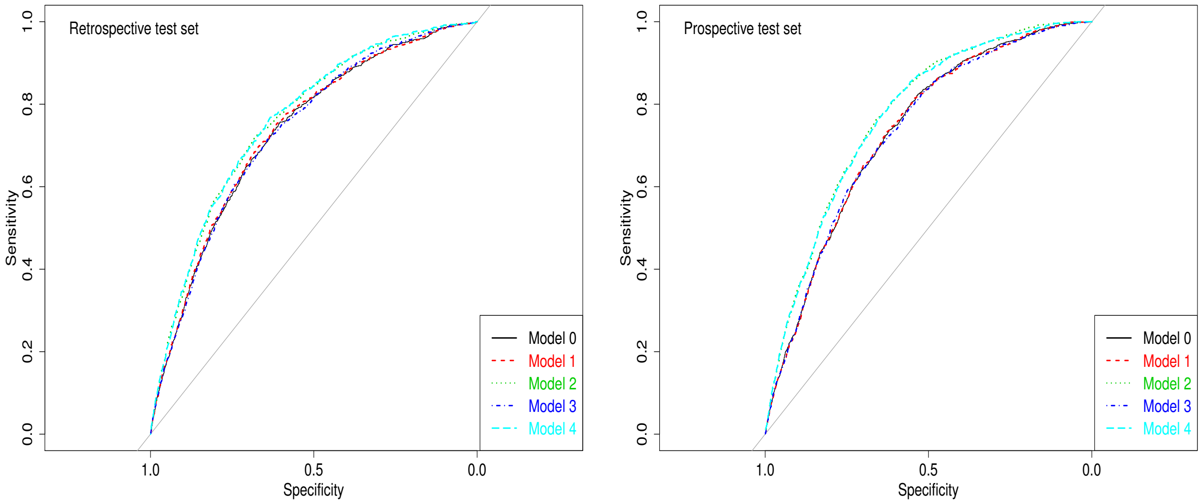
Upon fitting each of our five models, the parameters and spline functions were used to predict the probabilities of default for each account in the retrospective and prospective test sets, and these probabilities were used to construct the ROC curves for both test sets, separately. These curves are shown on Figure 8.

Table 5: Areas under the ROC curves.

	`Model0`	`Model1`	`Model2`	`Model3`	`Model4`
Retrospective test set	0.733	0.732	0.773	0.731	0.771
Prospective test set	0.731	0.734	0.757	0.736	0.763

The conclusion is consistent across both test sets: models with GAMs specification perform better than the standard model (i.e. `Model0`). In particular, the models with GAMs specification on behavioural variables (i.e. `Model2` and `Model4`) top the list. This is confirmed by the areas under the ROC curves in Table 5.

Figure 8: ROC curves.



5 Simulation exercise

In the previous section, we have illustrated the effectiveness of GAMs for credit risk data. However, the outputs presented were specific to the dataset being analysed. In this section, we undertake a short simulation exercise to investigate the ability of GAMs to enhance standard models. For computational reasons, we focus on three scale variables (*Age*, *%time with one outstanding payment* and *Consumer confidence*) and consider six scenarios. Each scenario is determined by the underlying shapes of the "true" marginal effect of these three variables. Our true marginal effects for the six scenarios are displayed in Table 6.

Table 6: True marginal effects for *Age* (x_1), *%time with one outstanding payment* (x_2) and *Consumer confidence* (x_3) in our six scenarios.

	For *Age*	For *Time with one outstanding*	For *Consumer confidence*
Scenario 1:	$\mathcal{S}(x_1) = -2 \times x_1$	$\mathcal{S}(x_2) = 1.25x_2$	$\mathcal{S}(x_3) = -2.5 \times x_3$
Scenario 2:	$\mathcal{S}(x_1) = -2x_1 - (x_1 - 30)^3$	$\mathcal{S}(x_2) = 1.25 \times x_2$	$\mathcal{S}(x_3) = -2.5 \times x_3$
Scenario 3:	$\mathcal{S}(x_1) = -2 \times x_1$	$\mathcal{S}(x_2) = -\exp(-7x_2 + 2.1)$	$\mathcal{S}(x_3) = -2.5 \times x_3$
Scenario 4:	$\mathcal{S}(x_1) = -2 \times x_1$	$\mathcal{S}(x_2) = 1.25 \times x_2$	$\mathcal{S}(x_3) = -2.5x_3 + 0.75 \times \sin(4\pi x_3)$
Scenario 5:	$\mathcal{S}(x_1) = -2x_1 - (x_1 - 30)^3$	$\mathcal{S}(x_2) = -\exp(-7x_2 + 2.1)$	$\mathcal{S}(x_3) = -2.5 \times x_3$
Scenario 6:	$\mathcal{S}(x_1) = -2x_1 - (x_1 - 30)^3$	$\mathcal{S}(x_2) = 1.25 \times x_2$	$\mathcal{S}(x_3) = -2.5x_3 + 0.75 \times \sin(4\pi x_3)$

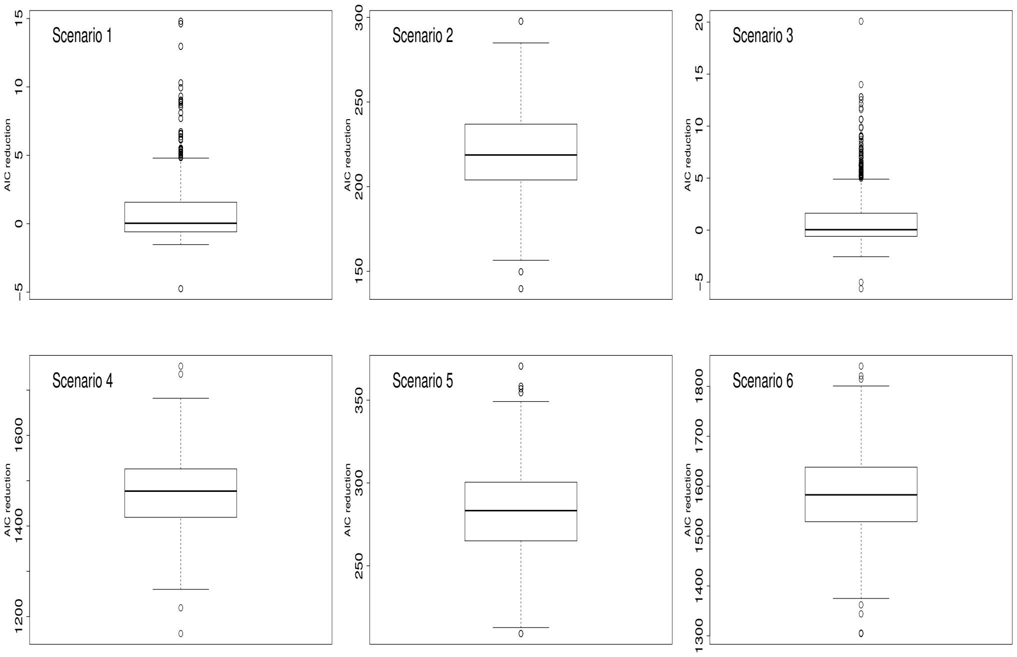
Under each scenario, we proceed as follows.

(i) Construct the linear predictor (using the marginal effects as specified in Table 6, with the baseline set to that of `Model0` fitted in Section 4) and calculate the conditional monthly default probabilities for each account in the training dataset.

(ii) Simulate the conditional default indicators, and fit two models to these simulated data: (a) the standard model without GAMs specification and (b) the flexible model with GAMs specification on *Age*, *%time with one outstanding payment* and *Consumer confidence*, simultaneously.

(iii) Repeat step (ii) 1000 times, and store the models summary statistics in each case.

Figure 9: Output summary of the simulation exercise. The vertical axis represents the reduction in AIC from the standard models without GAMs specification to the flexible models with GAMs specification.



A summary of the AIC statistics from this exercise is shown on Figure 9. On these graphics, the vertical axis represents the reduction in AIC from the standard models to the GAMs counterparts. Thus, positive numbers indicate that GAMs specification is broadly better than the standard linear specification. A general conclusion that emerges from this simulation exercise is that, in essentially all six scenarios, models with GAMs specification provide a better description of the data. In particular, the output from scenario 4 and scenario 6 highlight how using GAMs can yield a very large improvement when some of the underlying true marginal effects are far from linear.

6 Concluding remarks

Generalised Additive Models (GAMs) is a simple, yet, powerful technique for identifying hidden patterns in data. The main purpose of this work was to investigate if the standard survival models currently used in retail banking can be enhanced via GAMs. Thus, in the first half of the paper, we focussed on the parameterision of GAMs for discrete time survival data in the credit risk context, and described how these models can be implemented using frequentist and Bayesian MCMC methodologies. In the second half, we applied GAMs to a dataset of credit card accounts and to simulated datasets and found that, not only do GAMs significantly improve the overall quality of standard survival models, but also, using GAMs yield more accurate predictions on out-of-sample and out-of-time test sets.

Appendix

Figure A1: Marginal effects of *average wage earnings.*

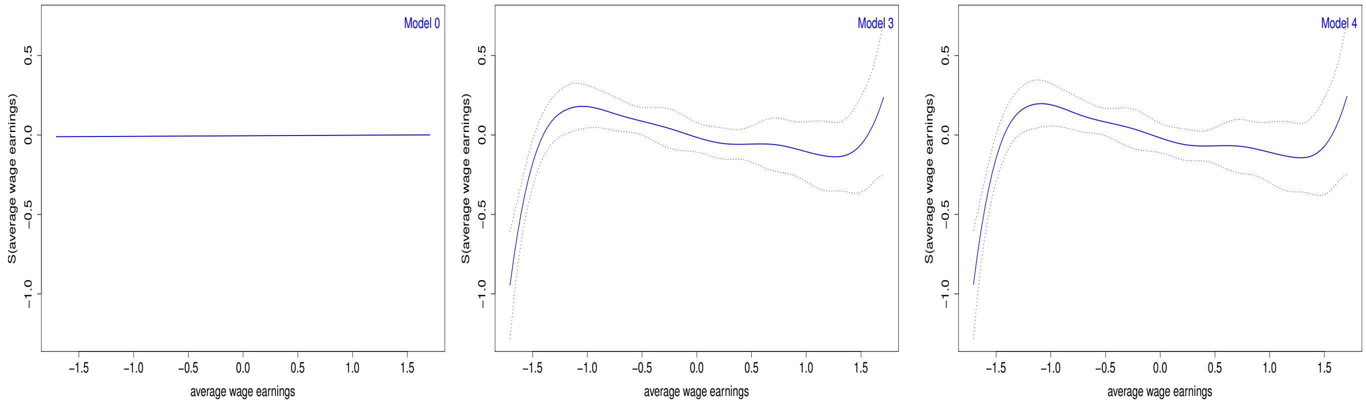
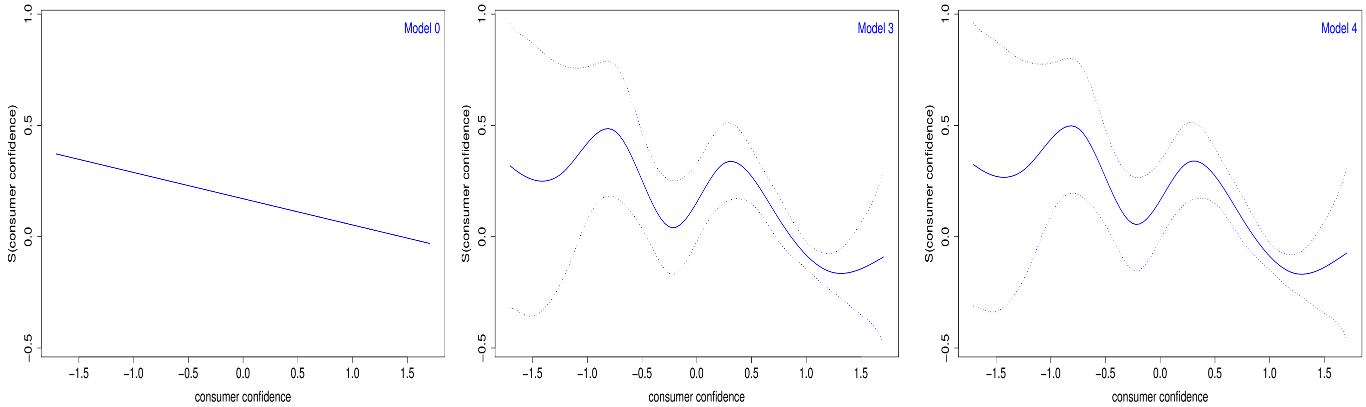


Figure A2: Marginal effects of number of *consumer confidence.*



Bibliography

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

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

Berg D. (2006) Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, **23**, 129–143.

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

Brezger A. and Kneib T. and Lang S. (2005) Bayesx: Analyzing bayesian structural additive regression models. *Journal of Statistical Software*, **14**, 1-22.

Brezger A. and Lang S. (2006) Generalized Structured Additive Regression Based on Bayesian P-Splines. *Computational Statistics and Data Analysis*, **50**, 947-991.

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

Crainiceanu C. and Ruppert D. and Wand M. P. (2005) Bayesian Analysis for Penalized Spline Regression Using WinBUGS *Journal of Statistical Software*, **14**.

Dakovica R. and Czadoa C. and Bergb D. (2010) Bankruptcy prediction in Norway: a comparison study. *Applied Economics Letters*, **17**, 1739–1746.

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

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. (2018) Dynamic survival models with varying coefficients for credit risks. *To appear.*

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.

Drexler M. and Ainsworth C. H. (2013) Generalized Additive Models Used to Predict Species Abundance in the Gulf of Mexico: An Ecosystem Modeling Tool. *PLoS ONE*, **8**, doi:10.1371/journal.pone.0064458

Fahrmeir L. and Lang L. (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society. Series C*, **50**, 201-220.

Friedman J. H. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19**, 1-67.

Gamerman D. (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57-68.

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

Kaplan E. L. and Meier P. (1958) Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*, **53**, 457-481.

Lang S. and Brezger A. (2004) Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183-212.

Liu W. and JP Morgan Chase and Vu C. and Acxiom and Cela J. (2009) Generalizations of Generalized Additive Model (GAM): A Case of Credit Risk Modeling.

Hastie T. J. and Tibshirani R. J. (1986). *Generalized Additive Models. Statistical Science*, **1**, 297-318.

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

Sapra A. K. (2013) Generalized additive models in business and economics. *International Journal of Advanced Statistics and Probability*, **1**, 64-81.

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

Therneau T. and Grambsch P. (2000). *Modeling Survival Data: Extending the Cox Model*. SpringerVerlag, New York.

Wood S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society. Series B*, **62**, 413-428.

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

Wood S. N. (2008) Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society. Series B*, **70**, 495–518.

Wood S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B*, **73**, 3-36.

Wood S. N. and Goude Y. and Shaw S. (2015) Generalized additive models for large datasets. *Journal of the Royal Statistical Society. Series C*, **64**, 139-155.

Wood S. N. and Li Z. and Shaddick G. and Augustin N. H. (2016) Generalized additive models for gigadata: modelling the UK black smoke network daily data. *Journal of the American Statistical Association*, **64**, 139-155.

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