SORT 45 (2) July-December 2021, 93-120 DOI: 10.2436/20.8080.02.111

Nonparametric estimation of the probability of default with double smoothing

Rebeca Peláez[1], Ricardo Cao[2] and Juan M. Vilar[2]

Abstract

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. An empirical study, based on modified real data, illustrates its practical application and a simulation study shows the performance of the proposed estimator and compares its behaviour with smoothed estimators only in the covariate. Asymptotic expressions for the bias and the variance of the probability of default estimator are found and asymptotic normality is proved.

MSC:62G05, 62G07, 62G08, 62G20, 62N02, 62P20.

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

1. Introduction

Credit risk is an important research area. 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 features of the credit and the clients. They usually build some linear combination (credit scoring) based on these informative variables and the probability of default is allowed to depend on this scoring

[1] Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

[2] Research Group MODES, Department of Mathematics, CITIC, University of A Coruña and ITMATI, A Coruña, Spain.

Received: February 2021

Accepted: June 2021

x, $PD(t|x)$. A common aproach in credit scoring is using logistic regression to build the index. The logistic model of credit scoring has been studied by Wiginton (1980), Srinivasan and Kim (1987), Steenackers and Goovaerts (1989), Thomas, Crook and Edelman (1992) and Samreen and Zaidi (2012), among others.

It can be deduced from the definition of the PD that it is a relevant measure in other fields apart from the financial one. For example, companies that provide energy services (electricity, gas), water, streaming services (TV, cinema, music), telephone or internet are interested in estimating the probability that a customer who receives their services at time t will leave the company before 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 and Djaïdja (2007) and Cao, Vilar and Devia (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 approach that only uses 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 work was expanded in the paper of Peláez, Cao and Vilar (2021b) 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, Cao and Vilar (2021a), 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. (2021a).

The remainder of this paper is organized as follows. In Section 2, the nonparametric estimator of the probability of default with double smoothing is defined, the doubly smoothed PD estimator based on Beran's estimator is applied to a set of modified real data and its asymptotic properties are presented. In Section 3, a simulation study shows the improvement obtained by using the double smoothing in several nonparametric estimators of the probability default. Finally, Section 4 contains some concluding remarks. Appendix A includes terminology, assumptions and detailed theoretical results. Appendix B includes a sketch of proof of the theoretical results.

2. Nonparametric PD estimator with double smoothing

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of (X, Z, δ) where X is the covariate, $Z = \min\{T, C\}$ is the follow-up time 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 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:

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

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. (2021b) used nonparametric estimators of the conditional survival function, $\widehat{S}_h(t|x)$ with $h = h_n$ being the smoothing parameter for the covariate, to obtain the corresponding nonparametric estimator of $PD(t|x)$ denoted by $\widehat{PD}_h(t|x)$.

In Peláez et al. (2021a) the following nonparametric estimator of the conditional survival function with double smoothing is proposed and studied:

$$\widetilde{S}_{h,g}(t|x) = 1 - \sum_{i=1}^{n} s_{(i)} \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right), \tag{2}$$

where $s_{(i)} = \widehat{S}_h(Z_{(i-1)}|x) - \widehat{S}_h(Z_{(i)}|x)$ with $i = 2, ..., n$ and $s_{(1)} = 1 - \widehat{S}_h(Z_{(1)}|x)$, $Z_{(i)}$ is 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, defined in (2), 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 on x by means of a weighted mean of the values that the estimator $\widehat{S}_h(u|x)$ 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 the doubly smoothed nonparametric estimator, $\widetilde{S}_{h,g}(t|x)$, obtaining the following nonparametric estimator of $PD(t|x)$:

$$\widetilde{PD}_{h,g}(t|x) = 1 - \frac{\widetilde{S}_{h,g}(t+b|x)}{\widetilde{S}_{h,g}(t|x)}. \tag{3}$$

Since $\widehat{S}_h(t|x)$ is any arbitrary conditional survival estimator, the probability of default estimator $\widetilde{PD}_{h,g}(t|x)$ is very general. From now on, this paper focuses on the doubly smoothed Beran's estimator $\widetilde{S}^B_{h,g}(t|x)$ associated through (2) with the classic Beran's estimator of the survival function given by

$$\widehat{S}^B_h(t|x) = \prod_{i=1}^{n}\left(1-\frac{I_{\{Z_i\leq t,\,\delta_i=1\}}w_{i,n}(x)}{1-\sum_{j=1}^{n}I_{\{Z_j<Z_i\}}w_{j,n}(x)}\right), \tag{4}$$

where

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

with $i=1,...,n$ and $h=h_n$ is the smoothing parameter for the covariable.

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

However, any other estimator of the conditional survival function could be considered to obtain the corresponding smoothed estimator defined in (2) and then, to estimate the probability of default through the expression given in (3). In particular, two other survival estimators are considered in this work: 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 local linear regression. Since the weighted local linear estimator presents problems when estimating probabilities, a constant fit is proposed in Peláez et al. (2021b). The VKA estimator was defined in Van Keilegom and Akritas (1999) and Van Keilegom, Akritas and Veraverbeke (2001). The expressions for both estimators are shown in Section 2 of Peláez et al. (2021a) and they are denoted by $\widehat{S}^{WNW}_h(t|x)$ and $\widehat{S}^{VKA}_h(t|x)$. Their smoothed versions are built according to Equation (2), obtaining the following smoothed survival estimators: $\widetilde{S}^{WNW}_{h,g}(t|x)$ and $\widetilde{S}^{VKA}_{h,g}(t|x)$. Replacing $\widetilde{S}_{h,g}(t|x)$ with $\widetilde{S}^{WNW}_{h,g}(t|x)$ and $\widetilde{S}^{VKA}_{h,g}(t|x)$ in Equation (3) gives the nonparametric smoothed estimators of $PD(t|x)$ denoted by $\widetilde{PD}^{WNW}_{h,g}(t|x)$ and $\widetilde{PD}^{VKA}_{h,g}(t|x)$.

2.1. Application to real data

In order to illustrate the use of these smoothed estimators in the context of credit risk, a real data set is analysed using the doubly smoothed Beran's estimator. The data consists of a sample of 10000 consumer credits from a Spanish bank registered between July 2004 and November 2006. They are also considered by Peláez et al. (2021b), where the PD is estimated using parametric and non-parametric methods for $S(t|x)$ which are not smoothed in t, as is the case in this paper. The data set provides the credit scoring computed for each borrower, the observed lifetime of the credit in months and the uncensoring indicator. To obtain each customer's credit scoring, the financial institution adjusted a scoring model on several informative variables collected in the dataset: gender, marital status, profession, place of residence, type of housing, age, employment

history and bank account balance. See Devia (2016) for more details. Due to confidentiality, the estimated coefficients of the original explanatory variables are not reported here. The resulting credit scoring is used as a covariate in this analysis. The sample censoring percentage is 92.8%; equivalently, the proportion os credits for which the default is observed is 7.2%. An intentionally biased subsample was obtained from the original sample, so as so as not to show the true solvency situation of the bank and thus preserve confidentiality.

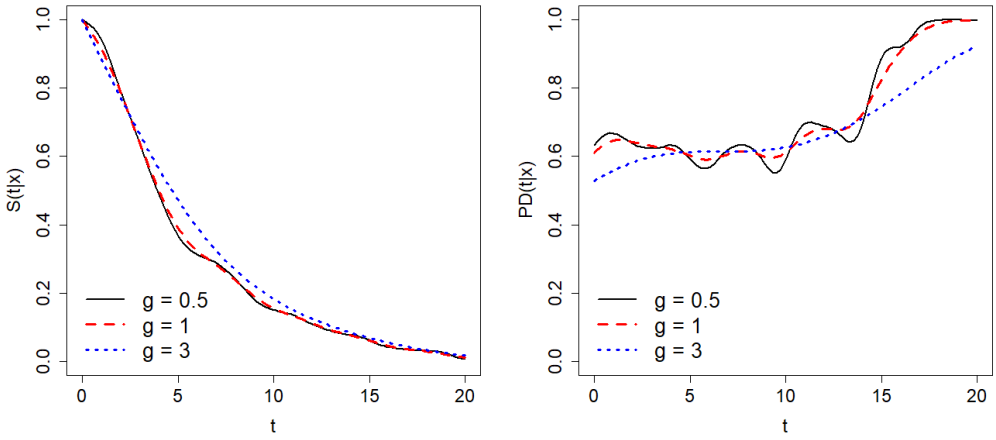

Figure 1. *Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.5$ by means of the smoothed Beran estimator on the consumer credits dataset for $h = 0.05$ and three different values of g.*

The probability of default for $x = 0.5$ at horizon $b = 5$ months 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 1 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 $g = 3$ are shown in Figure 2. 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.

Finally, sample quartiles of the credit scoring are considered for the group of clients with observed default (uncensored group) and the group with unobserved default (censored group). Figure 3 shows the PD estimation by means of the smoothed Beran's

estimator for these values of the credit scoring at horizon $b = 5$ months with $h = 0.05$ and $g = 3$.

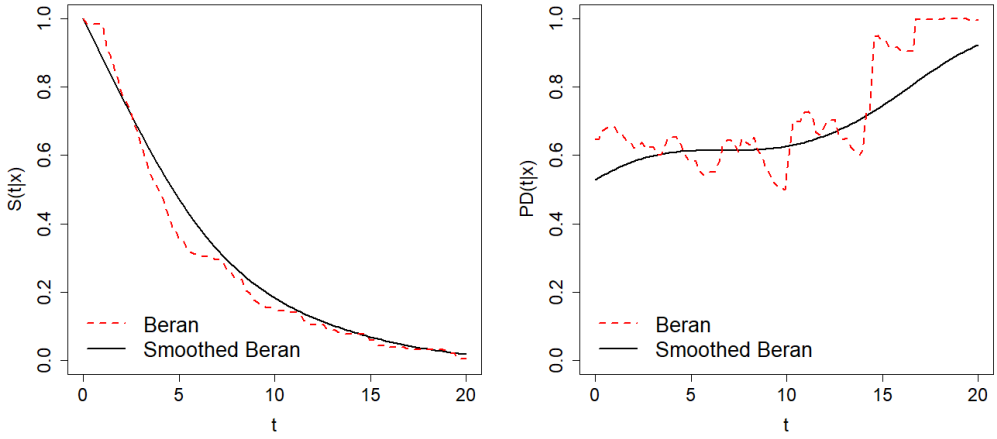

Figure 2. *Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.5$ by means of Beran's estimator (dashed line) and smoothed Beran's estimator (solid line) using the bandwidths $h = 0.05$ and $g = 3$ on the consumer credits dataset.*

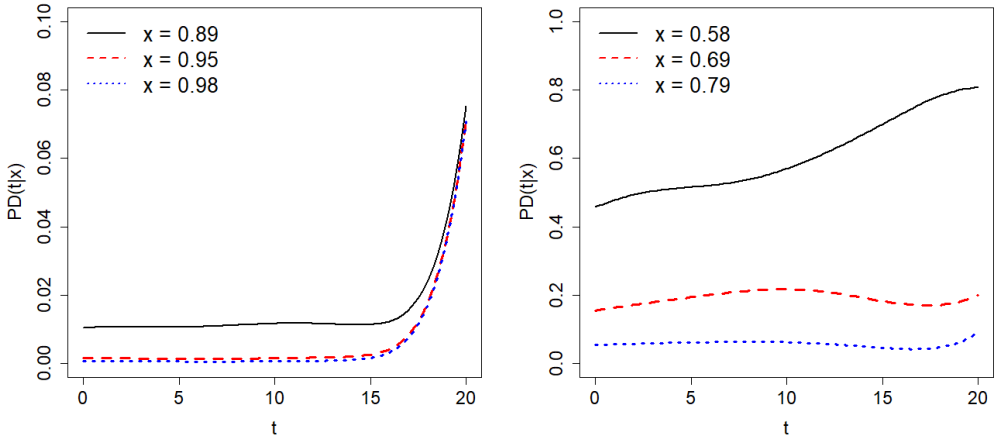

Figure 3. *Smoothed Beran's estimation of $PD(t|x)$ at horizon $b = 5$, for large (left) and small (right) values of the score x, using bandwidths $h = 0.05$ and $g = 3$. The large values chosen are the three sample quartiles of the score for nondefaulted credits, while the small values are the three sample quartiles of the score for the defaulted credits.*

2.2. Asymptotic results of the doubly smoothed Beran's PD estimator

Asymptotic properties of the smoothed Beran's estimator of the PD, $\widetilde{PD}^{B}_{h,g}(t|x)$, are obtained using the results for the smoothed Beran's survival estimator presented in Peláez

et al. (2021a). An intuitive idea of these results is shown here. The simplified expression of the asymptotic bias of $\widetilde{PD}^B_{h,g}(t|x)$ is as follows:

$$\mathrm{ABias}\big(\widetilde{PD}^B_{h,g}(t|x)\big) \quad = \quad c_1 h^2 + c_2 g^2,$$

and the asymptotic variance of $\widetilde{PD}^B_{h,g}(t|x)$ is given by

$$\mathrm{AVar}\big(\widetilde{PD}^B_{h,g}(t|x)\big) \quad = \quad c_3 \frac{1}{nh} + c_4 \frac{g}{nh} + c_5 \frac{h}{n},$$

for some real constants c_1, c_2, c_3, c_4 and c_5. For detailed expressions of these constants and the asymptotic normality of the estimator, see Appendix A. For proofs of these results see Appendix B.

It is difficult to use the theoretical bias and variance in an applied context in order to compare estimators or to obtain optimal smoothing parameters, since the constants c_1, c_2, c_3, c_4 and c_5 involved are complex and depend on too many population functions.

3. Simulation study

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

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

For Model 1, the time to occurrence of the event conditional on the covariate, $T|X=x$, follows a Weibull distribution with parameters $d=2$ and $A(x)^{-1/d}$ where $A(x)=1+5x$, and the censoring time conditional on the covariate, $C|X=x$, follows a Weibull distribution with parameters $d=2$ and $B(x)^{-1/d}$ where $B(x)=10+b_1x+20x^2$, 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 $0<t_1<\cdots<t_{n_t}$ of size n_t where $t_{n_t}+b=F^{-1}(0.95|x=0.6)$.

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 on the covariate, $T|X = x$ and an exponential distribution with parameter $\Delta(x) = 10 + d_1 x + 20x^2$, for some suitable values of d_1, for the censoring time conditional on the covariate, $C|X = x$. In this case, the conditional survival function, the probability of default and the censoring conditional probability are given by:

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

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

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

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. (2021a). This makes it possible to compare the results obtained in both studies.

First, the performance of Beran's PD estimator, $\widehat{PD}_h^B(t|x)$, and the smoothed Beran's PD estimator, $\widetilde{PD}_{h,g}^B(t|x)$, are compared.

The optimal bandwidth for $\widehat{PD}_h^B(t|x)$, h_1, is taken as the value which minimises a Monte Carlo approximation of the mean integrated squared error (MISE) given by

$$MISE_x(h) = E\left(\int \left(\widehat{PD}_h^B(t|x) - PD(t|x)\right)^2 dt\right)$$

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 $\widehat{PD}_h^B(t|x)$.

The smoothed PD estimator $\widetilde{PD}_{h,g}^B(t|x)$ 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 the optimal h_1 for Beran's estimator and approximating the optimal smoothing parameter g. The error to

minimise is

$$MISE_x(h_1,g) = E\left(\int \left(\widetilde{PD}^B_{h_1,g}(t|x) - PD(t|x)\right)^2 dt\right)$$

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 $\widetilde{PD}^B_{h,g}(t|x)$.

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 minimises some Monte Carlo approximation of

$$MISE_x(h,g) = E\left(\int \left(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\right)^2 dt\right)$$

based on $N = 100$ simulated samples. Then, the value of the *MISE* committed by $\widetilde{PD}^B_{h_2,g_2}(t|x)$ is approximated from $N = 1000$ simulated samples.

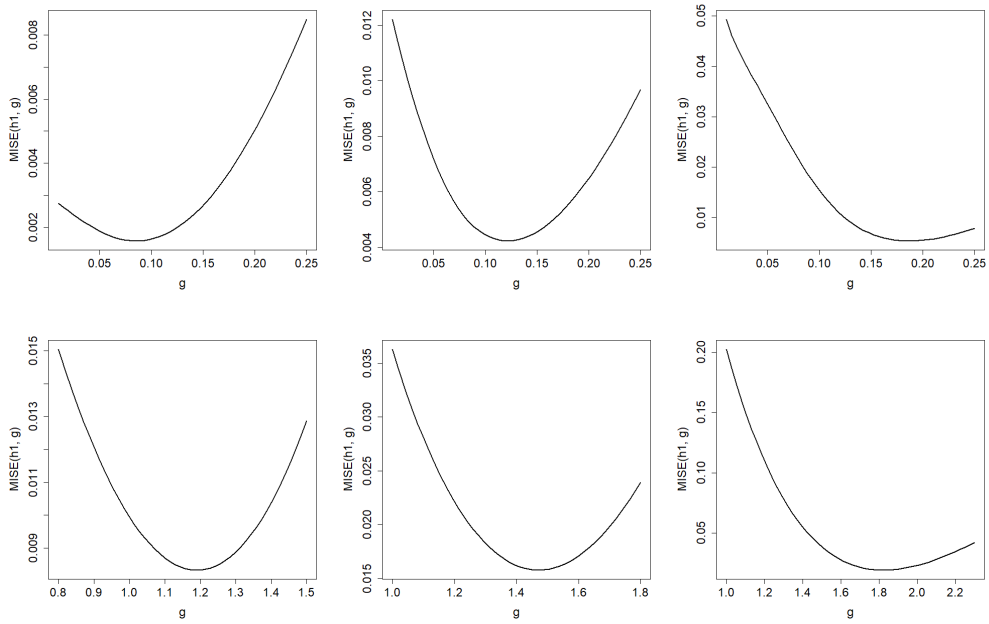

Figure 4. *$MISE_x(h_1,g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) with $P(\delta = 0|x) = 0.2$ (left), $P(\delta = 0|x) = 0.5$ (center) and $P(\delta = 0|x) = 0.8$ (right).*

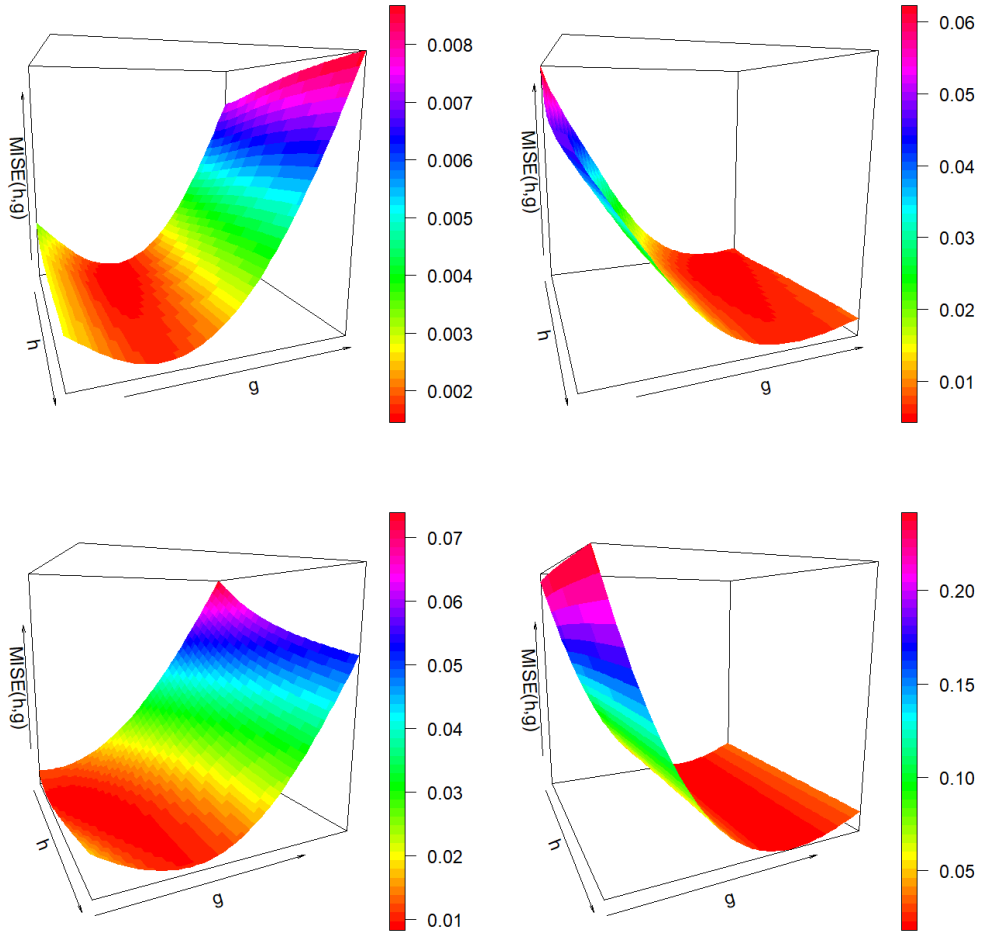

Figure 5. *$MISE_x(h,g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) with $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.8$ (right).*

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 they produce a fair comparison since the estimators are built using the best possible smoothing parameters.

The error curve $MISE_x(h_1,g)$, which is minimised to obtain the optimal time smoothing parameter according to Strategy 1, is shown in Figure 4 for each level of censoring conditional probability and each model. It follows from these graphs that the optimal bandwidth g is easily approximated by Strategy 1.

The function $MISE_x(h,g)$ for Models 1 and 2 for the lowest and highest censoring levels is shown in Figure 5. These plots show the two-dimensional functions to be min-

imised 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 the estimation error, whereas h seems not to affect much the quality of the estimator.

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 6 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 curves are similar and close for all values of h, mainly at the highest level of censoring conditional probability.

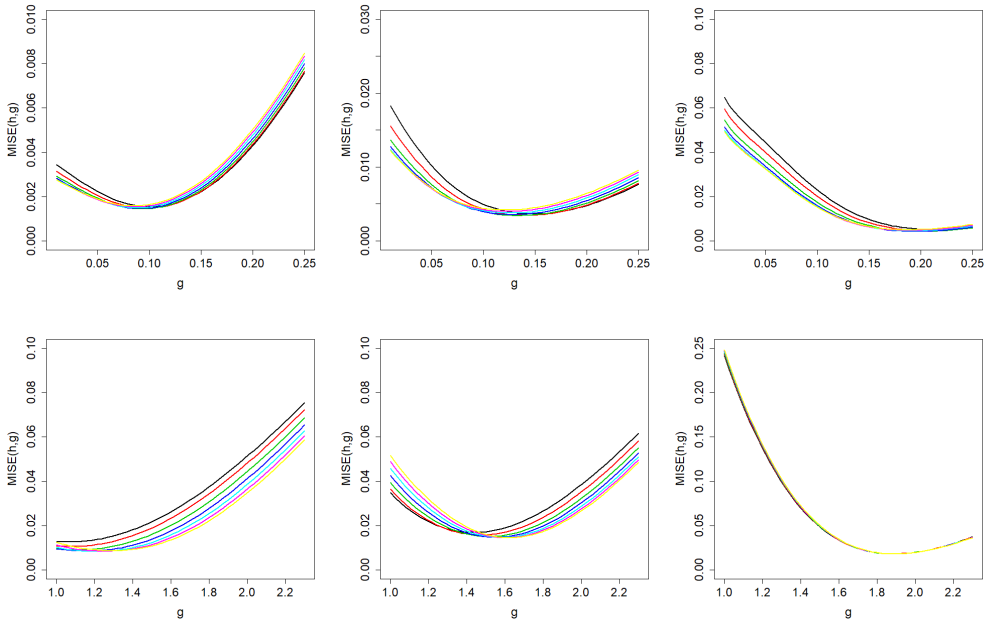

Figure 6. *$MISE_x(h,g)$ as a function of g, approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) for some fixed equispaced values of $h \in [0.1, 0.5]$ with $P(\delta = 0|x) = 0.2$ (left), $P(\delta = 0|x) = 0.5$ (center) 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:

$$R_i(x) = \frac{RMISE\left(\widetilde{PD}^B_{h_i,g_i}(\cdot|x)\right)}{RMISE\left(\widehat{PD}^B_{h_1}(\cdot|x)\right)},$$

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 double 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.

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.*

		Model 1			**Model 2**		
$\mathbf{P(\delta = 0 \mid x)}$		**0.2**	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**
$\widehat{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h_1}}$	$\mathbf{h_1}$	0.35714	0.34694	0.39796	0.10306	0.12265	0.14224
	RMISE	0.05437	0.11195	0.25738	0.27128	0.49813	0.67999
$\widetilde{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h_1,g_1}}$	$\mathbf{h_1}$	0.35714	0.34694	0.39796	0.10306	0.12265	0.14224
	$\mathbf{g_1}$	0.08347	0.12265	0.18633	1.18571	1.47755	1.82245
	RMISE	0.04065	0.06574	0.07246	0.25222	0.24154	0.20558
	$\mathbf{R_1}$	0.74765	0.58723	0.28153	0.92974	0.48489	0.30233
$\widetilde{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h_2,g_2}}$	$\mathbf{h_2}$	0.21429	0.15714	0.18980	0.10816	0.25918	1.00000
	$\mathbf{g_2}$	0.09327	0.13735	0.19612	1.21122	1.61020	1.90204
	RMISE	0.03845	0.05941	0.06208	0.09210	0.12350	0.13434
	$\mathbf{R_2}$	0.70719	0.53068	0.24120	0.33950	0.24793	0.19756

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 $\widetilde{PD}^{B}_{h,g}(t|x)$ 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 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 moderate or heavy, respectively. This reduction is bigger when using Strategy 2, reaching $75-80\%$.

A brief study not included here shows that the results of these simulations hold when the distribution of X is not uniform but a more realistic asymmetric distribution if X denotes the credit scoring.

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 spend 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 ones for estimating estimating the curve.

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.*

n	50	100	200	400	1200
Beran	0.01	0.01	0.01	0.02	0.03
SBeran	0.03	0.03	0.03	0.05	0.20

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

N	50	100	150	200
Strategy 1	3.01	4.28	5.40	7.32
Strategy 2	80.61	204.51	228.01	296.95

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 other estimators for the survival function are equally improved by applying this type of smoothing. With this aim, in a second simulation study, the behaviours of the smoothed Weighted Nadaraya-Watson estimator (SWNW), $\widetilde{PD}^{WNW}_{h,g}(t|x)$, and the smoothed Van Keilegom-Akritas estimator (SVKA), $\widetilde{PD}^{VKA}_{h,g}(t|x)$, are compared to each other as well as to the smoothed Beran's estimator.

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 $R^{\bullet}_{S}$ and $R^{\bullet}_{c}$ are defined as follows:

$$R^{\bullet}_{S}(x) = \frac{RMISE\left(\widetilde{PD}^{\bullet}_{h_1,g_1}(\cdot|x)\right)}{RMISE\left(\widehat{PD}^{\bullet}_{h_1}(\cdot|x)\right)},$$

$$R^{\bullet}_{c}(x) = \frac{RMISE\left(\widetilde{PD}^{\bullet}_{h_1,g_1}(\cdot|x)\right)}{RMISE\left(\widetilde{PD}^{B}_{h_2,g_2}(\cdot|x)\right)},$$

being $\bullet = B, WNW, VKA$ 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, $R_S^\bullet$ and $R_c^\bullet$ 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.*

	$\mathbf{P(\delta=0\|x)=0.2}$			$\mathbf{P(\delta=0\|x)=0.5}$			$\mathbf{P(\delta=0\|x)=0.8}$		
	SBeran	**SWNW**	**SVKA**	**SBeran**	**SWNW**	**SVKA**	**SBeran**	**SWNW**	**SVKA**
Model 1									
$\mathbf{h_1}$	0.35714	0.38776	0.25918	0.34694	0.90102	0.22857	0.39796	1.00000	0.23469
$\mathbf{g_1}$	0.08347	0.14020	0.06327	0.12265	0.20531	0.11653	0.18633	0.28367	0.19347
RMISE	0.04065	0.03513	0.06418	0.06574	0.03260	0.09957	0.07246	0.04705	0.09816
$\mathbf{R_S^\bullet}$	0.74765	0.50036	0.88744	0.58723	0.19457	0.76112	0.28153	0.14115	0.38976
$\mathbf{R_c^\bullet}$	1.05722	0.91365	1.66918	1.10655	0.54873	1.67598	1.16720	0.75789	1.58119
Model 2									
$\mathbf{h_1}$	0.10306	0.09143	0.04567	0.12265	0.10694	0.05380	0.14224	0.11857	0.12837
$\mathbf{g_1}$	1.18571	1.55102	1.44286	1.47755	1.77551	1.45714	1.82245	1.92857	1.52857
RMISE	0.25222	0.12628	0.49730	0.24154	0.13406	0.37621	0.20558	0.13375	0.11410
$\mathbf{R_S^\bullet}$	0.92974	0.33177	1.63226	0.48489	0.19828	0.88273	0.30233	0.16480	0.16868
$\mathbf{R_c^\bullet}$	2.73855	1.37112	5.39957	1.95580	1.08551	3.04623	1.53030	0.99561	0.84934

The values of $R_S^\bullet$ 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 $R_S^\bullet$, 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 $R_c^\bullet$ is useful to compare the behaviour of the three estimators with the behaviour of $\widetilde{PD}^B_{h_2,g_2}(t|x)$ (the smoothed Beran's estimator with Strategy 2). Since almost all the $R_c^\bullet$ 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 $R_c^\bullet$, 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 $R_c^\bullet$ less than 1, which indicates that the error it makes is occasionally smaller than the errror 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 computacional times os 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 5. *CPU time (in seconds) for estimating $PD(t|x)$ in a time grid of size 100 for every estimator and different sample sizes.*

n	Beran	SBeran	SWNW	SVKA
50	0.01	0.03	2.30	0.42
100	0.01	0.03	6.33	1.80
200	0.01	0.03	25.97	7.34
400	0.02	0.05	140.62	53.99
1200	0.03	0.20	1459.35	507.36

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 the slowest method and most affected by the sample size is the smoothed WNW estimator.

4. 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 scenarios analysed in 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. Nevertheless, the smoothed Beran's estimator of the PD turned out to have the most stable behaviour and to be the fastest of all. The selection of the smoothing parameters for the smoothed Beran's estimator is still an outstanding problem. The study of automatic methods probably based on the bootstrap is an appealing idea to be considered for future work.

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.

References

Allen, L. N. and Rose, L. C. (2006). Financial survival analysis of defaulted debtors. *Journal of the Operational Research Society 57*(6), 630–636.

Baba, N. and Goko, H. (2006). Survival analysis of hedge funds. *Bank of Japan, Working Papers Series* (06-E-05).

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. *Technical report, University of California*.

Billingsley, P. (1968). *Convergence of Probability Measure. Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics*, Volume 9. Wiley.

Cai, Z. (2003). Weighted local linear approach to censored nonparametric regression. In M. G. Akritas and D. N. Politis (Eds.), *Recent Advances and Trends in Nonparametric Statistics*, pp. 217–231.

Cao, R., Vilar, J. M. and Devia, A. (2009). Modelling consumer credit risk via survival analysis (with discussion). *Statistics and Operations Research Transactions 33*(1), 3–30.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics 17*(3), 1157–1167.

Devia, A. (2016). *Contribuciones al análisis estadístico del riesgo de crédito*. PhD Thesis, Universidade da Coruña.

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.

Hanson, S. G. and Schuermann, T. (2004). Estimating probabilities of default. *Staff Report Federal Reserve Bank of New York 190*, 923–947.

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*(3), 213–244.

Naraim, B. (1992). Survival analysis and the credit granting decision. In L. C. Thomas, J. N. Crook, and D. B. Edelman (Eds.), *Credit Scoring and Credit Control, Oxford University Press*, pp. 109–121.

Peláez, R., Cao, R. and Vilar, J. M. (2021a). Nonparametric estimation of the conditional survival function with double smoothing. Unpublished paper, URL: http://dm.udc.es/preprint/Pelaez_Cao_Vilar_Survival_estimator_double_smoothing_Nov2020.pdf.

Peláez, R., Cao, R. and Vilar, J. M. (2021b). Probability of default estimation in credit risk using a nonparametric approach. *TEST 30*, 383–405.

Samreen, A. and Zaidi, F. (2012). Design and development of credit scoring model for the commercial banks of pakistan: forecasting creditworthiness of individual borrowers. *International Journal of Business and Social Science 17*, 155–166.

Silverman, B. W. (1986). *Density Estimation for Statistics & Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Srinivasan, V. and Kim, Y. H. (1987). Credit granting: a comparative analysis of clasification procedures. *Journal of Finance 42*, 665–681.

Steenackers, A. and Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics 8*, 31–34.

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

Thomas, L. C., Crook, J. N. and Edelman, D. B. (1992). *Credit Scoring and Credit Control*. Oxford University Press.

Van Keilegom, I. and Akritas, M. (1999). Transfer of tail information in censored regression models. *The Annals of Statistics 27*(5), 1745–1784.

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 53*, 457–481.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behaviour. *Journal of Financial and Quantitative Analysis 15*, 757–770.

A. Asymptotic results of the doubly smoothed Beran's estimator of the PD

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

Let $R:\mathbb{R}\longrightarrow\mathbb{R}$ be any function and define the constants

$$c_R=\int R(t)^2dt,\quad d_R=\int t^2R(t)dt,$$

and the functions

$$R_l(u)=u^lR(u),\quad \mathbb{R}_l(u)=\int_{-\infty}^{u}R_l(t)dt. \tag{5}$$

Given any function $f:\mathbb{R}^k\longrightarrow\mathbb{R}$, its first derivatives with respect to the first and second variables are denoted as follows:

$$f'(x_1,...,x_k)=\frac{\partial f(x_1,...,x_k)}{\partial x_1},\quad \dot{f}(x_1,...,x_k)=\frac{\partial f(x_1,...,x_k)}{\partial x_2}$$

Correspondingly, the second derivatives with respect to the first or second variable are denoted by $f''(x_1,...,x_k)$ and $\ddot{f}(x_1,...,x_k)$, 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. (2021a), 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 $H(t)=P(Z\leq t)$ be the distribution function of Z and $H(t|x)$ 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 $l_{H(\cdot|x)}=\inf\{t/H(t|x)>0\}$ and $u_{H(\cdot|x)}=\inf\{t/H(t|x)=1\}$, for any $x\in I_c$, $0\leq l_{H(\cdot|x)}$, $0\leq u_{H(\cdot|x)}$

(d) There exist $l,u,\theta\in\mathbb{R}$ with $l<u$, satisfying $\inf\{1-H(u|x):x\in I_c\}\geq\theta>0$. Therefore $1-H(t|x)\geq\theta>0$ for every $(t,x)\in[l,u]\times I_c$.

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

A.5. Let $H_1(t) = P(Z \leq t, \delta = 1)$ be the subdistribution function of Z when $\delta = 1$. The corresponding density functions of $H(t)$ and $H_1(t)$ are bounded away from 0 in $[l,u]$.

A.6. Let $H_1(t|x)$ the conditional subdistribution function of $Z|X = x$ when $\delta = 1$. 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 in $[l,u] \times I_c$.

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. $\dot{H}'(t|x)$ and $\dot{H}_1'(t|x)$ respectively, exist and are continuous in $[l,u] \times I_c$.

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. (2021a), the asymptotic properties of the estimator $\widetilde{PD}^B_{h,g}(t|x)$ are obtained. The following are the functions required to state these results:

$$
\begin{aligned}
&\xi(Z,\delta,t,x) = \frac{1_{\{Z\leq t,\delta=1\}}}{1-H(Z|x)} - \int_0^t \frac{1_{\{u\leq Z\}}}{\big(1-H(u|x)\big)^2} dH_1(u|x),\\
&\eta(Z,\delta,t,x) = \int K(u)\big(1-F(t-gu|x)\big)\xi(Z,\delta,t-gu,x)du,\\
&\Phi_\xi(u,t,x) = E\big[\xi(Z_1,\delta_1,t,x)|X_1 = u\big],\\
&J(t|x) = \big(1-F(t|x)\big)L(t|x),\\
&L(t|x) = \int_o^t \frac{dH_1(z|x)}{\big(1-H(z|x)\big)^2},\\
&D_g(u,t_1,t_2,x) = \mathrm{Cov}\big[\eta(Z_1,\delta_1,t_1,x),\eta(Z_1,\delta_1,t_2,x)|X_1 = u\big]m(u),\\
&N(u,t_1,t_2,x) = E\big[\xi(Z_1,\delta_1,t_1,x)\xi(Z_1,\delta_1,t_2,x)\ X_1 = u\big],\\
&D(t,x) = \big(1-F(t|x)\big)^2\Big(m''(x)N(x,t,t,x) + m(x)N''(x,t,t,x) + 2m'(x)N'(x,t,t,x)\\
&\qquad\qquad -2c_K m(x)\Phi'_\xi(x,t,x)\Phi'_\xi(x,t,x)\Big),\\
&B_1(t,x) = \frac{d_K\big(1-F(t|x)\big)}{2m(x)}\big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)\big),\\
&B_2(t,x) = -\frac{1}{2}d_K F''(t|x),\\
&C_1(t_1,t_2,x) = \frac{2c_K}{m(x)}J(t_1|x)\big(1-F(t_2|x)\big)\mathbb{K} * K\left(\frac{t_2-t_1}{g}\right),
\end{aligned}
$$

$$
\begin{aligned}
C_2(t_1,t_2,x) &= \frac{c_K}{m(x)}\left(2J(t_1|x)f(t_2|x)\mathbb{K}*K_1\left(\frac{t_1-t_2}{g}\right)\right.\\
&\qquad \left.+2J'(t_1|x)\big(1-F(t_2|x)\big)\mathbb{K}*K_1\left(\frac{t_2-t_1}{g}\right)\right),\\
C_3(t_1,t_2,x) &= \frac{d_{K^2}}{m^2(x)}\Big(m(x)\big(1-F(t_1|x)\big)\big(1-F(t_2|x)\big)\Phi'_\xi(x,t_1,x)\Phi'_\xi(x,t_2,x)\\
&\qquad +\frac{1}{2}D''_g(x,t_1,t_2,x)\Big),\\
V_1(t,x) &= \frac{c_K}{m(x)}\big(1-F(t|x)\big)^2L(t|x),\\
V_2(t,x) &= \frac{c_K(c_{\mathbb{K}}-1)}{m(x)}\big(1-F(t|x)\big)^2L'(t|x),\\
V_3(t,x) &= \frac{d_{K^2}}{m^2(x)}\Big(m(x)\big(1-F(t|x)\big)^2\big(\Phi'_\xi(x,t,x)\big)^2+\frac{1}{2}D(t,x)\Big).
\end{aligned}
$$

Another assumption related to the differentiability of the above functions is required:

A.10 Let $(t,x)\in[l,u]\times I_c$. The first derivative of $L(u|x)$ with respect to u exists at (t,x). The second derivative of $m(u)$ exists et $u=x$. The second derivative of $S(u|x)$ exists at (t,x) and $(t+b,x)$. The second derivative of $\Phi_\xi(u,t,x)$ exists at (x,t,x). The second derivative of $J(u|x)$ exists at (t,x). The second derivative of $D_g(u,t_1,t_2,x)$ exists at $(x,t,t+b,x)$. The second derivatie of $N(u,t_1,t_2,x)$ exists at (x,t,t,x).

Theorem A.1. *Let $(t,x)\in[l,u]\times I_c$ be such that $S(t|x)>0$. Under assumptions A.1-A.10, expressions for the asymptotic bias of $\widetilde{PD}^B_{h,g}(t|x)$, $ABias\big(\widetilde{PD}^B_{h,g}(t|x)\big)$, and the asymptotic variance of $\widetilde{PD}^B_{h,g}(t|x)$, $AVar\big(\widetilde{PD}^B_{h,g}(t|x)\big)$, are the following:*

$$
\begin{aligned}
ABias\big(\widetilde{PD}^B_{h,g}(t|x)\big) &= \frac{\big(1-PD(t|x)\big)B_1(t,x)-B_1(t+b,x)}{S(t|x)}h^2\\
&\quad+\frac{\big(1-PD(t|x)\big)B_2(t,x)-B_2(t+b,x)}{S(t|x)}g^2,
\end{aligned}
$$

$$
\begin{aligned}
AVar\big(\widetilde{PD}^B_{h,g}(t|x)\big) &= \left(\frac{V_1(t+b,x)}{S(t|x)^2}-2\frac{S(t+b|x)C_1(t,t+b,x)}{S(t|x)^3}+\frac{S(t+b|x)^2V_1(t,x)}{S(t|x)^4}\right)\frac{1}{nh}\\
&\quad+\left(\frac{V_2(t+b,x)}{S(t|x)^2}-2\frac{S(t+b|x)C_2(t,t+b,x)}{S(t|x)^3}+\frac{S(t+b|x)^2V_2(t,x)}{S(t|x)^4}\right)\frac{g}{nh}\\
&\quad+\left(\frac{V_3(t+b,x)}{S(t|x)^2}-2\frac{S(t+b|x)C_3(t,t+b,x)}{S(t|x)^3}+\frac{S(t+b|x)^2V_3(t,x)}{S(t|x)^4}\right)\frac{h}{n}.
\end{aligned}
$$

Theorem A.2. *Under the assumptions of Theorem A.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* $\widetilde{PD}^B_{h,g}(t|x)$ *is given by*

$$\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\mu, s_0),$$

where

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

and

$$\begin{aligned}s_0^2 &= \frac{V_1(t+b,x)}{S(t|x)^2} - 4\frac{S(t+b|x)}{S(t|x)^3}\frac{c_K\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x)}{m(x)} \\ &\quad + \frac{S(t+b|x)^2 V_1(t,x)}{S(t|x)^4}.\end{aligned}$$

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* $\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\widetilde{\mu}, s_0)$*. with*

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

Assuming $n^{1/5}h \to 0$*,* $n^{1/5}g \to 0$ *and* $\dfrac{nh}{(\ln n)^3} \to \infty$*, the asymptotic distribution of the smoothed Beran's PD estimator is* $\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(0, s_0)$.

B. Proofs

Proofs of the results shown in Appendix A are done using results from papers Peláez et al. (2021b) and Peláez et al. (2021a).

Proof of Theorem A.1.

Denote $P = S(t+b|x)$, $Q = S(t|x)$ and $PD(t|x) = 1 - \dfrac{P}{Q}$. Similarly, $\widetilde{P} = \widetilde{S}^B_{h,g}(t+b|x)$, $\widetilde{Q} = \widetilde{S}^B_{h,g}(t|x)$ and $\widetilde{PD}^B_{h,g}(t|x) = 1 - \dfrac{\widetilde{P}}{\widetilde{Q}}$. As a consequence of the proof of Theorem 1 in Peláez et al. (2021b):

$$ABias\big(\widetilde{PD}^B_{h,g}(t|x)\big) = \alpha_1 + \alpha_2 + \alpha_3, \tag{6}$$

$$AVar\big(\widetilde{PD}^{B}_{h,g}(t|x)\big) = \beta_1 + \beta_2 + \beta_3, \tag{7}$$

where

$$\alpha_1 = \frac{P}{Q} - \frac{E(\widetilde{P})}{E(\widetilde{Q})}, \quad \alpha_2 = \frac{Cov(\widetilde{P},\widetilde{Q})}{E(\widetilde{Q})^2}, \quad \alpha_3 = -\frac{E\left[\frac{\widetilde{P}}{\widetilde{Q}}\big(\widetilde{Q}-E(\widetilde{Q})\big)^2\right]}{E(\widetilde{Q})^2} \tag{8}$$

and

$$\beta_1 = \frac{Var(\widetilde{P})}{E(\widetilde{Q})^2}, \quad \beta_2 = -2\frac{E(\widetilde{P})Cov(\widetilde{P},\widetilde{Q})}{E(\widetilde{Q})^3}, \quad \beta_3 = \frac{E(\widetilde{P})^2 Var(\widetilde{Q})}{E(\widetilde{Q})^4}. \tag{9}$$

The asymptotic expressions for the bias, the covariance and the variance of the survival estimator $\widetilde{S}^{B}_{h,g}(t|x)$ are obtained from Theorems 3 and 4 of Peláez et al. (2021a):

$$\text{Bias}\big(\widetilde{S}^{B}_{h,g}(t|x)\big) = B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2), \tag{10}$$

$$\begin{aligned}\text{Cov}\big(\widetilde{S}^{B}_{h,g}(t_1|x),\widetilde{S}^{B}_{h,g}(t_2|x)\big) &= C_1(t_1,t_2,x)\frac{1}{nh} + C_2(t_1,t_2,x)\frac{g}{nh} \\ &\quad + C_3(t_1,t_2,x)\frac{h}{n} + R_n(t,x),\end{aligned} \tag{11}$$

$$\text{Var}\big(\widetilde{S}^{B}_{h,g}(t|x)\big) = V_1(t,x)\frac{1}{nh} + V_2(t,x)\frac{g}{nh} + V_3(t,x)\frac{h}{n} + R_n(t,x), \tag{12}$$

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

Considering Equations (8)-(12), detailed expressions for α_1, α_2 and α_3 are obtained as follows:

$$\begin{aligned}\alpha_1 &= \frac{P}{Q} - \frac{P + B_1(t+b,x)h^2 + B_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)} \\ &= \frac{PQ + PB_1(t,x)h^2 + PB_2(t,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} + \\ &\quad \frac{-PQ - QB_1(t+b,x)h^2 - QB_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} \\ &= \frac{PB_1(t,x)h^2 - QB_1(t+b,x)h^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} + \\ &\quad \frac{PB_2(t,x)g^2 - QB_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} \\ &= \frac{\big(1 - PD(t|x)\big)B_1(t,x) - B_1(t+b,x)}{S(t|x)}h^2 \\ &\quad + \frac{\big(1 - PD(t|x)\big)B_2(t,x) - B_2(t+b,x)}{S(t|x)}g^2 + o(h^2) + o(g^2),\end{aligned} \tag{13}$$

$$\alpha_2 = \frac{C_1(t,t+b,x)}{S(t|x)^2}\frac{1}{nh} + \frac{C_2(t,t+b,x)}{S(t|x)^2}\frac{g}{nh} + \frac{C_3(t,t+b,x)}{S(t|x)^2}\frac{h}{n} + R_n(t,x), \tag{14}$$

$$\begin{aligned} \alpha_3 &= \frac{E\left[\frac{\widetilde{P}}{\widetilde{Q}}\left(\widetilde{Q}-E(\widetilde{Q})\right)^2\right]}{E(\widetilde{Q})^2} \leq \frac{\mathrm{Var}(\widetilde{Q})}{E(\widetilde{Q})^2} \\ &= \frac{V_1(t,x)}{S(t|x)^2}\frac{1}{nh} + \frac{V_2(t,x)}{S(t|x)^2}\frac{g}{nh} + \frac{V_3(t,x)}{S(t|x)^2}\frac{h}{n} + R_n(t,x). \end{aligned} \tag{15}$$

Plugging (13), (14) and (15) into (6) and using Assumption A.9,

$$\begin{aligned} \mathrm{ABias}\left(\widetilde{PD}^B_{h,g}(t|x)\right) &= \frac{\left(1-PD(t|x)\right)B_1(t,x) - B_1(t+b,x)}{S(t|x)}h^2 \\ &\quad + \frac{\left(1-PD(t|x)\right)B_2(t,x) - B_2(t+b,x)}{S(t|x)}g^2, \end{aligned}$$

and the bias part in Theorem A.1 is proved.

Now, expressions (9), (10), (11) and (12) lead to

$$\beta_1 = \frac{V_1(t+b,x)}{S(t|x)^2}\frac{1}{nh} + \frac{V_2(t+b,x)}{S(t|x)^2}\frac{g}{nh} + \frac{V_3(t+b,x)}{S(t|x)^2}\frac{h}{n} + R_n(t,x), \tag{16}$$

$$\begin{aligned} \beta_2 &= -2\frac{S(t+b,x)C_1(t,t+b,x)}{S(t|x)^3}\frac{1}{nh} - 2\frac{S(t+b,x)C_2(t,t+b,x)}{S(t|x)^3}\frac{g}{nh} \\ &\quad -2\frac{S(t+b,x)C_3(t,t+b,x)}{S(t|x)^3}\frac{h}{n} + R_n(t,x), \end{aligned} \tag{17}$$

$$\begin{aligned} \beta_3 &= \frac{S(t+b,x)^2V_1(t,x)}{S(t|x)^4}\frac{1}{nh} + \frac{S(t+b,x)^2V_2(t,x)}{S(t|x)^4}\frac{g}{nh} \\ &\quad + \frac{S(t+b,x)^2V_3(t,x)}{S(t|x)^4}\frac{h}{n} + R_n(t,x), \end{aligned} \tag{18}$$

and plugging Equations (16), (17) and (18) in (7) the variance part in Theorem A.1 is proved. ■

Proof of Theorem A.2

From Equations (1) and (3) follows:

$$\frac{\widetilde{S}^B_{h,g}(t+b|x)}{\widetilde{S}^B_{h,g}(t|x)} - \frac{S(t+b|x)}{S(t|x)} = -\left(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\right). \tag{19}$$

On the other hand, denoting $a_1 = \dfrac{1}{S(t|x)}$, $a_2 = -\dfrac{S(t+b|x)}{S(t|x)^2}$ and

$$C\left(\widetilde{S}^B_{h,g}(t|x)\right) = \frac{S(t|x)\left(\widetilde{S}^B_{h,g}(t+b|x) - S(t+b|x)\right) - S(t+b|x)\left(\widetilde{S}^B_{h,g}(t|x) - S(t|x)\right)}{\widetilde{S}^B_{h,g}(t|x)S(t|x)},$$

it holds

$$\begin{aligned}\frac{\widetilde{S}^{B}_{h,g}(t+b|x)}{\widetilde{S}^{B}_{h,g}(t|x)}-\frac{S(t+b|x)}{S(t|x)} &= a_1\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big)\\ &\quad +C\big(\widetilde{S}^{B}_{h,g}(t|x)\big)\left(1-\frac{\widetilde{S}^{B}_{h,g}(t|x)}{S(t|x)}\right),\end{aligned}$$

and considering (19):

$$\begin{aligned}PD(t|x) &- \widetilde{PD}^{B}_{h,g}(t|x)\\ &= a_1\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big)\\ &\quad +C\big(\widetilde{S}^{B}_{h,g}(t|x)\big)\left(1-\frac{\widetilde{S}^{B}_{h,g}(t|x)}{S(t|x)}\right).\end{aligned} \tag{20}$$

Since $\widetilde{S}^{B}_{h,g}(t|x)$ is a consistent estimator of $S(t|x)$, $\widetilde{S}^{B}_{h,g}(t|x)\xrightarrow{p}S(t|x)$. Thus,

$$1-\frac{\widetilde{S}^{B}_{h,g}(t|x)}{S(t|x)}\xrightarrow{p}0.$$

Therefore, the asymptotic distribution of $\sqrt{nh}\big(\widetilde{PD}^{B}_{h,g}(t|x)-PD(t|x)\big)$ is the same as the asymptotic distribution of the linear combination

$$a_1\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big).$$

From Lemma 1 in Peláez et al. (2021a), $\widetilde{S}^{B}_{h,g}(t|x)$ is split up into the following terms

$$\widetilde{S}^{B}_{h,g}(t|x) = S(t|x)+\sum_{i=1}^{n}\varphi_{n,i}(t,x)+B_2(t,x)g^2+R_n(t|x), \tag{21}$$

where $\varphi_{n,i}(t,x)=\dfrac{1}{nh}\dfrac{K\big((x-X_i)/h\big)}{m(x)}\eta(Z_i,\delta_i,t,x)$ 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),

$$\begin{aligned}&a_1\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big)\\ &\quad=\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x)+a_1B_2(t+b,x)g^2\sqrt{nh}+a_2B_2(t,x)g^2\sqrt{nh}+\widetilde{R}_n(t,x),\end{aligned} \tag{22}$$

where

$$\widetilde{\varphi}_{n,i}(t,x) = \sqrt{nh}\big(a_1\varphi_{n,i}(t+b,x)+a_2\varphi_{n,i}(t,x)\big) \tag{23}$$

and

$$\begin{aligned}\widetilde{R}_n(t,x) &= \sqrt{nh}\big(a_1R_n(t+b,x)+a_2R_n(t,x)\big)\\ &= \sqrt{nh}(a_1+a_2)O_p\left(\frac{\ln n}{nh}\right)^{3/4}+\sqrt{nh}(a_1+a_2)o(g^2)\\ &\quad +O_p\left(h^2+\frac{1}{\sqrt{nh}}\right)\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x).\end{aligned} \tag{24}$$

Since $h\to 0$ and $nh\to\infty$, the term $O_p\left(h^2+\frac{1}{\sqrt{nh}}\right)\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)$ in (24) is negligible with respect to $\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)$ in (22). Given that $g\to 0$, the term $\sqrt{nh}(a_1+a_2)o(g^2)$ in (24) is negligible with respect to $a_1B_2(t+b,x)g^2\sqrt{nh}+a_2B_2(t,x)g^2\sqrt{nh}$ 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 $\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)$ in (22) because $\frac{nh}{(\ln n)^3}=\frac{C_hn^{4/5}}{(\ln n)^3}\to\infty$. The variance of the dominant term in (22) is $O(1)$:

$$\begin{aligned}\mathrm{Var}\big(\textstyle\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)\big) &= n\mathrm{Var}\big(\widetilde{\varphi}_{n,1}(t,x)\big)\\ &= n^2h\Big(a_1^2\mathrm{Var}\big(\varphi_{n,1}(t+b,x)\big)+a_2^2\mathrm{Var}\big(\varphi_{n,1}(t,x)\big)\\ &\quad +2a_1a_2\mathrm{Cov}\big(\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\big)\Big).\end{aligned} \tag{25}$$

From the proof of Theorem 3 in Peláez et al. (2021a),

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

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

$$\begin{aligned}\lim_{n\to\infty}\mathbb{K}*K\left(\frac{b}{g}\right) &= \lim_{u\to\infty}\int_{-\infty}^{+\infty}K(y)\mathbb{K}(u-y)dy\\ &= \int_{-\infty}^{+\infty}\lim_{u\to\infty}\mathbb{K}(u-y)K(y)dy=\int_{-\infty}^{+\infty}K(y)dy=1.\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathrm{Cov}\big(\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\big)\\ &= \frac{2c_K}{m(x)n^2}\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x)\frac{1}{h}+O\left(\frac{g}{n^3h}\right)+o\left(\frac{1}{n^2h}\right).\end{aligned} \tag{26}$$

For $t_1=t_2$,

$$\begin{aligned}
\mathbb{K}*K\left(\frac{t_2-t_1}{g}\right) &= \mathbb{K}*K(0) = \int \mathbb{K}(u)K(-u)du \\
&= \int \mathbb{K}(u)K(u)du = \int K(u)\left(\int_{-\infty}^{u} K(v)dv\right)du = \int\int_{\{v\le u\}} K(u)K(v)dudv \\
&= \frac{1}{2}\left(\int\int_{\{v\le u\}} K(u)K(v)dudv + \int\int_{\{u\le v\}} K(v)K(u)dvdu\right) \\
&= \frac{1}{2}\int\int_{\mathbb{R}^2} K(u)K(v)dudv = \frac{1}{2}.
\end{aligned}$$

So,

$$\mathrm{Var}\big(\varphi_{n,1}(t,x)\big) = \frac{c_K}{m(x)n^2}\big(1-F(t|x)\big)^2\big)L(t|x)\frac{1}{h} + O\left(\frac{g}{n^3h}\right). \tag{27}$$

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

$$\begin{aligned}
&\mathrm{Var}\big(\textstyle\sum_{i=1}^n \widetilde{\varphi}_{n,i}(t,x)\big) \\
&\quad = a_1^2\frac{c_K}{m(x)}\big(1-F(t+b|x)\big)^2L(t+b|x) + a_2^2\frac{c_K}{m(x)}\big(1-F(t|x)\big)^2L(t|x) \\
&\quad +4a_1a_2\frac{c_K}{m(x)}\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x) + O\left(\frac{g}{n}\right) + o(1).
\end{aligned}$$

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

As the variables $\varphi_{n,i}(t,x)$ are independent and identically distributed for all $i = 1,...,n$, the variables $\widetilde{\varphi}_{n,i}(t,x)$ are also so. In addition, $\mathrm{Var}\big(\widetilde{\varphi}_{n,i}(t,x)\big)$ 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

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

where

$$\begin{aligned}
s_0^2 &= a_1^2\frac{c_K}{m(x)}\big(1-F(t+b|x)\big)^2L(t+b|x) + a_2^2\frac{c_K}{m(x)}\big(1-F(t|x)\big)^2L(t|x) \\
&\quad +4a_1a_2\frac{c_K}{m(x)}\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x).
\end{aligned} \tag{29}$$

We will now check Lindeberg's condition:

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

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

$$\mathbb{1}_{n,i} = \mathbb{1}\Big(\ \widetilde{\varphi}_{n,i}(t,x) - E[\widetilde{\varphi}_{n,i}(t,x)]\ > \varepsilon s_0\Big).$$

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

$$\begin{aligned}|\xi(Z,\delta,t,x)| &= \frac{1_{\{Z\le t,\delta=1\}}}{1-H(Z|x)} - \int_0^t \frac{dH_1(u|x)}{\big(1-H(u|x)\big)^2} \\ &\le \frac{1_{\{Z\le t,\delta=1\}}}{1-H(Z|x)} + \int_0^t \frac{dH_1(u|x)}{\big(1-H(u|x)\big)^2} \le \frac{1}{\theta} + \int_0^t \frac{dH_1(u|x)}{\theta^2} \\ &\le \frac{1}{\theta} + \frac{H(t|x)}{\theta^2} \le \frac{1}{\theta} + \frac{1}{\theta^2}\end{aligned}$$

and, consequently, η is also bounded:

$$\begin{aligned}|\eta(Z,\delta,t,x)| &\le \int K(u)\big(1-F(t-gu|x)\big)\left(\frac{1}{\theta}+\frac{1}{\theta^2}\right)du \\ &= \left(\frac{1}{\theta}+\frac{1}{\theta^2}\right)\left(\big(1-F(t|x)\big)+\frac{g^2}{2}d_K\big(1-F''(t|x)\big)\right)+O(g^2).\end{aligned}$$

Since η is bounded, K and $m(x)$ have compact support and $nh\to\infty$, $\big\{\widetilde{\varphi}_{n,i}(t,x), i=1,...,n, n\in\mathbb{N}$ 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 $n\ge n_0$ and accordingly,

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

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

Furthermore, from Theorem 3 in Peláez et al. (2021a),

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

so,

$$\begin{aligned}E\Big(\textstyle\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)\Big) &= nE\big(\widetilde{\varphi}_{n,1}(t,x)\big) \\ &= a_1n\sqrt{nh}E\big(\varphi_{n,1}(t+b,x)\big)+a_2n\sqrt{nh}E\big(\varphi_{n,1}(t,x)\big) \\ &= \sqrt{nh^5}\big(a_1B_1(t+b,x)+a_2B_1(t,x)+o(h^2)\big).\end{aligned}$$

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

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

where

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

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

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

where

$$\mu_1 = \mu_0 + C_h^{1/2} C_g^{4/2} \big(a_1 B_2(t+b,x) + a_2 B_2(t,x) \big).$$

Finally, using equation (20) with $a_1 = \dfrac{1}{S(t|x)}$ and $a_2 = -\dfrac{S(t+b|x)}{S(t|x)^2}$, the asymptotic distribution of the PD estimator holds:

$$\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\mu, s_0),$$

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

$$\begin{aligned}
\mu &= C_h^{5/2} \left(\frac{S(t+b|x)}{S(t|x)^2} B_1(t,x) - \frac{B_1(t+b,x)}{S(t|x)} \right) \\
&\quad + C_h^{1/2} C_g^{4/2} \left(\frac{S(t+b|x)}{S(t|x)^2} B_2(t,x) - \frac{B_2(t+b,x)}{S(t|x)} \right) \\
&= C_h^{5/2} \frac{\big(1 - PD(t|x)\big) B_1(t,x) - B_1(t+b,x)}{S(t|x)} \\
&\quad + C_h^{1/2} C_g^{4/2} \frac{\big(1 - PD(t|x)\big) B_2(t,x) - B_2(t+b,x)}{S(t|x)}
\end{aligned}$$

and

$$\begin{aligned}
s_0^2 &= \frac{1}{S(t|x)^2} \frac{c_K \big(1 - F(t+b|x)\big)^2 L(t+b|x)}{m(x)} + \frac{S(t+b|x)^2}{S(t|x)^4} \frac{c_K \big(1 - F(t|x)\big)^2 L(t|x)}{m(x)} \\
&\quad - 4 \frac{S(t+b|x)}{S(t|x)^3} \frac{c_K \big(1 - F(t|x)\big)\big(1 - F(t+b|x)\big) L(t|x)}{m(x)} \\
&= \frac{V_1(t+b,x)}{S(t|x)^2} - 4 \frac{S(t+b|x)}{S(t|x)^3} \frac{c_K \big(1 - F(t|x)\big)\big(1 - F(t+b|x)\big) L(t|x)}{m(x)} \\
&\quad + \frac{S(t+b|x)^2 V_1(t,x)}{S(t|x)^4}.
\end{aligned}$$

■