Computational Statistics and Data Analysis 189 (2024) 107853

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Probability of default estimation in credit risk using mixture cure models

Rebeca Peláez [a,*], Ingrid Van Keilegom [b], Ricardo Cao [c], Juan M. Vilar [c]

[a] *Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain*
[b] *Research Centre for Operations Research and Statistics (ORSTAT), KU Leuven, Leuven, Belgium*
[c] *Research Group MODES, Department of Mathematics, CITIC, Universidade da Coruña, A Coruña, Spain*

ARTICLE INFO

Article history:
Received 13 December 2022
Received in revised form 6 September 2023
Accepted 6 September 2023
Available online 14 September 2023

Keywords:
Censored data
Survival analysis
Nonparametric estimation
Kernel method

ABSTRACT

An estimator of the probability of default (PD) in credit risk is proposed. It is derived from a nonparametric conditional survival function estimator based on cure models. Asymptotic expressions for the bias and the variance, as well as the asymptotic normality of the proposed estimator are presented. A simulation study shows the performance of the nonparametric estimator compared with Beran's PD estimator and other semiparametric methods. Finally, an empirical study based on modified real data illustrates the practical behaviour.

1. Introduction

In the context of credit risks or credit scoring one is often interested in modelling and estimating the probability of default (PD) measuring the probability of an obligor to run into arrears on his/her credit obligation. A binary classification of customers into two categories (default or not default) is then required, which can be done using various statistical techniques ranging from purely parametric to fully nonparametric. However, a more refined analysis is possible, in which apart from this binary outcome (default or not default) one also takes the timing of default into account. The probability that a customer defaults before a given time point is of practical importance, since it can provide the bank with the ability to compute the profitability over a customer's lifetime and perform profit scoring. In this paper we will propose a novel method to estimate the probability of default (PD) in a time horizon $t+b$ from a maturity time t using nonparametric estimators. To estimate this probability, one commonly faces the problem that the time of default is censored to the right. This is because at the end of the study period some (or many) customers will not have defaulted, or some customers might be lost to follow up for various reasons in the course of the study period. As a result, appropriate estimators that take right censoring into account should be used. This has been recognised by Peláez et al. (2021a,b), who used nonparametric estimators of the PD based on Beran's estimator of the conditional survival function (Beran (1981)) given a set of covariates. This estimator is an extension of the Kaplan and Meier (1958) estimator to the regression context, where kernel smoothing and an appropriate bandwidth are used for the covariates. See also Naraim (1992), Stepanova and Thomas (2002), Roszbach (2003), Glennon and Nigro (2005), Allen and Rose (2006), Baba and Goko (2006), and Dirick et al. (2003), among others, for other contributions on the use of survival analysis in the context of credit scoring.

[*] Corresponding author.
E-mail address: rebeca.pelaez@uc3m.es (R. Peláez).

https://doi.org/10.1016/j.csda.2023.107853

In this paper we go one step further. In fact, the time to default does not only face a problem of right censoring. There is a second issue that should also be taken into account, and which is caused by the fact that some customers never default, that is, no matter how long you observe such individuals, they will never experience the event of interest. Hence, the survival function of the time to default will have a point mass at infinity. Survival models that take this feature into account are called cure models. We refer to Amico and Van Keilegom (2018), for an overview paper on this topic. Instead of working with the Beran estimator (Beran (1981)), we will therefore use another nonparametric estimator, that estimates separately the probability of no default (so the point mass at infinity), called the incidence, and the survival function for the defaulted customers, called the latency. For both quantities a kernel estimator (depending on possibly different bandwidths) will be used. This is useful, since different degrees of smoothness for the incidence and latency require different bandwidths in order to estimate the PD in an optimal way.

Cure survival models are nowadays well developed in the statistics and biostatistics literature, where the number of papers studying various aspects of cure models (on e.g. estimation, testing, prediction, model selection, among others) has increased a lot over the last 10 years. However in the area of credit risks cure models have not been used a lot so far, despite their natural applications. Notable exceptions are Beran and Djaïdja (2007), Dirick et al. (2019) and Dirick et al. (2015). In the latter paper an AIC variable selection procedure is proposed in the context of PD estimation based on cure models.

The remainder of this paper is organised as follows. In Section 2, the nonparametric estimator of the PD based on mixture cure models is proposed. Asymptotic properties of this PD estimator are presented in Section 3. Section 4 presents a bootstrap bandwidth selector for the bandwidths involved in the nonparametric estimator of the PD based on mixture cure models. In Section 5, a simulation study shows the behaviour of the nonparametric cure model estimator and a comparison with Beran's estimator and other semiparametric estimators. In Section 6, the PD estimators are applied to a set of modified real data. Finally, Section 7 contains some concluding remarks. Appendix A and Appendix B include the assumptions and detailed proofs of the theoretical results.

2. Probability of default estimator

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a random sample of (X, Z, δ) where X is the credit scoring, $Z = \min\{T, C\}$ is the observed maturity, T is the time to default, C is the time until the end of the study or the time until the anticipated cancellation on the credit and $\delta = I(T \leq C)$ is the uncensoring indicator. Let ν be a binary variable where $\nu = 0$ indicates if the individual belongs to the susceptible group (the individual will eventually experience the default if followed for long enough) and $\nu = 1$ indicates if the subject is cured (the individual will never experience the default). Therefore, $T = (1-\nu)T_0 + \nu\infty$, where T_0 denotes the survival time of an individual susceptible to default. According to these variables, the population is classified into three groups: those who are susceptible to default and censored ($\nu = 0$, $\delta = 0$), those who are susceptible to default and noncensored ($\nu = 0$, $\delta = 1$) and the group of cured individual who are not susceptible to default ($\nu = 1$, $\delta = 0$). The situation ν =1 and δ =1 is not feasible. In practice, distinguishing whether or not the censored individual was susceptible to experiencing the default (belongs to first or third group) is not possible without additional assumptions. In this context, the Law of Total Probability provides a useful decomposition of the conditional survival function as follows

$$\begin{aligned} S(t|x) &= P(T > t|\nu = 1, X = x)P(\nu = 1|X = x) \\ &\quad + P(T > t|\nu = 0, X = x)P(\nu = 0|X = x) = 1 - p(x) + S_0(t|x)p(x), \end{aligned}$$

where $p(x)$ is the probability of not being cured (susceptible to default) and $S_0(t|x)$ the conditional survival function of the uncured population. The functions $1 - p(x)$ and $S_0(t|x)$ are called the incidence and the latency, respectively.

Let $x \in I \subseteq \mathbb{R}$ be a fixed value of the covariate X (typically, the scoring), t covering certain interval $I_T \subseteq \mathbb{R}$ and b a horizon time (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

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

Replacing $S(t|x)$ with a conditional survival function estimator, $\widehat{S}_h(t|x)$, in (1), the following estimator for $\mathrm{PD}(t|x)$ is obtained:

$$\widehat{\mathrm{PD}}_h(t|x) = 1 - \frac{\widehat{S}_h(t+b|x)}{\widehat{S}_h(t|x)}, \tag{2}$$

where $h = h_n$ is the smoothing parameter for the covariable.

The aim is to find an appropriate survival estimator, $\widehat{S}_h(t|x)$, that captures the existence of a group of individuals not susceptible to default or cured, resulting in a good estimator of the probability of default, $\widehat{\mathrm{PD}}_h(t|x)$, in this context. For this purpose, a nonparametric survival estimator based on cure models is considered. Beran's estimator which, a priori, does not take into account the proportion of the curative population is also considered in this work to estimate the probability of default.

2.1. Beran's estimator

The estimator of the conditional survival function with censored data formulated in Beran (1981) is given by

$$\widehat{S}_h^{\mathrm{B}}(t|x) = \prod_{i=1}^{n}\left(1 - \frac{I_{\{Z_i \le t,\, \delta_i = 1\}} w_{h,i}(x)}{1 - \sum_{j=1}^{n} I_{\{Z_j < Z_i\}} w_{h,j}(x)}\right), \tag{3}$$

with

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

where K is a kernel function (typically a density function to be picked up by the user) and $h > 0$ is a smoothing parameter.

Replacing (3) in (2), we obtain Beran's estimator of the probability of default. It was previously used in Cao et al. (2009), Peláez et al. (2021b) and Peláez et al. (2021a).

2.2. Nonparametric cure model estimator

The nonparametric cure model estimator of the conditional survival function proposed by López-Cheda (2018) is given by

$$\widehat{S}_{h,g}^{\mathrm{NPCM}}(t|x) = 1 - \widehat{p}_h(x) + \widehat{p}_h(x)\widehat{S}_{0,g}(t|x). \tag{4}$$

The incidence estimator, $1 - \widehat{p}_h(x)$, is proposed by Xu and Peng (2014) and deeply studied in López-Cheda et al. (2017b). It corresponds to Beran's estimator evaluated at the highest uncensored lifetime:

$$1 - \widehat{p}_h(x) = \widehat{S}_h^{\mathrm{B}}\big(\max\{T_i : i = 1, \dots, n, \delta_i = 1\}|x\big).$$

The latency estimator, $\widehat{S}_{0,g}(t|x)$, proposed by López-Cheda et al. (2017a) is as follows:

$$\widehat{S}_{0,g}(t|x) = \frac{\widehat{S}_g^{\mathrm{B}}(t|x) - \big(1 - \widehat{p}_g(x)\big)}{\widehat{p}_g(x)},$$

where $g > 0$ is a smoothing parameter.

Replacing (4) in (2), we obtain the nonparametric cure model estimator (NPCM) of the probability of default:

$$\widehat{\mathrm{PD}}_{h,g}^{\mathrm{NPCM}}(t|x) = 1 - \frac{\widehat{S}_{h,g}^{NPCM}(t + b|x)}{\widehat{S}_{h,g}^{NPCM}(t|x)}. \tag{5}$$

Note that the particular case $h = g$ corresponds to Beran's estimator, which does not take into account a priori the existence of a group of cured individuals. In López-Cheda (2018) it was found by simulation that the bandwidths h and g are substantially different in practice, although they have the same convergence order. Choosing the best bandwidth h for incidence and the best bandwidth g for latency has a considerable effect on the estimation of the conditional survival curve in cure models and could have a considerable effect on the estimation of PD.

3. Asymptotic properties of the NPCM estimator

In this Section, asymptotic properties of the probability of default estimators are studied. Since Beran's estimator of the probability of default has been deeply studied in Peláez et al. (2021b) and details about its asymptotic properties can be found in that work, this section will focus on the NPCM estimator of the PD. The following notation is used.

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

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

and given any constant $a \in \mathbb{R}$,

$$\widetilde{c}_R(a) = \int R(at)R(t)dt. \tag{6}$$

Given any function $f : \mathbb{R}^k \longrightarrow \mathbb{R}$, its first derivative with respect to the first variable is denoted by: $f'(x_1, \dots, x_k) = \dfrac{\partial f(x_1, \dots, x_k)}{\partial x_1}$. Correspondingly, the second derivative with respect to the first variable is denoted by $f''(x_1, \dots, x_k)$.

The following functions are required to state the results. A number of notations used below are defined in Appendix A.

$$
\begin{aligned}
\xi(Z,\delta,t,x) &= \frac{1_{\{Z\leq t,\delta=1\}}}{1-H(Z|x)} - \int_0^t \frac{1_{\{u\leq Z\}}dH_1(u|x)}{\big(1-H(u|x)\big)^2},\\
\eta(Z,\delta,t,x) &= -\frac{S(t|x)}{p(x)}\xi(Z,\delta,t,x) - \frac{\big(1-p(x)\big)\big(1-S(t|x)\big)}{p^2(x)}\xi(Z,\delta,\infty,x),\\
\Phi(u,t,x) &= E\big[\xi(Z,\delta,t,x)|X=u\big], \quad \Phi_2(u,t,x) = E\big[\xi^2(Z,\delta,t,x)|X=u\big],\\
B_1(t,x) &= \frac{d_K\big(S_0(t|x)-1\big)\big(p(x)-1\big)}{2m(x)}\frac{\partial^2}{\partial u^2}\big(\Phi(u,t,x)m(u)\big)|_{u=x},\\
B_2(t,x) &= -\frac{d_k S(t|x)}{2m(x)}\frac{\partial^2}{\partial u^2}\big(\Phi(u,t,x)m(u)\big)|_{u=x}\\
&\quad -\frac{d_K\big(1-p(x)\big)\big(1-S(t|x)\big)}{2p(x)m(x)}\frac{\partial^2}{\partial u^2}\big(\Phi(u,\infty,x)m(u)\big)|_{u=x},\\
\widetilde{B}_1(t,x) &= -\frac{1}{S(t|x)}B_1(t+b,x) + \frac{S(t+b|x)}{S^2(t|x)}B_1(t,x),\\
\widetilde{B}_2(t,x) &= -\frac{1}{S(t|x)}B_2(t+b,x) + \frac{S(t+b|x)}{S^2(t|x)}B_2(t,x),\\
D(u,t_1,t_2,x) &= Cov\big[\xi(Z_1,\delta_1,t_1,x),\xi(Z_1,\delta_1,t_2,x)\big|X_1=u\big]m(u),\\
L(u,t_1,t_2,x) &= Cov\big[\xi(Z_1,\delta_1,t_1,x),\eta(Z_1,\delta_1,t_2,x)\big|X_1=u\big]m(u),\\
C_1(t_1,t_2,x) &= \frac{c_K S(t_1|x)S(t_2|x)}{p^2(x)}D(x,t_1,t_2,x) + \frac{c_K S(t_1|x)\big(1-S(t_2|x)\big)}{p^3(x)}D(x,t_1,\infty,x)\\
&\quad + \frac{c_K\big(1-S(t_1|x)\big)S(t_2|x)\big(1-p(x)\big)}{p^3(x)}D(x,\infty,t_2,x)\\
&\quad + \frac{c_K\big(1-p(x)\big)^2\big(1-S(t_1|x)\big)\big(1-S(t_2|x)\big)}{p^4(x)}\Phi_2(x,\infty,x)m(x),\\
V_1(t_1,t_2,x) &= \frac{\big(S_0(t_1|x)-1\big)\big(S_0(t_2|x)-1\big)\big(p(x)-1\big)^2}{m(x)}c_K\Phi_2(x,\infty,x),\\
V_2(t_1,t_2,x) &= \frac{p^2(x)C_1(t_1,t_2,x)}{m^2(x)},\\
V_3(t_1,t_2,x) &= \frac{\big(S_0(t_1|x)-1\big)\big(p(x)-1\big)p(x)}{m^2(x)}L(x,\infty,t_2,x)\\
&\quad + \frac{\big(S_0(t_2|x)-1\big)\big(p(x)-1\big)p(x)}{m^2(x)}L(x,t_1,\infty,x).
\end{aligned}
$$

The required assumptions are listed in Appendix A. They are standard in the literature and not very restrictive in this context. They were previously assumed by Peláez et al. (2021a) to prove the asymptotic properties for Beran's PD estimator, by López-Cheda et al. (2017a) and López-Cheda et al. (2017b) to prove the asymptotic properties of the incidence and latency estimators, and by Iglesias-Pérez and González-Manteiga (1999) and Dabrowska (1989) in the nonparametric conditional survival function estimation setup.

Assumptions A.1 and A.2 are about characteristics and independence of the variables involved. Assumptions A.3-A.12 are needed to bound some population functions. They require existence and continuity of population function derivatives. Kernel function requirements are covered in Assumption A.13 and bandwidth assumptions are included in A.14 and A.15. Assumption A.16 refers to the differentiability of the functions previously defined in this section.

Lemma 1 *(Almost sure representation of the NPCM estimator for the conditional survival function). Under Assumptions A.1-A.16, for fixed values $(t,x)\in[l,u]\times I$, defined in Appendix A,*

$$
\begin{aligned}
\widehat{S}_{h,g}^{NPCM}(t|x) - S(t|x) = \big(S_0(t|x)-1\big)\big(p(x)-1\big)\textstyle\sum_{i=1}^n w_{h,i}^A(x)\xi(Z_i,\delta_i,\infty,x)&\\
+p(x)\textstyle\sum_{i=1}^n w_{g,i}^A(x)\eta(Z_i,\delta_i,t,x) + R_n^1(t|x) \quad a.s.,&
\end{aligned} \tag{7}
$$

where $w_{h,i}^A(x) = \dfrac{1}{nh}\dfrac{K\big((x-X_i)/h\big)}{m(x)}$, *and* $\sup_{(t,x)\in[l,u]\times I}|R_n^1(t|x)| = O_p\left(\ln n\left(\dfrac{1}{nh}+\dfrac{1}{ng}\right)\right)^{3/4}$.

Theorem 1 *(Almost sure representation of the NPCM estimator for the PD). Under Assumptions A.1-A.16, for fixed values (t,x), $(t+b,x)\in[l,u]\times I$,*

$$
\widehat{PD}_{h,g}^{NPCM}(t|x) - PD(t|x) = \sum_{i=1}^n \Psi_{n,i}(t,x) + R_n^2(t|x) \quad a.s.,
$$

where

$$\Psi_{n,i}(t,x) = -\frac{1}{S(t|x)}\varphi_{n,i}(t+b,x) + \frac{S(t+b|x)}{S^2(t|x)}\varphi_{n,i}(t,x),$$
$$\varphi_{n,i}(t,x) = \big(S_0(t|x)-1\big)\big(p(x)-1\big)w^A_{h,i}(x)\xi(Z_i,\delta_i,\infty,x) + p(x)w^A_{g,i}(x)\eta(Z_i,\delta_i,t,x)$$

and

$$\sup_{(t,x)\in[l,u]\times I}|R_n^2(t|x)| = O_p\left(\ln n\left(\frac{1}{nh}+\frac{1}{ng}\right)\right)^{3/4}.$$

Theorem 2 *(Asymptotic bias and variance of the NPCM estimator for the PD). Under Assumptions A.1-A.16, for fixed values* (t,x), $(t+b,x)\in[l,u]\times I$, *the asymptotic expressions of the bias and the variance of the dominant term in the almost sure representation of* $\widehat{PD}^{NPCM}_{h,g}(t|x)$ *are the following:*

$$ABias\big(\widehat{PD}^{NPCM}_{h,g}(t|x)\big) = \widetilde{B}_1(t,x)h^2 + \widetilde{B}_2(t,x)g^2 + o(h^2) + o(g^2). \tag{8}$$

(i) *If* $C_{h,g} := \lim_{n\to\infty}\dfrac{h}{g}\in(0,\infty)$, *then*

$$AVar\big(\widehat{PD}^{NPCM}_{h,g}(t|x)\big) = \Big(\widetilde{V}_1(t+b,t,x) + C_{h,g}\widetilde{V}_2(t+b,t,x) + C_{h,g}\widetilde{c}_K(C_{h,g})\widetilde{V}_3(t+b,t,x)\Big)\frac{1}{nh} + o\left(\frac{1}{nh}\right) + O\left(\frac{h}{n}\right).$$

(ii) *If* $\lim_{n\to\infty}\dfrac{h}{g} = 0$, *then*

$$AVar\big(\widehat{PD}^{NPCM}_{h,g}(t|x)\big) = \widetilde{V}_1(t+b,t,x)\frac{1}{nh} + o\left(\frac{1}{nh}\right) + O\left(\frac{g}{n}\right).$$

(iii) *If* $\lim_{n\to\infty}\dfrac{g}{h} = 0$, *then*

$$AVar\big(\widehat{PD}^{NPCM}_{h,g}(t|x)\big) = \widetilde{V}_2(t+b,t,x)\frac{1}{ng} + o\left(\frac{1}{ng}\right) + O\left(\frac{h}{n}\right).$$

The functions $\widetilde{V}_i(t_1,t_2,x)$ *are defined as follows*

$$\widetilde{V}_i(t_1,t_2,x) = \frac{1}{S^2(t_2|x)}V_i(t_1,t_1,x) + \frac{S^2(t_1|x)}{S^2(t_2|x)}V_i(t_2,t_2,x) + 2\frac{S(t_1|x)}{S^2(t_2|x)}V_i(t_1,t_2,x),$$

where $i=1,2,3$ *and* $\widetilde{c}_K$ *is defined in* (6).

Theorem 3 *(Asymptotic normality of the NPCM estimator for the PD). Under Assumptions A.1-A.16, for fixed values* (t,x), $(t+b,x)\in[l,u]\times I$, *the limit distribution of* $\widehat{PD}^{NPCM}_{h,g}(t|x)$ *is the following:*

(i) *Assuming* $C_h := \lim_{n\to\infty} n^{1/5}h \in (0,\infty)$, $C_g := \lim_{n\to\infty} n^{1/5}g\in(0,\infty)$, *then*

$$\sqrt{nh}\big(\widehat{PD}^{NPCM}_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\mu,s),$$

where $\mu = C_h^{5/2}\widetilde{B}_1(t,x) + C_g^{5/2}\widetilde{B}_2(t,x)$ *and* $s^2 = \big(\widetilde{V}_1(t+b,t,x) + C_{h,g}\widetilde{V}_2(t+b,t,x) + C_{h,g}\widetilde{c}_K(C_{h,g})\widetilde{V}_3(t+b,t,x)\big)$.

(ii) *Assuming* $C_g := \lim_{n\to\infty} n^{1/5}g \in (0,\infty)$ *and* $\lim_{n\to\infty} n^{1/5}h = 0$, $\dfrac{(\ln n)^3}{nh}\to 0$ *and* $\left(\dfrac{\ln n}{ng}\right)^{3/4}(nh)^{1/2}\to 0$, *then*

$$\sqrt{nh}\big(\widehat{PD}^{NPCM}_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\mu,s),$$

where $\mu = C_g^{5/2}\widetilde{B}_2(t,x)$ *and* $s^2 = \widetilde{V}_1(t+b,t,x)$.

(iii) *Assuming* $C_h := \lim_{n\to\infty} n^{1/5}h \in (0,\infty)$, $\lim_{n\to\infty} n^{1/5}g = 0$, $\dfrac{(\ln n)^3}{ng} \to 0$ *and* $\left(\dfrac{\ln n}{nh}\right)^{3/4}(ng)^{1/2} \to 0$, *then*

$$\sqrt{ng}\left(\widehat{PD}^{NPCM}_{h,g}(t|x) - PD(t|x)\right) \xrightarrow{d} N(\mu, s),$$

where $\mu = C_h^{5/2}\widetilde{B}_1(t,x)$, $s^2 = \widetilde{V}_2(t+b,t,x)$ *and* $\widetilde{V}_i(t_1,t_2,x)$, $i=1,2,3$ *are defined in Theorem 2.*

Proofs of the results presented here are included in Appendix B.

The particular choice $h = g$ for the NPCM estimator corresponds to Beran's estimator. Therefore, the case $C_{h,g} = 1$ should give the same asymptotic bias and variance for Beran's and the NPCM estimators. Asymptotic expressions for the bias and variance of Beran's estimator are available in Cao et al. (2009) and Peláez et al. (2021b). It is clear that the order of these asymptotic expressions is the same for both estimators when $\lim_{n\to\infty} h/g = C_{h,g} \in (0,\infty)$. If we consider the particular case $C_{h,g} = 1$, then bandwidths h and g are asymptotically equal and so are the expressions for the bias and variance corresponding to Beran's and NPCM estimators.

4. Bandwidth selection

The choice of the smoothing parameters on which these estimators depend is certainly a point of crucial interest. The complexity of the asymptotic results shown in the previous section makes it difficult to obtain plug-in bandwidths, since they depend on too many parameters and population functions. For this reason, bootstrap-based bandwidth selectors are used.

An automatic selector based on a bootstrap procedure already exists in the literature for Beran's PD estimator. In Peláez et al. (2022), the obvious bootstrap method is combined with a smoothed bootstrap for the automatic selection of the bandwidth h of Beran's estimator, $\widehat{\mathrm{PD}}^{\mathrm{B}}_h(t|x)$, defined in Section 2.1.

There are two classic methods for bootstrap resampling in a censoring context: the obvious bootstrap and the simple bootstrap. In Li and Datta (2001), both methods are extended to the case where a covariate is involved, assuming there are no ties in the sample values of the covariate. In López-Cheda et al. (2017a) and López-Cheda et al. (2017b), automatic selectors were proposed for the bandwidths h and g on which the incidence and latency respectively depend. The proposed resampling algorithm is a simple weighted bootstrap, fixing the covariate, equivalent to the one presented in Li and Datta (2001). In Peláez et al. (2022) this method is combined with the smoothed bootstrap to approximate the bandwidth involved in Beran's estimator of the probability of default. In this paper, these techniques are extended to the case where there exists a cure fraction to approximate the smoothing parameters involved in the NPCM estimator. The algorithm for the bootstrap resampling is detailed below. For the sake of brevity, the NPCM estimator, $\widehat{\mathrm{PD}}^{\mathrm{NPCM}}_{h,g}(t|x)$, given in (5) is simply denoted by $\widehat{PD}_{h,g}(t|x)$ in this section.

Algorithm for bootstrap resampling based on the NPCM estimator (called BR). Let $I_1, I_2 \subseteq \mathbb{R}$ be intervals containing appropriate bandwidth values and let $(r,s) \in I_1 \times I_2$ be pilot bandwidths for the bootstrap resampling:

1. Obtain $U_1,\ldots,U_n$ iid with $U_i \sim U(0,1)$ for all $i = 1,\ldots,n$.
2. For each $i = 1,\ldots,n$, define

   $$X_i^* = X_{[nU_i]+1},$$

   where $[u]$ is the integer part of u and generate T_i^* from the NPCM estimator of the conditional distribution of T using the sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ and bandwidths (r,s), denoted by $\widehat{F}_{r,s}(t|X_i^*)$, and C_i^* from the NPCM estimator of the conditional distribution of C using the sample $\{(X_i, Z_i, 1-\delta_i)\}_{i=1}^n$ and bandwidths (r,s), denoted by $\widehat{G}_{r,s}(t|X_i^*)$.
   The estimators $\widehat{F}_{r,s}(t|X_i^*)$ and $\widehat{G}_{r,s}(t|X_i^*)$ are forced to be equal to one from the last observed lifetime ($\max\{Z_i : i = 1,\ldots,n\}$) onwards.
3. For each $i = 1,\ldots,n$, obtain

   $$Z_i^* = \min\{T_i^*, C_i^*\},$$
   $$\delta_i^* = I\left(T_i^* \le C_i^*\right).$$
4. Consider the bootstrap resample $\left\{(X_i^*, Z_i^*, \delta_i^*)\right\}_{i=1}^n$.

For the NPCM estimator, $\widehat{PD}_{h,g}(t|x)$, and a fixed x, the optimal two-dimensional bandwidth is the pair $(h,g) \in I_1 \times I_2 \subset \mathbb{R}^2$ that minimises the mean integrated squared error given by

$$\mathrm{MISE}_x(h,g) = E\left(\int\limits_{I_T}\left(\widehat{\mathrm{PD}}_{h,g}(t|x) - \mathrm{PD}(t|x)\right)^2 dt\right), \tag{9}$$

whose bootstrap approximation is

$$\mathrm{MISE}^*_x(h,g) = E\bigg(\int\limits_{I_T}\left(\widehat{\mathrm{PD}}^*_{h,g}(t|x) - \widehat{\mathrm{PD}}_{r,s}(t|x)\right)^2 dt\bigg), \tag{10}$$

where $\widehat{\mathrm{PD}}_{r,s}(t|x)$ is the NPCM estimator with pilot bandwidths $(r,s) \in I_1 \times I_2$ using the sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$ and $\widehat{\mathrm{PD}}^*_{h,g}(t|x)$ is the bootstrap NPCM estimator of PD with bandwidths (h,g) using the bootstrap resample $\left\{(X_i^*, Z_i^*, \delta_i^*)\right\}_{i=1}^n$.

The resampling distribution of $\widehat{\mathrm{PD}}^*_{h,g}(t|x)$ cannot be computed in a closed form, so the Monte Carlo method is used. The distribution of $\widehat{\mathrm{PD}}^*_{h,g}(t|x)$ is approximated by the empirical one of $\widehat{\mathrm{PD}}^{*,1}_{h,g}(t|x), \ldots, \widehat{\mathrm{PD}}^{*,B}_{h,g}(t|x)$, obtained from B bootstrap resamples using the bootstrap resampling algorithm (BR) explained above. Then, the bootstrap bivariate bandwidth, (h^*, g^*), is the minimiser of the Monte Carlo approximation of $\mathrm{MISE}^*_x(h,g)$ over a meshgrid of bandwidths $(h,g) \in I_1 \times I_2$ given by

$$\mathrm{MISE}^*_x(h,g) \simeq \frac{1}{B}\sum_{k=1}^{B}\bigg(\int\limits_{I_T}\left(\widehat{\mathrm{PD}}^{*,k}_{h,g}(t|x) - \widehat{\mathrm{PD}}_{r,s}(t|x)\right)^2 dt\bigg), \tag{11}$$

where $\widehat{\mathrm{PD}}_{r,s}(t|x)$ is the NPCM estimator with auxiliary bandwidths $(r,s) \in I_1 \times I_2$ using the sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$ and $\widehat{\mathrm{PD}}^{*,k}_{h,g}(t|x)$ is the bootstrap NPCM estimator of PD with bandwidths (h,g) using the k-th bootstrap resample $\left\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\right\}_{i=1}^n$. Likewise, the integral is approximated by a Riemann sum.

Concerning the auxiliary bandwidths, preliminary studies not shown here suggest that the pilot bandwidths defined by:

$$r = \frac{5}{6}\left(Q_X(0.975) - Q_X(0.025)\right)n^{-1/9}, \tag{12}$$

$$s = \frac{15}{4}\left(Q_Z(0.975) - Q_Z(0.025)\right)n^{-1/9}, \tag{13}$$

where $Q_X(u)$ is the u quantile of the sample $\left\{X_i\right\}_{i=1}^n$, are suitable choices in this context.

Note that, in López-Cheda et al. (2017a) and López-Cheda et al. (2017b), the authors propose the following pilot bandwidth for the incidence and the latency bandwidth selectors:

$$c\left(X_{(n)} - X_{(1)}\right)n^{-1/9}, \tag{14}$$

where $X_{(1)}$ and $X_{(n)}$ are the minimum and maximum values of the covariate X, respectively, and $c > 0$.

Regarding the auxiliary bandwidth $r \in I_1$, instead of a naive selector depending on $X_{(1)}$ and $X_{(n)}$, we use the quantiles $Q_X(0.025)$ and $Q_X(0.975)$ to avoid outliers. In addition, this bandwidth considers the variability of the covariate, $Q_X(0.975) - Q_X(0.025)$, and the sample size, n. The multiplicative constant $5/6$ is derived from several attempts in simulation.

The exponent of this sample size, $-1/9$, was heuristically determined by López-Cheda et al. (2017a) and López-Cheda et al. (2017b). The order $n^{-1/9}$ for this pilot bandwidth satisfies the conditions of Theorem 1 in Li and Datta (2001) and is the one obtained by Cao (1993) for the uncensored case in nonparametric density estimation. It should be noted that the bandwidth sequence $r = r(n)$ has to be typically asymptotically larger than $h = h(n)$.

The pilot bandwidth $s \in I_2$ given in (13) also follows the ideas of López-Cheda et al. (2017a) and López-Cheda et al. (2017b). Simulation studies in López-Cheda (2018) show that a good choice for the auxiliary bandwidth related to the latency would be to consider the same naive selector as for the incidence. Once again, the variability of the sample is taken into account, but we consider the quantiles $Q_Z(0.025)$ and $Q_Z(0.975)$ instead of the minimum and maximum to minimise the effect of outliers. The multiplicative constant $15/4$ is derived from several attempts in simulation.

Note that the proposed algorithm is also valid to obtain a bootstrap approximation of the optimal bandwidth for the estimation of $PD(t|x)$ for fixed values of $t \in I_T$ and $x \in I$ by replacing $MISE^*_x(h,g)$ by $MSE^*_{t,x}(h,g)$, which is the bootstrap analogue of the mean squared error given by

$$\mathrm{MSE}_{t,x}(h,g) = E\bigg(\left(\widehat{\mathrm{PD}}_{h,g}(t|x) - \mathrm{PD}(t|x)\right)^2\bigg).$$

5. Simulation study

A simulation study was conducted in order to compare the performance of the two proposed estimators of the probability of default. The study is focused on three different models. All three have a non zero probability of cure and the proportion of cured subjects and the survival distribution of uncured subjects are modelled separately. Therefore, they are mixture cure models.

In Model 1, the probability of cure $1-p(x)$ is a logistic function with the incidence given by

$$p(x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)},$$

where $\beta_0=1$ and $\beta_1=-1$. A uniform distribution $U(0,1)$ is considered for the credit scoring variable X. In the uncured population, the time to default conditional to the credit scoring, $T_0|_{X=x}$, follows a Weibull distribution with parameters d and $A(x)^{-1/d}$, with $d=2$ and $A(x)=1+5x$, $T_0|_{X=x}\sim\mathcal{W}(d,A(x)^{-1/d})$, and the censoring time conditional to the credit scoring, $C_0|_{X=x}$, follows a Weibull distribution with parameters d and $B(x)^{-1/d}$, with $B(x)=10-22x+20x^2$, $C_0|_{X=x}\sim\mathcal{W}(d,B(x)^{-1/d})$. Therefore, the latency is given by $S_0(t|x)=e^{-A(x)t^d}$. It is quite close to fulfil a proportional hazards model and an accelerated failure time model, since the polynomial $A(x)$ is a linear function which is reasonable close to the function $\exp(\gamma x)$ for some γ.

In this scenario, the conditional survival function and the probability of default are the following:

$$S(t|x)=1-p(x)+p(x)e^{-A(x)t^d},$$

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

In Model 2, the incidence is given by

$$p(x)=\frac{\exp(\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3)}{1+\exp(\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3)}, \tag{15}$$

where $\beta_0=15$, $\beta_1=-190/3$, $\beta_2=88$ and $\beta_3=-128/3$. A uniform distribution $U(0,1)$ is considered for the credit scoring variable X. In the uncured population, the time to default conditional to the credit scoring, $T_0|_{X=x}$, follows an exponential distribution with parameter $Q(x)=2+58x-160x^2+107x^3$, and the censoring time conditional to the credit scoring, $C_0|_{X=x}$, follows an exponential distribution with parameter $R(x)=10-\frac{55}{2}x+20x^2$. Then, the latency is given by $S_0(t|x)=e^{-Q(x)t}$. In this scenario, the conditional survival function and the probability of default are the following:

$$S(t|x)=1-p(x)+p(x)e^{-Q(x)t},$$

$$\mathrm{PD}(t|x)=1-\frac{1-p(x)+p(x)e^{-Q(x)(t+b)}}{1-p(x)+p(x)e^{-Q(x)t}}.$$

The incidence of this model is not a logistic function and the latency function does not fit a proportional hazards model nor an accelerated failure time model, since the polynomial $Q(x)$ is not monotone in x and, therefore, is far from an exponential function.

In Model 3, the incidence is given by (15) with $\beta_0=31$, $\beta_1=-398/3$, $\beta_2=184$ and $\beta_3=-256/3$. A uniform distribution $U(0,1)$ is considered for the credit scoring variable X. In the uncured population, the time to default conditional to the credit scoring, $T_0|_{X=x}$, follows a Weibull distribution with parameters $k_1(x)=\frac{5}{1000}+28x-16x^2$ and $B_1(x)=(\log(2))^{1/k_1(x)}$, $T_0|_{X=x}\sim\mathcal{W}\big(k_1(x),1/B_1(x)\big)$, and the censoring time conditional to the credit scoring, $C_0|_{X=x}$, follows a Weibull distribution with parameters $k_2(x)=1+8x$ and $B_2(x)=(\log(2))^{1/k_2(x)}$, $C_0|_{X=x}\sim\mathcal{W}\big(k_2(x),1/B_2(x)\big)$. Therefore, the latency is given by $S_0(t|x)=e^{-(B_1(x)t)^{k_1(x)}}$. In this scenario, the conditional survival function and the probability of default are the following:

$$S(t|x)=1-p(x)+p(x)e^{-(B_1(x)t)^{k_1(x)}},$$

$$\mathrm{PD}(t|x)=1-\frac{1-p(x)+p(x)e^{-(B_1(x)(t+b))^{k_1(x)}}}{1-p(x)+p(x)e^{-(B_1(x)t)^{k_1(x)}}}.$$

The incidence of this model is not a logistic function and the latency function does not fit a proportional hazards model nor an accelerated failure time model, since the shape parameter of the Weibull distribution, $k_1(x)$, depends on x.

The simulation analysis is conducted for different credit scoring values in each model. The unconditional probability of censoring of Models 1, 2 and 3 and the probabilities of censoring conditional on each chosen value of x are shown in Table 1.

Fig. 1 shows the theoretical probability of default of Models 1, 2 and 3 when the credit scoring is $x=0.5$.

Table 1
Unconditional and conditional probabilities of censoring in Models 1, 2 and 3.

	Model 1	Model 2	Model 3
$P(\delta=0)$	0.7715	0.6566	0.7068
$P(\delta=0\vert X=0.2)$	0.8357	0.3993	0.4832
$P(\delta=0\vert X=0.5)$	0.7095	0.6111	0.7454
$P(\delta=0\vert X=0.8)$	0.7305	0.8847	0.8705

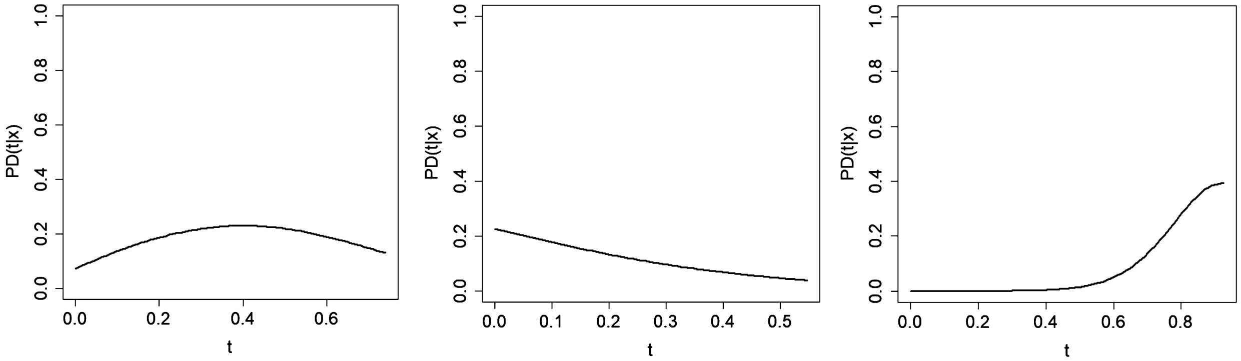

Fig. 1. Theoretical probability of default for Model 1 (left), Model 2 (centre) and Model 3 (right) when $x=0.5$.

The software for Beran's estimator was developed in R by the authors themselves. The nonparametric estimators of the incidence and latency required to compute the NPCM estimator are implemented in the R-Package *npcure* (see López-de Ullibarri et al. (2020)). Two other estimators are considered in this analysis as benchmark methods: the proportional hazards cure model estimator (PHCM) and the accelerated failure time cure model estimator (AFTCM).

The PHCM estimator and the AFTCM estimator both assume that the conditional survival function is defined by $S(t|x)=1-p(x)+p(x)S_0(t|x)$ with $1-p(x)$ fitting a logistic model and the latency $S_0(t|x)$ fitting a proportional hazards model or an accelerated failure time model, respectively. A comprehensive review of these models can be found in Peng and Yu (2021). The identifiability of the PHCM and the AFTCM, as well as the existing literature on these two models, can be consulted in Parsa and Van Keilegom (2022). The PHCM and AFTCM estimators are based on maximum likelihood techniques for the joint estimation of the incidence and latency regression parameters using the nonparametric form of the likelihood and an EM algorithm. The reader could check Sy and Taylor (2000) and Sy and Taylor (2001) for more details. Both methods are implemented in the R-Package *smcure* (see Cai et al. (2012)).

Model 1 fits Cox and AFT cure models with logistic cure probability, meanwhile Model 2 and 3 move away from semi-parametric models. Therefore, the PHCM and AFTCM methods are expected to have a reasonable behaviour in Model 1 but worse in Models 2 and 3.

The conditional survival function and the probability of default are estimated in a time grid of size n_t, $0<t_1<\cdots<t_{n_t}$, where $t_{n_t}+b=F_0^{-1}(0.95|x)$ with F_0 being the distribution function of the time variable in the uncured population and b is about 20% of the time grid. The size of the time grid is $n_t=100$. The sample size is $n=400$. The truncated Gaussian kernel is used for the covariable smoothing in Beran's estimator.

The optimal value of the bandwidth h, involved in Beran's estimator, is chosen as the value that minimises a Monte Carlo approximation of the MISE given by

$$\mathrm{MISE}_x(h)=E\left(\int\left(\widehat{\mathrm{PD}}_h^{\mathrm{B}}(t|x)-\mathrm{PD}(t|x)\right)^2dt\right),\tag{16}$$

based on the estimation for $N=100$ simulated samples for each value of h in a grid of $n_h=50$ possible values. Then, $N=300$ samples are simulated to approximate $\mathrm{MISE}_x(h)$.

The optimal bivariate bandwidth (h,g) involved in the NPCM estimator is chosen (from a meshgrid of 50 values of h and 50 values of g) as the pair that minimises a Monte Carlo approximation of the MISE given in (9) based on $N=100$ simulated samples. Then, $N=300$ simulated samples are used to approximate $\mathrm{MISE}_x(h,g)$.

Of course, these bandwidths cannot be used in practice, but this choice produces a fair comparison since the two estimators are constructed using their best possible bandwidths. In practice, automatic bandwidth selectors are used, as discussed in Section 2 and as it will be done in Section 5. The value of MISE and its square root, RMISE, are used as a measure of the estimation error committed by the PD estimators.

Tables 2-4 contain the optimal bandwidths and the square root of MISE (RMISE) for each estimator in Models 1, 2 and 3 when $x=0.2$, $x=0.5$ and $x=0.8$.

Table 2
Optimal bandwidth and RMISE for the probability of default estimators when $x = 0.2$, $x = 0.5$ and $x = 0.8$ in Model 1.

		Beran	NPCM	PHCM	AFTCM
$x = 0.2$	$h/(h, g)$	0.5224	(0.9265, 0.8714)	–	–
	RMISE	0.1351	0.1349	0.1391	0.0969
$x = 0.5$	$h/(h, g)$	0.5592	(1.0000, 0.7245)	–	–
	RMISE	0.0589	0.0589	0.0548	0.0507
$x = 0.8$	$h/(h, g)$	0.4306	(1.0000, 0.6878)	–	–
	RMISE	0.0377	0.0376	0.0457	0.0452

Table 3
Optimal bandwidth and RMISE for the probability of default estimators when $x = 0.2$, $x = 0.5$ and $x = 0.8$ in Model 2.

		Beran	NPCM	PHCM	AFTCM
$x = 0.2$	$h/(h, g)$	0.1082	(0.1276, 0.3755)	–	–
	RMISE	0.0890	0.0766	0.0939	0.1026
$x = 0.5$	$h/(h, g)$	0.1857	(0.4571, 0.3020)	–	–
	RMISE	0.0250	0.0252	0.0299	0.0305
$x = 0.8$	$h/(h, g)$	0.1469	(0.2633, 0.6327)	–	–
	RMISE	0.0668	0.0551	0.0519	0.0521

Table 4
Optimal bandwidth and RMISE for the probability of default estimators when $x = 0.2$, $x = 0.5$ and $x = 0.8$ in Model 3.

		Beran	NPCM	PHCM	AFTCM
$x = 0.2$	$h/(h, g)$	0.0776	(0.1602, 0.1786)	–	–
	RMISE	0.0684	0.0680	0.1016	0.1593
$x = 0.5$	$h/(h, g)$	0.3531	(0.9286, 0.5653)	–	–
	RMISE	0.0236	0.0238	0.0293	0.0581
$x = 0.8$	$h/(h, g)$	0.1786	(0.6143, 0.8143)	–	–
	RMISE	0.0169	0.0251	0.0283	0.0461

The NPCM estimator is performing very well in all scenarios. In general, it provides smaller errors than the semiparametric methods in Model 2 and 3. As expected, the behaviour of the AFTCM estimator is better under semiparametric Model 1, although the NPCM estimator is still competitive.

Beran's estimation error is similar to the NPCM estimation error in some cases. This is remarkable given that Beran's estimator does not consider the existence of a cured group in its definition, as the NPCM estimator does. Beran's estimator makes no assumptions about the survival function, but uses only the information provided by the data, being able to detect the nonzero tendency of the survival function and reflect it in the PD estimation.

The performance of the NPCM estimator is compared with Beran's estimator, $\widehat{\mathrm{PD}}_h^{\mathrm{B}}(t|x)$, when both are computed with bootstrap bandwidths. The bandwidth selectors presented in Section 4 are used. Models 1 and 2 when $x = 0.2$ are both considered for the study.

A number of $N = 300$ samples are simulated. For each simulated sample, the corresponding bootstrap bandwidths are approximated from $B = 500$ resamples, obtaining (h_j^*, g_j^*) with $j = 1, \ldots, N$. The mean values of the N bootstrap bandwidths defined by:

$$(\overline{h^*}, \overline{g^*}) = \left(\frac{1}{N}\sum_{j=1}^{N} h_j^*, \frac{1}{N}\sum_{j=1}^{N} g_j^*\right),$$

are included in Table 5.

For each sample, the estimation error of the NPCM estimator with the corresponding bootstrap bandwidth,

$$\mathrm{MISE}_x(h_j^*, g_j^*) = E\left(\int\limits_{I_T} \left(\widetilde{PD}_{h_j^*, g_j^*}(t|x) - PD(t|x)\right)^2 dt\right),$$

and its square root, $\mathrm{RMISE}_x(h_j^*, g_j^*)$, are approximated via Monte Carlo using 300 simulated samples. The mean of these estimation errors given by

Table 5
MISE and average bootstrap bandwidths and estimation errors of Beran's and the NPCM estimators of $PD(t|x)$ for Models 1 and 2 when $x = 0.2$.

		Beran	NPCM
Model 1	$h/(h,g)$	0.5224	(0.9265, 0.8714)
	RMISE_x	0.1351	0.1349
	$\overline{h^*}/(\overline{h^*},\overline{g^*})$	0.1282	(0.2036, 0.1944)
	$\overline{\text{RMISE}_x}$	0.2065	0.2180
Model 2	$h/(h,g)$	0.1082	(0.1276, 0.3755)
	RMISE_x	0.0890	0.0766
	$\overline{h^*}/(\overline{h^*},\overline{g^*})$	0.13756	(0.3103, 0.3119)
	$\overline{\text{RMISE}_x}$	0.1011	0.1020

Table 6
Computation time (in seconds) for the estimation of PD$(t|x)$ in time grid of size 100 and $x = 0.5$ for one sample of size n with Beran's estimator, the NPCM estimator, the PHCM estimator and the AFTCM estimator.

Sample size	$n = 100$	$n = 400$	$n = 800$	$n = 1600$	$n = 2400$
Beran	0.02	0.03	0.03	0.04	0.04
NPCM	0.02	0.02	0.02	0.02	0.02
PHCM	0.24	0.40	0.43	1.39	2.49
AFTCM	0.42	1.61	6.12	39.57	82.96

Table 7
Computation time (in minutes) for the approximation of the bootstrap bandwidths for $N = 1$ sample of size n using $B = 100$ bootstrap resamples to estimate PD$(t|x)$ in time grid of size 100 and $x = 0.5$ with Beran's estimator and the NPCM estimator.

Sample size	$n = 100$	$n = 400$	$n = 800$	$n = 1600$	$n = 2400$
Beran	2.28	4.53	20.37	156.48	455.90
NPCM	2.40	4.07	5.06	12.90	28.05

$$\overline{\text{RMISE}_x(h^*, g^*)} = \frac{1}{N}\sum_{j=1}^{N} \text{RMISE}_x\left(h_j^*, g_j^*\right)$$

is used as a measure of the estimation error made by the bootstrap two-dimensional bandwidth in the method.

In addition, for each model, the estimation error function of Beran's estimator given in (16) is approximated via Monte Carlo using 300 simulated samples. The bandwidth that minimises $\text{MISE}_x(h)$ is obtained and denoted by h_{MISE}. The value of the mean integrated squared error made by h_{MISE} and denoted by $\text{MISE}_x(h_{\text{MISE}})$ is computed.

In the simulation study, $N = 300$ simulated samples are used. For each sample, $B = 500$ bootstrap resamples are generated by using the resampling algorithm presented in Peláez et al. (2022) to approximate the bootstrap MISE function, $\text{MISE}_x^*(h)$ by the expression

$$\text{MISE}_x^*(h) \simeq \frac{1}{B}\sum_{k=1}^{B}\left(\int_{I_T}\left(\widehat{\text{PD}}_h^{*,k}(t|x) - \widehat{\text{PD}}_r(t|x)\right)^2 dt\right)$$

and obtain the bootstrap bandwidth associated to each simulated sample, h_j^*, $j = 1, 2, \ldots, N$. In addition, the estimation error of Beran's estimator with the corresponding bootstrap bandwidth, $\text{MISE}_x(h_j^*)$, is computed for each sample. The mean of the square root of these estimation errors, $\overline{\text{RMISE}_x(h^*)}$ is also considered for the comparison. The results are shown also in Table 5.

From the results shown in Table 5 it can be extrapolated that the observed differences between Beran's and the NPCM estimators of the probability of default are attenuated by using bootstrap bandwidths.

Since computational cost is an important aspect to be considered in the comparison of several estimators, a small study of the computation time is addressed in this section. Table 6 shows the computation times in seconds needed to estimate the PD for a single sample of different sizes with the four studied estimators. Table 7 shows the computation times in minutes needed to approximate the bootstrap bandwidths to estimate the PD for one simulated sample of different sizes with Beran's estimator and the NPCM estimator using $B = 100$ bootstrap resamples. The estimators based on PH cure model and AFT cure model do not depend on any smoothing parameter.

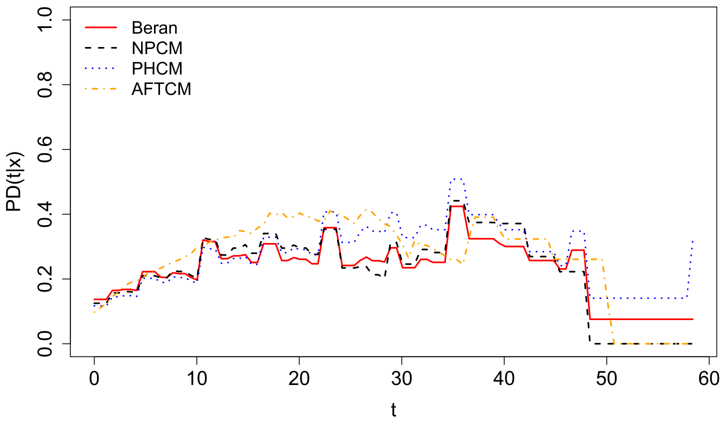

Fig. 2. $\mathrm{PD}(t|x = 0.85)$ estimated by Beran's estimator with bootstrap bandwidths (solid line), NPCM estimator with bootstrap bandwidths (dashed line), PHCM estimator (dotted line) and AFTCM estimator (dash-dotted line) for the German credit data.

According to Table 6, the NPCM estimator is the fastest of the four studied estimators. The NPCM estimator and Beran's estimator are barely affected by the increase in the sample size. Given the definitions of Beran's and the NPCM estimators, the differences in their computational costs are due to programming efficiency. The implementation of the NPCM estimator in the npcure package is based on the use of C++. The semiparametric methods are slower; in particular, the AFTCM estimator. It is the use of the EM algorithm to estimate the curve by the semiparametric methods that makes them slower. Nonparametric methods do not rely on the EM algorithm. However, the optimal bandwidth approximation is what slows down nonparametric methods as opposed to semiparametric methods, which do not depend on bandwidth parameters, as can be seen in Table 7. When analysing the times shown in Table 7 it is important to mention that the computation time increases linearly as the value of the number of resamples, B, increases.

An important advantage of the NPCM estimator over Beran's estimator is its computational efficiency. Both the estimator and its automatic bandwidth selector are less sensitive to the increase of the sample size than Beran's estimator, which leads to significantly shorter computational times.

6. Application to real data

In this section we apply the above PD estimators to the German Credit data set which is publicly available on the webpage http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) and was previously analysed in Strzalkowska-Kominiak and Cao (2013). This data set includes information of 1000 credits with a censoring ratio of 70.7%. The duration of the credits in months (Z) is available along with the amount of the credit in DM (X_1), the amount of money in the checking account in thousands of Deutsche Marks (X_2), the savings amount in thousands of Deutsche Marks (X_3) and years of employment (X_4). Let the credit scoring be denoted by $X = X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4$. Since some of the original covariates are ordinal (interval) variables, they are changed into numerical variables by following the criteria explained in Strzalkowska-Kominiak and Cao (2013) and the single-index method proposed there is used to estimate $(1, \theta_2, \theta_3, \theta_4)$, obtaining $X = X_1 + 3.2091X_2 + 0.2312X_3 + 2.1891X_4$. A distinction is made between credits for which default is observed and those that are censored. Censored credits correspond to cured credits that will never run into arrears, credits cancelled in advance or credits susceptible to default if the follow-up of the credit would be longer enough. The probability of default conditional on the credit scoring is estimated using the four estimators presented in this paper and the result is shown in Fig. 2. The estimations of these curves are obtained at $x = 0.85$. The default horizon b is 1 year, approximately 20% of the time range in the sample. The bandwidths involved in the estimators are chosen by automatic bootstrap selectors. The Beran estimation is computed with $h^* = 0.5000$ obtained by the selector proposed in Peláez et al. (2022). The bandwidths for the NPCM estimator are provided by the selector referred to in Section 2, $(h^*, g^*) = (0.3163, 0.0108)$.

7. Conclusion

A nonparametric estimator of the probability of default is proposed in this paper. This estimator takes into account the existence of a group of cured individuals who will never experience the default. It is based on the nonparametric survival estimator for mixture cure models proposed by López-Cheda (2018). The asymptotic bias and variance and the asymptotic normality of the NPCM probability of default estimator are proved. The simulation study carried out shows that the NPCM estimator is a very reasonable choice for estimating the probability of default, since it provides smaller estimation errors than classical methods, even in semiparametric models. The good behaviour of Beran's estimator, which was also included in the comparative study as another nonparametric method, is remarkable. Work is currently underway to reduce the computation times of the bandwidth selectors for the smoothing parameters involved in the above-mentioned estimators. Using cure models when the cure status is partially known is an appealing idea to be considered for future work. A nonparametric view along the lines similar to Safari et al. (2020) can be used.

Acknowledgements

This research has been supported by MICINN Grant PID2020-113578RB-100, by the Xunta de Galicia (Grupo de Referencia Competitiva ED431C-2020-14 and Centro Singular de Investigación de Galicia ED431G 2019/01), all of them through the ERDF and by the European Research Council (2016-2022, Horizon 2020 / ERC grant agreement No. 694409). Peláez, R. was sponsored by inMOTION Programme of grants for pre-doctoral stays Inditex-UDC 2021.

Appendix A. Assumptions

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 $H(t) = P(Z \le t)$ 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 and C are conditionally independent given $X = x$.

(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 $x \in I_c$, $0 \le l_{H(\cdot|x)}$, $0 \le u_{H(\cdot|x)} < \infty$.

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

A.4. Let $G(t) = P(C \le t)$ be the distribution function of C and $G(t|x)$ be the conditional distribution function of $C|X = x$. Let $\tau_G(x) = \sup\{t : G(t|x) < 1\}$, $\tau_{S_0}(x) = \sup\{t : S_0(t|x) > 0\}$ and $\tau_0 = \sup\{\tau_{S_0}(x) : x \in I\}$, then, $\tau_0 < \tau_G(x), \quad \forall x \in I$.

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

A.6. The first and second derivatives of m, $m'(x)$ and $m''(x)$, respectively, exist and are continuous on I_c.

A.7. Let $H_1(t|x)$ be the conditional subdistribution function of $Z|X = x$ when $\delta = 1$. The first derivatives with respect to t of the functions $S_0(t|x)$, $G(t|x)$, $H(t|x)$ and $H_1(t|x)$, i.e. $S_0'(t|x)$, $G'(t|x)$, $H'(t|x)$ and $H_1'(t|x)$ exist and are continuous on $[l, u] \times I_c$.

A.8. 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 on $[l, u] \times I_c$.

A.9. 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 on $[l, u] \times I_c$.

A.10. The functions $S_0(t|x)$, $H(t|x)$ and $G(t|x)$ have bounded second-order derivatives with respect to $x \in I_c$ given any value of $t \in [l, u]$.

A.11. The density function of T, $f(t)$ is bounded away from 0 on $[l, u]$.

A.12. $\displaystyle\int_0^\infty \frac{dH_1(t|x)}{\left(1 - H(t|x)\right)^2} < \infty \quad \forall x \in I.$

A.13. The kernel, K, is a symmetric, continuous and differentiable density function with compact support $[-1, 1]$ and the total variation of K is less than some $\lambda < \infty$.

A.14. The smoothing parameter $h = h_n$ satisfies $h \to 0$, $\dfrac{nh^5}{\ln n} = O(1)$ and $\dfrac{(\ln n)^3}{nh} \to 0$.

A.15. The smoothing parameter $g = g_n$ satisfies $g \to 0$, $\dfrac{ng^5}{\ln n} = O(1)$ and $\dfrac{(\ln n)^3}{ng} \to 0$.

A.16. Let $(t, x) \in [l, u] \times I_c$. The second derivative of $m(u)$ exists at $u = x$. The second derivative of $\Phi(u, t, x)$ exists at (x, t, x) and (x, ∞, x). The second derivative of $\Phi_2(u, t, x)$ exists at (x, t, x) and (x, ∞, x). The second derivative of $D(u, t_1, t_2, x)$ exists at $(x, t, t + b, x)$, (x, t, ∞, x) and (x, ∞, t, x). The second derivative of $L(u, t_1, t_2, x)$ exists at (x, t, ∞, x) and (x, ∞, t, x).

Appendix B. Proofs

Lemma 2. *Denote $\Phi(u, t, x) = E\left[\xi(Z, \delta, t, x)|X = u\right]$ with $\xi(Z, \delta, t, x)$ defined in Section 3. Under Assumptions A.13 and A.16, then*

$$E\left[K\left(\frac{x - X_1}{h}\right)\xi(Z_1, \delta_1, t, x)\right] = \frac{1}{2}h^3 \frac{\partial^2}{\partial u^2}\left(\Phi(u, t, x)m(u)\right)|_{u=x} + o(h^3).$$

Proof. Using a Taylor expansion for $\Phi(u,t,x)m(u)$ when $u = x - hv$ around $u = x$ and Assumption A.13:

$$E\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right] = \int_{-\infty}^{+\infty} K\left(\frac{x-u}{h}\right)\Phi(u,t,x)m(u)du$$
$$= \Phi(x,t,x)m(x)h + \frac{d_K}{2}\frac{\partial^2}{\partial u^2}\big(\Phi(u,t,x)m(u)\big)\big|_{u=x}h^3 + o(h^3).$$

Moreover, $\Phi(x,t,x) = 0 \quad \forall (t,x) \in [0,\infty) \times I$, since

$$\Phi(u,t,x) = E\big[\xi(Z,\delta,t,x)|X = u\big] = \int_0^t \frac{dH_1(z|u)}{1-H(z|x)} - \int_0^t \frac{1-H(v|u)}{\big(1-H(v|x)\big)^2}dH_1(v|x). \quad \square$$

Lemma 3. *Denote $\Phi_2(u,t,x) = E\big[\xi^2(Z,\delta,t,x)|X = u\big]$ with $\xi(Z,\delta,t,x)$ defined in Section 3. Under Assumptions A.13 and A.16, then*

$$Var\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right] = h\Phi_2(x,\infty,x)m(x)c_K$$
$$+h^3\frac{d_{K^2}}{2}\frac{\partial^2}{\partial u^2}\big(\Phi_2(u,\infty,x)m(u)\big)|_{u=x} + o(h^3).$$

Proof. First,

$$Var\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right]$$
$$= E\left[K^2\left(\frac{x-X_1}{h}\right)\xi^2(Z_1,\delta_1,t,x)\right] - E\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right]^2.$$

Using a Taylor expansion for $\Phi_2(u,t,x)m(u)$ when $u = x - hv$ around $u = x$ and Assumption A.13:

$$E\left[K^2\left(\frac{x-X_1}{h}\right)\xi^2(Z_1,\delta_1,t,x)\right] = \int_{-\infty}^{+\infty} K^2\left(\frac{x-u}{h}\right)\Phi_2(u,t,x)m(u)du$$
$$= c_K\Phi_2(x,t,x)m(x)h + \frac{d_{K^2}}{2}\frac{\partial^2}{\partial u^2}\big(\Phi_2(u,t,x)m(u)\big)\big|_{u=x}h^3 + o(h^3).$$

From Lemma 2, $E\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right]^2 = O(h^6)$. Then,

$$Var\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right]$$
$$= c_K\Phi_2(x,t,x)m(x)h + \frac{d_{K^2}}{2}\frac{\partial^2}{\partial u^2}\big(\Phi_2(u,t,x)m(u)\big)\big|_{u=x}h^3 + o(h^3). \quad \square$$

Lemma 4. *Denote $D(u,t_1,t_2,x) = Cov\big[\xi(Z_1,\delta_1,t_1,x),\xi(Z_1,\delta_1,t_2,x)|X_1 = u\big]$ and $B(u,t_1,t_2,x) = \Phi(u,t_1,x)\Phi(u,t_2,x)m(u)$. Under Assumptions A.13 and A.16, then*

$$Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_2,x)\right]$$
$$= c_K D(x,t_1,t_2,x)h + \frac{d_{K^2}}{2}\big(D''(x,t_1,t_2,x) + B''(x,t_1,t_2,x)\big)h^3 + o(h^3).$$

Proof. Using the Law of total covariance,

$$\begin{aligned} &Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_2,x)\right] \\ &\quad= E\left[Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_2,x)\big|X_1\right]\right] \\ &\quad\quad+E\left[K^2\left(\frac{x-X_1}{h}\right)\Phi(X_1,t_1,x)\Phi(X_1,t_2,x)\right] \\ &\quad\quad-E\left[K\left(\frac{x-X_1}{h}\right)\Phi(X_1,t_1,x)\right]E\left[K\left(\frac{x-X_1}{h}\right)\Phi(X_1,t_2,x)\right] = S_1+S_2-S_3. \end{aligned} \tag{B.1}$$

Using a Taylor expansion for $D(u,t_1,t_2,x)m(u)$ when $u=x-hv$ around $u=x$ and Assumption A.13:

$$S_1 = c_K D(x,t_1,t_2,x)h + \frac{d_{K^2}}{2}D''(x,t_1,t_2,x)h^3 + o(h^3).$$

Using a Taylor expansion for $B(u,t_1,t_2,x)$ when $u=x-hv$ around $u=x$ and Assumption A.13 and considering that $B(x,t_1,t_2,x)=0$ for all $t_1,t_2\in[0,\infty)$, since $\Phi(x,t,x)=0 \quad \forall(t,x)\in[0,\infty)\times I$:

$$S_2 = \frac{d_{K^2}}{2}B''(x,t_1,t_2,x)h^3 + o(h^3).$$

Finally, from Lemma 2, $E\left[K\left(\frac{x-X_1}{h}\right)\Phi(X_1,t,x)\right] = O(h^3)$. Then, $S_3 = O(h^6)$, and replacing S_1, S_2 and S_3 in (B.1), the lemma is proved. □

Proof of Lemma 1. Let us denote $\widehat{S}_{h,g}(t|x) := \widehat{S}^{\text{NPCM}}_{h,g}(t|x)$. According to the definition of the NPCM estimator in (4),

$$\begin{aligned} \widehat{S}_{h,g}(t|x) - S(t|x) &= 1-\widehat{p}_h(x)+\widehat{p}_h(x)\widehat{S}_{0,g}(t|x) - \Big(1-p(x)+p(x)S_0(t|x)\Big) \\ &= \big(S_0(t|x)-1\big)\big(\widehat{p}_h(x)-p(x)\big) + p(x)\big(\widehat{S}_{0,g}(t|x)-S_0(t|x)\big) \\ &\quad+\big(\widehat{p}_h(x)-p(x)\big)\big(\widehat{S}_{0,g}(t|x)-S_0(t|x)\big). \end{aligned} \tag{B.2}$$

From Theorem 3 in López-Cheda et al. (2017b) and Theorem 1 in López-Cheda et al. (2017a), the almost sure representations of the incidence and the latency nonparametric estimators are available:

$$\widehat{p}_h(x) - p(x) = \big(p(x)-1\big)\sum_{i=1}^n w^A_{h,i}(x)\xi(Z_i,\delta_i,\infty,x) + R_n(x), \tag{B.3}$$

$$\widehat{S}_{0,g}(t|x) - S_0(t|x) = \sum_{i=1}^n w^A_{g,i}(x)\eta(Z_i,\delta_i,t,x) + R_n(t|x), \tag{B.4}$$

with

$$\sup_{x\in I}|R_n(x)| = O\left(\frac{\ln n}{nh}\right)^{3/4} \quad \text{a.s.} \quad \text{and} \quad \sup_{(t,x)\in[l,u]\times I}|R_n(t|x)| = O\left(\frac{\ln n}{ng}\right)^{3/4} \quad \text{a.s.}$$

Replacing (B.3) and (B.4) in (B.2), the almost sure representation of the NPCM survival estimator is as follows:

$$\begin{aligned} &\widehat{S}_{h,g}(t|x) - S(t|x) = \\ &\quad= \big(S_0(t|x)-1\big)\big(p(x)-1\big)\sum_{i=1}^n w^A_{h,i}(x)\xi(Z_i,\delta_i,\infty,x) + p(x)\sum_{i=1}^n w^A_{g,i}(x)\eta(Z_i,\delta_i,t,x) \\ &\quad\quad+\big(S_0(t|x)-1\big)R_n(x) + p(x)R_n(t|x) + \big(\widehat{p}_h(x)-p(x)\big)\big(\widehat{S}_{0,g}(t|x)-S_0(t|x)\big). \end{aligned}$$

From Theorem 3 in López-Cheda et al. (2017b) and Theorem 3 in López-Cheda et al. (2017a), it follows that

$$\widehat{p}_h(x) - p(x) = O_p\left(\frac{1}{\sqrt{nh}}\right), \qquad \widehat{S}_{0,g}(t|x) - S_0(t|x) = O_p\left(\frac{1}{\sqrt{ng}}\right).$$

Then,

$$\big(\widehat{p}_h(x)-p(x)\big)\big(\widehat{S}_{0,g}(t|x)-S_0(t|x)\big) = O_p\left(\frac{1}{n\sqrt{hg}}\right)$$

and

$$\widehat{S}_{h,g}(t|x)-S(t|x)=\big(S_0(t|x)-1\big)\big(p(x)-1\big)\sum_{i=1}^{n}w_{h,i}^{A}(x)\xi(Z_i,\delta_i,\infty,x)$$
$$+p(x)\sum_{i=1}^{n}w_{g,i}^{A}(x)\eta(Z_i,\delta_i,t,x)+R_n^1(t|x),$$

where

$$R_n^1(t|x)=\big(S_0(t|x)-1\big)R_n(x)+p(x)R_n(t|x)+O_p\left(\frac{1}{n\sqrt{hg}}\right)$$
$$=O_p\left(\ln n\left(\frac{1}{nh}+\frac{1}{ng}\right)\right)^{3/4}. \quad \square$$

Proof of Theorem 1. Let us denote $\widehat{\mathrm{PD}}_{h,g}(t|x):=\widehat{\mathrm{PD}}_{h,g}^{\mathrm{NPCM}}(t|x)$ and $\widehat{S}_{h,g}(t|x):=\widehat{S}_{h,g}^{\mathrm{NPCM}}(t|x)$. Consider the function

$$W_{h,g}(t,t+b,x)=\frac{S(t|x)\big(\widehat{S}_{h,g}(t+b|x)-S(t+b|x)\big)-S(t+b|x)\big(\widehat{S}_{h,g}(t|x)-S(t|x)\big)}{\widehat{S}_{h,g}(t|x)S(t|x)}.$$

Since

$$\frac{\widehat{S}_{h,g}(t+b|x)}{\widehat{S}_{h,g}(t|x)}-\frac{S(t+b|x)}{S(t|x)}=-\big(\widehat{\mathrm{PD}}_{h,g}(t|x)-\mathrm{PD}(t|x)\big)$$

and

$$\frac{\widehat{S}_{h,g}(t+b|x)}{\widehat{S}_{h,g}(t|x)}-\frac{S(t+b|x)}{S(t|x)}=$$
$$=\frac{S(t|x)\big(\widehat{S}_{h,g}(t+b|x)-S(t+b|x)\big)-S(t+b|x)\big(\widehat{S}_{h,g}(t|x)-S(t|x)\big)}{\widehat{S}_{h,g}(t|x)S(t|x)}$$
$$=W_{h,g}(t,t+b,x)\left(\frac{\widehat{S}_{h,g}(t|x)}{S(t|x)}+1-\frac{\widehat{S}_{h,g}(t|x)}{S(t|x)}\right)$$
$$=\frac{1}{S(t|x)}\big(\widehat{S}_{h,g}(t+b|x)-S(t+b|x)\big)-\frac{S(t+b|x)}{S^2(t|x)}\big(\widehat{S}_{h,g}(t|x)-S(t|x)\big)$$
$$+W_{h,g}(t,t+b,x)\left(1-\frac{\widehat{S}_{h,g}(t|x)}{S(t|x)}\right),$$

we have

$$\widehat{\mathrm{PD}}_{h,g}(t|x)-\mathrm{PD}(t|x)=a_1\big(\widehat{S}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\big(\widehat{S}_{h,g}(t|x)-S(t|x)\big)$$
$$+W_{h,g}(t,t+b,x)\left(\frac{\widehat{S}_{h,g}(t|x)-S(t|x)}{S(t|x)}\right), \tag{B.5}$$

with $a_1=-\dfrac{1}{S(t|x)}$ and $a_2=\dfrac{S(t+b|x)}{S^2(t|x)}$.

Using the almost sure representation of $\widehat{S}_{h,g}(t+b|x)$ from Lemma 1 in (B.5) and considering the functions $\varphi_{n,i}(t|x)$ defined in the statement of Theorem 1, the almost sure representation of $\widehat{\mathrm{PD}}_{h,g}(t|x)$ is as follows:

$$\begin{aligned}\widehat{\mathrm{PD}}_{h,g}(t|x)-\mathrm{PD}(t|x)&=a_1\textstyle\sum_{i=1}^{n}\varphi_{n,i}(t+b|x)+a_2\sum_{i=1}^{n}\varphi_{n,i}(t|x)+R_n^2(t|x)\\&=\textstyle\sum_{i=1}^{n}\Psi_{n,i}(t,x)+R_n^2(t|x),\end{aligned} \tag{B.6}$$

where $\Psi_{n,i}(t,x)=a_1\varphi_{n,i}(t+b|x)+a_2\varphi_{n,i}(t|x)$ are independent and identically distributed for all $i=1,\dots,n$ and

$$R_n^2(t|x)=-\frac{1}{S(t|x)}R_n^1(t+b|x)+\frac{S(t+b|x)}{S^2(t|x)}R_n^1(t|x)+W_{h,g}(t,t+b,x)\left(\frac{\widehat{S}_{h,g}(t|x)-S(t|x)}{S(t|x)}\right).$$

From Equation (7) in Lemma 1, we have $\widehat{S}_{h,g}(t|x)-S(t|x)=\tau_1+\tau_2+\tau_3$ where

$$\tau_1=\big(S_0(t|x)-1\big)\big(p(x)-1\big)\sum_{i=1}^{n}w_{h,i}^{A}(x)\xi(Z_i,\delta_i,\infty,x),$$

$$\tau_2 = p(x)\sum_{i=1}^{n} w_{g,i}^{A}(x)\eta(Z_i,\delta_i,t,x),$$

$$\tau_3 = O_p\left(\ln n\left(\frac{1}{nh}+\frac{1}{ng}\right)\right)^{3/4}.$$

Lemmas 2 and 3 and straightforward but tedious calculations give $\tau_1 = O_p\left(h^2+\frac{1}{\sqrt{nh}}\right)$ and $\tau_2 = O_p\left(g^2+\frac{1}{\sqrt{ng}}\right)$. Since $\frac{nh}{(\ln n)^3}\to\infty$ and $\frac{ng}{(\ln n)^3}\to\infty$, τ_3 is negligible with respect to τ_1 and τ_2. Then,

$$W_{h,g}(t,t+b,x)\left(\frac{\widehat{S}_{h,g}(t|x)-S(t|x)}{S(t|x)}\right) = O_p\left(h^4+g^4+\frac{1}{nh}+\frac{1}{ng}\right).$$

Therefore,

$$R_n^2(t|x) = O_p\left(\ln n\left(\frac{1}{nh}+\frac{1}{ng}\right)\right)^{3/4} + O_p\left(h^4+g^4+\frac{1}{nh}+\frac{1}{ng}\right).$$

Using Assumptions A.14 and A.15, the second term in $R_n^2(t|x)$ is negligible with respect to $O_p\left(\ln n\left(\frac{1}{nh}+\frac{1}{ng}\right)\right)^{3/4}$ and Theorem 1 is proved. □

Proof of Theorem 2. According to the almost sure representation of $\widehat{\text{PD}}_{h,g}(t|x) := \widehat{\text{PD}}_{h,g}^{\text{NPCM}}(t|x)$, the asymptotic expression of the bias is obtained from its dominant term. Then,

$$\begin{aligned} E\left[\sum_{i=1}^{n}\Psi_{n,i}(t,x)\right] &= \sum_{i=1}^{n} E\left[\Psi_{n,i}(t,x)\right] = nE\left[\Psi_{n,1}(t,x)\right] \\ &= na_1E\left[\varphi_{n,1}(t+b,x)\right]+na_2E\left[\varphi_{n,1}(t,x)\right], \end{aligned} \tag{B.7}$$

with $a_1 = -\frac{1}{S(t|x)}$ and $a_2 = \frac{S(t+b|x)}{S^2(t|x)}$.

The expression of $E\left[\varphi_{n,1}(t,x)\right]$ in (B.7) is then calculated using Lemmas 2 and 3:

$$\begin{aligned} E\left[\varphi_{n,1}(t,x)\right] &= \left(S_0(t|x)-1\right)\left(p(x)-1\right)E\left[w_{h,1}^{A}(x)\xi(Z_1,\delta_1,\infty,x)\right] \\ &\quad + p(x)E\left[w_{g,i}^{A}(x)\eta(Z_1,\delta_1,t,x)\right] \\ &= B_1(t,x)\frac{h^2}{n}+B_2(t,x)\frac{g^2}{n}+o\left(\frac{h^2}{n}\right)+o\left(\frac{g^2}{n}\right). \end{aligned} \tag{B.8}$$

Replacing the expression (B.8) in (B.7), the bias part of the theorem is proved:

$$E\left[\sum_{i=1}^{n}\Psi_{n,i}(t,x)\right] = \widetilde{B}_1(t,x)h^2+\widetilde{B}_2(t,x)g^2+o(h^2)+o(g^2),$$

where $\widetilde{B}_1(t,x)$ and $\widetilde{B}_1(t,x)$ were defined in Section 3.

The asymptotic expression of the variance of $\widehat{\text{PD}}_{h,g}(t|x)$ is obtained from the variance of the dominant term of its almost sure representation:

$$\begin{aligned} Var\left[\sum_{i=1}^{n}\Psi_{n,i}(t,x)\right] &= \sum_{i=1}^{n} Var\left[\Psi_{n,1}(t,x)\right] = nVar\left[\Psi_{n,1}(t,x)\right] \\ &= na_1^2Var\left[\varphi_{n,1}(t+b,x)\right]+na_2^2Var\left[\varphi_{n,1}(t,x)\right] \\ &\quad +2na_1a_2Cov\left[\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\right]. \end{aligned} \tag{B.9}$$

To find the asymptotic expression of $Cov\left[\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\right]$,

$$
\begin{aligned}
&Cov\big[\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big]\\
&\quad= \big(S_0(t_1|x)-1\big)\big(S_0(t_2|x)-1\big)\big(p(x)-1\big)^2\frac{1}{n^2h^2m^2(x)}A_1\\
&\qquad+\big(S_0(t_1|x)-1\big)\big(p(x)-1\big)p(x)\frac{1}{n^2hgm^2(x)}A_2\\
&\qquad+\big(S_0(t_2|x)-1\big)\big(p(x)-1\big)p(x)\frac{1}{n^2hgm^2(x)}A_3+p^2(x)\frac{1}{n^2g^2m^2(x)}A_4.
\end{aligned}
\tag{B.10}
$$

First, from Lemma 3,

$$
A_1 = Var\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,\infty,x)\right] = h\Phi_2(x,\infty,x)m(x)c_K + O(h^3).
\tag{B.11}
$$

Second, using Lemmas 3 and 4,

$$
\begin{aligned}
A_4 &= Cov\left[K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right]\\
&= C_1(t_1,t_2,x)g+O(g^3).
\end{aligned}
\tag{B.12}
$$

In order to obtain asymptotic expressions of A_2 and A_3, an asymptotic expression for

$$
Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right]
$$

is obtained by distinguishing three different cases:

(i) If $C_{h,g} := \lim_{n\to\infty}\frac{h}{g}\in(0,\infty)$:

$$
\begin{aligned}
&Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right]\\
&\quad\simeq\ Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{h/C_{h,g}}\right)\eta(Z_1,\delta_1,t_2,x)\right]\\
&\quad=\ E\left[Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{h/C_{h,g}}\right)\eta(Z_1,\delta_1,t_2,x)\middle|X_1\right]\right]\\
&\qquad+E\left[K\left(\frac{x-u}{h}\right)K\left(C_{h,g}\frac{x-u}{h}\right)\Phi(X_1,t_1,x)\Phi_\eta(X_1,t_2,x)\right]\\
&\qquad-E\left[K\left(\frac{x-X_1}{h}\right)\Phi(X_1,t_1,x)\right]E\left[K\left(C_{h,g}\frac{x-u}{h}\right)\Phi_\eta(X_1,t_2,x)\right]\\
&\quad=\ S_1+S_2-S_3.
\end{aligned}
$$

Considering the function $L(u,t_1,t_2,x)$ and its Taylor expansion when $u=x-hv$ around $u=x$:

$$
\begin{aligned}
S_1 &= \int_{-\infty}^{+\infty}K\left(\frac{x-u}{h}\right)K\left(C_{h,g}\frac{x-u}{h}\right)L(u,t_1,t_2,x)du\\
&= h\int_{-\infty}^{+\infty}K(v)K(C_{h,g}v)\Big(L(x,t_1,t_2,x)-hvL'(x,t_1,t_2,x)+O(h^2)\Big)dv.
\end{aligned}
$$

Since K is symmetric, $K(C_{h,g}v)=K(-C_{h,g}v)$ and the function $K(v)K(C_{h,g}v)$ is also even. Consequently, $\int_{-\infty}^{+\infty}K(v)\times K(C_{h,g}v)vdv=0$. Then,

$$
S_1=\widetilde{c}_K(C_{h,g})L(x,t_1,t_2,x)h+O(h^3).
\tag{B.13}
$$

Defining $B_\eta(u,t_1,t_2,x)=\Phi(u,t_1,x)\Phi_\eta(u,t_2,x)m(u)$ and using a Taylor expansion for $B_\eta(u,t_1,t_2,x)$ when $u=x-hv$ around $u=x$ and considering that $B_\eta(x,t_1,t_2,x)=0$ for all $t_1,t_2\in[0,\infty)$, $x\in I$, since $\Phi(x,t,x)=0$ for all $(t,x)\in[0,\infty)\times I$:

$$\begin{aligned} S_2 &= \int_{-\infty}^{+\infty} K\left(\frac{x-u}{h}\right) K\left(C_{h,g}\frac{x-u}{h}\right)\Phi(u,t_1,x)\Phi_\eta(u,t_2,x)m(u)du \\ &= \widetilde{c}_K(C_{h,g})B_\eta(x,t_1,t_2,x)h + O(h^3) = O(h^3). \end{aligned} \tag{B.14}$$

From Lemma 2, $E\left[K\left(\frac{x-X_1}{h}\right)\Phi(X_1,t,x)\right] = O(h^3)$.
Now, using a Taylor expansion for $\Phi_\eta(u,t,x)m(u)$ when $u = x - hv$ around $u = x$,

$$E\left[K\left(C_{h,g}\frac{x-X_1}{h}\right)\Phi_\eta(X_1,t,x)\right] = \left(\int_{-\infty}^{+\infty} K\left(C_{h,g}v\right)dv\right)\Phi_\eta(x,t,x)m(x)h + O(h^3).$$

Considering the definition of the function $\eta(Z,\delta,t,x)$ given in Section 3 and Lemma 2, $\Phi_\eta(x,t,x) = 0$ for all $(t,x) \in [0,\infty) \times I$ and $E\left[K\left(C_{h,g}\frac{x-X_1}{h}\right)\Phi_\eta(X_1,t,x)\right] = O(h^3)$. Therefore,

$$S_3 = O(h^6). \tag{B.15}$$

Using the expressions of S_1 in (B.13), S_2 in (B.14) and S_3 in (B.15),

$$\begin{aligned} &Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right] \\ &= \widetilde{c}_K(C_{h,g})L(x,t_1,t_2,x)h + O(h^3). \end{aligned}$$

Therefore,

$$\begin{aligned} A_2 &= Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,\infty,x), K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right] \\ &= \widetilde{c}_K(C_{h,g})L(x,\infty,t_2,x)h + O(h^3) \end{aligned} \tag{B.16}$$

and

$$\begin{aligned} A_3 &= Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,\infty,x)\right] \\ &= \widetilde{c}_K(C_{h,g})L(x,t_1,\infty,x)h + O(h^3). \end{aligned} \tag{B.17}$$

Replacing (B.11), (B.12), (B.16) and (B.17) in (B.10) and assuming $\lim_{n\to\infty}\frac{h}{g} = C_{h,g}$, we have
$Cov\left[\varphi_{n,1}(t_1,x), \varphi_{n,1}(t_2,x)\right]$

$$\begin{aligned} &= \frac{\left(S_0(t_1|x)-1\right)\left(S_0(t_2|x)-1\right)\left(p(x)-1\right)^2}{m(x)} c_K \Phi_2(x,\infty,x)\frac{1}{n^2h} \\ &+ C_{h,g}\widetilde{c}_K(C_{h,g})\frac{\left(S_0(t_1|x)-1\right)\left(p(x)-1\right)p(x)}{m^2(x)}L(x,\infty,t_2,x)\frac{1}{n^2h} \\ &+ C_{h,g}\widetilde{c}_K(C_{h,g})\frac{\left(S_0(t_2|x)-1\right)\left(p(x)-1\right)p(x)}{m^2(x)}L(x,t_1,\infty,x)\frac{1}{n^2h} \\ &+ C_{h,g}\frac{p^2(x)C_1(t_1,t_2,x)}{m^2(x)}\frac{1}{n^2h} + o\left(\frac{1}{n^2h}\right) + O\left(\frac{h}{n^2}\right). \end{aligned}$$

Considering the functions V_1, V_2 and V_3, defined in Section 3:
$Cov\left[\varphi_{n,1}(t_1,x), \varphi_{n,1}(t_2,x)\right]$

$$\begin{aligned} &= \left(V_1(t_1,t_2,x) + C_{h,g}V_2(t_1,t_2,x) + C_{h,g}\widetilde{c}_K(C_{h,g})V_3(t_1,t_2,x)\right)\frac{1}{n^2h} \\ &+ o\left(\frac{1}{n^2h}\right) + O\left(\frac{h}{n^2}\right). \end{aligned} \tag{B.18}$$

Using Equation (B.18) with $t_1 = t_2 = t + b$ and $t_1 = t_2 = t$, the expressions of $Var\left[\varphi_{n,1}(t+b,x)\right]$ and $Var\left[\varphi_{n,1}(t,x)\right]$ are also available. Therefore, Case (i) of the Theorem is proved by replacing (B.18) in (B.9):

$$
\begin{aligned}
&Var\Big[\textstyle\sum_{i=1}^n \Psi_{n,i}(t,x)\Big]\\
&\quad= \big(\widetilde{V}_1(t+b,t,x) + C_{h,g}\widetilde{V}_2(t+b,t,x) + C_{h,g}\widetilde{c}_K(C_{h,g})\widetilde{V}_3(t+b,t,x)\big)\frac{1}{nh}\\
&\quad\quad +o\left(\frac{1}{nh}\right) + O\left(\frac{h}{n}\right).
\end{aligned}
$$

(ii) If $\lim_{n\to\infty}\dfrac{h}{g}=0$:
From Lemma 3 and Equation (B.12) when $t_1=t_2$, we have

$$
\begin{aligned}
&Var\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x)\right] = hc_K\Phi_2(x,t_1,x)m(x) + O(h^3),\\
&Var\left[K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right] = C_1(t_2,t_2,x)g + O(g^3).
\end{aligned}
$$

Then, using the Cauchy–Schwarz inequality:

$$
\begin{aligned}
&Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{g}\right)\eta(Z_1,\delta_1,t_2,x)\right]\\
&\quad\le \sqrt{hgc_K\Phi_2(x,t_1,x)m(x)C_1(t_2,t_2,x) + O(hg^3) + O(gh^3)}.
\end{aligned}
\tag{B.19}
$$

Therefore,

$$
A_2 = O\big((hg)^{1/2}\big), \quad A_3 = O\big((hg)^{1/2}\big).
\tag{B.20}
$$

Plugging (B.11), (B.12) and (B.20) in (B.10), we have

$$
\begin{aligned}
&Cov\big[\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big]\\
&\quad= \frac{\big(S_0(t_1|x)-1\big)\big(S_0(t_2|x)-1\big)\big(p(x)-1\big)^2}{m(x)}c_K\Phi_2(x,\infty,x)\frac{1}{n^2h}\\
&\quad\quad + \frac{p^2(x)C_1(t_1,t_2,x)}{m^2(x)}\frac{1}{n^2g} + O\left(\frac{h}{n^2}\right) + O\left(\frac{g}{n^2}\right) + O\left(\frac{\sqrt{hg}}{n^2hg}\right).
\end{aligned}
\tag{B.21}
$$

Assuming $\lim_{n\to\infty}\dfrac{h}{g}=0$ and considering the function $V_1(t_1,t_2,x)$, we have

$$
Cov\big[\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big] = V_1(t_1,t_2,x) + o\left(\frac{1}{n^2h}\right) + O\left(\frac{g}{n^2}\right).
\tag{B.22}
$$

Using the expression of $Cov\big[\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big]$ in (B.22) with $t_1=t_2=t+b$ and $t_1=t_2=t$, the expressions of $Var\big[\varphi_{n,1}(t+b,x)\big]$ and $Var\big[\varphi_{n,1}(t,x)\big]$ are also available. Therefore, Case (ii) of the Theorem is proved by replacing (B.22) in (B.9):

$$
Var\Big[\textstyle\sum_{i=1}^n \Psi_{n,i}(t,x)\Big] = \widetilde{V}_1(t+b,t,x)\dfrac{1}{nh} + o\left(\dfrac{1}{nh}\right) + O\left(\dfrac{g}{n}\right).
$$

(iii) From Equation (B.21) and assuming that $\lim_{n\to\infty} g/h = 0$, we have

$$
Cov\big[\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big] = V_2(t_1,t_2,x)\frac{1}{n^2g} + o\left(\frac{1}{n^2g}\right) + O\left(\frac{h}{n^2}\right).
\tag{B.23}
$$

Considering the expression of $Cov\big[\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big]$ in (B.23) with $t_1=t_2=t+b$ and $t_1=t_2=t$, the expressions of $Var\big[\varphi_{n,1}(t+b,x)\big]$ and $Var\big[\varphi_{n,1}(t,x)\big]$ are also available. Therefore, Case (iii) of the Theorem is proved by replacing (B.23) in (B.9):

$$
Var\Big[\textstyle\sum_{i=1}^n \Psi_{n,i}(t,x)\Big] = \widetilde{V}_2(t+b,t,x)\dfrac{1}{ng} + o\left(\dfrac{1}{ng}\right) + O\left(\dfrac{h}{n}\right). \quad \square
$$

Proof of Theorem 3. Denote $\widehat{\mathrm{PD}}_{h,g}(t|x) := \widehat{\mathrm{PD}}_{h,g}^{\mathrm{NPCM}}(t|x)$.

(ii) From Equation (B.6) in the proof of Lemma 1 we have

$$\sqrt{nh}\big(\widehat{\mathrm{PD}}_{h,g}(t|x)-\mathrm{PD}(t|x)\big)=\sqrt{nh}\sum_{i=1}^{n}\Psi_{n,i}(t,x)+\widetilde{R}_n^2(t|x), \tag{B.24}$$

where $\Psi_{n,i}(t,x)=a_1\varphi_{n,i}(t+b|x)+a_2\varphi_{n,i}(t|x)$ with $a_1=-\dfrac{1}{S(t|x)}$, $a_2=\dfrac{S(t+b|x)}{S^2(t|x)}$ and $\widetilde{R}_n^2(t|x)=\sqrt{nh}R_n^2(t|x)$. The variables $\Psi_{n,i}(t,x)$ are independent and identically distributed for all $i=1,\dots,n$.

From Theorem 3 in López-Cheda et al. (2017b) and Theorem 1 and Theorem 3 in López-Cheda et al. (2017a) and assuming $\lim_{n\to\infty}\dfrac{h}{g}\in(0,\infty)$, it follows that

$$\begin{aligned}\widetilde{R}_n^2(t|x)=\sqrt{nh}R_n^2(t|x)&=\sqrt{nh}O_P\left(\frac{\ln n}{nh}\right)^{3/4}+\sqrt{nh}O_P\left(\frac{\ln n}{ng}\right)^{3/4}\\&+\sqrt{nh}O_P\left(h^4+g^4+\frac{1}{nh}+\frac{1}{ng}\right).\end{aligned}$$

Under the assumptions of Theorem 3, $\dfrac{(\ln n)^3}{nh}\to 0$, $\left(\dfrac{\ln n}{ng}\right)^{3/4}(nh)^{1/2}\to 0$ and $nh\to\infty$, the remainder term $\widetilde{R}_n^2(t|x)$ is negligible with respect to the dominant term of (B.24).

On the other hand, from Case (i) of Theorem 2 and Equation (B.24), the variance of the dominant term is finite, since it is given by:

$Var\big[\sqrt{nh}\sum_{i=1}^{n}\Psi_{n,i}(t,x)\big]$

$$\begin{aligned}&=nh\Big(\widetilde{V}_1(t+b,t,x)+C_{h,g}\widetilde{V}_2(t+b,t,x)+C_{h,g}\widetilde{c}_K(C_{h,g})\widetilde{V}_3(t+b,t,x)\Big)\frac{1}{nh}\\&+nh\,o\left(\frac{1}{nh}\right)+nh\,O\left(\frac{h}{n}\right)=O(1).\end{aligned}$$

Therefore, the asymptotic distribution of $\sqrt{nh}\big(\widehat{\mathrm{PD}}_{h,g}(t|x)-\mathrm{PD}(t|x)\big)$ is the same as the asymptotic distribution of $\sqrt{nh}\sum_{i=1}^{n}\Psi_{n,i}(t,x)$. If Lindeberg's condition for triangular arrays (see Theorem 7.2 in Billingsley (1968)) is satisfied, then

$$\sum_{i=1}^{n}\Big(\sqrt{nh}\Psi_{n,i}(t,x)-E\big[\sqrt{nh}\Psi_{n,i}(t,x)\big]\Big)\xrightarrow{d}N(0,s), \tag{B.25}$$

where

$$s^2=\widetilde{V}_1(t+b,t,x)+C_{h,g}\widetilde{V}_2(t+b,t,x)+C_{h,g}\widetilde{c}_K(C_{h,g})\widetilde{V}_3(t+b,t,x).$$

Lindeberg's condition is now checked. It is given by

$$\lim_{n\to\infty}\frac{1}{s^2}E\left[\sum_{i=1}^{n}\Big(\sqrt{nh}\Psi_{n,i}(t,x)-E\big[\sqrt{nh}\Psi_{n,i}(t,x)\big]\Big)^2\mathbb{1}_{n,i}\right]=0 \tag{B.26}$$

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

$$\mathbb{1}_{n,i}=\mathbb{1}\Big(\big|\sqrt{nh}\Psi_{n,i}(t,x)-E[\sqrt{nh}\Psi_{n,i}(t,x)]\big|>\varepsilon s\Big).$$

One can define

$$\begin{aligned}\zeta_{n,i}(t,x)=\big(S_0(t|x)-1\big)\big(p(x)-1\big)\frac{K\big((x-X_i)/h\big)}{m(x)}\xi(Z_i,\delta_i,\infty,x)\\+p(x)\frac{K\big((x-X_i)/g\big)}{m(x)}\eta(Z_i,\delta_i,t,x).\end{aligned}$$

Then,

$$\Psi_{n,i}(t,x)=\frac{1}{nh}\chi_{n,i}(t,x),$$

where $\chi_{n,i}(t,x)=\left(-\dfrac{1}{S(t|x)}\zeta_{n,i}(t+b,x)+\dfrac{S(t+b|x)}{S^2(t|x)}\zeta_{n,i}(t,x)\right)$, which leads to

$$\mathbb{1}_{n,i} = \mathbb{1}\left(\left|\frac{1}{\sqrt{nh}}\chi_{n,i}(t,x) - E\left[\frac{1}{\sqrt{nh}}\chi_{n,i}(t,x)\right]\right| > \varepsilon s\right). \tag{B.27}$$

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

$$|\xi(Z,\delta,t,x)| \le \frac{1}{\theta} + \int_0^t \frac{dH_1(u|x)}{\theta^2} \le \frac{1}{\theta} + \frac{H(t|x)}{\theta^2} \le \frac{1}{\theta} + \frac{1}{\theta^2}$$

and, consequently, η is also bounded:

$$|\eta(Z,\delta,t,x)| \le \frac{S(t|x)}{p(x)}\left(\frac{1}{\theta}+\frac{1}{\theta^2}\right) + \frac{\big(1-p(x)\big)\big(1-S(t|x)\big)}{p^2(x)}\left(\frac{1}{\theta}+\frac{1}{\theta^2}\right).$$

On the one hand, $nh \to \infty$, then, $1/\sqrt{nh} \to 0$. On the other hand, η was proved to be bounded and K and m have compact support, according to Assumptions A.2 and A.13. Therefore, there exists $n_0 \in \mathbb{N}$ such that for all $i = 1, \ldots, n$, $\mathbb{1}_{n,i} = 0$ for all $n \ge n_0$ with $\mathbb{1}_{n,i}$ defined in (B.27). Consequently,

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

which proves Lindeberg's condition in (B.26).
Finally, assuming $h = C_h n^{-1/5}$ and $g = C_g n^{-1/5}$ and considering Equation (8), we have

$$\sqrt{nh}\sum_{i=1}^{n}\Psi_{n,i}(t,x) \xrightarrow{d} N(\mu, s),$$

where $\mu = C_h^{5/2}\widetilde{B}_1(t,x) + C_g^{5/2}\widetilde{B}_2(t,x)$.

(ii) Considering again (B.24), under the assumptions of Case (ii) in Theorem 3 and following the argument of the previous case, the remainder term $\widetilde{R}_n^2(t|x)$ is found to be negligible with respect to the dominant term in (B.24). Furthermore, the variance of this dominant term is finite, since, from the proof of Theorem 2,

$$Var\big[\sqrt{nh}\textstyle\sum_{i=1}^{n}\Psi_{n,i}(t,x)\big] = nh\left(\widetilde{V}_1(t+b,t,x)\dfrac{1}{nh} + o\left(\dfrac{1}{nh}\right) + O\left(\dfrac{h}{n}\right)\right) = O(1).$$

Therefore, the asymptotic distribution of $\sqrt{nh}\big(\widehat{\mathrm{PD}}_{h,g}(t|x) - \mathrm{PD}(t|x)\big)$ is the same as the asymptotic distribution of $\sqrt{nh}\sum_{i=1}^{n}\Psi_{n,i}(t,x)$. If Lindeberg's condition given in (B.26) is satisfied, then

$$\sum_{i=1}^{n}\left(\sqrt{nh}\Psi_{n,i}(t,x) - E\big[\sqrt{nh}\Psi_{n,i}(t,x)\big]\right) \xrightarrow{d} N(0,s), \tag{B.28}$$

where $s^2 = \widetilde{V}_1(t+b,t,x)$.
Lindeberg's condition is proved here following the same argument shown in the first case. Finally, assuming $g = C_g n^{-1/5}$ and $n^{1/5}h \to 0$ and considering Equation (8),

$$\sqrt{nh}\sum_{i=1}^{n}\Psi_{n,i}(t,x) \xrightarrow{d} N(\mu, s),$$

where $\mu = C_g^{5/2}\widetilde{B}_2(t,x)$.

(iii) Assuming $C_h := \lim_{n\to\infty} n^{1/5}h \in (0,\infty)$ and $\lim_{n\to\infty} n^{1/5}g = 0$:
Considering again (B.24), under the assumptions of Case (iii) in Theorem 3 and following the argument of the first case, the remainder term $\widetilde{R}_n^2(t|x)$ is found to be negligible with respect to the dominant term in (B.24). Furthermore, the variance of this dominant term is finite, since, from the proof of Theorem 2,

$$Var\big[\sqrt{ng}\textstyle\sum_{i=1}^{n}\Psi_{n,i}(t,x)\big] = ng\left(\widetilde{V}_2(t+b,t,x)\dfrac{1}{ng} + o\left(\dfrac{1}{ng}\right) + O\left(\dfrac{h}{n}\right)\right) = O(1).$$

Therefore, the asymptotic distribution of $\sqrt{ng}\big(\widehat{\mathrm{PD}}_{h,g}(t|x) - \mathrm{PD}(t|x)\big)$ is the same as the asymptotic distribution of $\sqrt{ng}\sum_{i=1}^{n}\Psi_{n,i}(t,x)$. If Lindeberg's condition given in (B.26) is satisfied, then

$$\sum_{i=1}^{n}\left(\sqrt{ng}\Psi_{n,i}(t,x)-E\left[\sqrt{ng}\Psi_{n,i}(t,x)\right]\right)\xrightarrow{d}N(0,s), \tag{B.29}$$

where $s^2=\widetilde{V}_2(t+b,t,x)$.

Lindeberg's condition is proved here following the same arguments used in the first case. Finally, assuming $h=C_h n^{-1/5}$ and $n^{1/5}g\to 0$ and considering Equation (8), we have

$$\sqrt{ng}\sum_{i=1}^{n}\Psi_{n,i}(t,x)\xrightarrow{d}N(\mu,s),$$

where $\mu=C_h^{5/2}\widetilde{B}_1(t,x)$. □

References

Allen, L.N., Rose, L.C., 2006. Financial survival analysis of defaulted debtors. J. Oper. Res. Soc. 57 (6), 630–636.

Amico, M., Van Keilegom, I., 2018. Cure models in survival analysis. Annu. Rev. Stat. Appl. 5 (1), 311–342.

Baba, N., Goko, H., 2006. Survival analysis of hedge funds. Bank of Japan Working Paper Series, 6-E-05.

Beran, J., Djaïdja, A., 2007. Credit risk modeling based on survival analysis with immunes. Stat. Methodol. 4 (3), 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, vol. 9. John Wiley and Sons, New York.

Cai, C., Zou, Y., Peng, Y., Zhang, J., 2012. An R-package for estimating semiparametric mixture cure models. https://cran.r-project.org/web/packages/smcure/smcure.pdf.

Cao, R., 1993. Bootstrapping the mean integrated squared error. J. Multivar. Anal. 45 (1), 137–160.

Cao, R., Vilar, J.M., Devia, A., 2009. Modelling consumer credit risk via survival analysis (with discussion). SORT 33 (1), 3–30.

Dabrowska, D.M., 1989. Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann. Stat. 17 (3), 1157–1167.

Dirick, L., Bellotti, T., Claeskens, G., Baesens, B., 2019. Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. J. Bus. Econ. Stat. 37 (1), 40–53.

Dirick, L., Claeskens, G., Baesens, B., 2003. Time to default in credit scoring using survival analysis: a benchmark study. J. Oper. Res. Soc. 68 (6), 652–665.

Dirick, L., Claeskens, G., Baesens, B., 2015. An Akaike information criterion for multiple event mixture cure models. Eur. J. Oper. Res. 241 (2), 449–457.

Glennon, D., Nigro, P., 2005. Measuring the default risk of small business loans: a survival analysis approach. J. Money Credit Bank. 37 (5), 923–947.

Iglesias-Pérez, M.C., González-Manteiga, W., 1999. Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. J. Nonparametr. Stat. 10 (3), 213–244.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53 (282), 457–481.

Li, G., Datta, S., 2001. A bootstrap approach to nonparametric regression for right censored data. Inst. Stat. Math. 53 (4), 708–729.

López-Cheda, A., 2018. Nonparametric inference in mixture cure models. PhD thesis. University of Coruña.

López-Cheda, A., Cao, R., Jácome, M.A., 2017a. Nonparametric latency estimation for mixture cure models. Test 26 (2), 353–376.

López-Cheda, A., Cao, R., Jácome, M.A., Van Keilegom, I., 2017b. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Comput. Stat. Data Anal. 105 (12), 144–165.

López-de Ullibarri, I., López-Cheda, A., Jácome, M.A., 2020. Nonparametric estimation in mixture cure models. https://cran.r-project.org/web/packages/npcure/npcure.pdf.

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

Parsa, M., Van Keilegom, I., 2022. Accelerated failure time vs Cox proportional hazards mixture cure models: David vs Goliath? Stat. Pap., 1–21.

Peláez, R., Cao, R., Vilar, J.M., 2021a. Nonparametric estimation of probability of default with double smoothing. SORT 45 (2), 93–120.

Peláez, R., Cao, R., Vilar, J.M., 2021b. Probability of default estimation in credit risk using a nonparametric approach. Test 30 (2), 383–405.

Peláez, R., Cao, R., Vilar, J.M., 2022. Bootstrap bandwidth selection and confidence regions for double smoothed default probability estimation. Mathematics 10 (9), 1523.

Peng, Y., Yu, B., 2021. Cure Models: Methods, Applications and Implementation. Chapman and Hall/CRC Biostatistics Series.

Roszbach, K., 2003. Bank lending policy, credit scoring and the survival of loans. Rev. Econ. Stat. 86 (4), 946–958.

Safari, W.C., López-de Ullibarri, I., Jácome, M.A., 2020. A product-limit estimator of the conditional survival function when cure status is partially known. Biom. J. 63 (5), 984–1005.

Stepanova, M., Thomas, L., 2002. Survival analysis methods for personal loan data. Oper. Res. 50 (2), 277–289.

Strzalkowska-Kominiak, E., Cao, R., 2013. Maximum likelihood estimation for conditional distribution single-index models under censoring. J. Multivar. Anal. 114 (7), 74–98.

Sy, J.P., Taylor, J.M.G., 2000. Estimation in a Cox proportional hazards cure model. Biometrics 56 (1), 227–236.

Sy, J.P., Taylor, J.M.G., 2001. Standard errors for the Cox proportional hazards cure model. Math. Comput. Model. 33 (12), 1237–1251.

Xu, J., Peng, Y., 2014. Nonparametric cure rate estimation with covariates. Can. J. Stat. 42 (1), 1–17.