SORT 33 (1) January-June 2009, 3-30

Modelling consumer credit risk via survival analysis

Ricardo Cao, Juan M. Vilar and Andrés Devia

Universidade da Coruña[*]

Abstract

Credit risk models are used by financial companies to evaluate in advance the insolvency risk caused by credits that enter into default. Many models for credit risk have been developed over the past few decades. In this paper, we focus on those models that can be formulated in terms of the probability of default by using survival analysis techniques. With this objective three different mechanisms are proposed based on the key idea of writing the default probability in terms of the conditional distribution function of the time to default. The first method is based on a Cox's regression model, the second approach uses generalized linear models under censoring and the third one is based on nonparametric kernel estimation, using the product-limit conditional distribution function estimator by Beran. The resulting nonparametric estimator of the default probability is proved to be consistent and asymptotically normal. An empirical study, based on modified real data, illustrates the three methods.

MSC: 62P20, 91B30, 62G05, 62N01

Keywords: Probability of default, Basel II, nonparametric regression, conditional survival function, generalized product-limit estimator.

1 Introduction

Determining the probability of default, *PD*, in consumer credits, loans and credit cards is one of the main problems to be addressed by banks, savings banks, savings cooperatives and other credit companies. This is a first step needed to compute the capital in risk of insolvency, when their clients do not pay their credits, which is called *default*. The risk coming from this type of situation is called *credit risk*, which has been the object of research since the middle of last century. The importance of credit risk, as part of

[*] Departamento de Matemáticas. Facultad de Informática. Universidade da Coruña. Campus de Elviña, s/n. A Coruña 15071, Spain

Received: November 2008

financial risk analysis, comes from the New Basel Capital Accord (Basel II), published in 1999 and revised in 2004 by the Basel Committee for Banking Supervision (BCBS). This accord consists of three parts, called pillars. They constitute a universal theoretical framework for the procedures to be followed by credit companies in order to guarantee minimal capital requirements, called *statistical provisions for insolvency* (SPI).

Pillar I of the new accord establishes the parameters that play some role in the credit risk of a financial company. These are the probability of default, *PD*, the exposition after default, *EAD*, and the loss given default, *LGD*. The quantitative methods that financial entities can use are those used for computing credit risk parameters and, more specifically, for computing *PD*. These are the standard method and the internal ratings based method (IRB). Thus, credit companies can elaborate and use their own credit qualification models and, by means of them, conclude the Basel implementation process, with their own estimations of SPI.

There is an extensive literature on quantitative methods for credit risk, since the classical *Z*-score model introduced by Altman (1968). Nowadays there exist plenty of approaches and perspectives for modelling credit risk starting from *PD*. Most of them have provided better predictive powers and classification error rates than Altman's discriminant model, for credit solicitors (*application scoring*), as well as for those who are already clients of the bank (*behavioural scoring*). This is the case of logistic regression models, artificial neural networks (*ANN*), support vector machines (*SVM*), as well as hybrid models, as mixtures of parametric models and *SVM*. For the reader interested in a more extended discussion on the evolution of these techniques over the past 30 years we mention the work by Altman and Saunders (1998), Saunders (1999), Crouhy et al. (2000), Hand (2001), Hamerle et al. (2003), Hanson and Schuermann (2004), Wang et al. (2005), and Chen et al. (2006).

The main aim of this paper is to introduce an alternative approach for modelling credit risk. More specifically, we will focus on estimating *PD* for consumer credits and personal credits using survival analysis techniques.

The idea of using survival analysis techniques for constructing credit risk models is not new. It started with the paper by Narain (1992) and, later, was developed by Carling et al. (1998), Stepanova and Thomas (2002), Roszbach (2003), Glennon and Nigro (2005), Allen and Rose (2006), Baba and Goko (2006), Malik and Thomas (2006) and Beran and Djaïdja (2007). A common feature of all these papers is that they use parametric or semiparametric regression techniques for modelling the time to default (*duration models*), including exponential models, Weibull models and Cox's proportional hazards models, which are very common in this literature. The model established for the time to default is then used for modelling *PD* or constructing the scoring discriminant function.

In this paper we propose a basic idea to estimate *PD*, which is performed by three different methods. The first one is based on Cox's proportional hazards model, the second one uses generalized linear models, while the third one consists in using a random design nonparametric regression model. In all the cases, some random right censoring mechanism appears in the model, so survival analysis techniques are natural tools to be used.

The conditional survival function used for modelling credit risk opens an interesting perspective to study default. Rather than looking at default or not, we look at the time to default, given credit information of clients (endogenous covariates) and considering the indicators for the economic cycle (exogenous covariates). Thus, the default risk is measured via the conditional distribution of the random variable time to default, T, given a vector of covariates, X. The variable T is not fully observable due to the censoring mechanism.

In practice, since the proportion of defaulted credits is small, the proportion of censored data is large, which may cause poor performance of the statistical methods. On the other hand, the sample size is typically very large. This alleviates somehow the problem of the large proportion of censoring.

In order to estimate empirically the conditional distribution function of the time to default, we use the generalized product-limit estimator by Beran (1981). This estimator has been extensively studied by Dabrowska (1987), Dabrowska (1989), González-Manteiga and Cadarso-Suárez (1994), Van Keilegom and Veraverbeke (1996), Iglesias-Pérez and González-Manteiga (1999), Li and Datta (2001), Van Keilegom et al. (2001) and Li and Van Keilegom (2002), among other authors.

The rest of the paper proceeds as follows. Section 2 presents some conditional functions, often used in survival analysis, and explains how they can be used for credit risk analysis. The estimation of the probability of default is considered in Section 3, under different models: Cox's proportional hazards model, generalized linear models and a nonparametric model. Special attention is given to the study of the theoretical properties of the nonparametric estimator for *PD*, denoted by $\widehat{PD}^{NPM}$. Its asymptotic bias and variance, as well as uniform consistency and asymptotic normality are stated in Section 4. An application to a real data set, with a brief discussion about the empirical results obtained, is presented in Section 5. Finally, Section 6 contains the proofs of the results included in Section 4.

2 Conditional survival analysis in credit risk

The use of survival analysis techniques to study credit risk, and more particularly to model *PD*, can be motivated via Figure 1. It presents three common situations that may occur in practice when a credit company observes the "lifetime" of a credit.

Let us consider the interval $[0,\tau]$ as the horizon of the study. Case (a) shows a credit with default before the endpoint of the time under study (τ). In this case, the lifetime of the credit, T, which is the time to default of the credit, is an observable variable. Cases (b) and (c) show two different situations. In both of them it is not possible to observe the time instant when a credit enters into default, which causes a lack of information coming

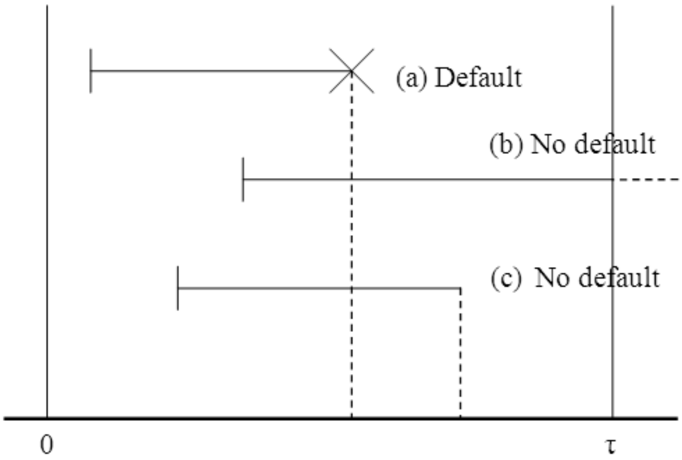

Figure 1: *Time to default in consumer credit risk.*

from right censoring. In case (b) it is only the time from the start of the credit to the end of the study, while (c) accounts for a situation where anticipated cancellation or the end of the credit occurs before default.

The available information to model the *PD* is a sample of n iid random variables $\{(Y_1,X_1,\delta_1),\ldots,(Y_n,X_n,\delta_n)\}$, of the random vector $\{Y,X,\delta\}$, where $Y=\min\{T,C\}$ is the observed maturity, T is the time to default, C is the time to the end of the study or anticipated cancellation of the credit, $\delta = I(T\le C)$ is the indicator of noncensoring (default) and X is a vector of explanatory covariates. In this survival analysis setting we will assume that there exists an unknown relationship between T and X. We will also assume that the random variables T and C are conditionally independent given X.

In the previous setup it is possible to characterize completely the conditional distribution of the random variable T using some common relations in survival analysis. Thus the conditional survival function, $S(t|x)$, the conditional hazard rate, $\lambda(t|x)$, the conditional cumulative hazard function, $\Lambda(t|x)$, and the conditional cumulative distribution function, $F(t|x)$, are related as follows:

$$S(t|x) = P(T>t|X=x) = \int_t^\infty f(u|x)du$$

$$\lambda(t|x) = \lim_{\Delta t\to 0} \frac{P(t\le T<t+\Delta t|T\ge t, X=x)}{\Delta t} = \frac{f(t|x)}{S(t|x)}$$

$$\Lambda(t|x) = \int_0^t \lambda(u|x)du = \int_0^t \frac{f(t|x)}{S(t|x)}du$$

$$S(t|x) = e^{-\Lambda(t|x)}$$

$$F(t|x) = 1 - S(t|x)$$

3 Probability of default in consumer portfolio

In the literature devoted to credit risk analysis there are not many publications on modelling the credit risk in consumer portfolios or personal credit portfolios. Most of the research deals with measuring credit risk by *PD* modelling in portfolios of small, medium and large companies, or even for financial companies. There exist, however, several exceptions. In the works by Carling et al. (1998), Stepanova and Thomas (2002) and Malik and Thomas (2006), the lifetime of a credit is modelled with a semiparametric regression model, more specifically with Cox's proportional hazards model.

In the following we present three different approaches to model the probability of default, *PD*, using conditional survival analysis. All the models are based on writing *PD* in terms of the conditional distribution function of the time to default. Thus *PD* can be estimated, using this formula, either by (i) Cox's proportional hazards model, where the estimator of the survival function is obtained by solving the partial likelihood equations in Cox's regression model, which gives $\widehat{PD}^{PHM}$, by (ii) a generalized linear model, with parameters estimated by the maximum likelihood method, which gives $\widehat{PD}^{GLM}$, or by (iii) using the nonparametric conditional distribution function estimator by Beran, which gives the nonparametric estimator of the default probability, $\widehat{PD}^{NPM}$.

3.1 Modelling the probability of default via the conditional distribution function

Following Basel II, credit scoring models are used to measure the probability of default in a time horizon $t+b$ from a maturity time, t. A typical value is $b = 12$ (in months). Thus, the following probability has to be computed:

$$\begin{aligned} PD(t|x) &= P(t \le T < t+b | T \ge t, X = x) \\ &= \frac{P(T < t+b | X = x) - P(T \le t | X = x)}{P(T \ge t | X = x)} \\ &= \frac{F(t+b|x) - F(t|x)}{1 - F(t|x)} = 1 - \frac{S(t+b|x)}{S(t|x)} \end{aligned} \tag{1}$$

where t is the observed maturity for the credit and x is the value of the covariate vector, X, for that credit.

3.2 Proportional hazards model

In this section, a semiparametric approach to perform the study of *PD* is given. Here we use Cox's proportional hazards approach to model the conditional survival function $S(t|x)$. The key in this method rests on the estimation of the cumulative conditional hazard function, $\Lambda(t|x)$, using maximum likelihood.

We follow the idea introduced by Narain (1992) for the estimation of $S(t|x)$, but we apply it in the definition of *PD*, as we have stated above in formula (1). The objective is to build a conditional model for the individual $PD(t|x)$, which is defined in terms of $\Lambda(t|x)$. In order to describe $\widehat{PD}^{PHM}$, we define the following expressions relative to Cox's regression theory.

The estimator of the conditional hazard rate function is defined as:

$$\hat{\lambda}(t|x) = \hat{\lambda}_0(t) \exp\left(x^{\mathsf{T}} \hat{\boldsymbol{\beta}}\right),$$

where $\hat{\lambda}_0(t)$ is an estimator of the hazard rate baseline function, $\lambda_0(t)$, and $\hat{\boldsymbol{\beta}}$ is an estimator of the parameter vector, $\boldsymbol{\beta}$.

Thus, under the assumption of a proportional hazards model, *PD* is estimated by:

$$\widehat{PD}^{PHM}(t|x) = \frac{\hat{F}_{\hat{\boldsymbol{\beta}}}(t+b|x) - \hat{F}_{\hat{\boldsymbol{\beta}}}(t|x)}{1 - \hat{F}_{\hat{\boldsymbol{\beta}}}(t|x)} = 1 - \frac{\hat{S}_{\hat{\boldsymbol{\beta}}}(t+b|x)}{\hat{S}_{\hat{\boldsymbol{\beta}}}(t|x)}, \tag{2}$$

where

$$1 - \hat{F}_{\hat{\boldsymbol{\beta}}}(t|x) = \hat{S}_{\hat{\boldsymbol{\beta}}}(t|x) = \exp\left(-\hat{\Lambda}(t|x)\right)$$

The estimation method for this model consists of two steps. In the first step the cumulative baseline hazard function, $\Lambda_0(t)$, is estimated by:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{1\{Y_i \le t, \delta_i = 1\}}{\sum_{j=1}^{n} 1\{Y_j \ge Y_i\}},$$

then the parameter $\boldsymbol{\beta}$ is estimated by

$$\hat{\boldsymbol{\beta}}^{PHM} = \arg\max_{\beta} L(\boldsymbol{\beta}),$$

where the partial likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\left(x_i^{\mathsf{T}} \boldsymbol{\beta}\right)}{\left(\sum_{j=1}^{n} 1_{\{Y_j > Y_i\}} \exp\left(x_j^{\mathsf{T}} \boldsymbol{\beta}\right)\right)}$$

Thus, the conditional cumulative hazard function estimator is given by

$$\hat{\Lambda}(t|x) = \int_0^t \hat{\lambda}(s|x)ds = \exp\left(x^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{PHM}\right)\hat{\Lambda}_0(t).$$

The asymptotic properties of this estimator can be found, for instance, in the book by Fleming and Harrington (1991). As a consequence of these, similar properties can be obtained for the estimator of the default probability defined in (2).

Remark 1 *Narain (1992) and many other authors defined the probability of default as the complement of the conditional survival function evaluated at the forecast horizon, $1 - S(t+b|x)$. According to this, the formulation by Narain does not take into account the fact that the credit should not be into default at maturity t.*

3.3 Generalized linear model

A generalized linear model can be assumed for the lifetime distribution:

$$P(T \leq t|X = x) = F_\theta(t|x) = g\left(\theta_0 + \theta_1 t + \theta^{\mathsf{T}} x\right),$$

where $\theta = (\theta_2, \theta_3, \ldots, \theta_{p+1})^{\mathsf{T}}$ is a p-dimensional vector and g is a known link function, like the logistic or the probit function. Thus, this model characterizes the conditional distribution of the lifetime of a credit, T, in terms of the unknown parameters. Once this parameters are estimated, an estimator of the conditional distribution function is obtained, $F_{\hat{\boldsymbol{\theta}}}$, and, finally, an estimator of PD can be computed by plugging this estimator in equation (1), i.e.

$$\widehat{PD}^{GLM}(t|x) = \frac{F_{\hat{\boldsymbol{\theta}}}(t+b|x) - F_{\hat{\boldsymbol{\theta}}}(t|x)}{1 - F_{\hat{\boldsymbol{\theta}}}(t|x)} = 1 - \frac{S_{\hat{\boldsymbol{\theta}}}(t+b|x)}{S_{\hat{\boldsymbol{\theta}}}(t|x)},$$

where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{GML}$ is the maximum likelihood estimator of the parameter vector.

Let us consider the one-dimensional covariate case. Then $\theta = \theta_2$ and the conditional distribution given by the model is $F(t|x) = g(\theta_0 + \theta_1 t + \theta_2 x)$, with density $f(t|x) = \theta_1 g'(\theta_0 + \theta_1 t + \theta_2 x)$. Since we are given a random right censored sample, the conditional likelihood function is a product of terms involving the conditional density, for the uncensored data, and the conditional survival function, for the censored data:

$$L(Y, X, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i|X_i)^{\delta_i}\left(1 - F(Y_i|X_i)\right)^{1-\delta_i},$$

where Y_i is the maturity of the i-th credit and δ_i is the indicator of default for the i-th credit. Thus, the log-likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \ln(L(Y,X,\boldsymbol{\theta})) = \sum_{i=1}^{n}\left[\delta_i \ln(f(Y_i|X_i)) + (1-\delta_i)\ln(1-F(Y_i|X_i))\right] \\ &= \sum_{i=1}^{n}\left[\delta_i \ln(\theta_1 g'(\theta_0+\theta_1 Y_i+\theta_2 X_i)) + (1-\delta_i)\ln(1-g(\theta_0+\theta_1 Y_i+\theta_2 X_i))\right] \\ &= \sum_{i=1}^{n}\delta_i\left[\ln(\theta_1)+\ln(g'(\theta_0+\theta_1 Y_i+\theta_2 X_i))\right] + \sum_{i=1}^{n}(1-\delta_i)\ln(1-g(\theta_0+\theta_1 Y_i+\theta_2 X_i))\end{aligned}$$

Finally, the estimator is found as the maximizer of the log-likelihood function:

$$\hat{\boldsymbol{\theta}}^{GML} = \arg\max_{\theta}\, \ell(\boldsymbol{\theta}).$$

The works by Jorgensen (1983) and McCullagh and Nelder (1989) deal with generalized linear models in a regression context. These models can be adapted to the conditional distribution function setup.

3.4 Nonparametric conditional distribution estimator

First of all a nonparametric estimator of the conditional distribution function is obtained. This estimator, say $\hat{S}_h(t|x)$, is used to derive an estimator of $PD(t|x)$, say $\widehat{PD}^{NPM}(t|x)$, for the desired values of t and x.

Since we have a sample of right censored data for the lifetime distribution of a credit, we use the estimator proposed by Beran (1981) for the conditional survival function of T given $X = x$:

$$\hat{S}_h(t|x) = \prod_{i=1}^{n}\left(1 - \frac{1_{\{Y_i \le t, \delta_i = 1\}} B_{ni}(x)}{1-\sum_{j=1}^{n} 1_{\{Y_j < Y_i\}} B_{nj}(x)}\right),$$

where Y_i is the observed lifetime of the i-th credit, δ_i is the indicator of observing default of the i-th credit (uncensoring) and X_i is the vector of explanatory covariates for the i-th credit. The terms $B_{ni}(x)$ are Nadaraya-Watson nonparametric weights:

$$B_{ni}(x) = \frac{K((x-X_i)/h)}{\sum_{j=1}^{n} K((x-X_j)/h)},\ 1 \le i \le n,$$

and $h \equiv h_n$ is the smoothing parameter that tends to zero as the sample size tends to infinity.

To estimate the probability of default at time t given a covariate vector x, we replace, in (1), the theoretical value of the conditional survival function by its estimator $\hat{S}_h$:

$$\widehat{PD}^{NPM}(t|x) = \frac{\hat{F}_h(t+b|x) - \hat{F}_h(t|x)}{1-\hat{F}_h(t|x)} = 1 - \frac{\hat{S}_h(t+b|x)}{\hat{S}_h(t|x)} \tag{3}$$

The asymptotic properties of this estimator will be studied in the next section.

4 Asymptotic results for the nonparametric approach

The asymptotic properties for the nonparametric estimator of the default probability, $\widehat{PD}^{NPM}$, have been obtained from the analogous properties for the conditional distribution function estimator under censoring, already obtained by Dabrowska (1989), Iglesias-Pérez and González-Manteiga (1999), Van Keilegom and Veraverbeke (1996) and Van Keilegom et al. (2001).

Using equation (3) the asymptotic bias, variance and mean squared error of the estimator $\widehat{PD}^{NPM}$ can be obtained via some expansions. Consistency and asymptotic normality can also be derived.

To simplify our notation, let us define $\varphi(t|x) = PD(t|x)$ and $\hat{\varphi}_n(t|x) = \widehat{PD}^{NPM}(t|x)$. Then, the nonparametric estimator of the default probability function is

$$\hat{\varphi}_n(t|x) = 1 - \frac{\hat{S}_h(t+b|x)}{\hat{S}_h(t|x)}. \tag{4}$$

Before stating the asymptotic results concerning $\hat{\varphi}_n$ we need to present some definitions and assumptions. Most of these assumptions were already required by Iglesias-Pérez and González-Manteiga (1999) and Dabrowska (1989) to obtain their results.

The function $G(t|x) = P(C \leq t|X = x)$ is the conditional distribution of the censoring random variable given the covariate X and $H(t|x) = P(Y \leq t|X = x)$ is the conditional distribution of the observed lifetime of the credit given the covariate X. The random lifetime, T, and the censoring time, C, are conditionally independent given the covariate X. As a consequence, $1-H(t|x) = (1-F(t|x))(1-G(t|x))$. The conditional subdistribution function of the observed lifetime for default credits is denoted by $H_1(t|x) = P(Y \leq t, \delta = 1|X = x) = \int_0^t (1-G(u|x))dF(u|x)$. Similarly, for nondefaulted credits, $H_0(t|x) = P(Y \leq t, \delta = 0|X = x) = \int_0^t (1-F(u|x))dG(u|x)$. The distribution function and the density function of the covariate X are denoted by $M(x)$ and $m(x)$. The set $\Omega_X = \{x \in \mathbb{R}^+ : m(x) > 0\}$ will denote the support of m. The lower and upper endpoints of the support of any distribution function L will be denoted by $\underline{\tau}_L = \inf\{t : L(t) > 0\}$ and $\overline{\tau}_L = \sup\{t : L(t) < 1\}$.

The following assumptions are needed for the asymptotic results.

A.1 The kernel K is a symmetric density function with support $[-1,1]$ and bounded variation.

A.2 Let us consider Ω_X, the support of the density m, and let $I = [x_1, x_2]$ be an interval contained in Ω_X, such that there exist $\alpha, \beta, \delta > 0$ with $\alpha\delta \leq \beta\delta < 1$,

$$\alpha \leq \inf\left\{m(x) : x \in I_\delta\right\} \leq \sup\left\{m(x) : x \in I_\delta\right\} \leq \beta,$$

where $I_\delta = [x_1 - \delta, x_2 + \delta]$. Then the functions $m'(x)$ and $m''(x)$ are continuous and bounded in the set I_δ.

A.3 There exist positive real numbers θ and τ_H^*, such that

$$0 < \theta \leq \inf_{0 \leq t \leq \tau_H^*} \left\{1 - H(t|x) : x \in I_\delta\right\}$$

A.4 The functions $H'(t|x) = \frac{\partial H(t|x)}{\partial x}$, $H''(t|x) = \frac{\partial^2 H(t|x)}{\partial x^2}$, $H_1'(t|x) = \frac{\partial H_1(t|x)}{\partial x}$ and $H_1''(t|x) = \frac{\partial^2 H_1(t|x)}{\partial x^2}$ exist, are continuous and bounded in $(t,x) \in [0, +\infty) \times I_\delta$.

A.5 The functions $\dot{H}(t|x) = \frac{\partial H(t|x)}{\partial t}$, $\ddot{H}(t|x) = \frac{\partial^2 H(t|x)}{\partial t^2}$, $\dot{H}_1(t|x) = \frac{\partial H_1(t|x)}{\partial t}$, $\ddot{H}_1(t|x) = \frac{\partial^2 H_1(t|x)}{\partial t^2}$ exist, are continuous and bounded in $(t,x) \in [0, \tau_H^*) \times I_\delta$.

A.6 The functions $\dot{H}'(t|x) = \frac{\partial^2 H(t|x)}{\partial x \partial t} = \frac{\partial^2 H(t|x)}{\partial t \partial x}$, $\dot{H}_1'(t|x) = \frac{\partial^2 H_1(t|x)}{\partial x \partial t} = \frac{\partial^2 H_1(t|x)}{\partial t \partial x}$ exist, are continuous and bounded in $(t,x) \in [0, \tau_H^*) \times I_\delta$.

A.7 The smoothing parameter h satisfies $h \longrightarrow 0$, $(\ln n)^3/nh \longrightarrow 0$ and $nh_n^5/\ln n = O(1)$, when $n \longrightarrow \infty$.

The consistency and asymptotic normality of the nonparametric estimator $\widehat{\varphi}_n$ are stated in the next two theorems. The proofs of these results are given in Section 6.

Theorem 1 *Fix some t and x for which $0 < \varphi(t|x) < 1$. Under the assumptions A.1-A.7, $\hat{\varphi}_n(t|x)$ is a strongly consistent estimator of the default probability function, $\varphi(t|x)$. Moreover if $b < \tau_H^*$ and $\inf_{x \in I} S(\tau_H^*|x) > 0$, the consistency is uniform in $(t,x) \in [0, \tau_H^* - b] \times I$, i.e.*

$$\sup_{t \in \left[0, \tau_H^* - b\right]} \sup_{x \in I} |\hat{\varphi}_n(t|x) - \varphi(t|x)| \longrightarrow 0 \; a.s.$$

Theorem 2 *Assume conditions A.1-A.7. Then the mean squared error of the nonparametric estimator for the default probability is*

$$MSE\left(\hat{\varphi}_n(t|x)\right) = b(t|x)^2 h^4 + \frac{1}{nh} v(t|x) + o\left(h^4 + \frac{1}{nh}\right), \tag{5}$$

where

$$b(t|x) = \frac{1}{2}c_K\left(1-\varphi(t|x)\right)B_H\left(t,t+b|x\right), \tag{6}$$

$$v(t|x) = \frac{d_K D_H\left(t,t+b|x\right)}{m\left(x\right)}\left(1-\varphi(t|x)\right)^2, \tag{7}$$

$c_K = \int K\left(u\right)^2 du$, $d_K = \int u^2 K\left(u\right) du$,

$$\begin{aligned} B_H\left(t,t+b|x\right) &= \int_t^{t+b}\left[\ddot{H}(s|x)+2\frac{m'\left(x\right)}{m\left(x\right)}\dot{H}(s|x)\right]dH_1(s|x) \\ &\quad+\left(1+2\frac{m'\left(x\right)}{m\left(x\right)}\right)\int_t^{t+b}\frac{d\dot{H}_1(s|x)}{1-H(t|x)}, \end{aligned} \tag{8}$$

$$D_H\left(t|x\right) = \int_0^t \frac{dH_1\left(s|x\right)}{\left(1-H\left(s|x\right)\right)^2}. \tag{9}$$

Furthermore, if $nh^5 \to c \in (0,\infty)$*, the limit distribution of* $\hat{\varphi}_n(t|x)$ *is given by*

$$\sqrt{nh}\left(\hat{\varphi}_n(t|x)-\varphi\left(t|x\right)\right) \xrightarrow{d} N\left(c^{1/2}b(t|x), v(t|x)\right).$$

Remark 2 *As a consequence, the bandwidth that minimizes the dominant terms of the MSE in (5) is*

$$h_0 = \left(\frac{v(t|x)}{4b(t|x)^2}\right)^{1/5} n^{-1/5}. \tag{10}$$

5 Application to real data

In this section we apply the estimation methods given in Section 3 to a real data set. Our goal is to show the results obtained from the application of the three models to the estimation of default probabilities in a sample of consumer loans. An empirical comparison between the models is given through the descriptive statistics as well as the estimated default rate curves. In all cases, the curves were constructed taking into account the recommendations from the Basel statements, i.e., *PD* estimates with maturity of one year forward.

The data consist of a sample of 25 000 consumer loans from a Spanish bank registered between July 2004 and November 2006. To preserve confidentiality, the data were selected in order to provide a large distortion in the parameters describing the true solvency situation of the bank.

The sample represents two populations, non-defaulted loans and defaulted loans, where the observed cumulative default rate was 7.2%. The variables treated here are the following:

Y = maturity or loan lifetime. Here, maturity means time to default (T), when time is uncensored or time to withdrawal (C), in any other case. Time was measured in months.

X = scoring (credit ratio) observed for each debtor. Its range lies inside the interval $[0, 100]$. In this paper, X is an univariate endogenous measure of propensity to default. The closer to zero the better the credit behaviour.

δ = default indicator (uncensoring indicator). This variable takes value 1 if loan defaults or 0 if not.

Figure 2 shows that the scoring characteristics of debtors are clearly different in the two groups (defaulted and non-defaulted). The moment-based characteristics like the kurtosis (2.68 and 4.29) and the skewness (0.51 and 1.37) of these two subsamples are very different each other and they also reflect non-normal distributions. A large proportion (about 75%) of debtors belonging to the sample of non-defaulted loans show a credit scoring varying between 0.0 and 11.07. This whole range is below the first quartile (approximately 20.93) of the scoring in the group of defaulted loans.

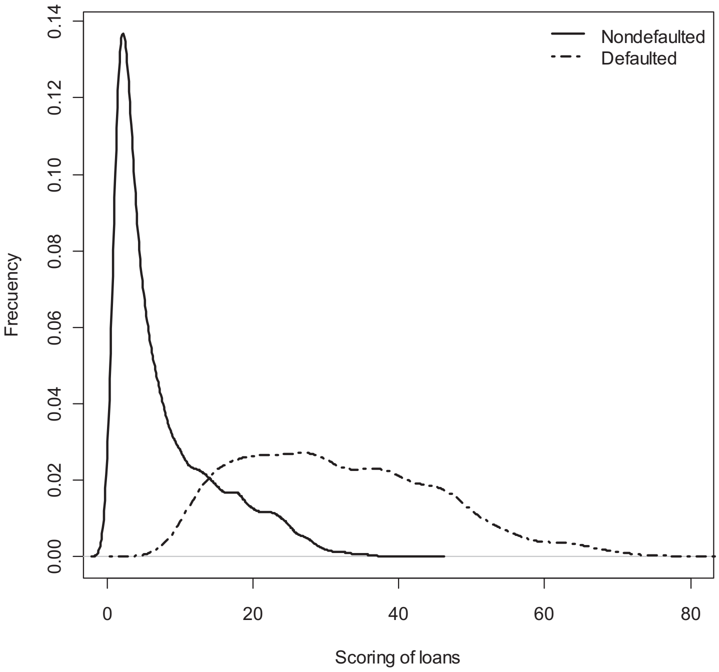

Figure 2: *Kernel density estimates for defaulted and non-defaulted loans.*

Table 1: *Descriptive statistics for maturity and covariate (X) in defaulted loans (DL), non-defaulted loans (NDL) and aggregated loans (AL).*

Sample		min.	1st. Q.	median	mean	3rd. Q.	max.
DL	maturity (T)	0.033	2.933	5.500	7.458	11.15	24.767
	X	8.398	20.295	30.066	31.817	41.167	77.819
NDL	maturity (C)	0.000	6.767	11.367	13.455	20.033	29.500
	X	0.150	2.412	4.857	7.688	11.070	43.920
AL	maturity (Y)	0.000	6.500	10.870	13.020	19.570	29.500
	X	0.150	2.540	5.440	9.425	13.405	77.819

The data show that the random variable X is a reasonable predictor to study loan default. This is also evident when observing the descriptive statistics for both groups of loans in Table 1.

Figure 3 shows curves for the empirical default rates obtained directly from the sample. These curves can be thought as the result of a naïve nonparametric estimator for the unconditional default rates curves. The study of this estimator is not the goal of this paper. Focusing the attention in the right panel in Figure 3, it is clear that the unconditional estimates of *PD* become constant when the loan maturity gets large. Naive approximations to *PD* do not behave well because of the lack of sensitivity to variations in the scoring characteristics of debtors. This result show that the unconditional approach may not be used to predict *PD* because the estimation accuracy on the right tail seems to be poor. This fact motivates the use of the conditional framework to obtain more realistic estimations for *PD*.

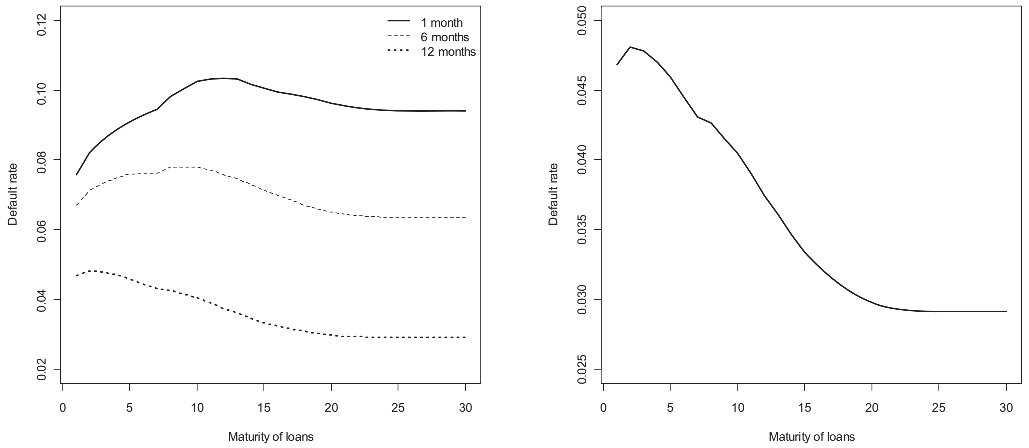

Figure 3: *Empirical default rates with different forecasting periods. Left panel shows default rates curves for 1, 6, and 12 months forward horizons, while the right panel shows the particular case of a default rate curve with 1 year forward horizon, which is a very common decision tool in credit risk analysis.*

5.1 Empirical results and discussion

The plots included in this section give a graphical description of the estimators proposed in this paper concerned with the conditional approach in consumer credit risk. All these results show that a reasonable improvement can be achieved when a survival analysis approach is used to model the credit quality in terms of “lifetime of loans”.

5.1.1 Results for the proportional hazards model

Estimating the *PD* under the proportional hazards model presents clear differences with the results for the unconditional setting (Figure 3). It is easy to see that a clear disadvantage of using an unconditional approach is that the *PD* forecasts do not change with X. The conditional approach gives more realistic estimates using the scoring information, which is a reasonable covariate, as was established at the beginning of this section. The covariate X explains the propensity to defaults in loan obligations. Figure 4 shows that the *PD* estimates increase as the customer scoring increases.

A careful look at Figure 4 shows that the estimator of *PD* is zero when the time to default gets close to the maximum of the maturity observed in the sample (approximately 25). This effect on the *PD* curve is due to the heavy censoring and the lack of depth in the sample. As a consequence, the accuracy of the estimator at the right tail of *PD* is poor. Nevertheless, Cox’s proportional hazards model gives more realistic default probabilities than the unconditional approximation (see Figure 3) when the bank previously knows the scoring characteristics of the portfolio customers.

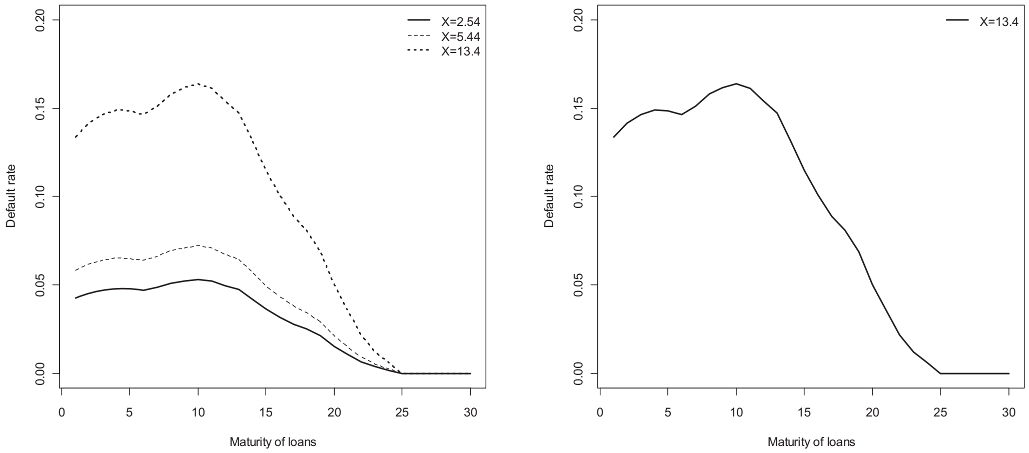

Figure 4: *$\widehat{PD}$ with maturity 1 year forward given $X = 2.54,\ 5.44,\ 13.4$ (left panel) and given the mean value of X (right panel).*

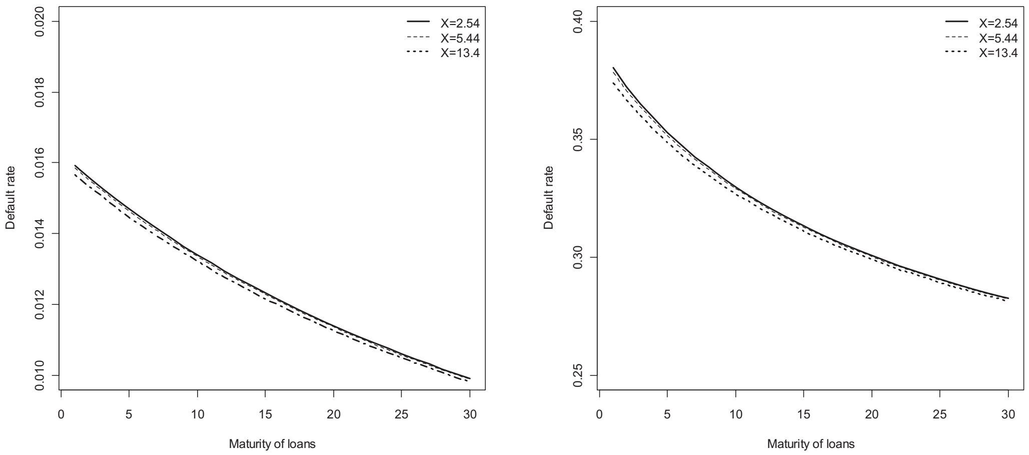

Figure 5: *$\widehat{PD}$ estimated with the Pareto (left panel) and Snedecor's F (right panel) link function.*

5.1.2 Results for the generalized linear model

Figure 5 show the results obtained for the *PD* estimated with the *GLM* model using two parametric links: Pareto and Snedecor's F. The range of the estimated *PD* lies within the interval $[0.0, 0.016]$ when the link function is Pareto and grows up to the interval $[0.0, 0.378]$ when the link function is $F_{10,50}$, as it can be seen in Table 2. The *PD* curves obtained with this model are exponentially decreasing, as expected, but in this case it seems that there is no a significantly contribution of the variable X in the accuracy of the estimated default probability curves. Furthermore, the estimated curves are all above the range of the observed default rate with maturity one year forward. The results achieved by using these two parametric links do not fit as well as expected to the data, when compared to the empirical default rate curves depicted in Figure 3. In spite of this, the *GLM* method may be useful to study the *PD* horizon in the long run.

Other link distributions belonging to the exponential family have been used to fit these data via *GLM*. The normal distribution, the Weibull distribution and the Cauchy distribution were used, among others. The results obtained were even worse than those presented in Figure 5 above.

5.1.3 Results for the nonparametric estimator

The results for the nonparametric method presented in (3) are collected in this subsection. In practice, we have used a k-nearest neighbour (*KNN*) type of bandwidth, which consists in fixing some positive integer k and define the parameter as follows:

$$h = h^{KNN}(x) = d(x, X_{[k]})$$

where $d(x, X_{[k]})$ is the k-th order statistic of the sample of distances from x to those X_i with $\delta_i = 1$. In other terms $h^{KNN}(x)$ is the k-th smallest distance from x to the uncensored observations of the X sample.

Figures 6-7 show the behaviour of the nonparametric estimator introduced in Section 3. In Figure 6 the value of the number of nearest neighbours has remained fixed ($k = 100$) and the estimator $\widehat{PD}(t|x)$ has been computed for three different values of X ($x = 2.54$, 5.44, 13.4). The reverse situation is showed in Figure 7, i.e., the curves $\widehat{PD}(t|x)$ were obtained for two fixed values of X ($x = 9.43$, 20) and varying the number of nearest neighbours ($k = 100$, 300, 500).

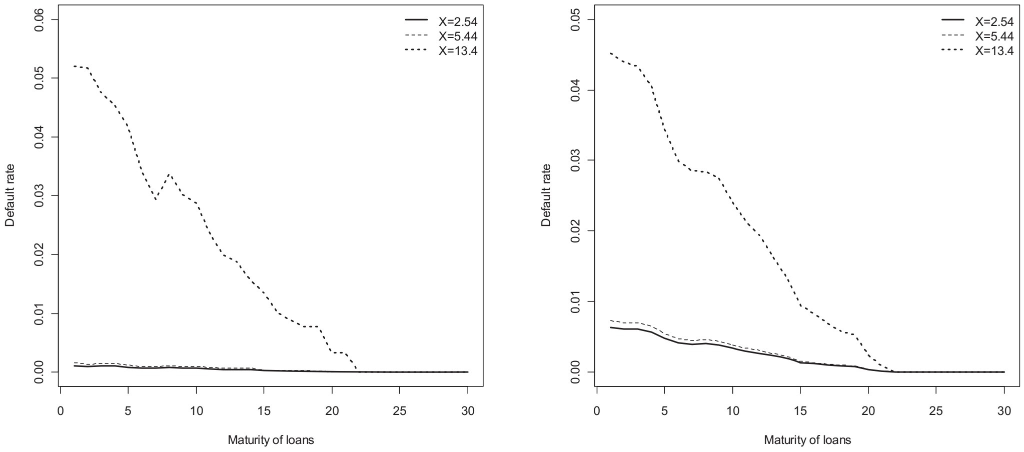

Figure 6: *$\widehat{PD}$ with fixed bandwidth parameter $k = 100$ (left panel) and $k = 400$ (right panel) given three scoring values.*

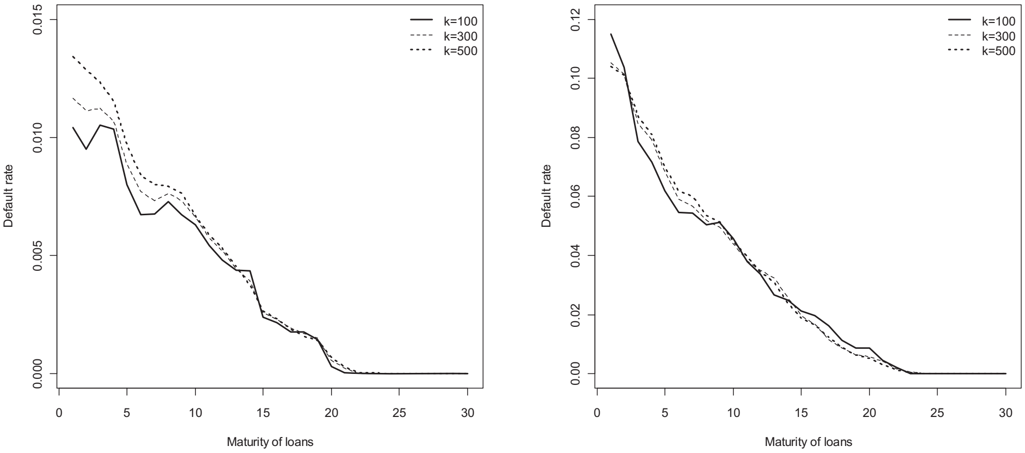

Figure 7: *$\widehat{PD}$ with three different bandwidth parameters, given $X = 9.43$ (left panel) and given $X = 20$ (right panel).*

The first two curves show much smaller values for PD when the values of X are close to or below the first quartile of the distribution. For $k = 100$ (see Figure 6, left panel) there is an apparent undersmoothing effect for the estimated default probability curve. The situation improves in the right panel of Figure 6. There, since $k = 400$, the $\widehat{PD}$ is much smoother. The estimates of the PD have a large sensitivity to small changes in the scoring variable. As a consequence the PD can be overestimated at the beginning of loan lifetime. A possible reason for this is the heavy censoring that usually affects consumer credit loans.

Table 2: *Descriptive statistics for the empirical default rates (EDR) and for the PD estimates obtained by Cox's proportional hazards model (PHM), the generalized linear model (GLM) and the nonparametric model (NPM).*

Model			min.	1st. Q.	median	mean	3rd. Q.	max.
		Maturity (months)						
		1	0.0758	0.0940	0.0947	0.0952	0.0994	0.1033
EDR		6	0.0636	0.0638	0.0692	0.0698	0.0755	0.0779
		12	0.0292	0.0292	0.0329	0.0359	0.0424	0.0481
		x						
		2.54	0.0000	0.0045	0.0343	0.0286	0.0138	0.0159
PHM		5.44	0.0000	0.0062	0.0467	0.0389	0.0138	0.0159
		13.4	0.0000	0.0147	0.1080	0.0891	0.0136	0.0157
	Link	x						
		2.54	0.0099	0.0110	0.0122	0.0125	0.0139	0.0159
	Pareto	5.44	0.0010	0.0109	0.0122	0.0124	0.0138	0.0159
GLM		13.4	0.0098	0.0108	0.0121	0.0123	0.0136	0.0157
		2.54	0.2826	0.2950	0.3120	0.3183	0.3370	0.4468
	$F_{10,50}$	5.44	0.2823	0.2946	0.3114	0.3176	0.3361	0.3784
		13.4	0.2815	0.2935	0.3098	0.3156	0.3336	0.3738
	k	x						
	100	2.54	0.0000	0.0000	0.0002	0.0004	0.0007	0.0012
	100	5.44	0.0000	0.0000	0.0003	0.0005	0.0010	0.0015
	100	13.4	0.0000	0.0000	0.0118	0.0175	0.0300	0.0520
	400	2.54	0.0000	0.0000	0.0012	0.0021	0.0039	0.0064
	400	5.44	0.0000	0.0001	0.0014	0.0024	0.0045	0.0073
NPM	400	13.4	0.0000	0.0001	0.0089	0.0152	0.0282	0.0452
	100	9.43	0.0000	0.0000	0.0023	0.0037	0.0067	0.0105
	300	9.43	0.0000	0.0000	0.0024	0.0040	0.0073	0.0117
	500	9.43	0.0000	0.0001	0.0025	0.0042	0.0079	0.0134
	100	20	0.0000	0.0005	0.0205	0.0301	0.0509	0.1149
	300	20	0.0000	0.0009	0.0183	0.0302	0.0514	0.1054
	500	20	0.0000	0.0006	0.0177	0.0306	0.0531	0.1040

Figure 7 includes the default probability conditional to just a single value of X, using three different levels of smoothness. Visual inspection of Figure 7 shows that, for a fixed bandwidth, the larger the scoring, the smoother the estimated *PD* curve. It is also clear that the variability of the *PD* reduces when the scoring gets large.

5.1.4 Comparison

A summary with a descriptive comparison of the three models is given in Table 2. Fixed values for the covariate X (first, second and third quartiles) were used for the conditional distributions. Of course, the empirical default rate does not depend on the value of X.

Although no goodness-of-fit tests have been applied for the proposed models, the results of the estimation can be checked by simple inspection of Figures 4–7 and the descriptive statistics collected in Table 2. The results for each model can be compared with those of the aggregated default rates in the whole portfolio. Such values should be considered as a reference value for the three models.

6 Proofs

Proof of Theorem 1

Recall equations (1) and (4). Let us write

$$\begin{aligned} \varphi(t|x) &= 1-\frac{P}{Q}, \\ \hat{\varphi}_n(t|x) &= 1-\frac{\hat{P}}{\hat{Q}}, \end{aligned}$$

with $P = S(t+b|x)$, $Q = S(t|x)$, $\hat{P} = \widehat{S}_h(t+b|x)$ and $\hat{Q} = \widehat{S}_h(t|x)$. Using Theorem 2 in Iglesias-Pérez and González-Manteiga (1999) we have

$$\left(\hat{P}, \hat{Q}\right) \longrightarrow (P, Q) \text{ a.s.}$$

Since the function $g(x,y) = \frac{x}{y}$ is continuous in (P,Q), then we obtain

$$\hat{\varphi}_n(t|x) \longrightarrow \varphi(t|x) \text{ a.s.}$$

and the first part of the proof is concluded.

For the second part of the proof we use Corollary 2.1 in Dabrowska (1989) to obtain

$$\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}\left|\hat{S}_h\left(t+b|x\right)-S\left(t+b|x\right)\right| \to 0 \text{ a.s.} \tag{11}$$

$$\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}\left|\hat{S}_h\left(t|x\right)-S\left(t|x\right)\right| \to 0 \text{ a.s.} \tag{12}$$

We now use the following identity:

$$\frac{1}{z}=1-(z-1)+\cdots+(-1)^p(z-1)^p+(-1)^{p+1}\frac{(z-1)^{p+1}}{z}, \tag{13}$$

that is valid for any $p\in\mathbb{N}$. Applying (13) with $p=1$ and $\frac{1}{z}=\frac{Q}{\hat{Q}}$ we obtain:

$$\begin{aligned}1-\hat{\varphi}_n(t|x) &= \frac{\hat{P}}{\hat{Q}}=\frac{\hat{P}}{Q}\frac{Q}{\hat{Q}}\\ &= \frac{\hat{P}}{Q}\left[1-\left(\frac{\hat{Q}}{Q}-1\right)+\frac{Q}{\hat{Q}}\left(\frac{\hat{Q}}{Q}-1\right)^2\right]\\ &= \frac{\hat{P}}{Q}-\frac{\hat{P}\left(\hat{Q}-Q\right)}{Q^2}+\frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-Q\right)^2}{Q^2},\end{aligned}$$

thus

$$|(1-\hat{\varphi}_n(t|x))-(1-\varphi(t|x))|\leq A_1+A_2+A_3 \tag{14}$$

where

$$\begin{aligned}A_1 &= \frac{\left|\hat{P}-P\right|}{Q},\\ A_2 &= \frac{\hat{P}\left|\hat{Q}-Q\right|}{Q^2},\\ A_3 &= \frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-Q\right)^2}{Q^2}.\end{aligned}$$

On the other hand if $x\in I$ and $t\leq\tau_H^*-b$,

$$A_1\leq\frac{\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}\left|\hat{S}_h\left(t+b|x\right)-S\left(t+b|x\right)\right|}{\inf_{x\in I}S(\tau_H^*|x)}, \tag{15}$$

$$A_2 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*-b\right]} \sup\limits_{x\in I} \left|\hat{S}_h(t|x) - S(t|x)\right|}{\inf\limits_{x\in I} S(\tau_H^*|x)^2}, \tag{16}$$

$$A_3 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*-b\right]} \sup\limits_{x\in I} \left|\hat{S}_h(t|x) - S(t|x)\right|^2}{\inf\limits_{x\in I} S(\tau_H^*|x)^2}. \tag{17}$$

Finally using (11) and (12) in (15), (16) and (17), equation (14) gives

$$\sup_{t\in\left[0,\tau_H^*-b\right]} \sup_{x\in I} |\hat{\varphi}_n(t|x) - \varphi(t|x)| \to 0 \text{ a.s.}$$

and the proof is concluded.

Proof of Theorem 2

To study the bias, we use (13) for $p = 1$ and $\frac{1}{z} = \frac{E(\hat{Q})}{\hat{Q}}$ to obtain:

$$\begin{aligned} 1-\hat{\varphi}_n(t|x) &= \frac{\hat{P}}{\hat{Q}} = \frac{\hat{P}}{E(\hat{Q})}\frac{E(\hat{Q})}{\hat{Q}} \\ &= \frac{\hat{P}}{E(\hat{Q})}\left[1-\left(\frac{\hat{Q}}{E(\hat{Q})}-1\right)+\frac{E(\hat{Q})}{\hat{Q}}\left(\frac{\hat{Q}}{E(\hat{Q})}-1\right)^2\right] \\ &= \frac{\hat{P}}{E(\hat{Q})} - \frac{\hat{P}(\hat{Q}-E(\hat{Q}))}{(E(\hat{Q}))^2} + \frac{\hat{P}}{\hat{Q}}\frac{(\hat{Q}-E(\hat{Q}))^2}{(E(\hat{Q}))^2}. \end{aligned} \tag{18}$$

As a consequence

$$E(1-\hat{\varphi}_n(t|x)) = A_1 + A_2 + A_3, \tag{19}$$

with

$$A_1 = \frac{E(\hat{P})}{E(\hat{Q})}, \tag{20}$$

$$A_2 = -\frac{Cov(\hat{P},\hat{Q})}{(E(\hat{Q}))^2}, \tag{21}$$

$$A_3 = \frac{E\left[\frac{\hat{P}}{\hat{Q}}(\hat{Q}-E(\hat{Q}))^2\right]}{(E(\hat{Q}))^2}. \tag{22}$$

Theorem 2 and Corollary 3 in Iglesias-Pérez and González-Manteiga (1999) give

$$E\left(\hat{P}\right) = P\left(1-\frac{1}{2}c_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)\right), \tag{23}$$

$$E\left(\hat{Q}\right) = Q\left(1-\frac{1}{2}c_K A_H\left(t|x\right)h^2+o\left(h^2\right)\right), \tag{24}$$

where

$$\begin{aligned} A_H\left(t|x\right) &= \int_0^t \left[\ddot{H}(s|x)+2\frac{m'(x)}{m(x)}\dot{H}(s|x)\right]dH_1(s|x) \\ &\quad + \left(1+2\frac{m'(x)}{m(x)}\right)\int_0^t \frac{d\dot{H}_1(s|x)}{1-H(t|x)}. \end{aligned} \tag{25}$$

Recall expressions (8) and (25). Then equations (23) and (24) can be used to find asymptotic expressions for (20) and (21):

$$\begin{aligned} A_1 &= \frac{P\left(1-\frac{1}{2}c_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)\right)}{Q\left(1-\frac{1}{2}c_K A_H\left(t|x\right)h^2+o\left(h^2\right)\right)} \\ &= \left(1-\varphi(t|x)\right)\frac{1-\frac{1}{2}c_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)}{1-\frac{1}{2}c_K A_H\left(t|x\right)h^2+o\left(h^2\right)} \\ &= \left(1-\varphi(t|x)\right)\left[1-\frac{1}{2}c_K\left(A_H\left(t+b|x\right)-A_H\left(t|x\right)\right)h^2\right]+o\left(h^2\right) \\ &= \left(1-\varphi(t|x)\right)-\frac{1}{2}c_K B_H\left(t,t+b|x\right)\left(1-\varphi(t|x)\right)h^2+o\left(h^2\right), \end{aligned} \tag{26}$$

$$A_2 = -\frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2} = O\left(\frac{1}{nh}\right). \tag{27}$$

Finally, since $1-\hat{\varphi}_n(t|x) = \dfrac{\hat{P}}{\hat{Q}} \in [0,1]$, the term (22) can be easily bounded:

$$0 \le A_3 \le \frac{Var\left[\hat{Q}\right]}{\left(E\left(\hat{Q}\right)\right)^2} = O\left(\frac{1}{nh}\right). \tag{28}$$

Using (26), (27), (28) and (6) in (19) we get

$$E\left(\hat{\varphi}_n(t|x)\right)-\varphi(t|x) = b(t|x)h^2+o\left(h^2\right). \tag{29}$$

To deal with the variance we use (13) for $p = 3$ and $\frac{1}{z} = \frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}$ to obtain:

$$\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2} = 1 + \sum_{i=1}^{3} (-1)^i \left(\frac{\hat{Q}^2 - E\left(\hat{Q}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} \right)^i + \left(\frac{\hat{Q}^2 - E\left(\hat{Q}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} \right)^4 \frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}. \qquad (30)$$

On the other hand

$$\hat{Q}^2 - \left(E\left(\hat{Q}\right)\right)^2 = \left[\hat{Q} - E\left(\hat{Q}\right)\right]^2 + 2E\left(\hat{Q}\right)\left[\hat{Q} - E\left(\hat{Q}\right)\right]$$

gives

$$\begin{aligned} \left(\frac{\hat{Q}^2 - \left(E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} \right)^i &= \sum_{j=0}^{i} \binom{i}{j} \left[\frac{\left(\hat{Q} - E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} \right]^j \left[\frac{2E\left(\hat{Q}\right)\left[\hat{Q} - E\left(\hat{Q}\right)\right]}{\left(E\left(\hat{Q}\right)\right)^2} \right]^{i-j} \\ &= \sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j}\left(\hat{Q} - E\left(\hat{Q}\right)\right)^{j+i}}{\left(E\left(\hat{Q}\right)\right)^{j+i}} \qquad (31) \end{aligned}$$

Substituting (30) in (31) we obtain:

$$\begin{aligned} \frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2} &= 1 + \sum_{i=1}^{3} (-1)^i \sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j}\left(\hat{Q} - E\left(\hat{Q}\right)\right)^{j+i}}{\left(E\left(\hat{Q}\right)\right)^{j+i}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} \frac{2^{4-j}\left(\hat{Q} - E\left(\hat{Q}\right)\right)^{j+4}}{\left(E\left(\hat{Q}\right)\right)^{j+4}} \frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}. \qquad (32) \end{aligned}$$

Equation (32) is useful to obtain an expansion for the second moment:

$$\begin{aligned} E\left[\left(1 - \hat{\varphi}_n(t|x)\right)^2\right] &= E\left(\frac{\hat{P}^2}{\hat{Q}^2}\right) = E\left(\frac{\hat{P}^2}{\left(E\left(\hat{Q}\right)\right)^2} \frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2} \right) \\ &= \frac{E\left[\left(\hat{P} - E\left(\hat{P}\right)\right)^2\right]}{\left(E\left(\hat{Q}\right)\right)^2} + \frac{E\left(\hat{P}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} \\ &\quad + \sum_{i=1}^{3} (-1)^i \sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j} E\left[\hat{P}^2 \left(\hat{Q} - E\left(\hat{Q}\right)\right)^{j+i}\right]}{\left(E\left(\hat{Q}\right)\right)^{j+i+2}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} \frac{2^{4-j} E\left[\frac{\hat{P}^2}{\hat{Q}^2} \left(\hat{Q} - E\left(\hat{Q}\right)\right)^{j+4}\right]}{\left(E\left(\hat{Q}\right)\right)^{j+4}}. \qquad (33) \end{aligned}$$

Defining, for $i, j = 0, 1, \ldots$, the notation

$$A_{ij} = E\left[\left(\hat{P} - E\left(\hat{P}\right)\right)^i \left(\hat{Q} - E\left(\hat{Q}\right)\right)^j\right], \tag{34}$$

$$B_{ij} = E\left[\hat{P}^i \left(\hat{Q} - E\left(\hat{Q}\right)\right)^j\right], \tag{35}$$

$$C_i = \left(E\left(\hat{Q}\right)\right)^i, \tag{36}$$

$$D_{ij} = E\left[(1 - \hat{\varphi}_n(t|x))^i \left(\hat{Q} - E\left(\hat{Q}\right)\right)^j\right] \tag{37}$$

and using

$$A_{2j} = B_{2j} - 2B_{10}A_{1j} + B_{10}^2 A_{0j},$$

expression (33) can be rewritten as

$$\begin{aligned} E\left[(1 - \hat{\varphi}_n(t|x))^2\right] &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3} (-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{B_{2\ i+j}}{C_{j+i+2}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2\ j+4}}{C_{j+4}} \\ &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} \\ &\quad + \sum_{i=1}^{3} (-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{A_{2\ i+j} + 2B_{10}A_{1\ i+j} - B_{10}^2 A_{0\ i+j}}{C_{j+i+2}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2\ j+4}}{C_{j+4}} \end{aligned} \tag{38}$$

It is easy, but long and tedious, to show that

$$E\left[\left(\hat{P} - E\left(\hat{P}\right)\right)^i\right] = o\left(\frac{1}{nh}\right), \text{ for } i \geq 3,$$

$$E\left[\left(\hat{Q} - E\left(\hat{Q}\right)\right)^i\right] = o\left(\frac{1}{nh}\right), \text{ for } i \geq 3.$$

Now recalling (34), (35), (36) and (37), and using Cauchy-Schwartz inequality and straight forward bounds, it can be proven that

$$A_{01} = A_{10} = 0, \tag{39}$$

$$A_{ij} = o\left(\frac{1}{nh}\right), \text{ whenever } i + j \geq 3, \tag{40}$$

$$B_{ij} = o\left(\frac{1}{nh}\right), \text{ for } j \geq 3, \tag{41}$$

$$D_{ij} = o\left(\frac{1}{nh}\right), \text{ for } j \geq 3. \tag{42}$$

Using (39), (40), (41) and (42) in (38), we conclude:

$$\begin{aligned} E\left[(1-\hat{\varphi}_n(t|x))^2\right] &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} - \frac{4B_{10}A_{11}}{C_3} - \frac{3B_{10}^2A_{02}}{C_4} + o\left(\frac{1}{nh}\right) \\ &= \frac{Var\left(\hat{P}\right)}{\left(E\left(\hat{Q}\right)\right)^2} + \frac{E\left(\hat{P}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} - \frac{4E\left(\hat{P}\right)Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3} \\ &\quad + \frac{3E\left(\hat{P}\right)^2 Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^4} + o\left(\frac{1}{nh}\right) \end{aligned} \tag{43}$$

On the other hand, plugging (18) in the term A_3 of expression (19), using (39), (40), (41) and (42) and some simple algebra gives:

$$\begin{aligned} E\left(1-\hat{\varphi}_n(t|x)\right) &= \frac{B_{10}}{C_1} - \frac{A_{11}}{C_2} + \frac{A_{12}+B_{10}A_{02}}{C_3} - \frac{A_{13}+B_{10}A_{03}}{C_4} + \frac{D_{14}}{C_4} \\ &= \frac{E\left(\hat{P}\right)}{E\left(\hat{Q}\right)} - \frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2} \\ &\quad + \frac{E\left(\hat{P}\right)Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3} + o\left(\frac{1}{nh}\right) \end{aligned} \tag{44}$$

The residual term $R_n'(y|x)$ in Theorem 2 of Iglesias-Pérez and González-Manteiga (1999) was proved to be uniformly negligible almost surely. A uniform rate for the moments of $R_n'(y|x)$ can be also obtained similarly. As a consequence of this, Theorem 2 and Corollary 4 in Iglesias-Pérez and González-Manteiga (1999) are applicable to obtain asymptotic expressions for the covariance structure of the process $\hat{S}_h(\cdot|x)$. This can be used to find and asymptotic expression for variances of $\hat{P}$ and $\hat{Q}$:

$$Var\left(\hat{P}\right) = \frac{1}{nh}v_1(t+b|x) + o\left(\frac{1}{nh}\right), \tag{45}$$

$$Var\left(\hat{Q}\right) = \frac{1}{nh}v_1(t|x) + o\left(\frac{1}{nh}\right), \tag{46}$$

$$Cov\left(\hat{P},\hat{Q}\right) = \frac{1}{nh}v_2(t,t+b|x) + o\left(\frac{1}{nh}\right), \tag{47}$$

where

$$v_1(t|x) = \frac{(1-F(t|x))^2}{m(x)} d_K C_H(t|x), \tag{48}$$

$$v_2(t,s|x) = \frac{(1-F(t|x))(1-F(s|x))}{m(x)} d_K C_H(t \wedge s|x), \tag{49}$$

$$C_H(t|x) = \int_0^t \frac{dH_1(s|x)}{(1-H(s|x))^2}. \tag{50}$$

Now using the orders found in (45), (46) and (47) in expressions (43) and (44) gives:

$$\begin{aligned} Var(\hat{\varphi}_n(t|x)) &= Var(1-\hat{\varphi}_n(t|x)) = \frac{Var(\hat{P})}{\left(E(\hat{Q})\right)^2} - \frac{2E(\hat{P})\,Cov(\hat{P},\hat{Q})}{\left(E(\hat{Q})\right)^3} \\ &\quad + \frac{\left(E(\hat{P})\right)^2 Var(\hat{Q})}{\left(E(\hat{Q})\right)^4} + o\left(\frac{1}{nh}\right). \end{aligned}$$

Finally, the asymptotic expressions (23), (24), (45), (46) and (47) and the definitions (48), (49), (50), (9) and (7) can be used to conclude:

$$\begin{aligned} Var(\hat{\varphi}_n(t|x)) &= \frac{1}{nh}\frac{v_1(t+b|x)}{(S(t|x))^2} - \frac{2}{nh}\frac{v_2(t,t+b|x)\,S(t+b|x)}{(S(t|x))^3} + \\ &\quad \frac{1}{nh}\frac{v_1(t|x)(S(t+b|x))^2}{(S(t|x))^4} + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{d_K C_H(t|x)}{m(x)}\frac{(S(t+b|x))^2 - 2(S(t+b|x))^2 + (S(t+b|x))^2}{(S(t|x))^2} \\ &\quad + \frac{1}{nh}\frac{d_K\left[C_H(t+b|x) - C_H(t|x)\right]}{m(x)}\left(\frac{S(t+b|x)}{S(t|x)}\right)^2 + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{d_K D_H(t,t+b|x)}{m(x)}(1-\varphi(t|x))^2 + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}v(t|x) + o\left(\frac{1}{nh}\right). \end{aligned} \tag{51}$$

Finally collecting expressions (29) and (51) we conclude (5). The formula for the asymptotic optimal bandwidth, (10), can be easily derived from (5).

To prove the last part of Theorem 2, we use Corollaries 3 and 4 in Iglesias-Pérez and González-Manteiga (1999) to show that

$$\sqrt{nh}\left[\left(\hat{P},\hat{Q}\right)^t - (P,Q)^t\right] \xrightarrow{\mathrm{d}} N_2(\mathbf{b},\mathbf{V}),$$

where

$$\begin{aligned}
\mathbf{b} &= (b_1, b_2)^t = -c^{1/2}\frac{1}{2}c_K \left(A_H(t+b|x)P, A_H(t|x)Q\right)^t, \\
\mathbf{V} &= \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} = \begin{pmatrix} v_1(t+b|x) & v_2(t,t+b|x) \\ v_2(t,t+b|x) & v_1(t|x) \end{pmatrix}.
\end{aligned}$$

Now applying the continuous function $g(u,v) = \frac{u}{v}$ to the sequence of the bivariate random variable above and using the delta method, simple but long and tedious algebra gives

$$\sqrt{nh}\left(\frac{\hat{P}}{\hat{Q}} - \frac{P}{Q}\right) \xrightarrow{\mathrm{d}} N\left(\mu, \sigma^2\right), \tag{52}$$

with

$$\begin{aligned}
\mu &= \left(\frac{\partial g(u,v)}{\partial u}, \frac{\partial g(u,v)}{\partial v}\right)\Big|_{(u,v)=(P,Q)} \mathbf{b} \\
&= \frac{1}{Q}b_1 - \frac{P}{Q^2}b_2 = -c^{1/2}\frac{1}{2}c_K\frac{P}{Q}\left(A_H(t+b|x) - A_H(t|x)\right) \\
&= -c^{1/2}b(t|x), \\
\sigma^2 &= \left(\frac{\partial g(u,v)}{\partial u}, \frac{\partial g(u,v)}{\partial v}\right)\Big|_{(u,v)=(P,Q)} \mathbf{V} \left(\frac{\partial g(u,v)}{\partial u}, \frac{\partial g(u,v)}{\partial v}\right)^t\Big|_{(u,v)=(P,Q)} \\
&= \frac{1}{Q^2}v_1(t+b|x) - \frac{2P}{Q^3}v_2(t,t+b|x) + \frac{P^2}{Q^4}v_1(t|x) \\
&= v(t|x).
\end{aligned}$$

This concludes the proof by substituting $\frac{\hat{P}}{\hat{Q}} = 1 - \hat{\varphi}_n(t|x)$ and $\frac{P}{Q} = 1 - \varphi(t|x)$ in (52).

Acknowledgements

This research was partially supported by Grants 07SIN01205PR from Xunta de Galicia (Spain) and MTM2009-00166 from Ministerio de Ciencia e Innovación (Spain).

7 References

Allen, L. N. and Rose, L. C. (2006). Financial survival analysis of defaulted debtors, *Journal of Operational Research Society*, 57, 630-636.

Altman, E. I. (1968). Finantial ratios, discriminant analysis, and the prediction of corporate bankruptcy, *Journal of Finance*, 23, 589-611.

Altman, E. I. and Saunders, A. (1998). Credit risk measurement: developments over the last 20 years, *Journal of Banking and Finance*, 21, 1721-1742.

Baba, N. and Goko, H. (2006). Survival analysis of hedge funds, Bank of Japan, Working Papers Series No. 06-E-05.

Basel Comitee on Banking Supervison (1999). International convergence of capital measurement and capital standards, Bank for International Settlements.

Basel Comitee on Banking Supervision (2004). International convergence of capital measurement and capital standards: a revised framework, Bank for International Settlements.

Beran, R. (1981). Nonparametric regression with randomly censored survival data, Unpublished technical report, University of California, Berkeley.

Beran, J. and Djaïdja, A. K. (2007). Credit risk modeling based on survival analysis with inmunes, *Statistical Methodology*, 4, 251-276.

Carling, K., Jacobson, T. and Roszbach, K. (1998). Duration of consumer loans and bank lending policy: dormancy versus default risk, Working Paper Series in Economics and Finance No. 280, Stockholm School of Economics.

Chen, S., Härdle, W. K. and Moro, R. A. (2006). Estimation of default probabilities with support vector machines, Center of Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Germany, SFB 649 discussion paper No. 2006-077.

Crouhy, M., Galai, D. and Mark, R. (2000). A comparative analysis of current credit risk models, *Journal of Banking and Finance*, 24, 59-117.

Dabrowska, D. (1987). Non-parametric regression with censored survival time data, *Scandinavian Journal of Statistics*, 14, 181-197.

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate, *Annals of Statistics*, 17, 1157-1167.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.

Glennon, D. and Nigro, P. (2005). Measuring the default risk of small business loans: a survival analysis approach, *Journal of Money, Credit, and Banking*, 37, 923-947.

González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications, *Journal of Nonparametric Statistics*, 4, 65-78.

Hamerle, A., Liebig, T. and Rösch, D. (2003). Credit risk factor modeling and the Basel II IRB Approach, Deutsche Bundesbank Discussion Paper Series 2, Banking and Financial Supervision, document No. 02/2003.

Hand, D. J. (2001). Modelling consumer credit risk, *IMA Journal of Management Mathematics*, 12, 139-155.

Hanson, S. and Schuermann, T. (2004). Estimating probabilities of default, Federal Reserve Bank of New York, Staff Report No. 190.

Iglesias Pérez, M. C. and González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications, *Journal of Nonparametric Statistics*, 10, 213-244.

Jorgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models, *Biometrika*, 70, 19-28.

Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data, *Annals of the Institute of Statistical Mathematics*, 53, 708-729.

Li, G. and Van Keilegom, I. (2002). Likelihood ratio confidence bands in non-parametric regression with censored data, *Scandinavian Journal of Statistics*, 29, 547-562.

Malik, M. and Thomas L. (2006). Modelling credit risk of portfolio of consumer loans, University of Southampton, School of Management Working Paper Series No. CORMSIS-07-12.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* , 2nd. Ed., Chapman and Hall, London.

Narain, B. (1992). Survival analysis and the credit granting decision. In: Thomas L., Crook, J. N. and Edelman, D. B. (eds.). *Credit Scoring and Credit Control*. OUP: Oxford, 109-121.

Roszbach, K. (2003). Bank lending policy, credit scoring and the survival of loans, Sverriges Riksbank Working Paper Series No. 154.

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

Saunders, A. (1999). *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms*, John Wiley & Sons, New York.

Van Keilegom, I. and Veraverbeke, N. (1996). Uniform strong results for the conditional Kaplan-Meier estimators and its quantiles, *Communications in Statistics: Theory and Methods*, 25, 2251-2265.

Van Keilegom, I., Akritas, M. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study, *Computational Statistics & Data Analysis*, 35, 487-500.

Wang, Y., Wang, S. and Lai, K. K. (2005). A fuzzy support vector machine to evaluate credit risk, *IEEE Transactions on Fuzzy Systems*, 13, 820-831.

Discussion of “Modelling consumer credit risk via survival analysis” by Ricardo Cao, Juan M. Vilar and Andrés Devia

Noël Veraverbeke
Centrum Voor Statistiek
Universiteit Hasselt, Belgium

The present paper deals with the estimation of the probability of default (PD) which is a very important parameter in many models for consumer credit risk in the literature.

If T denotes the time to default of a client, then it is immediately clear that in many cases T will not be observed, due to the ending of the observation period or the occurrence of some other event that happens earlier in time. This perfectly fits into the classical model of right random censoring in survival analysis. Here the observations are $Y = \min(T,C)$ and $\delta = I(T \leq C)$ where T is the time to default and C is the censoring time.

Classical survival analysis tools like Kaplan-Meier estimation and Cox estimation allow to obtain estimates for the distribution function of T. Moreover it is also possible to incorporate a vector X of explanatory variables and to estimate the conditional distribution function of T, given that $X = x$.

Since the probability of default just the conditional residual life distribution function (see Veraverbeke (2008)), it can be expressed as a simple function of the conditional distribution function and different estimation methods of the latter lead to different estimators for the PD.

Three methods are explored in this paper. The first is based on Cox's proportional hazards regression model, the second on a generalized linear model and the third on Beran's (1981) nonparametric product limit estimator for the conditional distribution function. For the third method, some new asymptotic properties are derived for the conditional residual life distribution function estimator. The illustration with real data clearly shows that the covariate information is essential and that methods 1 and 3 give a good fit.

I want to congratulate the authors for their contribution to this field of modelling credit risk using regression techniques from survival analysis. The results are very promising and I hope to see further work in that direction. My comments/questions below are meant to stimulate this.

1) It would be interesting to explore the use of time-dependent covariates. In particular, how could this be done for the nonparametric method?
2) The theoretical results and also the real data application are shown for one single covariate. Is the extension to more than one covariate straightforward?

3) An assumption throughout is the conditional independence of T and C, given X. But there are more and more examples in survival analysis where this is questionable. See, for example, Zheng and Klein (1995), Braekers and Veraverbeke (2005). How realistic is the independence assumption in credit risk modelling and how could this assumption possibly be relaxed?
4) Is it possible to generalize the asymptotic normality result in Theorem 2 in order to obtain practical confidence bands for the default rate curves?
5) The third method relies on a good choice for the bandwidth. Is there a suggestion for an optimal choice?

It was a pleasure for me to be invited as a discussant for this interesting paper.

References

Braekers, R. and Veraverbeke, N. (2005). Copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics*, 33, 429-447.

Veraverbeke, N. (2008). Conditional residual life under random censorship. In: B. C. Arnold, U. Gather, S. M. Bendre (eds). *Advances in Statistics: Felicitation Volume in Honour of B. K. Kale*. MacMillan India, New Dehli, pp.174-185.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82, 127-138.

Jean-Philippe Boucher
Département de Mathématiques
Université du Québec à Montréal, Québec, Canada

The paper deals the default intensity of consumer to determine the probability of defaults. Because the authors use the fact that a large proportion of consumers does not have default, they use censored models in the estimation.

I would like to point out to authors some important details. In consumer credit data, the amount of information from real data is very small. Indeed, depending on the definition of what is a *default*, we can suppose that a default can arise continuously. However the default will only be observable on a small period of time since consumer only pays his debt at the beginning of each month. Consequently, we must deal with even less information than what is assumed in the paper. For the Cox proportional hazard model, Malik and Thomas (2006) worked with a modified likelihood function when dealing with this situation.

This problem shares similarities with insurance data. Indeed, with aggregate insurance data, it is impossible to know at what time insureds had their accident (see for example Boucher and Denuit (2007)). A major difference between credit and claim count analysis is the fact that a default of credit happens only once, while it is possible to see more than one claim in a single insurance period. However, even with this difference, for a parametric approach such as the GLM model proposed by the authors, it is possible to construct credit risk models based on models of Boucher and Denuit (2007).

Conceptually, let τ be the waiting time between the beginning of the loan and the default. Let $I(t)$ be the indicator of a default during the interval $[0,t]$. Hence,

$$\begin{aligned} P(I(t)=0) &= P(\tau > t) \qquad (1)\\ P(I(t)=1) &= 1 - P(\tau > t) \end{aligned}$$

For a loan of one year, we only have up to 12 partial informations on the credit default. Consequently, we then observed intervals $[0, \frac{1}{12}]$, $]\frac{1}{12}, \frac{2}{12}]$, $]\frac{2}{12}, \frac{3}{12}]$, $\ldots$, $]\frac{11}{12}, \frac{12}{12}]$.

In count data, duration dependence occurs when the outcome of an experiment depends on the time that has elapsed since the last success (Winkelmann (2003)). Then, the occurrence of an event modifies the expected waiting time to the next occurrence of the event. For credit risk, a positive (negative) duration dependence would mean that the probability of default decreases (increases) over time. Consequently, the true probability depends on which interval the default happens and can be expressed:

$$P(I(1) = y) \; = \; \begin{cases} P(\tau \leq \frac{1}{12}) & \text{for a default } y \text{ occuring in } [0, \frac{1}{12}] \\ P(\frac{1}{12} < \tau \leq \frac{2}{12}) & \text{for a default } y \text{ occuring in }]\frac{1}{12}, \frac{2}{12}] \\ ... & \\ P(\frac{11}{12} < \tau \leq \frac{12}{12}) & \text{for a default } y \text{ occuring in }]\frac{11}{12}, \frac{12}{12}] \\ 1 - P(\tau > \frac{12}{12}) & \text{for a non-registered default} \end{cases}, \qquad (2)$$

By comparison with this last equation, for an individual i, the contribution of conditional likelihood function of the authors, involving the conditional density, was written as $f(\tau)^{\delta}(1-F(\tau))^{1-\delta}$, where $\delta = 1$ if an individual did a default. Less information is available with (2) since differences in cumulative distributions is used rather than density functions.

Except when working with the Exponential distribution that is known to be memoryless, it is not possible to simply express all the probabilities of a default as:

$$P(I(t) = 0) = P(\tau > \frac{1}{12}),$$

for all possible values of t because it is only valid for the first interval. Indeed, for illustration, let us assume that τ is Gamma distributed, with density

$$f(\tau; \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\lambda \tau}, \qquad (3)$$

where $\Gamma(\cdot)$ is the gamma function, $\alpha > 0$ and $\lambda > 0$. Note that the Gamma hazard function is increasing for $\alpha > 1$ and is decreasing for $\alpha < 1$ (and shows constant hazard of $\alpha = 1$). Thus, for $\alpha \leq 1$, the model exhibits positive duration, while $\alpha \geq 1$ implies negative duration (and does not show duration dependence for $\alpha = 1$, from which we find the Exponential distribution). It can be interesting to see how well real data can estimate these parameters. Indeed, with the use of (2), the model can be expressed by using this useful notation:

$$P(t_a < \tau \leq t_b) = \frac{1}{\Gamma(\alpha)} \int_{t_a}^{t_b} \lambda^{\alpha} \nu^{\alpha-1} e^{-\lambda \nu} d\nu$$

where the integral is known as an incomplete gamma function. This probability can be evaluated using integrations approximations or asymptotic expansions (Abramowitz and Stegun (1968)).

It would be interesting to apply the unusual non-parametric approach of the authors using equation (0.2).

References

Abramowitz, M. and Stegun, I. A. (1968). *Handbook of Mathematical Functions*. National Bureau of Standards, Applied Mathematics Series Nr. 55, Washington, D.C.

Boucher, J.-P. and Denuit, M. (2007). Duration Dependence Models for Claim Counts. *Deutsche Gesellschaft fur Versicherungsmathematik (German Actuarial Bulletin)*, 28, 29-45.

Malik, M. and Thomas, L. (2006). Modeling Credit Risk of Portfolio of Consumer Loans. *University of Southampton, School of Management, Working Paper, Series No. CORMSIS-07-12*.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer-Verlag, Berlin, 4th ed.

Jan Beran
Department of Mathematics and Statistics
University of Konstanz, Germany

Since Basel II, the modeling of credit risks has become an important practical issue that has to be addressed in a mathematically tractable manner while taking into account particular characteristics of the market and available data. One general approach discussed in the literature is modelling probability of default (PD) by applying survival analysis. The idea is quite natural, since in the financial context the time until default can be interpreted directly as a survival time and such data are readily available. As in usual survival analysis, the observed times until default are partially censored. In spite of the obvious analogy to the biological context, survival analysis may not be very well known among practitioners in finance. The paper by Cao, Vilar and Devia is therefore a welcome contribution.

The authors essentially discuss three methods of estimation:

1. Cox's proportional hazard model;
2. generalized linear models; and
3. nonparametric conditional distribution estimation.

For the third method, the asymptotic mean squared error and a formula for the asymptotically optimal bandwidth h_o are given. While 1 and 2 and their properites are well known, the asymptotic result for the third method appears to be new. From the practical point of view, the question is which of the three methods perform best when applied to real data, and also whether there may be any alternative methods that even outperform any of these. Before answering this question, one needs to define a criterion for judging the performance. In the paper here, empirical and estimated PDs are compared. Thus, the criterion is simply to what extent a model fits the data. More interesting would be to use predictive out of sample criteria and also financial risk measures. Furthermore, the fitted PDs reported in table 2 are of varying quality. One may therefore ask whether any of the models considered here reflect the underlying mechanism with sufficient accuracy. In particular, the perfomance of standard models in survival analysis depends on the amount of censoring. Typically for credit default data, a large majority (often more than 95%) of the observations are censored. In such situations, maximum likelihood estimates based on unimodal distributions tend to be highly biased. For this reason, Beran and Djaïdja (2007) adopted an idea originally introduced by Maller and Zhou (1996) in a medical context. Observations are assumed to come from a mixture distributions consisting of a usually large proportion p of "immunes" and a smaller proportion $1-p$ of clients who may default. Thus, the time until a randomly chosen client number i defaults can be written as

$$Y_i = \varsigma_i \cdot \infty + (1-\varsigma_i)W_i$$

where $P(\varsigma_i = 1) = 1 - P(\varsigma_i = 0) = p$ and W_i is a continuous distribution $F_W(.;\lambda)$ with density $f_W(.;\lambda)$ $(\lambda \in \Lambda \subseteq \mathbb{R}^k)$ on $\mathbb{R}_+$. Conditionally on the censoring constants c_i, the maximum likelihood estimate of $\theta = (p,\lambda)$ is obtained from observed survival times $x_i = y_i \wedge c_i$ by maximizing

$$L(\theta) = n_1 \log(1-p) + \sum_{i \in I} \log f_W(y_i;\lambda) + \sum_{i \in I^c} \log\left[1-(1-p)F_W(c_i;\lambda)\right]$$

where $I = \{i : y_i \leq c_i\}$ and $n_1 = |I|$. In practice estimates of PDs and prediction of defaults turned out to be much more accurate in the case of retail clients where defaults are (or used to be) very rare. It may therefore be worth the effort to see whether the same applies to the consumer loans considered in this discussion paper.

References

Beran, J. and Djaïdja (2007). Credit risk modeling based on survival analysis with immunes. *Statistical Methodology*, 4, 251-276.

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Rejoinder

First of all we would like to thank the discussants for their kind words and suggestions concerning this paper. According to the topics mentioned in the comments we will organize this rejoinder in four sections. These sections deal with other censoring models, predictive criteria, bandwidth selection and extensions to other settings.

1 Other censoring models

As mentioned by Prof. Beran, some other alternative models are available for heavy censoring situations like in credit risk. In this rejoinder we will adopt the approach by Maller and Zhou (1996) and Beran and Djaïdja (2007) for the generalized linear model presented in the paper. Using the notation of Subsection 3.3, we have considered the model

$$F(t|x) = (1-p)\,g\,(\theta_0 + \theta_1 t + \theta_2 x)\,, \tag{1}$$

where $p \in (0,1)$ is the proportion of credits that are immune to default and F is any of the two parametric distributions considered in Subsection 5.1.2 of the paper. Using equation (1), the log-likelihood function in Subsection 3.3 results in

$$\begin{aligned}\ell\,(\theta_0,\theta_1,\theta_2,p) \;=\; & [\ln(1-p) + \ln\theta_1]\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\delta_i \ln g'\,(\theta_0+\theta_1 Y_i+\theta_2 X_i)\\ & + \sum_{i=1}^{n}(1-\delta_i)\ln\left[1-(1-p)\,g\,(\theta_0+\theta_1 Y_i+\theta_2 X_i)\right]\end{aligned}$$

For dealing with the high complexity of the model and the minimization of the log-likelihood equations, we have used a differential evolution based program called *DEoptim*, implemented in R. See, for instance, Price et al. (2005) for details about this numerical optimization approach.

Figure 1 shows the estimated *PD* using these heavy censoring models when conditioning to $X = 5.44$, the median value of the covariate. The $\widehat{PD}$ curves for the *GLM* and the modified *GLM* (*MGLM*) are shown in a range of maturity times given by the depth of the sample. The *GLM* curves in the left panel are those presented in Figure 5 of the paper. Using the same link functions, the heavy censoring models with

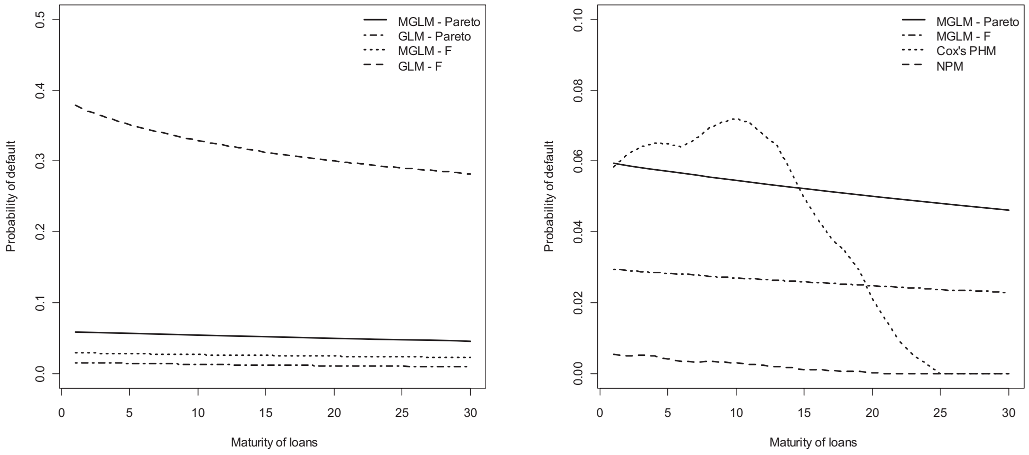

Figure 1: *Left panel: $\widehat{PD}$ curves for the GLM and the MGLM. Right panel: $\widehat{PD}$ curves for the MGLM, Cox's proportional hazard model and the nonparametric approach. All these $\widehat{PD}$ curves were obtained conditioning on $X = 5.44$.*

single parameter Pareto and Snedecor's F distributions are also plotted in the left panel. The estimated parameters were $\alpha = 0.6$ and $p = 0.01$ for the Pareto distribution and 8.553 and 0.525 for the degrees of freedom of Snedecor's F distribution with $p = 0.01$. The right panel plots a graphical comparison of the *MGLM*, Cox's proportional hazard model and the nonparametric approach.

The results obtained with the *GLM* approach are not good in general, and the modified version proposed in equation (1) did not produced a significant improvement in the estimated *PD* for our data set. The $\widehat{PD}$ curve computed with the F link fits better than that with the Pareto link function for a range of covariate values. Thus, in the following we will only present results concerning the *MGLM* approach with the Snedecor's F link function. The estimated default probabilities with both links were extremely large for those values of X smaller than 1, or extremely small for values of X larger than the third quartile (28.2703).

An alternative way to deal with heavy censoring, not considered here, is to use the transfer tail models introduced by Van Keilegom and Akritas (1999) and Van Keilegom, Akritas and Veraverbeke (2001). This consists in using nonparametric regression residuals to transfer tail information from regions of light censoring to regions of heavy censoring in conditional distribution function estimation.

The possible discrete nature of the defaults, mentioned by Prof. Boucher, gives rise to an interval censored model for the time to default (see his equation (2)). This censoring model is very useful when the defaults are reported in multiples of a given time unit (e.g., a month). This is not the case for our data set with 1800 defaults corresponding to 576 different values. The highest frequency of these values is only 15 and the average frequency of these 576 different values is only 3.156.

Our data set has been facilitated by a financial company. This company records the contract date and sends a payment order on a fixed date in the second month following the contract formalization date. This fixed date may change from month to month. When a client does not make one of these payments and this situation is maintained for more than 90 days, the 91st day after the due payment date is considered as the default time. However there are even a few exceptions in which default may be considered even before than four months from the contract date. For all these reasons it is virtually impossible, at least for this data set, that default times occur in multiples of one month.

Nevertheless, there may exist practical situations where defaults exhibit a discrete nature. In these cases the nonparametric estimator given by Beran (1981) can be extended to interval-censored response lifetimes. The idea is to adapt the estimator proposed by Turnbull (1976) to the conditional setting, in a general framework of censoring and truncation (which includes interval censoring). This adaptation could be very similar to the one used in Beran (1981) to extend the Kaplan-Meier estimator to a conditional setup.

As Professor Veraverbeke points out, one could consider more general censoring models that allow for some sort of conditional dependence between the censoring time, C, and the life time, T, of a credit. The hypothesis of conditional independence is very common in survival analysis and it is also very convenient in credit risk applications.

In principle, when the censoring times come from time from contract formalization to end of the study, the conditional independence assumption seems a natural one. However, this is not the only source of censored data. For instance credit cancellation, which also causes censoring, may be correlated to possible time to default. Unfortunately it is often very difficult to test such an assumption from real data. This is because most of the times there is no available information about jointly observed values of (C,T). As Professor Veraverbeke mentions, copula models are useful tools for constructing more flexible models that allow for conditional dependence. An interesting future study would be to extend the results on nonparametric estimation of default probability to copula models as those proposed in Braekers and Veraverbeke (2005).

2 Predictive criteria

As Professor Beran explains in his report interesting model adequacy tests for a financial firms are based on predictive criteria. The estimated probability of default can be used to classify a credit in default or nondefault. Using the three methods proposed in the paper and fixing a maturity time of $t = 5$ months and a forecast time horizon of $b = 12$ months, the estimated PD has been computed for every single credit of a real loan portfolio.

Starting from the sample of credits alive at time t, the two subsamples of defaulted and non-defaulted credits at time $t + b$ have been considered. In order to study the discrimination power of the three models, we have considered the pertaining estimated

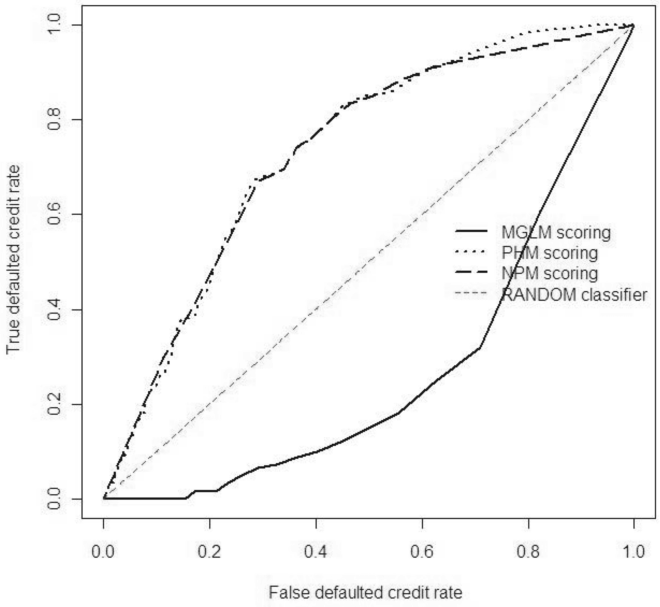

Figure 2: *ROC curves for the three PD approaches: MGLM, Cox's PHM, and NPM.*

PD and computed the *ROC* curves. This tool has been used in financial setups by Thomas (2000), Stein (2005), Blöchlinger and Leippold (2006) and Engelmann (2006), among others. The area under the *ROC* curve (*AUC*), which is a measure of the the discrimination power of the methods, has also been computed.

The study was performed by just dividing our data set of size 25000 in a training sample of size 20000 and a test sample of size 5000. The choice of these two samples was made at random. The test sample was split up into defaulted and non-defaulted credits. The *PD* estimates, obtained for the three approaches using the training sample, were applied to the test sample and the out-of-sample *ROC* curves are plotted in Figure 2. The areas under these curves and their confidence intervals are collected in Table 1.

Figure 2 shows a surprisingly poor discrimination power of the *MGLM* model. This is also reflected by the *AUC* values in Table 1. An open question is how important is the choice of the link function in order to produce much better results. The performance of Cox's proportional hazard model and the nonparametric approach is very comparable. Their discrimination power (measured via the *AUC*) is about 74%.

A first conclusion is that the modification of the original *GLM* approach was not able to produce the expected improvement in the original *GLM* setting, but it may be interesting to study the problem of choice of the link function in a deeper way. On the other hand *PD* estimates obtained by Cox's proportional hazards model and the nonparametric approach provide quite powerful discrimination between default and non-default credits.

Table 1: *Area under the ROC curves for the three approaches computed by using the validation sample.*

Model	AUC	95% asymptotic confidence interval	
$MGLM$	0.265	0.234	0.297
Cox's PHM	0.735	0.703	0.766
NPM	0.738	0.706	0.770

3 Bandwidth selection

As Professor Veraverbeke points out, the nonparametric approach relies on a good choice for the bandwidth. Direct plug-in methods for the selection of the smoothing parameter require the estimation of plenty of population functions involved in equation (10): $H_1(t|x)$, $H(t|x)$, $\dot{H}(t|x)$, $\ddot{H}(t|x)$, $m(x)$, $m'(x)$ and $\varphi(t|x)$. This turns out to be a tedious procedure. Furthermore, since the method is based on an asymptotic expression, it may not produce accurate results for samples with a moderate number of uncensored data. See, for instance Cao, Janssen and Veraverbeke (2001) for similar ideas in a different context.

A good alternative for bandwidth selection in this context is the bootstrap method. This method can be used to find a bootstrap analogue of the mean squared error of $\varphi(t|x) = PD(t|x)$ (see, for instance, Cao (1993) for the use of the bootstrap for estimation of the mean integrated squared error in a different context). This method would require the use of two pilot bandwidths, g_1 and g_2, for estimating $F(t|x)$ and $G(t|x)$ and a pilot bandwidth, g_3, for the density m. The method proceeds as follows:

1. Compute, $\hat{F}_{g_1}(t|x)$, Beran's estimator of $F(t|x)$ and $\hat{G}_{g_2}(t|x)$, Beran's estimator of $G(t|x)$.
2. Estimate $m(x)$ by $\hat{m}_{g_3}(x)$.
3. Draw a sample $(X_1^*, X_2^*, \ldots, X_n^*)$ from $\hat{m}_{g_3}(x)$.
4. For every $i = 1, 2, \ldots, n$, draw T_i^* from $\hat{F}_{g_1}(t|x)$ and C_i^* from $\hat{G}_{g_2}(t|x)$.
5. Compute, for every $i = 1, 2, \ldots, n$, $Y_i^* = \min\{T_i^*, C_i^*\}$ and $\delta_i^* = \mathbf{1}_{\{T_i^* \leq C_i^*\}}$.
6. Use the sample $\{(Y_1^*, \delta_1^*, X_1^*), (Y_2^*, \delta_2^*, X_2^*), \ldots, (Y_n^*, \delta_n^*, X_n^*)\}$ to compute $\hat{\varphi}_h^*(t|x)$, the bootstrap analogue of $\hat{\varphi}_h(t|x)$.
7. Approximate the mean squared error of $\hat{\varphi}_h(t|x)$ by its bootstrap version:

$$MSE_{t,x}^*(h) = E^*\left[(\hat{\varphi}_h^*(t|x) - \hat{\varphi}_{g_1}(t|x))^2\right].$$

8. This bootstrap MSE can be approximated by drawing a large number, B, of bootstrap replications following steps 4-6 and computing

$$\frac{1}{B}\sum_{j=1}^{B}\left(\hat{\varphi}_h^{*j}(t|x)-\hat{\varphi}_{g_1}(t|x)\right)^2.$$

9. Finally the bootstrap bandwidth, $h^*_{MSE,t,x}$, is the minimizer of $MSE^*_{t,x}(h)$ in h.

Since this resampling plan may be very time consuming, a possible way to make this approach feasible for very large sample sizes (like $n = 25000$) is the following. Fix some smaller subsample size (for instance $m = 2500$), i.e., $n = \lambda m$, with λ typically large (in this example $\lambda = 10$). Use the bootstrap resampling plan to get a bootstrap bandwidth, $h^*_{MSE,m,t,x}$, for sample size m. Based on the asymptotic formula (10), in the paper, obtain $h^*_{MSE,n,t,x} = \lambda^{-1/5}h^*_{MSE,m,t,x}$.

Consistency and practical behaviour of this bootstrap method is left for future work.

4 Extensions to other settings

Professor Veraverbeke raises the question of extension of the nonparametric default probability estimator to the multiple covariate case. We believe that this extension is rather straightforward, as it is for the conditional distribution estimator. From the theoretical viewpoint, it is expected that the convergence rate gets worse when the dimension of the covariate vector increases. In fact, it is very likely that the *PD* nonparametric estimator is worthless for covariates of dimension larger to 3 or 4, except for huge sample sizes (curse of dimensionality). A possible way to overcome this problem is to use the dimension reduction ideas proposed by Hall and Yao (2005) to produce a semiparametric estimator of the default probability that is free of the curse of dimensionality. At the same time, the projection of the covariate vector obtained by such a procedure would probably be interpretable as a kind of overall scoring that accounts for propensity of credits to default.

The time-dependent covariate case mentioned by Professor Veraverbeke can be treated using ideas of McKeague and Utikal (1990), who extended Beran's estimator to time-dependent covariates. Last, but not least, although convergence of the default probability process could be studied and used to derive asymptotic theory for confidence bands, in our opinion this is out of the scope of the present paper. On the other hand we believe that, for practical reasons, financial companies are more interested (for prediction) in the estimation of the default probability at a given maturity and with fixed values of the covariates, than in a confidence band.

We would like to finish this rejoinder by thanking, once again, the discussants for their suggestions and comments. We are also grateful to the Editor in Chief of SORT, Montserrat Guillén, for her kind invitation to write this paper and for her efficiency along the editing process. Her support has helped us a lot to improve the quality of this paper.

References

Beran, J. and Djaïdja, A. K. (2007). Credit risk modeling based on survival analysis with inmunes, *Statistical Methodology*, 4, 251-276.

Beran, R. (1981). Nonparametric regression with randomly censored survival data, Unpublished technical report, University of California, Berkeley.

Blöchlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring, *Journal of Banking and Finance*, 30, 851-873.

Braekers, R. and Veraverbeke, N. (2005). Copula-graphic estimator for the conditional survival function under dependent censoring, *Canadian Journal of Statistics*, 33, 429-447.

Cao, R. (1993). Bootstrapping the mean integrated squared error, *Journal of Multivariate Analysis*, 45, 137-160.

Cao, R., Janssen, P. and Veraverbeke, N. (2001). Relative density estimation and local bandwidth selection with censored data, *Computational Statistics* & *Data Analysis*, 36, 497-510.

Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. *The Basel Risk Parameters: Estimation, Validation, and Stress Testing.* Springer, New York.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction, *The Annals of Statistics*, 33, 1404-1421.

Maller, R.A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*, Wiley, New York.

McKeague, I. W. and Utikal, K. (1990). Inference for a nonlinear counting process regression model, *The Annals of Statistics*, 18, 1172-1187.

Price, K., Storn, R. and Lampinen, J. (2005). *Differential Evolution – a Practical Approach to Global Optimization*. Springer, New York.

Stein, R. (2005). The relationship between default prediction and lending profits: integrating the ROC analysis and loan pricing, *Journal of Banking and Finance*, 29, 1213-1236.

Thomas, L. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16, 149-172.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B*, 38, 290-295.

Van Keilegom, I. and Akritas, M. G. (1999). Transfer of tail information in censored regression models, *The Annals of Statistics*, 27, 1745-1784.

Van Keilegom, I., Akritas, M. G. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study, *Computational Statistics & Data Analysis*, 35, 487-500.