In this paper, a general nonparametric estimator of the probability of default is proposed and studied. It is derived from an estimator of the conditional survival function for censored data obtained with a double smoothing, on the covariate and on the variable of interest. Asymptotic expressions for the bias and the variance of the probability of default estimator are found and asymptotic normality is proved. A simulation study shows the performance of the proposed estimator and compares its behaviour with smoothed estimators only in the covariate. An empirical study, based on modified real data, illustrates its practical application.

Keywords: Censored data; kernel method; probability of default; risk analysis; survival analysis

Acknowledgements

This research has been supported by MINECO Grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015, ED431C-2020-14 and Centro Singular de Investigación de Galicia ED431G 2019/01), all of them through the ERDF.
1 Introduction

Credit risk is an important area of investigation. It is useful for financial companies to assess the risk of insolvency caused by unpaid loans. Estimating the probability of default on consumer credits, loans and credit cards is one of the main problems that banks, savings banks, savings cooperatives and other credit companies must address. For a fixed time, \( t \), and a horizon time, \( b \), the probability of default (PD) can be defined as the probability that a credit that has been paid until time \( t \) becomes unpaid not later than time \( t + b \). To estimate the PD, banks and financial institutions typically use informative covariates of the credit and the clients usually building some linear combination (scoring) based on these informative covariates and the \( PD(t|x) \) is allowed to depend on this scoring \( x \). It can be deduced from the definition of the PD that it is a relevant measure in other fields appart from the financial one. For example, companies that provide energy services (electricity, gas), water, streaming services (TV, cinema, music), telephone, internet... are interested in estimating the probability that a customer who receives their services at time \( t \) will remain with the company at time \( t + b \).

There is an extensive literature in which survival analysis methods are used for solving credit risk problems. Among others, we mention the work of Naraim (1992), Stepanova and Thomas (2002), Hanson and Schuermann (2004), Glennon and Nigro (2005), Allen and Rose (2006), Baba and Goko (2006), Beran (2007) and Cao et al. (2009). A common feature of all these papers is the use of parametric or semiparametric regression techniques for modelling the time to default, including exponential models, Weibull models and Cox’s proportional hazards models, which are typical in this literature. Nonparametric curve estimation is a flexible method that only use the information that the data provides without making assumptions about the shape of the curve. Therefore, it is very convenient in this context. Following this idea, Cao et al. (2009) proposed a PD estimator using Beran’s estimator for the conditional survival function, Beran (1981). This study was expanded in the paper of Peláez et al. (2020b) who studied four nonparametric estimators of the probability of default in credit risk derived from estimators of the conditional survival function for censored data.

In the recent work, Peláez et al. (2020a), a general nonparametric estimator of the
conditional survival function with double smoothing is proposed and studied. This survival estimator is not only smoothed in the covariate but also in the time variable. A large simulation study shows there that the estimator with double smoothing improves on the corresponding nonparametric estimator of the survival function which is smoothed only in the covariate. Here, a general nonparametric estimator of the PD with double smoothing is proposed and studied. It is derived from the survival estimator with double smoothing studied in Peláez et al. (2020a).

The remainder of this paper is organized as follows. The nonparametric estimator of the probability of default with double smoothing is defined in Section 2. Asymptotic properties of the proposed nonparametric estimator associated with Beran’s estimator of the conditional survival function are presented in Section 3. In Section 4 a simulation study shows the improvement obtained by using the double smoothing in several nonparametric estimators of the probability default. In Section 5, the proposed nonparametric estimator of the PD is applied to a set of modified real data. Section 6 contains some concluding remarks. Finally, Section 7 includes a sketch of the proof of the theoretical results presented in Section 3.

2 PD nonparametric estimator with double smoothing

Let \( \{(X_i, Z_i, \delta_i)\}_{i=1}^n \) be a simple random sample of \((X, Z, \delta)\) where \( X \) is the covariate, \( Z = \min\{T, C\} \) is the observed variable, \( T \) is the time to occurrence of the event, \( C \) is the censoring time and \( \delta = I_{\{T \leq C\}} \) is the uncensoring indicator. It is assumed that an unknown relationship between \( T \) and \( X \) exists. In credit risk, usually, \( X \) is the credit scoring, \( Z \) is the observed maturity, \( T \) is the time to default and \( C \) is the time until the end of the study or the time until the anticipated cancellation of the credit. The distribution function of \( T \) is denoted by \( F(t) \) and the survival function by \( S(t) \). The functions \( F(t|x) \) and \( S(t|x) \) are the conditional distribution and survival functions of \( T \) given \( X = x \) evaluated at \( t \). In this context, let \( x \) be a fixed value of the covariate \( X \) and \( b \) any fixed value (typically, \( b = 12 \) in months). Then the probability of default in a time horizon \( t + b \) from a maturity time,
t, is defined as follows

\[
PD(t|x) = P(T \leq t + b | T > t, X = x) = \frac{F(t + b | x) - F(t | x)}{1 - F(t | x)} = 1 - \frac{S(t + b | x)}{S(t | x)}.
\]  

(1)

Therefore, an estimator of \(PD(t|x)\) could be obtained by replacing \(S(t+b|x)\) and \(S(t|x)\) in (1) with appropriate estimators. Following this idea, Cao et al. (2009) and Peláez et al. (2020b) used nonparametric estimators of the conditional survival function, \(\hat{S}_h(t|x)\) with \(h = h_n\) being the smoothing parameter for the covariate, to obtain the corresponding nonparametric estimator of \(PD(t|x)\) denoted by \(\hat{PD}_h(t|x)\).

In Peláez et al. (2020a) the following nonparametric estimator of the conditional survival function with double smoothing is proposed and studied

\[
\tilde{S}_{h,g}(t|x) = 1 - \sum_{i=1}^{n} s_{(i)} \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right)
\]  

(2)

where \(s_{(i)} = \hat{S}_h(Z_{(i-1)} | x) - \hat{S}_h(Z_{(i)} | x)\) whith \(Z_{(i)}\) the \(i\)-th element of the sorted sample of \(Z\), \(\mathbb{K}(t)\) is the distribution function of a kernel \(K\), \(\mathbb{K}(t) = \int_{-\infty}^{t} K(u)du\), and \(g = g_n\) is the smoothing parameter for the time variable. This survival estimator is not only smoothed in the covariate but also in the time variable. It is based on the idea of estimating the survival function in a point \(t\) conditional to \(x\) by means of a weighted mean of the values that the estimator \(\hat{S}_h(t|x)\) takes in points near \(t\).

Estimating the probability of default, \(PD(t|x)\), by means of the nonparametric estimator of the conditional survival function with double smoothing is the aim of this paper. For this purpose, \(S(t|x)\) in (1) is replaced by a doubly smoothed nonparametric estimator, \(\tilde{S}_{h,g}(t|x)\), obtaining the following nonparametric estimator of \(PD(t|x)\):

\[
\tilde{PD}_{h,g}(t|x) = 1 - \frac{\tilde{S}_{h,g}(t + b | x)}{\tilde{S}_{h,g}(t | x)}.
\]  

(3)

Since \(\hat{S}_h(t|x)\) is any arbitrary conditional survival estimator, the probability of default estimator \(\tilde{PD}_{h,g}(t|x)\) is very general. From now on, this paper focuses on the doubly smoothed Beran’s estimator \(\tilde{S}_{h,g}^B(t|x)\) associated through (2) with the classic Beran’s esti-
mator of the survival function given by
\[
\hat{S}_B^h(t|x) = \prod_{i=1}^n \left( 1 - \frac{I_{\{Z_i \leq t, \delta_i = 1\}} w_{i,n}(x)}{1 - \sum_{j=1}^n I_{\{Z_j < Z_i\}} w_{n,j}(x)} \right)
\]  
(4)

where
\[
w_{i,n}(x) = \frac{K((x - X_i)/h)}{\sum_{j=1}^n K((x - X_j)/h)}
\]
with \(i = 1, \ldots, n\) and \(h = h_n\) is the smoothing parameter for the covariable.

Using \(\hat{S}_{h,g}^B(t|x)\) in (3), the smoothed probability of default estimator based on Beran’s survival estimator is obtained. It is denoted by \(\hat{PD}_{h,g}^B(t|x)\).

3 Asymptotic results of the doubly smoothed Beran’s estimator of the PD

Asymptotic properties of the smoothed Beran’s estimator of the PD, \(\hat{PD}_{h,g}^B(t|x)\), are analysed in this section. The following notation is used.

Let \(R : \mathbb{R} \rightarrow \mathbb{R}\) be any function and define the constants
\[
c_R = \int R(t)^2 dt, \quad d_R = \int t^2 R(t) dt,
\]
and the functions
\[
R_t(u) = u^t R(u), \quad \mathbb{R}_t(u) = \int_{-\infty}^u R(t) dt.
\]

Given any function \(f : \mathbb{R}^k \rightarrow \mathbb{R}\), its first derivatives with respect to the first and second variables are denoted as follows:
\[
f'(x_1, \ldots, x_k) = \frac{\partial f(x_1, \ldots, x_k)}{\partial x_1}, \quad \hat{f}(x_1, \ldots, x_k) = \frac{\partial f(x_1, \ldots, x_k)}{\partial x_2}
\]
Correspondingly, the second derivatives with respect to the first or second variable are denoted by \(f''(x_1, \ldots, x_k)\) and \(\hat{f}(x_1, \ldots, x_k)\), respectively. Finally, let \(f * g\) be the convolution of any two functions \(f\) and \(g\).

The required assumptions are listed below. They are standard in the literature and not too restrictive in this context. They were previously assumed in Peláez et al. (2020a), Dabrowska (1989) and Iglesias-Pérez and González-Manteiga (1999) in the nonparametric conditional survival function estimation setup.
A.1. $X$, $T$, $C$ are absolutely continuous random variables.

A.2. The density function of $X$, $m$, has support $[0, 1]$.

A.3. Let $H(t|x)$ be the conditional distribution function of $Z|X = x$,

(a) Let $I = [x_1, x_2]$ be an interval contained in the support of $m$ such that,

$$0 < \gamma = \inf \{m(x) : x \in I_c\} < \sup \{m(x) : x \in I_c\} = \Gamma < \infty$$

for some $I_c = [x_1 - c, x_2 + c]$ with $c > 0$ and $0 < c \Gamma < 1$.

(b) For any $x \in I$, the random variables $T|_{X=x}$ and $C|_{X=x}$ are independent.

(c) Denoting $a_{H(t|x)} = \inf \{t/H(t|x) > 0\}$ and $b_{H(t|x)} = \inf \{t/H(t|x) = 1\}$, for any $x \in I_c$, $0 \leq a_{H(t|x)}$, $0 \leq b_{H(t|x)}$.

(d) There exist $a, b, \theta \in \mathbb{R}$ with $a < b$, satisfying $\inf \{1 - H(b|x) : x \in I_c\} \geq \theta > 0$.

Therefore $1 - H(t|x) \geq \theta > 0$ for every $(t, x) \in [a, b] \times I_c$.

A.4. The first and second derivatives of $m$, $m'(x)$ and $m''(x)$, respectively, exist and are continuous at any $x \in I_c$.

A.5. Let $H_1(t|x) = P(T \leq t, \delta = 1|X = x)$ be the conditional subdistribution function when $\delta = 1$. The corresponding density functions of $H(t)$ and $H_1(t)$ are bounded away from 0 in $[a, b]$.

A.6. The first and second derivatives with respect to $t$ of the functions $H(t|x)$ and $H_1(t|x)$, i.e. $H'(t|x)$, $H'_1(t|x)$, $H''(t|x)$ and $H''_1(t|x)$, exist and are continuous at any $(t, x) \in [a, b] \times I_c$.

A.7. The second partial derivatives first with respect to $x$ and second with respect to $t$ of the functions $H(t|x)$ and $H_1(t|x)$, i.e. $\hat{H}''(t|x)$ and $\hat{H}''_1(t|x)$ respectively, exist and are continuous at any $(t, x) \in [a, b] \times I_c$.

A.8. The kernel, $K$, is a symmetric, continuous and differentiable density function with compact support $[-1, 1]$. 
A.9. The smoothing parameters \( h = h_n \) and \( g = g_n \) satisfy \( h \to 0 \), \( g \to 0 \), and \( nh^3 \to \infty \) and \( nhg^2 \to \infty \) when \( n \to \infty \).

Using the asymptotic results for the smoothed Beran’s estimator of the conditional survival function given in Peláez et al. (2020a), the asymptotic properties of the estimator \( \widetilde{PD}_{h,g}^B(t|x) \) are obtained. The following are the functions required to state these results:

\[
\xi(Z, \delta, t, x) = \frac{1\{Z \leq t, \delta = 1\}}{1 - H(Z|x)} - \int_0^t \frac{1\{u \leq t\}}{(1 - H(u|x))^2} dH_1(u|x)
\]

\[
\eta(Z, \delta, t, x) = \int K(u)(1 - F(t - gu|x))\xi(Z, \delta, t - gu, x)du
\]

\[
\Phi_\xi(u, t, x) = E[\xi(Z_1, \delta_1, t, x)|X_1 = u]
\]

\[
\Phi_\eta(u, t, x) = E[\eta(Z_1, \delta_1, t, x)|X_1 = u]
\]

\[
J(t|x) = (1 - F(t|x))L(t|x)
\]

\[
L(t|x) = \int_0^t \frac{dH_1(z|x)}{(1 - H(z|x))^2}.
\]

\[
D(u, t_1, t_2, x) = \text{Cov}[\eta(Z_1, \delta_1, t_1, x), \eta(Z_1, \delta_1, t_2, x)|X_1 = u]m(u)
\]

\[
B_1(t, x) = \frac{d_K}{2m(x)}(2\Phi_\eta(x, t, x)m'(x) + \Phi''_\eta(x, t, x))
\]

\[
B_2(t, x) = -\frac{1}{2}d_KF''(t|x),
\]

\[
C_1(t_1, t_2, x) = \frac{2c_K}{m(x)}J(t_1|x)(1 - F(t_2|x))K * K\left(\frac{t_2 - t_1}{g}\right)
\]

\[
C_2(t_1, t_2, x) = \frac{c_K}{m(x)}\left(2J(t_1|x)f(t_2|x)K * K\left(\frac{t_1 - t_2}{g}\right)
\right.
\]

\[
\left. + 2J'(t_1|x)(1 - F(t_2|x))K * K\left(\frac{t_2 - t_1}{g}\right)\right)
\]

\[
C_3(t_1, t_2, x) = \frac{d_{K^2}}{m^2(x)}\left(m(x)(1 - F(t_1|x))(1 - F(t_2|x))\Phi_\xi(x, t_1, x)\Phi_\xi(x, t_2, x)
\right.
\]

\[
\left. + \frac{1}{2}D''(x, t_1, t_2, x)\right)
\]

\[
V_1(t, x) = \frac{c_K}{m(x)}(1 - F(t|x))^2L(t|x),
\]

\[
V_2(t, x) = \frac{c_K(c_K - 1)}{m(x)}(1 - F(t|x))^2L'(t|x),
\]

\[
V_3(t, x) = \frac{d_{K^2}}{m^2(x)}\left(m(x)(1 - F(t|x))^2(\Phi_\xi(x, t, x))^2 + \frac{1}{2}D''(x, t, t, x)\right),
\]

**Theorem 1.** Under assumptions A.1-A.9, expressions for the asymptotic bias of \( \widetilde{PD}_{h,g}^B(t|x) \), \( ABias(\widetilde{PD}_{h,g}^B(t|x)) \), and the asymptotic variance of \( \widetilde{PD}_{h,g}^B(t|x) \), \( AVar(\widetilde{PD}_{h,g}^B(t|x)) \), are the
Assuming $h > 0$, $g > 0$, and $h, g \to \infty$, the limit distribution of $\sqrt{nh}(\hat{PD}_{h,g}(t|x) - PD(t|x)) \to N(\mu, s_0)$, where

$$
\mu = C_h^{5/2} \frac{(1 - PD(t|x)) B_1(t, x) - B_1(t + b, x)}{S(t|x)} + C_h^{1/2} C_g^{4/2} \frac{(1 - PD(t|x)) B_2(t, x) - B_2(t + b, x)}{S(t|x)}
$$

and

$$
s_0^2 = \frac{V_1(t + b, x)}{S(t|x)^2} - 2 \frac{S(t + b|x) C_1(t, t + b, x)}{S(t|x)^3} + \frac{S(t + b|x)^2 V_1(t, x)}{S(t|x)^4}.
$$

Remark 1. Assuming $C_h := \lim_{n \to \infty} n^{1/5} h > 0$, but $n^{1/5} g \to 0$, the asymptotic distribution of the smoothed Beran’s PD estimator is $\sqrt{nh}(\hat{PD}_{h,g}(t|x) - PD(t|x)) \to N(\tilde{\mu}, s_0)$, with

$$
\tilde{\mu} = C_h^{5/2} \frac{(1 - PD(t|x)) B_1(t, x) - B_1(t + b, x)}{S(t|x)}
$$

Assuming $n^{1/5} h \to 0$, $n^{1/5} g \to 0$ and $\frac{nh}{(\ln n)^3} \to \infty$, the asymptotic distribution of the smoothed Beran’s PD estimator is $\sqrt{nh}(\hat{PD}_{h,g}(t|x) - PD(t|x)) \to N(0, s_0)$. 

Theorem 2. Under the assumptions of Theorem 1 and assuming $C_h := \lim_{n \to \infty} n^{1/5} h > 0$ and $C_g := \lim_{n \to \infty} n^{1/5} g > 0$, the limit distribution of $\sqrt{nh}(\hat{PD}_{h,g}(t|x) - PD(t|x)) \to N(\mu, s_0)$, where

$$
\mu = C_h^{5/2} \frac{(1 - PD(t|x)) B_1(t, x) - B_1(t + b, x)}{S(t|x)} + C_h^{1/2} C_g^{4/2} \frac{(1 - PD(t|x)) B_2(t, x) - B_2(t + b, x)}{S(t|x)}
$$

and

$$
s_0^2 = \frac{V_1(t + b, x)}{S(t|x)^2} - 2 \frac{S(t + b|x) C_1(t, t + b, x)}{S(t|x)^3} + \frac{S(t + b|x)^2 V_1(t, x)}{S(t|x)^4}.
$$
4 Simulation study

Intuitively, the improvement coming from smoothing in the time variable in the conditional survival function estimator will lead to a similar benefit for nonparametric PD estimators. The aim of this section is to make sure of this by simulation.

Two models are considered and three different censoring scenarios are distinguished for each model. Model 1 is a kind of proportional hazards model, while Model 2 moves away from this parametric model. The covariate $X$ follows a $U(0,1)$ distribution in both models.

For Model 1, the time to occurrence of the event conditional to the covariate, $T_{|X=x}$, follows a Weibull distribution with parameters $d = 2$ and $A(x) = 1 + 5x$, and the censoring time conditional to the covariate, $C_{|X=x}$, follows a Weibull distribution with parameters $d = 2$ and $B(x) = 10 + b_1x + 20x^2$ for some suitable values of $b_1$. The conditional survival function, the probability of default and the censoring conditional probability of this model are the following:

$$S(t|x) = e^{-A(x)t^d},$$

$$PD(t|x) = 1 - \frac{e^{-A(x)(t+b)^d}}{e^{-A(x)t^d}},$$

$$P(\delta = 0|X = x) = \frac{B(x)}{A(x) + B(x)}.$$

Setting $x = 0.6$, the chosen values are $b_1 = -27$, $b_1 = -22$ and $b_1 = -2$, so that the censoring probability is 0.2, 0.5 and 0.8, respectively. The time horizon is $b = 0.1$ (20\% of the time range) and the estimation is obtained in a time grid $0 < t_1 < \cdots < t_{nt}$ of size $nt$ where $t_{nt} + b = F^{-1}(0.95|x = 0.6)$.

Model 2 considers an exponential distribution with parameter $\Gamma(x) = 2 + 58x - 160x^2 + 107x^3$ for the time to occurrence of the event conditional to the covariate, $T_{|X=x}$ and an exponential distribution with parameter $\Delta(x) = 10 + d_1x + 20x^2$ for some suitable values of $d_1$ for the censoring time conditional to the covariate, $C_{|X=x}$. In this case, the conditional survival function, the probability of default and the censoring conditional probability are given by:

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

$$PD(t|x) = 1 - e^{-\Gamma(x)b},$$

$$P(\delta = 0|X = x) = \frac{B(x)}{A(x) + B(x)}.$$
\[
P(\delta = 0|X = x) = \frac{\Delta(x)}{\Gamma(x) + \Delta(x)}.
\]

Setting \( x = 0.8 \), the chosen values are \( d_1 = -113/4, d_1 = -55/2 \) and \( d_1 = -123/5 \), so that the censoring conditional probability is 0.2, 0.5 and 0.8, respectively. The time horizon is \( b = 0.7 \) (20\% of the time range) and the PD is estimated in a time grid \( 0 < t_1 < \cdots < t_{n_t} \) of size \( n_t \) where \( t_{n_t} + b = F^{-1}(0.95|x = 0.8) \).

The standard Gaussian kernel truncated in the range \([-50, 50]\) is used for both covariate and time variable smoothing. The sample size is \( n = 400 \), and the size of the lifetime grid is \( n_t = 100 \). The boundary effect is corrected using the reflexion principle proposed in Silverman (1986).

These models were previously used in the simulation study of Peláez et al. (2020a). This makes it possible to compare the results obtained in both studies.

First, the performance of the Beran’s PD estimator, \( \hat{PD}_h(t|x) \), and the smoothed Beran’s PD estimator, \( \tilde{PD}_{h,g}(t|x) \), are compared.

The optimal bandwidth for \( \hat{PD}_h(t|x) \), \( h_1 \), is taken as the value which minimizes a Monte Carlo approximation of the MISE given by

\[
MISE_x(h) = E\left( \int (\hat{PD}_h(t|x) - PD(t|x))^2 dt \right)
\]

based on \( N = 100 \) simulated samples. The value of \( MISE_x(h) \) using this smoothing parameter is approximated from \( N = 1000 \) simulated samples and used, along with its square root \((RMISE)\), as a measure of the estimation error which is committed by \( \hat{PD}_h(t|x) \).

The smoothed PD estimator \( \tilde{PD}_{h,g}(t|x) \) depends on two bandwidths: \( h \) that measures the smoothing degree introduced in the covariate and \( g \) that measures the smoothing in the time variable. Two strategies are used in order to obtain these smoothing parameters.

**Strategy 1** It consists in fixing the covariate smoothing parameter to the optima \( h_1 \) and approximating the smoothing parameter \( g \). The error to minimize is

\[
MISE_x(h_1, g) = E\left( \int (\tilde{PD}_{h_1,g}(t|x) - PD(t|x))^2 dt \right)
\]

considered as a function of the bandwidth \( g \). It is approximated from \( N = 100 \) simulated samples in a grid of 50 \( g \) values and the bandwidth which provides the smaller error is
chosen as \( g_1 \). Then, \( N = 1000 \) samples are simulated to approximate \( MISE_x(h_1, g_1) \) which is the measure of the estimation error of \( \tilde{P}D_{h,g}^{B}(t|x) \).

**Strategy 2** The optimal bandwidth \((h_2, g_2)\) is chosen (from a meshgrid of 50 values of \( h \) and 50 values of \( g \)) as the pair which minimizes some Monte Carlo approximations of

\[
MISE_x(h, g) = E \left( \int \left( \tilde{P}D_{h,g}^{B}(t|x) - PD(t|x) \right)^2 dt \right)
\]

based on \( N = 100 \) simulated samples. Then, the value of the \( MISE \) committed by \( \tilde{P}D_{h_2,g_2}^{B}(t|x) \) is approximated from \( N = 1000 \) simulated samples.

The main advantage of using Strategy 1 is its lower computational cost, but it provides rather worse results than Strategy 2. It should be noted that neither the bandwidth obtained with Strategy 1 nor Strategy 2 can be used in practice but their choice produces a fair comparison since the estimators are built using the best possible smoothing parameters.

The error curve \( MISE_x(h_1, g) \) which is minimized to obtain the optimal time smoothing parameter according to Strategy 1 is shown in Figure 1 for each level of censoring conditional probability and each model. It follows from this graphs that the optimal bandwidth \( g \) is easily approximated by Strategy 1.
Figure 1: \(MISE_x(h_1, g)\) function approximated via Monte Carlo for the smoothed Beran’s estimator using \(N = 100\) simulated samples from Model 1 (top) and Model 2 (bottom) with \(P(\delta = 0|x) = 0.2\) (left), \(P(\delta = 0|x) = 0.5\) (center) and \(P(\delta = 0|x) = 0.8\) (right).

The function \(MISE_x(h, g)\) for Models 1 and 2 for the lowest and highest censoring levels is shown in Figure 2. These graphs show the two-dimensional functions to be minimized using Strategy 2. The red area is the region where this minimum is reached and its coordinates provide the optimal smoothing bandwidths. It is clear that the choice of the time bandwidth \((g)\) notably affects the estimation error.
Figure 2: $MISE_x(h, g)$ function approximated via Monte Carlo for the smoothed Beran’s estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) with $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.8$ (right).

On the contrary, the value of $g$ for which the smallest error is committed does not seem to depend too much on the value of the covariate bandwidth ($h$). Figure 3 shows $MISE_x(h, g)$ as a function of $g$ for some fixed values of $h$ within the interval where the optimum is reached. The obtained curves are similar and close for all values of $h$, mainly
at the highest level of censoring conditional probability.

Figure 3: $\text{MISE}_x(h, g)$ function approximated via Monte Carlo for the smoothed Beran’s estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) for some fixed equispaced values of $h \in [0.1, 0.5]$ with $P(\delta = 0|x) = 0.2$ (left), $P(\delta = 0|x) = 0.5$ (center) and $P(\delta = 0|x) = 0.8$ (right).

The optimal bandwidths and the estimation errors that are committed by Beran’s estimator and the smoothed Beran’s estimator with both Strategies 1 and 2 for each model are shown in Table 1. The value of $R_i$ is defined as follows

$$R_i(t|x) = \frac{\text{RMISE}(\hat{PD}_{h_i, g_i}(t|x))}{\text{RMISE}(\hat{PD}_{h_i}(t|x))},$$

with $i = 1, 2$ depending on whether Strategy 1 or 2 is used. They help to compare the behaviour of the estimators and quantify the improvement of the smoothing over the smoothed estimator only in the covariate. The closer to 0 the value of $R_1$ or $R_2$, the greater the improvement with respect to Beran’s estimator. The relation between $R_1$ and $R_2$ ($R_1$ greater than $R_2$ or viceversa) also informs which of the two strategies reduces the error most.
In all cases, $RMISE$ values are lower for the smoothed Beran’s estimator with both Strategies 1 and 2 than for Beran’s estimator and this difference becomes bigger when increasing the censoring conditional probability. The estimator $\hat{PD}_{h,g}(t|x)$ with optimal bandwidth $(h_2, g_2)$ (Strategy 2) provides more accurate estimations than the others, since the relation $0 < R_2 < R_1 < 1$ is satisfied for all cases.

When the censoring conditional probability is 0.2 or 0.5 in Model 1, the time smoothing with the Strategy 1 reduces the error by about 35% and this improvement is about 60% when the conditional probability of censoring is 0.8. This improvement increases by an additional $5 - 10\%$ when using Strategy 2. The error reduction in Model 2 with respect to the nonsmoothed PD estimator is more significant, reaching 50% and 70% when using Strategy 1 and censoring is regular or heavy, respectively. This reduction is bigger when using Strategy 2, reaching $75 - 80\%$.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2</td>
<td>0.5</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>$P(\delta = 0</td>
<td>x)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{PD}_{h_1}$</td>
<td>$h_1$</td>
<td>0.35714 0.34694 0.39796</td>
<td>0.10306</td>
<td>0.12265 0.14224</td>
</tr>
<tr>
<td>RMISE</td>
<td></td>
<td>0.05437 0.11195 0.25738</td>
<td>0.27128</td>
<td>0.49813 0.67999</td>
</tr>
<tr>
<td>$\hat{PD}_{h_1,g_1}$</td>
<td>$h_1$</td>
<td>0.35714 0.34694 0.39796</td>
<td>0.10306</td>
<td>0.12265 0.14224</td>
</tr>
<tr>
<td>RMISE</td>
<td></td>
<td>0.08347 0.12265 0.18633</td>
<td>1.18571</td>
<td>1.47755 1.82245</td>
</tr>
<tr>
<td>$R_1$</td>
<td></td>
<td>0.74765 0.58723 0.28153</td>
<td>0.92974</td>
<td>0.48489 0.30233</td>
</tr>
<tr>
<td>$\hat{PD}_{h_2,g_2}$</td>
<td>$h_2$</td>
<td>0.21429 0.15714 0.18980</td>
<td>0.10816</td>
<td>0.25918 1.00000</td>
</tr>
<tr>
<td>RMISE</td>
<td></td>
<td>0.09327 0.13735 0.19612</td>
<td>1.21122</td>
<td>1.61020 1.90204</td>
</tr>
<tr>
<td>$R_2$</td>
<td></td>
<td>0.03845 0.05941 0.06208</td>
<td>0.09210</td>
<td>0.12350 0.13434</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.70719 0.53068 0.24120</td>
<td>0.33950</td>
<td>0.24793 0.19756</td>
</tr>
</tbody>
</table>

Table 1: Optimal bandwidths, $RMISE$, $R_1$ and $R_2$ of the PD estimation for Beran’s estimator, the smoothed Beran’s estimator with Strategy 1 and the smoothed Beran’s estimator with Strategy 2 in each level of censoring conditional probability for Model 1 and Model 2.
The computation time of both estimators should be considered in the comparison. Table 2 shows the CPU times (in seconds) that Beran’s estimator and the smoothed Beran’s estimator spends on estimating the probability of default curve in a 100-point time grid and a fixed value of $x$ for different values of the sample size. The smoothing parameters are fixed to the optimal values before estimating the curve.

<table>
<thead>
<tr>
<th>n</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>1200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beran</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>SBeran</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 2: CPU time (in seconds) for estimating $PD(t|x)$ in a time grid of size 100 for each estimator and different sample sizes.

The result of the analysis shows that the second smoothing increases the CPU time and the Beran’s PD estimator with double smoothing is more affected by the increase in sample size than Beran’s estimator.

It is expected that the two strategies used to find the optimal bandwidths will also have different computational efficiency. Table 3 shows the CPU time (in minutes) for each strategy and several number of trials to check this. For both strategies the size of each sample is $n = 400$ and the PD is estimated in a time grid of size 100. The number of simulated samples to approximate the $MISE$ by Monte Carlo is the parameter that varies in order to compare the CPU time of each strategy. Strategy 1 has a clear computational advantage over Strategy 2, since Strategy 2 is significantly slower.

<table>
<thead>
<tr>
<th>N</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy 1</td>
<td>3.01</td>
<td>4.28</td>
<td>5.40</td>
<td>7.32</td>
</tr>
<tr>
<td>Strategy 2</td>
<td>80.61</td>
<td>204.51</td>
<td>228.01</td>
<td>296.95</td>
</tr>
</tbody>
</table>

Table 3: CPU time (in minutes) for approximating the optimal bandwidth $(h, g)$ for $\tilde{PD}_{h,g}^B(t|x)$ with Strategies 1 and 2.

Since the improvement in statistical efficiency that the time variable smoothing provides to Beran’s PD estimator has been verified, it is interesting to check if other PD estimators based on estimators for the survival function are equally improved by applying this type of
smoothing. With this aim, in a second simulation study, the behaviour of different doubly smoothed PD estimators is compared to each other as well as to the smoothed Beran’s estimator.

Three nonparametric estimators of the conditional survival function are considered to obtain the corresponding smoothed estimator given by (2) and then, to estimate the probability of default through the expression given in (3). These survival estimator are Beran’s estimator, the Weighted Nadaraya-Watson estimator (WNW) and the Van Keilegom-Akritas estimator (VKA). The WNW estimator was built following the idea of Cai (2003), where the survival estimator is based on a local lineal regression. Since the weighted local lineal estimator presents problems when estimating probabilities, a constant fit is proposed in Peláez et al. (2020b). The VKA estimator was defined in Van Keilegom and Akritas (1999) and Van Keilegom et al. (2001). The expressions for both estimators are shown in Section 2 of Peláez et al. (2020a) and they are denoted by \( \tilde{S}_{WNW}^{h}(t|x) \) and \( \tilde{S}_{VKA}^{h}(t|x) \). Their smoothed versions are built according to Equation (2), obtaining the following smoothed survival estimators: \( \tilde{S}_{h,g}^{WNW}(t|x) \) and \( \tilde{S}_{h,g}^{VKA}(t|x) \). It is enough to replace \( \tilde{S}_{h,g}(t|x) \) with \( \tilde{S}_{h,g}^{B}(t|x) \), \( \tilde{S}_{h,g}^{WNW}(t|x) \) and \( \tilde{S}_{h,g}^{VKA}(t|x) \) in Equation (3) to get the three nonparametric smoothed estimators of PD \( PD(t|x) \) which are denoted by \( \tilde{PD}_{h,g}^{B}(t|x) \), \( \tilde{PD}_{h,g}^{WNW}(t|x) \) and \( \tilde{PD}_{h,g}^{VKA}(t|x) \). These doubly smoothed PD estimators are studied here by simulation.

Since the computational times of these estimators are pretty high, only Strategy 1 is used to look for the optimal smoothing parameters, since Strategy 2 would further increase the computation time of the simulations.

In order to quantify the improvement that the smoothing provides to the PD estimators and compare the performance of the three estimators, the ratios \( R_{S}^{•} \) and \( R_{c}^{•} \) are defined as follows

\[
R_{S}^{•}(t|x) = \frac{RMISE(\tilde{PD}_{h_{1},g_{1}}^{•}(t|x))}{RMISE(\tilde{PD}_{h_{1}}^{•}(t|x))}
\]

\[
R_{c}^{•}(t|x) = \frac{RMISE(\tilde{PD}_{h_{1},g_{1}}^{•}(t|x))}{RMISE(\tilde{PD}_{h_{2},g_{2}}^{B}(t|x))}
\]

being \( • = B, WNW, VKA \) and they are included in Table 4 along with the approximation of the optimal smoothing parameters and the error committed by each estimator.
Table 4: Optimal bandwidths, $RMISE$, $R_S^*$ and $R_c^*$ of the PD estimation for the smoothed Beran’s estimator, the smoothed WNW estimator and the smoothed VKA estimator with Strategy 1 for each level of censoring conditional probability for Models 1 and 2.

The values of $R_S^*$ report the influence of the smoothing. The smaller the value, the better the estimation obtained with the smoothed estimator compared to the corresponding nonsmoothed estimator. Since its value is less than 1 in almost all cases of Models 1 and 2, it is confirmed that the smoothing in the time variable is an improvement of any of the estimators, mainly when censoring is heavy. In addition, the smaller the value of $R_S^*$, the greater the improvement that smoothing provides to the estimator. In this line, the doubly smoothed WNW estimator is the estimator whose error is reduced the most.

The value of $R_c^*$ is useful to compare the behaviour of the three estimators with the behaviour of $\tilde{PD}_{h_2,g_2}^B(t|x)$ (the smoothed Beran’s estimator with Strategy 2). Since almost all the $R_c^*$ values obtained are greater than 1, it can be concluded that the smoothed Beran’s estimator with Strategy 2 provides more accurate estimations of the probability of default. Moreover, the closer to 1 the value of $R_c^*$, the better the estimators. Thus, in general terms, the smoothed Beran’s estimator with Strategy 1 is the second best option for estimating the probability of default.

In some cases the smoothed WNW estimator presents an $R_c^*$ less than 1, which indicates...
that the error it makes is occasionally smaller than the error committed by the smoothed Beran’s estimator with Strategy 2. Therefore, the smoothed WNW estimator appears to be competitive with Beran’s in some contexts.

It is also appropriate to analyse the differences between the computational times of these techniques. Table 5 shows the CPU time (in seconds) that is needed by each estimator to obtain the probability of default curve in a time grid of size 100 and a fixed value of $x$ for different values of the sample size.

```
<table>
<thead>
<tr>
<th>n</th>
<th>Beran</th>
<th>SBeran</th>
<th>SWNW</th>
<th>SVKA</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.03</td>
<td>2.30</td>
<td>0.42</td>
</tr>
<tr>
<td>100</td>
<td>0.01</td>
<td>0.03</td>
<td>6.33</td>
<td>1.80</td>
</tr>
<tr>
<td>200</td>
<td>0.01</td>
<td>0.03</td>
<td>25.97</td>
<td>7.34</td>
</tr>
<tr>
<td>400</td>
<td>0.02</td>
<td>0.05</td>
<td>140.62</td>
<td>53.99</td>
</tr>
<tr>
<td>1200</td>
<td>0.03</td>
<td>0.20</td>
<td>1459.35</td>
<td>507.36</td>
</tr>
</tbody>
</table>
```

Table 5: CPU time (in seconds) for estimating $PD(t|x)$ in a time grid of size 100 for every estimator and different sample sizes.

The time variable smoothing clearly implies an increase of the CPU time. The three doubly smoothed PD estimators which were considered have higher CPU times than Beran’s estimator. It should be noted that the smoothed Beran’s estimator is least affected by the increase of the sample size and it is the fastest of the three doubly smoothed estimators. The CPU time of the smoothed VKA increases very fast with the sample size but he slowest and most affected by the sample size increase is the smoothed WNW estimator.

## 5 Application to real data

The smoothed Beran’s estimator is now applied to a real data set. The data consists of a sample of 10000 consumer credits from a Spanish bank registered between July 2004 and November 2006. They were previously used in Cao et al. (2009) and Peláez et al. (2020b) and a more exhaustive study of them can be found in Peláez et al. (2020b). The data set provides the credit scoring observed for each borrower, the observed lifetime of
the credit in months and the uncensoring indicator. The sample censoring percentage is 92.8%; equivalently, the proportion of credits for which the default is observed is 7.2%. This value has been modified so as not to show the true solvency situation of the bank and thus preserve confidentiality.

The probability of default for $x = 0.5$ at horizon $b = 5$ month is estimated in a time grid along the interval $[0, 25]$ using the smoothed Beran’s estimator. The estimation is obtained with some different possible values of the time variable smoothing parameter, while the covariate bandwidth is fixed to a reasonable value ($h = 0.05$), since it has a very slight influence on the estimation. Figure 4 shows the results.

Beran’s estimation and the smoothed Beran’s estimation of the conditional survival function and the $PD$ for $h = 0.05$ and $g = 1$ are shown in Figure 5. Although the survival estimations are very similar with both estimators, it can be seen how the roughness of the curve estimation is reduced and the jumps are removed when using the smoothed Beran’s estimator. This is even more remarkable when estimating the probability of default.

According to the smoothed Beran’s estimation, the probability of default has an increasing tendency. It follows from it that the higher the debt maturity, the higher the probability of falling into default for an individual with this credit scoring.
Figure 4: Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.8$ by means of the smoothed Beran estimator on the consumer credits dataset for three different values of $g$.

Figure 5: Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.8$ by means of Beran’s estimator (dashed line) and smoothed Beran’s estimator (solid line) on the consumer credits dataset.
6 Conclusions

A general doubly smoothed estimator of the probability of default is proposed in this paper. Asymptotic properties of the smoothed PD estimator based on the smoothed Beran’s estimator for the survival function are proved and its asymptotic distribution is found. This doubly smoothed Beran’s estimator of the PD showed a remarkably good behavior in the analyzed scenarios of the simulation study. The time variable smoothing results in a significant improvement of the PD estimator, since the estimation error \(MISE\) is reduced, mainly when using Strategy 2 for approximating the optimal bandwidth. However, the computational time is increased. These same evidences were observed in any of the smoothed PD estimators studied by simulation. Although the smoothed Beran’s estimator of the PD turned out to have the most stable behaviour and to be the fastest of all.

7 Proofs

Proofs of the results shown in Section 3 are done using results from papers Peláez et al. (2020b) and Peláez et al. (2020a).

Proof of Theorem 1.

Denote \(P = S(t + b|x), Q = S(t|x)\) and \(PD(t|x) = 1 - \frac{P}{Q}\). Similarly, \(\tilde{P} = \tilde{S}_{h,g}^B(t + b|x), \tilde{Q} = \tilde{S}_{h,g}^B(t|x)\) and \(\tilde{PD}_{h,g}^B(t|x) = 1 - \frac{\tilde{P}}{\tilde{Q}}\). As a consequence of the proof of Theorem 1 in Peláez et al. (2020b):

\[
\begin{align*}
ABias(\tilde{PD}_{h,g}^B(t|x)) &= \alpha_1 + \alpha_2 + \alpha_3, \\
AVar(\tilde{PD}_{h,g}^B(t|x)) &= \beta_1 + \beta_2 + \beta_3,
\end{align*}
\]

where

\[
\begin{align*}
\alpha_1 &= \frac{P}{Q} - \frac{E(\tilde{P})}{E(\tilde{Q})}, & \alpha_2 &= \frac{Cov(\tilde{P}, \tilde{Q})}{E(\tilde{Q})^2}, & \alpha_3 &= -\frac{E\left[\frac{\tilde{P}}{\tilde{Q}}(\tilde{Q} - E(\tilde{Q}))^2\right]}{E(\tilde{Q})^2} \\
\beta_1 &= \frac{Var(\tilde{P})}{E(\tilde{Q})^2}, & \beta_2 &= -\frac{2E(\tilde{P})Cov(\tilde{P}, \tilde{Q})}{E(\tilde{Q})^3}, & \beta_3 &= \frac{E(\tilde{P})^2Var(\tilde{Q})}{E(\tilde{Q})^4}.
\end{align*}
\]
The asymptotic expressions for the bias, the covariance and the variance of the survival estimator \( \tilde{S}_{h,g}(t|x) \) are obtained from Theorems 3 and 4 of Peláez et al. (2020a):

\[
\text{Bias}(\tilde{S}_{h,g}(t|x)) = B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2),
\]

\[
\text{Cov}(\tilde{S}_{h,g}(t_1|x), \tilde{S}_{h,g}(t_2|x)) = C_1(t_1,t_2,x) \frac{1}{nh} + C_2(t_1,t_2,x) \frac{g}{nh} + C_3(t_1,t_2,x) \frac{h}{n} + R_n(t,x),
\]

\[
\text{Var}(\tilde{S}_{h,g}(t|x)) = V_1(t,x) \frac{1}{nh} + V_2(t,x) \frac{g}{nh} + V_3(t,x) \frac{h}{n} + R_n(t,x)
\]

where \( R_n(t,x) = o \left( \frac{g^2}{nh} + \frac{h}{n} \right) \).

Considering Equations (8)-(12), detailed expressions for \( \alpha_1 \), \( \alpha_2 \) and \( \alpha_3 \) are obtained as follows:

\[
\alpha_1 = \frac{P Q - P B_1(t+b,x)h^2 - B_2(t+b,x)g^2 + o(h^2) + o(g^2) - P Q + PB_1(t,x)h^2 + PB_2(t,x)g^2 + o(h^2) + o(g^2)}{Q \left( B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2) \right)}
\]

\[
= \frac{P B_1(t,x)h^2 - QB_1(t+b,x)h^2 + o(h^2) + o(g^2)}{Q \left( B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2) \right)} + \frac{P B_2(t,x)g^2 - QB_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q \left( B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2) \right)}
\]

\[
= \frac{1 - PD(t|x) B_1(t,x) - B_1(t+b,x) h^2}{S(t|x)} + \frac{1 - PD(t|x) B_2(t,x) - B_2(t+b,x) g^2 + o(h^2) + o(g^2)}{S(t|x)},
\]

\[
\alpha_2 = \frac{C_1(t,t+b,x)}{S(t|x)^2} \frac{1}{nh} + \frac{C_2(t,t+b,x)}{S(t|x)^2} \frac{g}{nh} + \frac{C_3(t,t+b,x)}{S(t|x)^2} \frac{h}{n} + R_n(t,x),
\]

\[
\alpha_3 = \frac{E \left[ \frac{\tilde{Q}(\tilde{Q}) - E(\tilde{Q})^2}{\tilde{Q}^2} \right]}{E(\tilde{Q})^2} \leq \frac{\text{Var}(\tilde{Q})}{E(\tilde{Q})^2}
\]

\[
= \frac{V_1(t,x)}{S(t|x)^2 nh} + \frac{V_2(t,x) g}{S(t|x)^2 nh} + \frac{V_3(t,x) h}{S(t|x)^2 n} + R_n(t,x).
\]
Plugging (13), (14) and (15) into (6) and using Assumption A.9,
\[
\text{ABias}(\tilde{PD}^B_{h,g}(t|x)) = \frac{(1 - PD(t|x))B_1(t, x) - B_1(t + b, x)}{S(t|x)} h^2 + \frac{(1 - PD(t|x))B_2(t, x) - B_2(t + b, x)}{S(t|x)} g^2.
\]
and the bias part in Theorem 1 is proved.

Now, expressions (9), (10), (11) and (12) lead to
\[
\beta_1 = \frac{V_1(t + b, x)}{S(t|x)^2} \frac{1}{nh} + \frac{V_2(t + b, x)}{S(t|x)^2} \frac{g}{nh} + \frac{V_3(t + b, x)}{S(t|x)^2} \frac{h}{n} + R_n(t, x) \tag{16}
\]
\[
\beta_2 = -2 \frac{S(t + b, x)C_1(t, t + b, x)}{S(t|x)^3} \frac{1}{nh} - 2 \frac{S(t + b, x)C_2(t, t + b, x)}{S(t|x)^3} \frac{g}{nh}
- 2 \frac{S(t + b, x)C_3(t, t + b, x)}{S(t|x)^3} \frac{h}{n} + R_n(t, x) \tag{17}
\]
\[
\beta_3 = \frac{S(t + b, x)^2V_1(t, x)}{S(t|x)^4} \frac{1}{nh} + \frac{S(t + b, x)^2V_2(t, x)}{S(t|x)^4} \frac{g}{nh}
+ \frac{S(t + b, x)^2V_3(t, x)}{S(t|x)^4} \frac{h}{n} + R_n(t, x) \tag{18}
\]
and plugging Equations (16), (17) and (18) in (7) the variance part in Theorem 1 is proved.

\[
\square
\]

**Proof of Theorem 2**

From Equations (1) and (3) follows
\[
\frac{\tilde{S}^B_{h,g}(t + b|x)}{\tilde{S}^B_{h,g}(t|x)} - \frac{S(t + b|x)}{S(t|x)} = -\left(\tilde{PD}^B_{h,g}(t|x) - PD(t|x)\right). \tag{19}
\]

On the other hand, denoting \(a_1 = \frac{1}{S(t|x)}\), \(a_2 = -\frac{S(t + b|x)}{S(t|x)^2}\) and
\[
C(\tilde{S}^B_{h,g}(t|x)) = \frac{S(t|x)\left(\tilde{S}^B_{h,g}(t + b|x) - S(t + b|x)\right) - S(t + b|x)\left(\tilde{S}^B_{h,g}(t|x) - S(t|x)\right)}{\tilde{S}^B_{h,g}(t|x)S(t|x)},
\]
it holds
\[
\frac{\tilde{S}^B_{h,g}(t + b|x)}{\tilde{S}^B_{h,g}(t|x)} - \frac{S(t + b|x)}{S(t|x)} = a_1\left(\tilde{S}^B_{h,g}(t + b|x) - S(t + b|x)\right) + a_2\left(\tilde{S}^B_{h,g}(t|x) - S(t|x)\right)
+ C(\tilde{S}^B_{h,g}(t|x)) \left(1 - \frac{\tilde{S}^B_{h,g}(t|x)}{S(t|x)}\right),
\]
\[
20.
\]
and considering (19):

\[ PD(t|x) = \frac{\tilde{PD}^B_{h,g}(t|x)}{PD(t|x)} \]
\[ = a_1(\tilde{S}^B_{h,g}(t + b|x) - S(t + b|x)) + a_2(\tilde{S}^B_{h,g}(t|x) - S(t|x)) + C(\tilde{S}^B_{h,g}(t|x)) \left( 1 - \frac{\tilde{S}^B_{h,g}(t|x)}{S(t|x)} \right). \]  \hspace{1cm} (20)

Since \( \tilde{S}^B_{h,g}(t|x) \) is a consistent estimator of \( S(t|x) \), \( \tilde{S}^B_{h,g}(t|x) \xrightarrow{p} S(t|x) \), thus,

\[ 1 - \frac{\tilde{S}^B_{h,g}(t|x)}{S(t|x)} \xrightarrow{p} 0. \]

Therefore, the asymptotic distribution of \( \sqrt{n}h(\tilde{PD}^B_{h,g}(t|x) - PD(t|x)) \) is the same as the asymptotic distribution of the linear combination

\[ a_1\sqrt{n}h(\tilde{S}^B_{h,g}(t + b|x) - S(t + b|x)) + a_2\sqrt{n}h(\tilde{S}^B_{h,g}(t|x) - S(t|x)). \]

From Lemma 1 in Peláez et al. (2020a), \( \tilde{S}^B_{h,g}(t|x) \) is split up into the following terms

\[ \tilde{S}^B_{h,g}(t|x) = S(t|x) + \sum_{i=1}^{n} \varphi_{n,i}(t,x) + B_2(t,x)g^2 + R_n(t|x), \]  \hspace{1cm} (21)

where \( \varphi_{n,i}(t,x) = \frac{1}{nh} K((x - X_i)/h) \eta(Z_i, \delta_i, t, x) \) are independent and identically distributed random variables for all \( i = 1, ..., n \) and \( R_n(t|x) \) is negligible with respect to the other terms:

\[ R_n(t|x) = O_p\left( \frac{\ln n}{nh} \right)^{3/4} + o(g^2) + O_p\left( h^2 + \frac{1}{\sqrt{nh}} \right) \sum_{i=1}^{n} \varphi_{n,i}(t,x). \]

Using (21),

\[ a_1\sqrt{n}h(\tilde{S}^B_{h,g}(t + b|x) - S(t + b|x)) + a_2\sqrt{n}h(\tilde{S}^B_{h,g}(t|x) - S(t|x)) \]
\[ = \sum_{i=1}^{n} \tilde{\varphi}_{n,i}(t,x) + a_1B_2(t + b, x)g^2\sqrt{nh} + a_2B_2(t, x)g^2\sqrt{nh} + \tilde{R}_n(t,x) \]  \hspace{1cm} (22)

where

\[ \tilde{\varphi}_{n,i}(t,x) = \sqrt{nh}(a_1\varphi_{n,i}(t + b, x) + a_2\varphi_{n,i}(t, x)) \]  \hspace{1cm} (23)

and

\[ \tilde{R}_n(t, x) = \sqrt{nh}(a_1R_n(t + b, x) + a_2R_n(t, x)) \]
\[ = \sqrt{nh}(a_1 + a_2)O_p\left( \frac{\ln n}{nh} \right)^{3/4} + \sqrt{nh}(a_1 + a_2)o(g^2) \]
\[ + O_p\left( h^2 + \frac{1}{\sqrt{nh}} \right) \sum_{i=1}^{n} \tilde{\varphi}_{n,i}(t,x). \]  \hspace{1cm} (24)
Since $h \to 0$ and $nh \to \infty$, the term $O_p \left( h^2 \frac{1}{\sqrt{nh}} \right)$ \( \sum_{i=1}^{n} \tilde{r}_{n,i}(t, x) \) in (24) is negligible with respect to $\sum_{i=1}^{n} \tilde{r}_{n,i}(t, x)$ in (22). Given that $g \to 0$, the term $\sqrt{nh}(a_1 + a_2)O_p \left( \frac{\ln n}{nh} \right)^{3/4}$ in (24) is negligible with respect to $a_1 B_2(t + b, x)g^2 \sqrt{nh} + a_2 B_2(t, x)g^2 \sqrt{nh}$ in (22). Finally, the term $\sqrt{nh}(a_1 + a_2)O_p \left( \frac{\ln n}{nh} \right)^{3/4}$ in (24) is negligible with respect to $\sum_{i=1}^{n} \tilde{r}_{n,i}(t, x)$ in (22) because $\frac{nh}{(\ln n)^3} = \frac{C_h n^{3/5}}{(\ln n)^3} \to \infty$. The variance of the dominant term in (22) is $O(1)$:

$$\text{Var}(\sum_{i=1}^{n} \tilde{r}_{n,i}(t, x)) = n \text{Var}(\tilde{r}_{n,1}(t, x)) = n^2 h \left( a_1^2 \text{Var}(\varphi_{n,1}(t + b, x)) + a_2^2 \text{Var}(\varphi_{n,1}(t, x)) \right) + 2a_1a_2 \text{Cov}(\varphi_{n,1}(t + b, x), \varphi_{n,1}(t, x)).$$

From the proof of Theorem 3 in Peláez et al. (2020a),

$$\text{Cov}(\varphi_{n,1}(t_1, x), \varphi_{n,1}(t_2, x)) = \frac{2c_K}{m(x) n^2} \left( 1 - F(t_1|x) \right) \left( 1 - F(t_2|x) \right) L(t_1|x) \mathbb{K} * K \left( \frac{t_2 - t_1}{g} \right) \frac{1}{h} + O \left( \frac{g}{n^3 h} \right).$$

In particular, for $t_1 = t$, $t_2 = t + b$, $\mathbb{K} * K \left( \frac{t_2 - t_1}{g} \right) = \mathbb{K} * K \left( \frac{b}{g} \right)$ and

$$\lim_{n \to \infty} \mathbb{K} * K \left( \frac{b}{g} \right) = \lim_{u \to \infty} \int_{-\infty}^{+\infty} K(y) \mathbb{K}(u - y) dy$$

$$= \int_{-\infty}^{+\infty} \lim_{u \to \infty} \mathbb{K}(u - y) K(y) dy = \int_{-\infty}^{+\infty} K(y) dy = 1.$$

Consequently,

$$\text{Cov}(\varphi_{n,1}(t + b, x), \varphi_{n,1}(t, x)) = \frac{2c_K}{m(x) n^2} \left( 1 - F(t|x) \right) \left( 1 - F(t + b|x) \right) L(t|x) \frac{1}{h} + O \left( \frac{g}{n^3 h} \right) + o \left( \frac{1}{n^2 h} \right).$$

(26)

For $t_1 = t_2$,

$$\mathbb{K} * K \left( \frac{t_2 - t_1}{g} \right) = \mathbb{K} * K(0) = \int \mathbb{K}(u) K(-u) du$$

$$= \int \mathbb{K}(u) K(u) du = \int K(u) \left( \int_{-\infty}^{u} K(v) dv \right) du = \int \int_{\{v \leq u\}} K(u) K(v) dv du$$

$$= \frac{1}{2} \left( \int \int_{\{v \leq u\}} K(u) K(v) dv du + \int \int_{\{u \leq v\}} K(v) K(u) dv du \right)$$

$$= \frac{1}{2} \int \int_{\mathbb{R}^2} K(u) K(v) dv du = \frac{1}{2}.$$
We will now check Lindeberg’s condition: 

\[
\text{Var}(\varphi_{n,1}(t, x)) = \frac{cK}{m(x)n^2} (1 - F(t|x))^2 L(t|x) \frac{1}{h} + O\left(\frac{g}{n^3h}\right). \tag{27}
\]

Replacing (26) and (27) in (25),

\[
\text{Var}\left(\sum_{i=1}^{n} \tilde{\varphi}_{n,i}(t, x)\right) = a_1^2 \frac{cK}{m(x)} (1 - F(t + b|x))^2 L(t + b|x) + a_2^2 \frac{cK}{m(x)} (1 - F(t|x))^2 L(t|x) + 4a_1a_2 \frac{cK}{m(x)} (1 - F(t|x))(1 - F(t + b|x)) L(t|x) + O\left(\frac{g}{n}\right) + o(1).
\]

Thus, \(\text{Var}\left(\sum_{i=1}^{n} \tilde{\varphi}_{n,i}(t, x)\right) = O(1)\) and the linear combination can be expressed as (22) with \(\tilde{R}_n(t, x)\) negligible with respect to the term \(\sum_{i=1}^{n} \tilde{\varphi}_{n,i}(t, x)\). Therefore, we proceed to analyse the asymptotic distribution of \(\sum_{i=1}^{n} \tilde{\varphi}_{n,i}(t, x)\).

As the variables \(\varphi_{n,i}(t, x)\) are independent and identically distributed for all \(i = 1, ..., n\), the variables \(\tilde{\varphi}_{n,i}(t, x)\) are also so. In addition, \(\text{Var}(\tilde{\varphi}_{n,i}(t, x))\) exists and it is finite for all \(i = 1, ..., n\). In this scenario, if Lindeberg’s condition for triangular arrays (see Theorem 7.2 in Billingsley (1968)) is satisfied, then

\[
\sum_{i=1}^{n} \left(\tilde{\varphi}_{n,i}(t, x) - E[\tilde{\varphi}_{n,i}(t, x)]\right) \xrightarrow{d} N(0, s_0).
\tag{28}
\]

where

\[
s_0^2 = a_1^2 \frac{cK}{m(x)} (1 - F(t + b|x))^2 L(t + b|x) + a_2^2 \frac{cK}{m(x)} (1 - F(t|x))^2 L(t|x) + 4a_1a_2 \frac{cK}{m(x)} (1 - F(t|x))(1 - F(t + b|x)) L(t|x) \tag{29}
\]

We will now check Lindeberg’s condition:

\[
\lim_{n \to \infty} \frac{1}{s_0^2} E\left[\sum_{i=1}^{n} \left(\tilde{\varphi}_{n,i}(t, x) - E[\tilde{\varphi}_{n,i}(t, x)]\right)^2 \mathbb{1}_{n,i}\right] = 0 \tag{30}
\]

for every \(\varepsilon > 0\), where \(\mathbb{1}_{n,i}\) denotes the indicator function given by

\[
\mathbb{1}_{n,i} = \mathbb{1}\left(\left|\tilde{\varphi}_{n,i}(t, x) - E[\tilde{\varphi}_{n,i}(t, x)]\right| > \varepsilon s_0\right).
\]

Using assumption A.3d, \(\xi(Z, \delta, t, x)\) is found out to be bounded:

\[
|\xi(Z, \delta, t, x)| = \frac{1_{\{Z \leq t, \delta = 1\}}}{1 - H(Z|x)} - \int_{0}^{t} \frac{dH_1(u|x)}{(1 - H(u|x))^2} \leq \frac{1_{\{Z \leq t, \delta = 1\}}}{1 - H(Z|x)} + \int_{0}^{t} \frac{dH_1(u|x)}{(1 - H(u|x))^2} \leq \frac{1}{\overline{\theta}} + \frac{1}{\overline{\theta}^2}
\]

27
and, consequently, $\eta$ is also bounded:

$$|\eta(Z, \delta, t, x)| \leq \int K(u)(1 - F(t - gu|x))(\frac{1}{\delta} + \frac{1}{\delta^2})du$$

$$= \left(\frac{1}{\delta} + \frac{1}{\delta^2}\right)\left((1 - F(t|x)) + \frac{g^2}{2} d_K(1 - F''(t|x))\right) + O(g^2).$$

Since $\eta$ is bounded, $K$ and $m(x)$ have compact support and $nh \to \infty$, $\{\bar{\varphi}_{n,i}(t, x), i = 1, ..., n, n \in \mathbb{N}\}$ is a sequence of random variables which is bounded by a convergent to zero sequence. Hence, there exists $n_0 \in \mathbb{N}$ such that for all $i = 1, ..., n$, \(\mathbb{1}_{n,i} = 0\) for all $n \geq n_0$ and consequently,

$$\lim_{n \to \infty} \frac{1}{n} E\left[\sum_{i=1}^{n} (\bar{\varphi}_{n,i}(t, x) - E[\bar{\varphi}_{n,i}(t, x)])^2 \mathbb{1}_{n,i}\right] = 0,$$

which proves Lindeberg’s condition given in (30).

Furthermore, from Theorem 3 in Peláez et al. (2020a),

$$E(\varphi_{n,1}(t, x)) = B_1(t, x)\frac{h^2}{n} + o\left(\frac{h^2}{n}\right),$$

so,

$$E\left(\sum_{i=1}^{n} \bar{\varphi}_{n,i}(t, x)\right) = nE(\bar{\varphi}_{n,1}(t, x)) = a_1n\sqrt{nh}E(\varphi_{n,1}(t + b, x)) + a_2n\sqrt{nh}E(\varphi_{n,1}(t, x))$$

$$= \sqrt{nh}^5 (a_1B_1(t + b, x) + a_2B_1(t, x) + o(h^2)).$$

Therefore, taking into account that $h = C_hn^{-1/5}$ we have

$$\sum_{i=1}^{n} \bar{\varphi}_{n,i}(t, x) \xrightarrow{d} N(\mu_0, s_0).$$

where

$$\mu_0 = C_h^{5/2}(a_1B_1(t + b, x) + a_2B_1(t, x)).$$

Consequently, recalling (22) and assuming $g = C_gn^{-1/5},$

$$a_1\sqrt{nh}(\bar{S}_{h,g}(t + b|x) - S(t + b|x)) + a_2\sqrt{nh}(\bar{S}_{h,g}(t|x) - S(t|x)) \xrightarrow{d} N(\mu_1, s_0)$$

where

$$\mu_1 = \mu_0 + C_h^{1/2}C_g^{1/2}(a_1B_2(t + b, x) + a_2B_2(t, x)).$$
Finally, using equation (20) with $a_1 = \frac{1}{S(t|x)}$ and $a_2 = -\frac{S(t + b|x)}{S(t|x)^2}$, the asymptotic distribution of the PD estimator holds,

$$\sqrt{n h} (\widetilde{PD}_{B_{h,g}}(t|x) - PD(t|x)) \overset{d}{\to} N(\mu, s_0)$$

where $\mu = -\mu_1$. Then,

$$\mu = C^{5/2}_h \left( \frac{S(t + b|x)}{S(t|x)^2} B_1(t, x) - \frac{B_1(t + b, x)}{S(t|x)} \right)$$

$$+ C^{1/2}_h C^{1/2}_g \left( \frac{S(t + b|x)}{S(t|x)^2} B_2(t, x) - \frac{B_2(t + b, x)}{S(t|x)} \right)$$

$$= C^{5/2}_h \left( 1 - PD(t|x) \right) B_1(t, x) - \frac{B_1(t + b, x)}{S(t|x)}$$

$$+ C^{1/2}_h C^{1/2}_g \left( 1 - PD(t|x) \right) B_2(t, x) - \frac{B_2(t + b, x)}{S(t|x)}$$

and

$$s_0^2 = \frac{1}{S(t|x)^2} \frac{c_K(1 - F(t + b|x))^2 L(t + b|x)}{m(x)} + \frac{S(t + b|x)^2}{S(t|x)^4} \frac{c_K(1 - F(t|x))^2 L(t|x)}{m(x)}$$

$$- 4 \frac{S(t + b|x)}{S(t|x)^3} \frac{c_K(1 - F(t|x))(1 - F(t + b|x)) L(t|x)}{m(x)}$$

$$= \frac{V_1(t + b, x) m(x)}{S(t|x)^2} - 2 \frac{S(t + b|x) C_1(t, t + b, x)}{S(t|x)^3} + \frac{S(t + b|x)^2 V_1(t, x)}{S(t|x)^4}.$$

References


