Nonparametric estimation of the probability of default in credit risk

Rebeca Peláez Suárez

Doctoral Thesis UDC/2022

Advisors: Ricardo Cao, Juan M. Vilar

Doctoral Programme in Statistics and Operations Research

Nonparametric estimation of the probability of default in credit risk

Rebeca Peláez Suárez

Doctoral Thesis UDC/2022

Advisors: Ricardo Cao, Juan M. Vilar

Doctoral Programme in Statistics and Operations Research

Department of Mathematics

Faculty of Computer Science

The public defense of the Doctoral Thesis entitled "Nonparametric estimation of the probability of default in credit risk", developed by Rebeca Peláez Suárez and supervised by Dr. Ricardo Cao Abad and Juan M. Vilar Fernández, was held on _____, at the Faculty of Computer Sciences of A Coruña University, with the examining committee: Dr. José Antonio Vilar Fernández (President), Dr. Montserrat Guillén Estany (Secretary) and Dr. Stefan Sperlich (Board member).

A Coruña, _____.

PhD Committee:

José Antonio Vilar Fernández Montserrat Guillén Estany

Stefan Sperlich

Advisors:

Ricardo Cao Abad Juan M. Vilar Fernández

PhD candidate:

Rebeca Peláez Suárez

"No es una poesía gota a gota pensada.
No es un bello producto. No es un fruto perfecto."
Gabriel Celaya, "Cantos íberos", 1955.

Agradecimientos

Juan, Ricardo, sin vosotros este día no habría llegado. Gracias, Ricardo, por contagiar tu amor por la investigación. Gracias, Juan, por tu sabiduría y tus consejos. Ambos me disteis la oportunidad de empezar esta aventura y me habéis guiado a lo largo de ella. No puedo pensar en nadie mejor de quien aprender y a quien admirar.

Ingrid, me gustaría agradecerte tu cálida acogida en Leuven, tu supervisión durante los meses que estuve allí y todas tus aportaciones a la elaboración de esta tesis. *Ingrid, I would like to thank you for your warm welcome in Leuven, for your supervision during the months I was there and for all your contributions to the preparation of this thesis.* También quiero agradeceros a los miembros del tribunal de seminario de tesis, Antonio Vaamonde, Alberto Rodríguez-Casal y José Antonio Vilar, vuestro apoyo a lo largo del proceso y vuestras sugerencias para la versión final de esta memoria.

Gracias a aquellos que me motivaron a dar el primer paso de esta aventura. Y me remonto, además, al primero de todos, Roberto, mi profesor de estadística en bachillerato, por animarme a estudiar matemáticas y por ver, antes que nadie, que el mundo de la investigación podía ser para mí.

Gracias a los que han estado a mi lado durante estos años, especialmente a ti, Sandra, pilar fundamental desde que nos conocimos. Gracias también a los que he tenido la suerte de encontrarme en el camino del doctorado, a mis compañeros de laboratorio y a mis incansables compañeros de congresos.

Y, por supuesto, "tengo un hogar del que nadie podría echarme; una guarida a buen recaudo de paredes indestructibles". El esfuerzo, la dedicación y las ganas se quedan cortas sin un hogar hecho de personas increíbles que te acompañan a lo largo de todo el camino.

Funding

The PhD candidate's research was sponsored by the Spanish Grant for Predoctoral Research Trainees RD 103/2019. The work has been partially carried out during a stay at KU Leuven (Belgium), which was supported by inMOTION Programme of grants for pre-doctoral stays Inditex-UDC 2021.

This research has also been partially supported by MINECO Grants MTM2017-82724-R and PID2020-113578RB-100, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and ED431C-2020-14 and Centro Singular de Investigación de Galicia ED431G/01 and ED431G 2019/01), all of them through the ERDF.

Abstract

Financial institutions are interested in knowing the probability that their clients declare themselves unable to pay the debts incurred by granting a credit. The aim of this work is to propose models to estimate this probability, called probability of default (PD), using the information provided by the credit scoring.

The PD conditional on the credit scoring can be written as a transformation of the conditional survival function of the variable "time to default". This property is used to propose new PD estimators, based on nonparametric estimators of the survival function. The time to default faces a right-censoring problem, since in the study of a set of loans, it is not possible to observe default for all of them. Consequently, censored data techniques and survival analysis are used. Given the possible existence of individuals not susceptible to default, mixture cure models are also discussed in this work.

The asymptotic expression for the mean squared error and the asymptotic normality of the proposed estimators are obtained. Automatic bootstrap selectors are proposed for the smoothing parameters on which the estimators depend. The performance of the proposed techniques is analysed and compared with existing semiparametric approaches through simulation studies and illustrated by analysing bank loan data.

Resumen

A las entidades financieras les interesa conocer la probabilidad de que sus clientes se declaren incapaces de hacer frente a las deudas contraídas con la concesión de un crédito. El objetivo de este trabajo es proponer modelos para estimar esta probabilidad, denominada probabilidad de mora (PD), utilizando la información proporcionada por la puntuación crediticia.

La PD condicionada a la puntuación crediticia puede escribirse como una transformación de la función de supervivencia condicional de la variable "tiempo hasta la mora". Esta propiedad se utiliza para proponer nuevos estimadores de la PD, basados en estimadores no paramétricos de la función de supervivencia. El tiempo hasta el impago se enfrenta a un problema de censura por la derecha, pues en el estudio de un conjunto de créditos, no es posible observar la mora para todos ellos. Consecuentemente, se utilizan técnicas de datos censurados y análisis de supervivencia. Ante la posible existencia de individuos no susceptibles a la mora, los modelos de curación de tipo mixtura también se discuten en este trabajo.

Se obtiene la expresión asintótica para el error cuadrático medio y la normalidad asintótica de los estimadores propuestos. Se proponen selectores automáticos bootstrap para los parámetros de suavizado de los que dependen los estimadores. El comportamiento de las técnicas propuestas se analiza y se compara con enfoques semiparamétricos existentes mediante estudios de simulación y se ilustra mediante el análisis de datos de préstamos bancarios.

Resumo

Ás entidades financeiras interesalles coñecer a probabilidade de que os seus clientes declárense incapaces de facer fronte ás débedas contraídas coa concesión dun crédito. O obxectivo deste traballo é propoñer modelos para estimar esta probabilidade, denominada probabilidade de morosidade (PD), utilizando a información proporcionada pola puntuación crediticia.

A PD condicionada á puntuación crediticia pode escribirse como unha transformación da función de supervivencia condicional da variable "tempo ata a morosidade". Esta propiedade utilízase para propoñer novos estimadores da PD, baseados en estimadores non paramétricos da función de supervivencia. O tempo ata a falta de pagamento enfróntase a un problema de censura pola dereita, pois no estudo dun conxunto de créditos, non é posible observar a falta de pagamento para todos eles. Consecuentemente, utilízanse técnicas de datos censurados e análise de supervivencia. Ante a posible existencia de individuos non susceptibles á morosidade, os modelos de curación de tipo mestura tamén se discuten neste traballo.

Obtense a expresión asintótica para o erro cuadrático medio e a normalidade asintótica dos estimadores propostos. Propóñense selectores automáticos bootstrap para os parámetros de suavizado dos que dependen os estimadores. O comportamento das técnicas propostas analízase e compárase con enfoques semiparamétricos existentes mediante estudos de simulación e ilústrase mediante a análise de datos de préstamos bancarios.

Contents

Introduction

Credit risk is the possibility of economic loss arising from the default on obligations assumed by the counterparties of a contract due to insolvency or inability to pay. The concept is associated with financial institutions and banks, but may be extended to companies, financial markets and organisations in other sectors. The granting of credits, both individual and corporate, is one of the main activities of banks and financial institutions, so there is a clear interest in preventing financial loss as a consequence of the debtor defaulting on debt and interest payments. Therefore, banks often require certain guarantees or impose additional clauses depending on the client's risk profile. For example, they may charge higher interest rates to riskier customers or impose a debt limit on companies to which they have granted a credit.

The financial crisis that began to surface in 2007 and then spread to the world economy as a whole is a clear example of the fundamental role played by risk management in the financial sector. The so-called subprime or junk mortgages that had been granted in the United States in the preceding years were the initial focal point of the crisis. These loans were granted mainly for the purchase of housing, with high bank fees. They were granted to people with insufficient solvency and, therefore, with a higher level of risk of default than the average for other loans. Debtors accepted the conditions hoping that house prices would continue to increase and that they would be able to refinance the initial mortgage on better terms. When interest rates began to rise and, as a consequence, house prices began to go down, subprime defaults began to spread and many institutions did not have enough reserves to cope with a liquidity crisis.

This mortgage crisis resulted in a number of financial collapses, bank nationalisations, interventions by the central banks of the main developed economies, a fall in stock market prices and a worldwide economic downturn. It is considered to have triggered the Great Recession internationally, including the Spanish property bubble.

The reasons why financial institutions would have offered and granted these subprime mortgages are complex, but root of this crisis, and similar crises throughout history, lies in inadequate regulation and imperfect supervision. Avoiding recessions of this type therefore requires accurate risk measurement and adequate supervision as the starting point.

The Basel Committee on Banking Supervision is a benchmark in this field, as it has been responsible for the prudential regulation of banks and their solvency worldwide since the 1980s (Basel Committee on Banking Supervision (1999, 2001b,a, 2004, 2005b,a)). The committee was established in 1975 by the central bank presidents of the eleven member countries of the Group of Ten (G10) at that time. It is currently formed by the G10 countries together with Luxembourg and Spain. The recommendations on banking regulation issued by this committee from its birth to date are collected in the so-called Basel Accords: Basel I, Basel II and Basel III.

The constant internationalisation of banking, the globalisation of the financial markets of the most advanced countries and the lack of financial regulation led banks from different countries to compete with each other under different operating rules and often insufficient solvency levels, increasingly endangering financial stability. In view of this situation, in 1988 in Basel, Switzerland, the first Basel Committee agreement is created. Basel I demands a minimum level of capital from banks depending on the risks they face. The accord establishes the concept of "regulatory capital" as the minimum amount of own funds that a financial institution must have to cover credit, market and exchange rate risks in order to cope with losses arising from potential defaults, without collapsing. This agreement was created as a recommendation and the signatory countries were free to implement it. However, it came into force in more than 100 countries.

In order to determine how much liquidity is needed to cover potential losses, a good assessment of the risk faced by the banking system is necessary. Financial institutions are subject to different types of risk depending on their economic activity (Pyle (1997); Bessis (2002); Saunders and Cornett (2008)); some of them are credit risk, market risk, operational risk, interest rate risk, liquidity risk, foreign exchange risk and sovereign risk, with credit risk being one of the most relevant. As long as Basel I was in force, regulatory capital was calculated on the basis of risk weights provided by the relevant central banks. Specifically, the accord required own funds to be greater than 8% of risk assets, covering credit risk, trading risk and foreign exchange risk. The limitation of this agreement was therefore to assume that all credits have the same probability of default, ignoring the credit quality of the different borrowers.

Throughout the years, different risk measurement and management techniques have been used and have progressed towards more sophisticated statistical-financial approaches. Furthermore, technological advances have allowed credit risk to be calculated in ever more precise ways. An example of this is the second Basel accord, Basel II, which is proposed in 2004 with the purpose of establishing a more refined credit risk-sensitive regulatory capital concept than Basel I. This arrangement makes the use of new techniques indispensable in various aspects: from the estimation of a client's solvency in the granting of credit up to the use of complex statistical models for the calculation of the risk associated with investments. Several historical precedents (Girón (1998); Bank for International Settlements (2009); Cecchetti et al. (2009)) have proved, however, that the industry has not always made a good use of statistical models. This is the reason why the new Basel Accord also addresses the process of supervision and transparency.

Although Basel I and Basel II focus in one way or another on the capital that banks must hold to avoid failure in the event of significant losses, they do not consider the possibility of a "bank run". A bank run occurs when large groups of depositors simultaneously withdraw their money from banks based on fears that the institution will become insolvent and their deposits or savings will not be covered by the bank's

capital. The Basel III reform, introduced in 2010 and triggered by the 2007 financial crisis, recognises the problem of bank panic and requires different levels of capital for different types of bank deposits and other loans. Basel III does not replace the guidelines established in Basel I and Basel II, but complements them.

Basel II implements different mechanisms to measure the quality of borrowers and allows banks to use internal metrics to evaluate their risk. The Basel II framework operates under three pillars:

Pillar 1: Capital adequacy requirements

This pillar sets out how credit risk is measured and how regulatory capital is calculated.

Pillar 2: Supervisory review

National supervisory institutions must validate the methods used to estimate the parameters required under Pillar 1, as well as the sufficiency of own funds levels to cope with an economic crisis. These supervisory authorities, typically the corresponding central banks, have the competence to increase the level of prudence required of the banks under their jurisdiction.

Pillar 3: Market discipline

The agreement establishes transparency rules, demanding the regular publication of information about their exposure to different risks and the adequacy of their own funds.

The so-called Pillar 1 is the core of the accord and, so far, is the main way in which financial institutions manage credit risk. Regulatory capital under Pillar 1 of Basel II must exceed 8% of assets at risk, taking into account credit, market, foreign exchange and operational risks. Credit risk is defined as the expected losses due to counterparty default in a transaction and these depend on a number of variables, known as risk factors. However, the expected loss (EL) can be considered to depend essentially on whether the counterparty defaults or not, the amount at risk at the

time of default and the recovery that can be obtained in the event of default. The Pillar 1 credit risk loss model is therefore a multiplicative relationship between these three basic factors: probability of default (PD), exposure at default (EAD) and loss given default (LGD). This relationship can be expressed by the following equation:

$$EL = PD \times EAD \times LGD$$

The parameters involved in this formula are further described in the following paragraphs.

The probability of default (PD) is the probability that a customer, after a certain period of time from the formalisation of his contract with the bank, will declare himself unable to pay the credit he enjoys. In practice, a contract is considered to be in default if it is more than 90 days overdue. This probability depends on the credit score, a rating that the bank assigns to customers or prospective customers with the intention of assessing their ability to pay the potential debt they may acquire with the bank through a loan. Credit scoring is therefore a way to quantify the creditworthiness of the client.

The portion of the debt that is exposed to the risk of loss when the default occurs is called exposure at default (EAD). This is the maximum loss that can be incurred and is a priori unknown at the time of default. For example, the exposure of a derivative will depend on the value of several market factors; the EAD of a credit card will depend on the extent of the clients's drawdown at the time of default. The proportion of the debt that the institution eventually expects to lose once the borrower defaults on its contractual obligations, i.e. the percentage of EAD that is not expected to be recoverable, is the lost given default (LGD).

Estimation of the probability of default for each client is a key element in this credit risk model and is, indeed, the subject of this study.

Widely used default probability estimation techniques include logistic regression, discriminant analysis and Cox proportional hazards models, among others. A simple approach adopted by many banks is using external rating agencies to estimate PDs based on historical default experience. Logistic regression based on a historical series

of defaults is the technique commonly used for small companies. In retail defaults, credit scoring is often used as a euphemism for the probability of default, which is the true objective of the lender.

In 1992, Naraim published his work "Survival analysis and the credit granting decision". There, he advocated the use of survival analysis in the context of credit risk. Specifically, he argued for the possibility of analysing all credit transactions involving predictor variables in which time to the occurrence of an event is the variable of interest by means of survival analysis.

In our context, the variable of interest to which Naraim refers is the time to default. This variable is not fully observable, since 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. The credit scoring plays the role of the predictor variable. The existing analogy pointed out by Naraim (see Naraim (1992)), between the "time to default" and the "time to the event of interest", which is common in biometric models, is obvious. In this analogy lies the motivation for applying survival analysis techniques to credit risk problems.

The approach of Naraim (1992) was further explored by Banasik et al. (1999) and abundant literature has since been developed using survival analysis in credit risk. To name a few papers, survival analysis makes it possible to obtain confidence intervals for the probability of default in Hanson and Schuermann (2004); the time to default distribution function is estimated using a hazard model in Glennon and Nigro (2005) and the Kaplan-Meier estimator is used to estimate the time to default survival function in Allen and Rose (2006).

In Naraim (1992), the proposal is a Cox proportional hazard model to estimate the conditional survival function $S(t|x)$. Cao et al. (2009) start from this and, writing the probability of default in terms of conditional survival function, they get an estimator of the PD. A second alternative given in Cao et al. (2009) is to assume a generalized linear model for the lifetime distribution under censoring. These approaches use parametric or semiparametric models for the time to default

as a function of the credit scoring and the credit lifetime. The use of nonparametric curve estimation for this purpose is however convenient. These are flexible methods that use only the information that the data provide without making assumptions about the shape of the curve. This is already the approach used in the third model in Cao et al. (2009) who proposed to estimate the probability of default using Beran's estimator for the conditional survival function.

Following this research line, in this thesis we propose nonparametric survival models that allow estimating the probability of default or PD conditionally on the credit scoring for personal credits. We work with credits requested by individuals (Basel Committee on Banking Supervision (2001b)) such as personal credits, consumer loans, credit cards or mortgage loans. The latter are not specifically dealt with in this thesis, but the results obtained here could be extended to mortgage loans following the ideas of Beran and Djaïdja (2007).

The event of interest to be observed is the fall into default, which is partially determined by the covariate X, denoting the credit scoring. 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, $PD(t|x)$, is allowed to depend on this scoring x. A common aproach in credit scoring is using logistic regression to build the index. The logistic model for credit scoring has been studied by Wiginton (1980), Srinivasan and Kim (1987), Steenackers and Goovaerts (1989), Thomas et al. (1992) and Samreen et al. (2013), among others. Therefore, throughout this study, credit scoring is assumed to be known and one-dimensional.

It is only possible to know the time it takes for a client to default, T, when the default happens during the credits follow-up time; otherwise, the data is censored and the observed time is the time to censoring, C. Figure 1 illustrates the censoring mechanism that can affect the lifetimes of a portfolio of personal loans. Assuming that the observation time is the interval $[0, \tau]$, at the end of the study, there are three possible situations:

(a) The credit is defaulted: The time instant at which the default occurs is within the interval $[0, \tau]$ and can therefore be observed. In this case, $T \leq C$ and the lifetime of the credit is uncensored.

(b) The credit is active and being paid: The credit has not yet defaulted when the observation period ends. The default, if it occurs, cannot be observed. In this case, $T > C$ and the lifetime of the credit is right-censored.

(c) The credit is not being followed up: The credit has either been paid off or cancelled before the end of the study. In either case, no default is observed. Then, $T > C$ and the lifetime of the credit is considered to be a right-censored data.

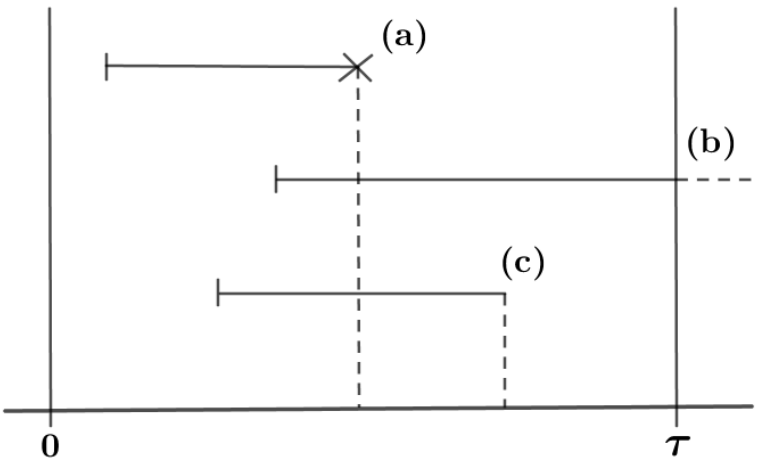

Figure 1: Mechanism for censoring the life of a credit when the follow-up period is $[0, \tau]$.

In this scenario, the available information is a simple random sample of the triplet (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 of the credit and $\delta = I(T \leq C)$ is the uncensoring indicator. The distribution function of T is denoted by $F(t)$ and its survival function by $S(t)$.

Let $M(x)$ and $m(x)$ be the distribution and density functions of the covariate X. It is assumed that an unknown relationship between T and X exists. Let x be a fixed value of the covariate X (typically, the scoring) 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

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

It is worth mentioning that the function defined in (1) is a relevant measure in other fields appart from the financial one. This curve is important in many other contexts where survival analysis can be used to quantify the probability that the event of interest occurs not much later than t, given that it didn't happen before t. For example, companies that provide energy services, streaming services, 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 $t + b$.

This thesis addresses the problem of estimating the probability of default under the conditions presented above.

Chapter 1 is devoted to introduce the reader some methodological tools that are needed: survival analysis, nonparametric curve estimation and bandwidth selection based on bootstrap.

In Chapter 2, four nonparametric estimators of the probability of default in credit risk are proposed and compared. They are derived from estimators of the conditional survival function for censored data. Asymptotic expressions for the bias and the variance of these probability of default estimators are derived from similar properties for the conditional survival function estimators. A simulation study shows the performance of these four estimators. Finally, an empirical study,

based on modified real data, illustrates their practical behaviour.

In Chapter 3, a conditional survival function estimator for censored data is studied. It is based on a double smoothing technique: both the covariate and the time variable are smoothed. Asymptotic expressions for the bias and the variance and the asymptotic normality of the smoothed survival estimator derived from Beran's estimator are found. A simulation study shows the performance of the smoothed Beran's estimator of the conditional survival function and compares it with the smoothed one only in the covariate. The influence of the smoothing parameters involved in both estimators is also studied.

In Chapter 4, resampling methods are used to solve two problems related to the Beran's estimator and the double smoothed Beran's estimator of the conditional survival function for censored data. The bootstrap is used for the automatic choice of the necessary smoothing parameter in the computation of Beran's estimator and the two smoothing parameters used in the smoothed Beran's estimator. Bootstrap algorithms for the calculation of confidence regions of the conditional survival function for censored data are proposed. Extensive simulation studies show the good behavior of the proposed bootstrap algorithms. Finally, the proposed techniques are used to estimate the survival function of the time that COVID-19 patients remain hospitalized in ward or in the Intensive Care Unit.

A doubly smoothed estimator of the probability of default is proposed and studied in Chapter 5. It is derived from the doubly smoothed estimator of the conditional survival function proposed in Chapter 3. Asymptotic expressions for the bias and the variance of the probability of default estimator are found and asymptotic normality is proved. A simulation study shows the performance of the proposed estimator and compares its behaviour with smoothed estimators only in the covariate. An empirical study, based on modified real data, illustrates its practical application.

In Chapter 6, a practical way to choose the smoothing parameters involved in the probability of default estimators is proposed. Resampling methods based on bootstrap techniques are proposed to approximate the bandwidths which Beran's

and the smoothed Beran's estimators of the PD depend on. Bootstrap algorithms for the calculation of confidence regions of the probability of default are also proposed. Extensive simulation studies show the good behaviour of presented algorithms. The bandwidth selector and the confidence region algorithm are applied to the German credit data set to analyse the probability of default conditional on the credit scoring.

An estimator for the probability of default that considers the existence of a cured individuals group is proposed in Chapter 7. It is derived from a nonparametric conditional survival function estimator based on mixture 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 nonparamtric estimator compared with Beran's and the smoothed Beran's PD estimator and other semiparametric methods. Finally, an empirical study of the German credit data set illustrates the practical behaviour.

Some concluding remarks and comments regarding future lines of research are given in Chapter 8.

The main chapters that compose this work have been published in international scientific journals or are currently under review, so the reader may wish to explore them independently. To implement the methods proposed in this thesis several R packages have been developed by ourselves. Although they have not been uploaded to the CRAN yet, our plan is to do it in the very near future.

Chapter 1

State of Art

1.1 Survival analysis

Survival analysis is a collection of statistical procedures to describe and study data when the variable of interest is the time until an event occurs. The impossibility of observing the event of interest in all subjects is inherent in these techniques. Reasons such as the study ending or subjects leaving the study before experiencing the event of interest make a proportion of survival times of interest unknown. Censoring is therefore a distinguishing feature of survival analysis.

The term *failure* is used to specify the occurrence of the event of interest and the *lifetime* refers the length of the time from the beginning of the study until the occurrence of the event. In the classical biomedical applications, lifetime may represent the survival time of a living organism or the time until a disease is cured. Nevertheless, these techniques can be applied to data from different areas, for example, times to default in a financial context.

Individuals in the study may be subject to different types of censoring. We highlight the following:

Right censoring occurs when the study ends before all individuals has experi-

enced the event of interest. The lifetime is considered to be *left censored* if the failure happens some time before the follow-up period, i.e., the event of interest has already ocurred for the individual before the observed time. An individual is *interval censored* if the event is known to occur at a certain specified time interval, but the exact time of occurrence is unknown.

Depending on the way in which the duration of the experiment is limited in order to obtain the data, the most frequent censoring types are the following:

Type I censoring: The study has a duration, C, established a priori. The survival time of an individual will be observed if it is less than or equal to that pre-set value. Otherwise, the corresponding observation will have a censored value C.

Type II censoring: The trial ends at the time of the k-th failure. That instant will be the observed time of all individuals who have not yet failed at that time.

Random censoring: The censoring variable is assumed to be independent of the variable of interest. Times corresponding to individuals who have not yet experienced the event of interest at the end of the study or who have experienced other circumstances independent of the event that caused them to drop out of the study are considered censored.

Let us denote the time to occurrence of the event by T and denote the censoring time by C. In the precense of random right censoring, it is only possible to observe the pair (Z, δ) where $Z = \min\{T, C\}$ is the observed time and $\delta = I(T \leq C)$ is the uncensoring indicator. The distribution function of T is denoted by $F(t) = P(T \leq t)$ and its survival function is $S(t) = P(T > t) = 1 - F(t)$. The ditribution function of C is denoted by $G(t) = P(C \leq t)$. The distribution function of the observed time Z is denoted by $H(t) = P(Z \leq t)$. Under the assumption that T and C are independent, it is easily proved that $1 - H(t) = \Big(1 - F(t)\Big)\Big(1 - G(t)\Big)$.

Estimating and comparing conditional survival functions of different groups and assesing the relationship between covariates and times until the event are the main purposes of survival analysis. Parametric models are often assumed for lifetimes

in survival analysis. Distributions such as exponential, Weibull or log-normal have been playing an important role in modelling failure times. Cox (1972) revolutionised survival analysis through his semiparametric regression model for the hazard function (proportional hazards model or Cox Model), which depends arbitrarily on time and parametrically on covariates. Another method for semiparametric survival estimation is the accelerated failure time models proposed by Kalbfleisch and Prentice (1980) and location-scale models proposed by Lawless (1982). Nonparametric methods to estimate the cumulative survival function from lifetime data are detailed in next section.

1.2 Nonparametric curve estimation

Nonparametric curve estimation has been one of the most active fields of statistics in recent decades. A proof of this is the long list of studies and papers on this topic. Nonparametric methods require few assumptions regarding the underlying distribution of the data (Hollander et al. (1999)).

In statistical inference it is often necessary to know some of the curves that characterise the distribution of the population under study. These curves are not usually known in practice and a useful approach to approximate them is to assume some parametric or semiparametric model. However, there are examples where parametric models are not found to properly describe the data generation process. In these cases, it is of interest to estimate such curves in a flexible way, assuming, at most, continuity and differentiability conditions for the underlying curve. Nonparametric estimators let the data "speak for themselves" and relative mild assumptions, relative simplicity and relative insensitive to outlying observations are some of the advantages that nonparametric methods enjoy. One of the curves that we will be interested in estimating throughout this thesis is the conditional survival function.

In an uncensored context, counting the proportion of subjects alive at time t from a random sample $\{T_i\}_{i=1}^n$ gives the empirical survival estimation (Andersen

et al. (1993)) at that time:

$$S_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t). \tag{1.1}$$

In Kaplan and Meier (1958), this survival estimator in (1.1) is extended to an incomplete data scenario. Let C be the censoring variable and $\{(Z_i, \delta_i)\}$ be a random right censoring sample satisfying $Z_i = T_i$ when $\delta_i = 1$ and $Z_i = C_i$ when $\delta_i = 0$, for all $i = 1, \ldots, n$. Then, the nonparametric maximum likelihood estimator of the survival function estimator for censored data is given by

$$\widehat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{\delta_{(i)}}{n - i + 1}\right) \tag{1.2}$$

where $\delta_{(i)}$ is the concomitant of the ordered sample of T, $T_{(1)} < \cdots < T_{(n)}$. The Kaplan-Meier estimator, also known as the product-limit estimator, is the survival function estimator mostly used for random right censored data. It is a stair-step function with jumps at the uncensored observations and weights which depend on the number of censored observations among them. The Kaplan-Meier estimator assumes independence between the survival and censoring times. If the assumption of independence does not hold, the estimator may be biased (see Kaplan and Meier (1958)).

The generalised Kaplan-Meier estimator to the case where a continuous covariate, X, is involved was introduced by Beran (1981). Let $\{(X_i, Z_i, \delta_i)\}$ be a random right censored sample of (X, Z, δ) and denote by $S(t|x)$ the conditional survival function of the time T. The conditional product-limit estimator of the conditional survival function is given by

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

where

$$w_{h,i}(x) = \frac{K\Big((x - X_i)/h\Big)}{\sum_{j=1}^{n} K\Big((x - X_j)/h\Big)}, \quad i = 1, \ldots, n, \tag{1.3}$$

are the Nadaraya-Watson weights with K being a kernel function and $h > 0$ being a smoothing parameter. This estimator has been deeply studied in the literature (see Dabrowska (1989), González-Manteiga and Cadarso-Suárez (1994), Van Keilegom and Veraverbeke (1997)).

1.3 Bootstrap methodology

A common problem in a nonparametric context is to study a specific characteristic of the distribution of some statistics, but making no assumptions about its shape. Therefore, important achievements of nonparametric methods have been made by introducing resampling techniques such as jackknife or bootstrap. Efron (1979) introduced the bootstrap method to approximate the sampling distribution of a statistic, $R = R(\mathbf{T}, F)$, which depends on the population distribution F and the observed sample $\mathbf{T} = \{T_i\}_{i=1}^n$. The idea is to approximate the distribution of R by the resampling distribution of $R^* = R(\mathbf{T}^*, \widehat{F})$, where $\widehat{F}$ is an estimator of the underlying distribution, for example the empirical distribution, and $\mathbf{T}^* = \{T_i^*\}_{i=1}^n$ is a random sample obtained from $\widehat{F}$ often called resample. Since $\widehat{F}$ can be computed from the observed data, resamples can be obtained by simulation. Then, the Monte Carlo method can be used to approximate the resampling distribution of R^* just repeating this procedure an appropriate number of times.

Bootstrap methods based on this idea do not make any assumptions about the data generation process, which is their main advantage. This results in a high computational cost, since, in general, it is necessary to resort to the Monte Carlo method (see Efron (1979), Hall (1992) and Efron and Tibshirani (1993), among others). In Efron and Tibshirani (1993) a bootstrap method which provides a straightforward nonparametric way to estimate the standard error and to construct confidence intervals of a parameter.

Some schemes of bootstrap resampling techniques for independent data are described in the following paragraphs.

Uniform bootstrap

The uniform bootstrap (or naive bootstrap) is a resampling method where the (unknown) population distribution is replaced by the empirical distribution given by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t).$$

1. For each $i = 1, \ldots, n$, sample T_i^* from $F_n(t)$, i.e., $P^*(T_i^* = T_j) = \frac{1}{n}$, for all $j = 1, \ldots, n$.

2. Consider the bootstrap resample $\{T_1^*, \ldots, T_n^*\}$.

Smoothed bootstrap

The smoothed bootstrap is the resampling technique that assumes the distribution function $F(t)$ to be continuous and incorporates this information into the resampling method. Since a continuous distribution function has an associated density function, the smoothed bootstrap is based on resampling from a density function estimator. An appropriate estimator of the density function proposed by Parzen (1962) and Rosenblatt (1956) is as follows:

$$\widehat{f}_h(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - T_i}{h}\right)$$

where K is a kernel function and $h > 0$ is a smoothing parameter, called bandwidth, which determines the size of the neighbourhood used to compute the estimation.

The smoothed bootstrap method proceeds as follows:

1. From the sample $\{T_1, \ldots, T_n\}$, compute the Parzen-Rosenblatt estimator, $\widehat{f}_h(t)$ with smoothing parameter $h > 0$.

2. For each $i = 1, ..., n$, sample T_i^* from the density estimation $\widehat{f}_h(t)$.

3. Consider the bootstrap resample $\{T_1^*, \ldots, T_n^*\}$.

A cursory analysis of the density estimator leads to the conclusion that Step 2 of the previous algorithm can be replaced by the following:

3. For each $i = 1, ..., n$, draw $U_i \sim U(0,1)$ and V_i with density K and obtain

$$T_i^* = T_{[nU_i]+1} + hV_i,$$

where $[u]$ is the integer part of u.

Bootstrap with censored data

In Efron (1981) two equivalent resampling methods, simple bootstrap and obvious bootstrap, adapted to censored data are proposed. See also the works of Reid (1981) and Akritas (1986).

The simple bootstrap consists of the following steps:

1. Obtain the two-dimensional empirical distribution function $F_n^{Z,\delta}$ from the sample $\{(Z_i, \delta_i)\}_{i=1}^n$.

2. For each $i = 1, \ldots, n$, draw (Z_i^*, δ_i^*) from the empirical distribution function, that is,

$$P^*\Big((Z_i^*, \delta_i^*) = (Z_j, \delta_j)\Big) = \frac{1}{n},$$

for all $j = 1, \ldots, n$.

3. Consider the bootstrap resample $\{(Z_i^*, \delta_i^*)\}_{i=1}^n$.

In order to detail the obvious bootstrap algorithm it is necessary to introduce the Kaplan-Meier estimator of the distribution function of the survival and censoring times. According to (1.2), they are given by

$$\widehat{F}(t) = 1 - \prod_{T_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)}}$$

and

$$\widehat{G}(t) = 1 - \prod_{T_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{1-\delta_{(i)}}.$$

Then, the obvious bootstrap consists of the following steps:

1. Obtain the Kaplan-Meier estimation of the distribution functions of the survival time, $\widehat{F}(t)$, and the censoring time, $\widehat{G}(t)$.

2. For each $i = 1, \ldots, n$, draw independent bootstrap observations T_i^* from the distribution function $\widehat{F}$ and C_i^* from the distribution function $\widehat{G}$.

3. For each $i = 1, \ldots, n$, define

$$Z_i^* = \min\{T_i^*, C_i^*\}$$

and

$$\delta_i^* = I(T_i^* \leq C_i^*).$$

4. Consider the bootstrap resample $\{(Z_i^*, \delta_i^*)\}_{i=1}^n$.

The obvious bootstrap and the simple bootstrap are equivalent when there are no ties between censored and uncensored observations (see Efron (1981)). Then, the distribution of the resample (T^*, δ^*) is the same for both methods. However, the obvious bootstrap is computationally more expensive.

Bootstrap with covariates

In Li and Datta (2001) a bootstrap method for nonparametric regression with right censored data is proposed. The method is used to obtain confidence intervals and bands for the conditional survival function. They present two equivalent resampling algorithms for bootstraping the generalised Kaplan-Meier estimator: the simple weighted bootstrap and the obvious bootstrap. The equivalence of the two methods is obtained in a straightforward way by applying the arguments in Efron (1981) for the unconditional setting. These two bootstrap methods are detailed in the following paragraphs:

Simple weighted bootstrap with covariates

1. Generate $\{X_1^*, \ldots, X_n^*\}$ from the empirical distribution of $\{X_i\}_{i=1}^n$.

2. For each $i = 1, \ldots, n$, generate the pair (Z_i^*, δ_i^*) from the weighted empirical distribution $\widehat{F}_h(u, v|X_i^*)$ where

$$\widehat{F}_h(u, v|x) = \sum_{i=1}^{n} w_{h,i}(x) I(Z_i \leq u, \delta_i \leq v),$$

   with $w_{h,i}(x)$ defined in (1.3).

3. Consider the bootstrap resample $\{(X_i^*, Z_i^*, \delta_i^*)\}_{i=1}^n$.

Obvious bootstrap with covariates

1. Generate $\{X_1^*, \ldots, X_n^*\}$ from the empirical distribution of $\{X_i\}_{i=1}^n$.

2. For each $i = 1, \ldots, n$, generate T_i^* from the Beran's estimator of the conditional distribution of T using the sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ denoted by $\widehat{F}_h(t|X_i^*)$ and C_i^* from the Beran's estimator of the conditional distribution of C using the sample $\{(X_i, Z_i, 1-\delta_i)\}_{i=1}^n$ denoted by $\widehat{G}_h(t|X_i^*)$.

   The estimators $\widehat{F}_h(t|X_i^*)$ and $\widehat{G}_h(t|X_i^*)$ 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

$$Z_i^* = \min\{T_i^*, C_i^*\}, \quad \delta_i^* = I\Big(T_i^* \leq C_i^*\Big).$$

4. Consider the bootstrap resample $\{(X_i^*, Z_i^*, \delta_i^*)\}_{i=1}^n$.

1.4 Bandwidth selection in nonparametric curve estimation

One of the crucial issues in nonparametric estimation, and in particular in kernel estimation, is the selection of the smoothing parameters. In general, an excessively

large smoothing parameter will increase the bias of the estimator, but choosing an excessively small value will increase the variance. The selection of an appropriate parameter is therefore a matter of finding a balance between bias and variance. For this purpose, bandwidth selection methods in nonparametric curve estimation often look for a minimal value of the estimation error that is made when approximating the curve by the smoothed estimator. The mean squared error (MSE) at a fixed point (t, x) where the curve is to be estimated or the mean integrated squared error (MISE) over a time interval for a fixed value of x can be considered measures of the estimation error and be defined as functions of the bandwidth. Bootstrap techniques are, naturally, a useful tool in approximating any of these estimation errors (see, among others, López-Cheda et al. (2017b); Cao (1993); Barbeito and Cao (2019)).

In this thesis, a global criterion for the estimation error is chosen. Therefore, the considered error function is the mean integrated squared error as a function of the bandwidth h, $MISE(h)$. Bootstrap bandwidth selection consists of estimating the function $MISE(h)$ by means of bootstrap resampling and obtaining the bandwidth h that minimises its bootstrap version. Given the sample $\mathbf{T} = \{T_i\}_{i=1}^n$ with distribution function F, we can consider the random variable

$$R_h(\mathbf{T}, F) = \int \left(\widehat{\theta}_h(t) - \theta(t)\right)^2 dt$$

and the MISE function defined by $MISE(h) = E(R_h(\mathbf{T}, F))$ where $\widehat{\theta}_h(t)$ is some kernel estimator of a curve of interest $\theta(t)$. The distribution of $R_h(\mathbf{T}, F)$ is approximated by the sampling distribution of

$$R_h^*(\mathbf{T}^*, \widehat{F}) = \int \left(\widehat{\theta}_h^*(t) - \widehat{\theta}(t)\right)^2 dt$$

and, consequently, the bootstrap version of the MISE function is as follows

$$MISE^*(h) = E^*(R_h^*(\mathbf{T}^*, \widehat{F})).$$

This bootstrap MISE can be approximated by the Monte Carlo method using B resamples:

$$MISE^*(h) = E^*(R_h^*(\mathbf{T}^*, \widehat{F})) \simeq \frac{1}{B}\sum_{j=1}^{B} R_h^*(\mathbf{T}_j^*, \widehat{F})$$

where $\mathbf{T}_j^*$ is the j-th bootstrap resample.

Applying Monte Carlo approximation is computationally expensive, specially if the function to be minimised is difficult to compute. It is, however, a straightforward way of obtaining bootstrap resamples, regardless of the data generating process.

1.5 Cure models

Thanks to the medical advances that have taken place in recent decades, the survival and quality of life of patients with various diseases has improved remarkably. As a consequence, clinical studies analysing the evolution of these patients suffer from the fact that a large proportion of patients do not experience the event of interest in the follow-up period. These scenarios present a proportion of subjects who can be considered cured or not susceptible to the event of interest. At this point, survival models that explicitly incorporate the possibility that a subject may never experience the event of interest arise. These are the so-called cure models.

Cure models were originally proposed to model long-term survival of cancer patients. However, they can be applied to any survival context where a group of individuals is assumed not to experience the event of interest, no matter how long they are followed. For example, a financial context where a proportion of borrowers will not default during the loan term.

Figure 1.1 shows the differences between the survival functions of a standard survival analysis model and a model with a proportion of cured subjects. We can appreciate the nonzero tendency of the survival function under a cure model. This plateau in the right tail informs about the proportion of nonsusceptible individuals.

There are two main classes of cure models: mixture and nonmixture cure models. Nonmixture cure models were firstly due to Haybittle (1959, 1965). In Yakovlev and Tsodikov (1996) a proportional hazards cure model is proposed. A semiparametric maximum likelihood estimation for nonmixture cure models is presented in Liu and

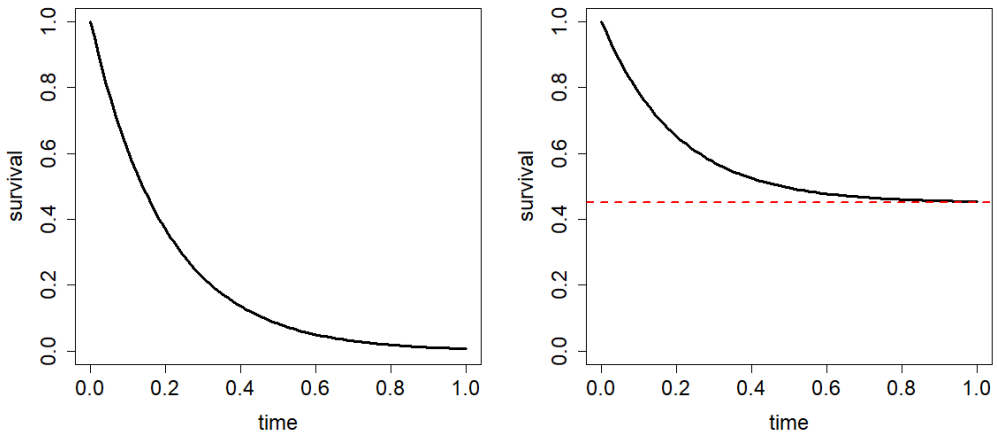

Figure 1.1: Standard survival function (left) and survival function with a fraction of cured population (right).

Shen (2009) using the expectation-maximisation method for interval censored data.

Mixture cure models were proposed by Boag (1949) and they consider the following useful decomposition of the conditional survival function

$$S(t|x) = 1 - p(x) + p(x)S_0(t|x),$$

where $1 - p(x)$ is the probability of being cured (nonsusceptible to default) and $S_0(t|x)$ the conditional survival function of the uncured population. The functions $p(x)$ and $S_0(t|x)$ are called the incidence and the latency, respectively. A detailed review of this model is provided by Maller and Zhou (1996) and Corbière et al. (2009). There, the incidence is assumed to be a logistic function and the latency is parametrically estimated. Since the conditional survival function cannot always be well approximated using parametric or semiparametric methods, R. A. Maller (1992) proposed a consistent nonparametric estimator of the incidence without handling covariates. In Laska and Meisner (1992) a nonparametric estimator of the cure rate that allows discrete covariates is presented. In López-Cheda et al. (2017a) and López-Cheda et al. (2017b), nonparametric estimators of the incidence and the latency which also consider the effects of continuous covariates are proposed and deeply studied.

Chapter 2

Nonparametric estimation of the probability of default

2.1 Introduction

Since the work by Naraim (1992), abundant literature has been developed using survival analysis in credit risk. To name a few papers, in Hanson and Schuermann (2004) survival analysis makes it possible to obtain confidence intervals for the probability of default; in Glennon and Nigro (2005) the time to default distribution function is estimated using a hazard model and in Allen and Rose (2006) the Kaplan-Meier estimator is used to estimate the time to default survival function.

In Naraim (1992), the proposal is a Cox proportional risk model to estimate the conditional survival function $S(t|x)$. Cao et al. (2009) start from this and, writing the probability of default in terms of conditional survival function, they get an estimator of the PD. A second alternative given in Cao et al. (2009) is to assume a generalized linear model for the lifetime distribution under censoring: $P(T \leq t|X = x) = F_\theta(t|x) = g(\theta_0 + \theta_1 t + \theta_2 x)$, where g is an unknown link function and $\theta = (\theta_0, \theta_1, \theta_2)$. These approaches use parametric or semiparametric models for the time to default as a function of the credit scoring and the credit lifetime. The

use of nonparametric curve estimation for this purpose is however convenient. These are flexible methods that use only the information that the data provide without making assumptions about the shape of the curve. This is already the approach used in the third model in Cao et al. (2009) who proposed to estimate the probability of default using Beran's estimator for the conditional survival function. Here, four nonparametric estimators of the probability of default are defined. Their asymptotic properties are studied and their performance is evaluated and compared by means of a simulation study. Finally, the four nonparametric estimators of the probability of default are applied to a set of modified real data.

The content of this chapter has been published in Peláez et al. (2021b).

2.2 Nonparametric PD estimators

Let $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^{n}$ be the right censored random sample of (X, Z, δ) where X_i represents the covariate, $Z_i = \min\{T_i, C_i\}$ the observed lifetime and $\delta_i = I(T_i \leq C_i)$ the censoring indicator, where $T_i \geq 0$ and $C_i \geq 0$ are the time to occurrence of the event and the censoring time for the i-th individual of the sample with $i = 1, \ldots, n$. Let x be a fixed value of the covariate X and b a horizon time, then the probability of default in a time horizon $t + b$ from a maturity time t is defined as follows

$$PD(t|x) = 1 - \frac{S(t+b|x)}{S(t|x)}. \tag{2.1}$$

Replacing $S(t|x)$ with a nonparametric estimator, $\widehat{S}_h(t|x)$, in (2.1), the following estimator of the probability of default is obtained:

$$\widehat{PD}_h(t|x) = 1 - \frac{\widehat{S}_h(t+b|x)}{\widehat{S}_h(t|x)}, \tag{2.2}$$

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

In this work, the following four nonparametric estimators of the conditional survival function are used to estimate the probability of default through the expression in (2.2).

2.2.1 Beran's estimator

The estimator of the conditional survival function with censored data formulated in Beran (1981) was already used in Cao et al. (2009) to obtain a probability of default estimator. Beran's estimator is given by

$$\widehat{S}_h^B(t|x) \;=\; \prod_{i=1}^{n}\left(1-\frac{I(Z_i\le t,\ \delta_i=1)w_{h,i}(x)}{1-\sum_{j=1}^{n} I(Z_j<Z_i)w_{h,j}(x)}\right), \tag{2.3}$$

where the weights are

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

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

The Beran's estimator of the probability of default denoted by $\widehat{PD}_h^B(t|x)$ is obtained by replacing $\widehat{S}_h(t|x)$ in (2.2) with the conditional survival estimator $\widehat{S}_h^B(t|x)$ in (2.3) as follows:

$$\widehat{PD}_h^B(t|x) \;=\; 1-\frac{\widehat{S}_h^B(t+b|x)}{\widehat{S}_h^B(t|x)}. \tag{2.4}$$

2.2.2 Weighted local linear (WLL) estimator

In Cai (2003) a nonparametric estimator of the regression function for censored lifetime response variable is proposed using local polynomial fitting. Without loss of generality, an arbitrary function, τ, and the variable $V=\tau(T)$ can be used to establish the following nonparametric regression model

$$V=\tau(T)=r(X)+\varepsilon, \tag{2.5}$$

where $r(x)=E(V|X=x)$ is the regression function of V given X and ε is the error variable that satisfies $E(\varepsilon|X)=0$ and $Var(\varepsilon|X)=\sigma^2(X)$.

In order to estimate the conditional survival function, $V_t=\tau_t(T)=I(T>t)$ is chosen, so that

$$r(x)=E(V_t|X=x)=E(I(T>t)|X=x)=P(T>t|X=x)=S(t|X=x),$$

and the estimator of the regression function $r(x)$ will be an estimator of $S(t|x)$ for a fixed value t.

Let $\{(X_{[i]}, Z_{(i)}, \delta_{[i]})\}_{i=1}^n$ be a random sample which is sorted according to the values $\{Z_i\}_{i=1}^n$ of the population (X, Z, δ) and let $X_{[i]}$, $\delta_{[i]}$ be the concomitants of $\{Z_i\}_{i=1}^n$. Consider the functions

$$\begin{aligned} S_{n,l}(x) &= \sum_{i=1}^{n} (X_{[i]} - x)^l \, w_{[i],h} W_{[i],n}, \\ T_{n,l}(x) &= \sum_{i=1}^{n} \tau(Z_{(i)}) (X_{[i]} - x)^l \, w_{[i],h} W_{[i],n}, \end{aligned}$$

for $l = 0, 1, 2$, where $w_{[i],h} = K_h(X_{[i]} - x)$ with K a kernel function, $K_h(u) = K(u/h)/h$ and $h = h_n$ a smoothing parameter, are the covariate weights and

$$W_{[i],n} = \frac{\delta_{[i]}}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_{[j]}},$$

are the Kaplan-Meier censoring weights.

The weighted local linear regression estimator (WLL) proposed in Cai (2003) is given by

$$\widehat{S}_h^{WLL}(t|x) = \widehat{r}_h^{WLL}(x) = \frac{S_{n,2}(x) T_{n,0}(x) - S_{n,1}(x) T_{n,1}(x)}{S_{n,2}(x) S_{n,0}(x) - S_{n,1}^2(x)}. \tag{2.6}$$

and it provides an estimator of the conditional survival function. This estimator was used to estimate the conditional distribution function under censoring along with Beran's estimator in Gannoun et al. (2007).

Replacing $\widehat{S}_h(t|x)$ in (2.2) with the conditional survival estimator $\widehat{S}_h^{WLL}(t|x)$, the weighted local lineal estimator of the PD denoted by $\widehat{PD}_h^{WLL}(t|x)$ is defined.

2.2.3 Weighted Nadaraya-Watson (WNW) estimator

The WLL conditional survival estimator presents two problems: it does not always take values within the interval $[0, 1]$ and it is not always isotonic (non increasing). Both problems become worse when considering the PD estimator, $\widehat{PD}_h^{WLL}(t|x)$. Figure 2.1 shows an example of this.

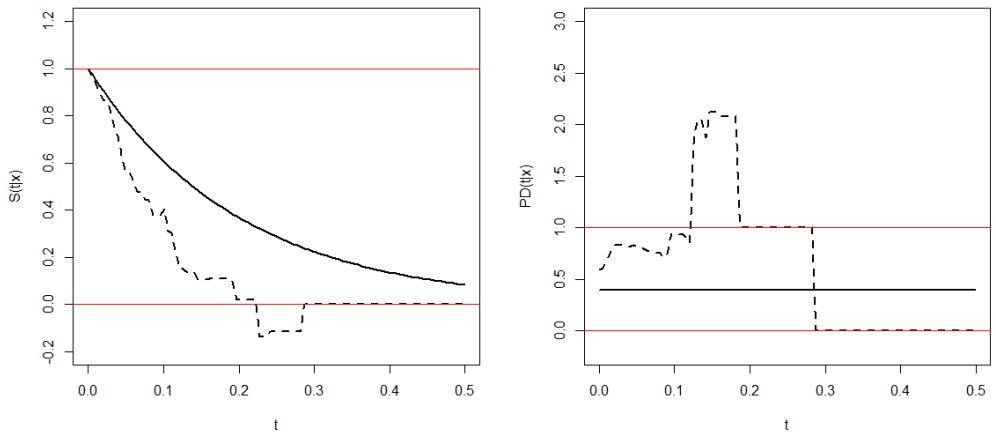

Figure 2.1: Theoretical curves (solid lines) and their estimations obtained by the WLL estimator (dashed lines) for the conditional survival function $S(t|x)$ (left) and probability of default $PD(t|x)$ (right) for one sample from (X, Z, δ) where $X \sim U(0,1)$, $T|_{X=x} \sim Exp(1+5x)$ and $C|_{X=x} \sim (10 - 7/2x + 20x^2)$.

The first problem can be solved by restricting the values that $\widehat{S}_h^{WLL}(t|x)$ takes to the interval $[0,1]$ but the second one cannot. For this reason, a weighted local constant estimator is proposed. It is obtained by replacing the local linear regression weights with the Nadaraya-Watson weights. It does not present the problems that $\widehat{S}_h^{WLL}(t|x)$ does. Its expression is the following one:

$$\widehat{S}_h^{WNW}(t|x) = \widehat{r}_h^{WNW}(x) = \frac{\sum_{i=1}^n \tau(Z_{(i)}) w_{[i],h} W_{[i],n}}{\sum_{i=1}^n w_{[i],h} W_{[i],n}}, \tag{2.7}$$

using the same notation as for (2.5).

Replacing $\widehat{S}_h(t|x)$ in (2.2) with the conditional survival estimator $\widehat{S}_h^{WNW}(t|x)$, the weighted Nadaraya-Watson estimator of the PD denoted by $\widehat{PD}_h^{WNW}(t|x)$ is defined.

2.2.4 Van Keilegom-Akritas (VKA) estimator

In Van Keilegom and Akritas (1999) and Van Keilegom et al. (2001) a nonparametric estimator of the conditional survival function is proposed. It presents a better

behaviour than Beran's estimator in the right tail of the distribution in a heavy censoring context. It is introduced here to study if this property is inherited by the corresponding PD estimator.

In order to define this estimator, the following nonparametric regression model is assumed:

$$T = r(X) + \sigma(X)\varepsilon,$$

where $r(x) = E(T|X = x)$ is the unknown regression curve; $\sigma(x)$ is the conditional standard deviation, enabling a possible heteroscedastic model, and ε is the error variable.

Note that

$$P(T \leq t|X = x) = P(r(X) + \sigma(X)\varepsilon \leq t|X = x) = P\left(\varepsilon \leq \frac{t - r(x)}{\sigma(x)}\right),$$

so,

$$F(t|x) = F_\varepsilon\left(\frac{t - r(x)}{\sigma(x)}\right),$$

where F_ε denotes the distribution function of the error variable ε. This relationship between the conditional distribution function of T and F_ε suggests the following estimator for $F(t|x)$ (and hence for $S(t|x)$).

Let $\widehat{r}(x)$ and $\widehat{\sigma}(x)$ be consistent estimators of $r(x)$ and $\sigma(x)$, respectively, and let $\widehat{F}_\varepsilon$ be the Kaplan-Meier estimator of F_ε. The estimator of the conditional distribution function $F(t|x)$ according to this model is:

$$\widehat{F}(t|x) = \widehat{F}_\varepsilon\left(\frac{t - \widehat{r}(x)}{\widehat{\sigma}(x)}\right).$$

Thus, the estimator of the conditional survival function of Van Keilegom-Akritas is given by

$$\widehat{S}_h^{VKA}(t|x) = 1 - \widehat{F}_\varepsilon\left(\frac{t - \widehat{r}(x)}{\widehat{\sigma}(x)}\right). \tag{2.8}$$

In Van Keilegom and Akritas (1999), without loss of generality these location and scale functions are considered to define $r(x)$ and $\sigma(x)$:

$$r(x) \;=\; \int_0^1 F^{-1}(s|x)J(s)ds, \tag{2.9}$$

$$\sigma^2(x) \;=\; \int_0^1 F^{-1}(s|x)^2 J(s)ds - r^2(x), \tag{2.10}$$

where $F^{-1}(s|x) = \inf\{t : F(t|x) \geq s\}$ is the conditional quantile function of T given x and $J(s)$ is such that $\int_0^1 J(s)ds = 1$. When choosing $J(s) = 1$, $\forall s \in [0,1]$, expressions (2.9) and (2.10) turn out to be $E(T|X = x)$ and $Var(T|X = x)$, respectively.

Considering the Beran's estimator of $F(t|x)$, $\widetilde{F}_h(t|x)$, with bandwidth $h = h_n$, the corresponding estimator for $r(x)$ and $\sigma(x)$ are obtained by

$$\widehat{r}(x) \;=\; \int_0^1 \widetilde{F}^{-1}(s|x)J(s)ds,$$

$$\widehat{\sigma}^2(x) \;=\; \int_0^1 \widetilde{F}^{-1}(s|x)^2 J(s)ds - \widehat{r}^2(x).$$

Finally, considering the Kaplan-Meier estimator for F_ε based on the regression residuals

$$\widehat{E}_i = \frac{Z_i - \widehat{r}(X_i)}{\widehat{\sigma}(X_i)},\ i = 1, \ldots, n,$$

all the elements required in (2.8), to obtain $\widehat{F}(t|x)$, are available.

The Van Keilegom-Akritas estimator of the probability of default denoted by $\widehat{PD}_h^{VKA}(t|x)$ is obtained by replacing $\widehat{S}_h(t|x)$ in (2.2) with the conditional survival estimator $\widehat{S}_h^{VKA}(t|x)$.

2.3 Asymptotic results

It is known that many estimators of the conditional distribution function (and, therefore, of the conditional survival function) enjoy desirable properties for their bias, variance and asymptotic normality. It is interesting to obtain similar properties for the probability of default estimators.

The theoretical results shown in this section allow to obtain, under general conditions, asymptotic properties for a PD estimator, based on these properties for the corresponding estimator of the conditional survival function.

Let $\widehat{S}(t|x)$ be an estimator of the conditional survival function, $S(t|x)$, and let $\widehat{PD}(t|x)$ be its corresponding estimator of the probability of default at horizon b. The necessary conditions to prove the asymptotic properties of the PD estimator are the following:

C.1 The estimator of $PD(t|x)$ is a transformation of the conditional survival estimator of the following form $\widehat{PD}(t|x) = 1 - \dfrac{\widehat{S}(t+b|x)}{\widehat{S}(t|x)}$

C.2 The bias and the covariance of $\widehat{S}(t|x)$ admit the following asymptotic expressions:

$$B(t|x) := Bias\left(\widehat{S}(t|x)\right) = B_0(t|x)h^2 + o(h^2),$$
$$C(t_1,t_2|x) := Cov\left(\widehat{S}(t_1|x), \widehat{S}(t_2|x)\right) = C_0(t_1,t_2|x)\frac{1}{nh} + o\left(\frac{1}{nh}\right),$$

for any t, t_1 and t_2. As a consequence, defining $V(t|x) := Var\left(\widehat{S}(t|x)\right)$ and $V_0(t|x) := C_0(t,t|x)$ we have $V(t|x) = V_0(t|x)\dfrac{1}{nh} + o\left(\dfrac{1}{nh}\right)$.

C.3 The terms

$$E\left(\left(\widehat{S}(t_1|x) - E\left(\widehat{S}(t_1|x)\right)\right)^i \left(\widehat{S}(t_2|x) - E\left(\widehat{S}(t_2|x)\right)\right)^{3-i}\right) = o\left(\frac{1}{nh}\right),$$

for $i = 0, 1, 2, 3$.

Theorem 2.1. *Assume Conditions C.1-C.3. Asymptotic expressions of bias and variance for the estimator $\widehat{PD}(t|x)$ are the following:*

$$\begin{aligned} Bias\left(\widehat{PD}(t|x)\right) &= \frac{(1-PD(t|x))B_0(t|x) - B_0(t+b|x)}{S(t|x)}h^2 + o(h^2) + O\left(\frac{1}{nh}\right) \\ Var\left(\widehat{PD}(t|x)\right) &= \left[\frac{V_0(t+b|x)}{S(t|x)^2} - \frac{2S(t+b|x)C_0(t,t+b|x)}{S(t|x)^3}\right. \\ &\quad \left. + \frac{S(t+b|x)^2 V_0(t|x)}{S(t|x)^4}\right]\frac{1}{nh} + o\left(\frac{1}{nh}\right) \end{aligned}$$

The asymptotic properties of Beran's estimator for the conditional survival function were proven in both Dabrowska (1989) and Iglesias-Pérez and González-Manteiga (1999) under certain assumptions. From them, the expressions of the bias and the

variance of the estimator $\widehat{PD}_h^B(t|x)$ can be found by using Theorem 2.1. This was done in Cao et al. (2009). The asymptotic bias and variance of the WLL estimator of the survival function are proven in Cai (2003) under suitable conditions. Van Keilegom and Akritas (1999) gave necessary conditions for the asymptotic expressions of bias and variance of the VKA estimator. It is enough to recover the expressions for $B_0(t|x)$, $C_0(t, t+b|x)$ and $V_0(t|x)$ from the above articles and use Theorem 2.1 to obtain the asymptotic properties of the corresponding estimators for the probability of default. The expressions obtained in most of the cases are complex and depend on too many parameters. It is then difficult to use them in order to compare estimators or to obtain optimal smoothing parameters.

Proofs of these results can be found in Section 2.6.

2.4 Simulation study

A simulation study was conducted in order to compare the performance of the four proposed estimators for the probability of default. The study is focused on three models, one with Weibull lifetime and censoring time distributions and two models with exponential distributions.

Model 1

For Model 1, a $U(0,1)$ distribution is considered for the credit scoring, X. The time to default conditional to the credit scoring, $T|_{X=x}$, follows an exponential distribution of parameter $P(x) = a_0 + a_1 x$,

$$T|_{X=x} \sim Exp\Big(P(x)\Big),$$

and the censoring time conditional to the credit scoring, $C|_{X=x}$, follows an exponential distribution with parameter $Q(x) = b_0 + b_1 x + b_2 x^2$,

$$C|_{X=x} \sim Exp\Big(Q(x)\Big).$$

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

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

$$PD(t|x) = 1 - e^{-P(x)b}.$$

Let $H_0(t|x) = P(Z \leq t, \delta = 0|X = x)$ be the conditional subdistribution function of Z when $\delta = 0$. The censoring conditional probability is obtained as follows:

$$\begin{aligned} P(\delta = 0|X = x) &= H_0(\infty|X = x) = \int_0^\infty \Big(1 - F(u|x)\Big) dG(u|x) \\ &= \int_0^\infty e^{-P(x)t} d\Big(1 - e^{-Q(x)t}\Big) = \int_0^\infty Q(x) e^{-P(x)t} e^{-Q(x)t} dt \\ &= -\frac{Q(x)}{P(x) + Q(x)} e^{-(P(x)+Q(x))t} \Big]_0^\infty = \frac{Q(x)}{P(x) + Q(x)} \end{aligned}$$

and the censoring unconditional probability is given by

$$P(\delta = 0) = \int_{-\infty}^{+\infty} P(\delta = 0|X = x) m(x) dx,$$

where $m(x)$ is the density function of the covariate X.

Note that if $P(x)$ is large, then the mean lifetime of the credit $(1/P(x))$ is small and the probability of conditional censoring too; whereas if $P(x)$ is small the censoring conditional probability is large, which is compatible with the mean of the credit's lifetime also being large. It is clear that the censoring probability of an observation in this model is determined by the choice of the coefficients of the polynomials P and Q. In this model, the polynomials chosen are: $P(x) = 1 + 5x$ and $Q(x) = 10 + b_1 x + 20x^2$. Having set the value of the credit scoring, $x = 0.8$, the value of b_1 is chosen so that the censoring conditional probability is 0.2, 0.5 and 0.8. The resulting values are shown in Table 2.1 along with the corresponding both conditional and unconditional censoring probabilities.

b_1	$P(\delta = 0 \mid X = 0.8)$	$P(\delta = 0)$
$-431/16$	0.2	0.438
$-89/4$	0.5	0.613
$-7/2$	0.8	0.816

Table 2.1: Values of b_1 and the associated censoring probabilities for Model 1.

In addition, Figures 2.2, 2.3 and 2.4 show the plots of the resulting polynomials $P(x)$ and $Q(x)$ and the plot of $P(\delta = 0|X = x)$ in each case. It can be seen that $P(x), Q(x) \geq 0 \ \forall x \in [0, 1]$.

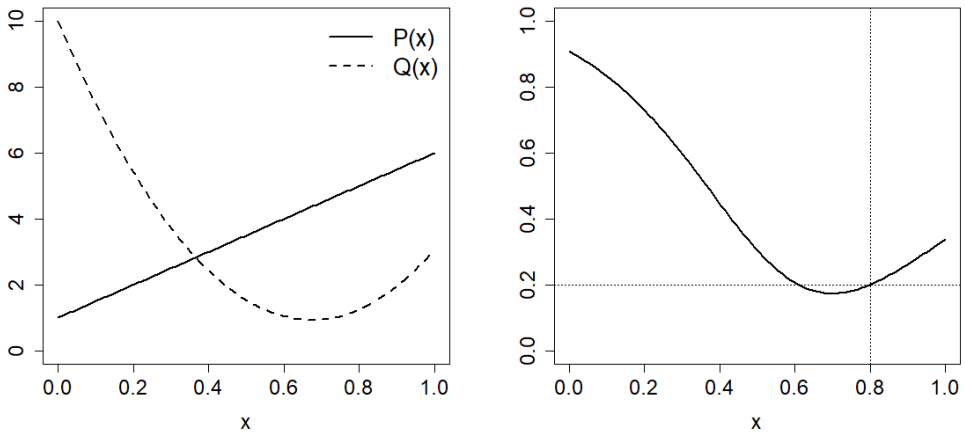

Figure 2.2: $P(x)$ (solid line) and $Q(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.8) = 0.2$ in Model 1.

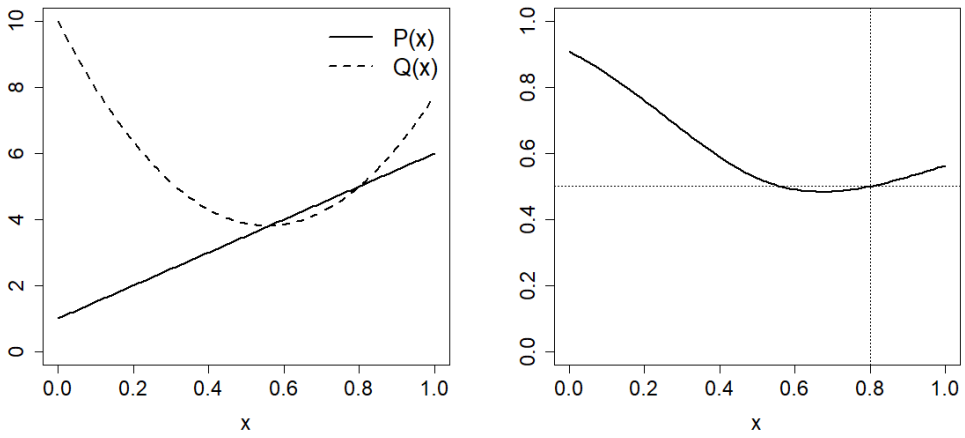

Figure 2.3: $P(x)$ (solid line) and $Q(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.8) = 0.5$ in Model 1.

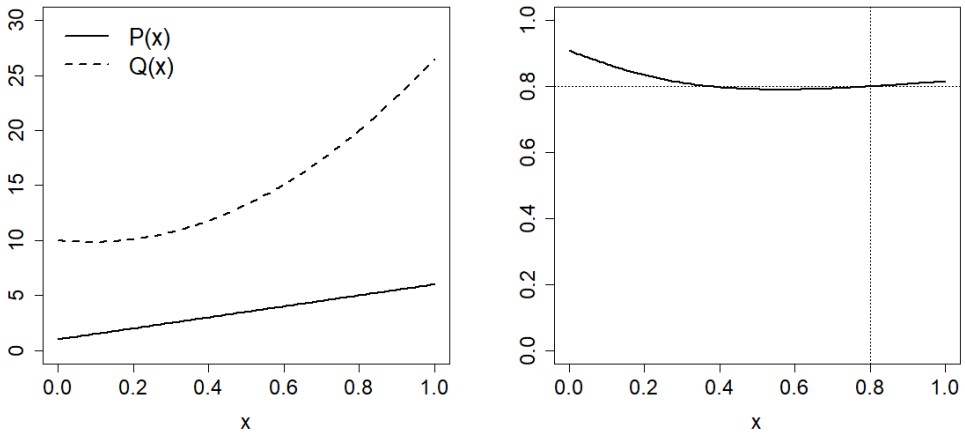

Figure 2.4: $P(x)$ (solid line) and $Q(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.8) = 0.8$ in Model 1.

The conditional survival function and the probability of default in Model 1 are estimated in a time grid of size n_T, $0 < t_1 < \cdots < t_{n_T}$, where $t_{n_T} + b = F^{-1}(0.95|x)$ for the value of the covariable $x = 0.8$. For the previously set parameters, one has $b = 0.1$ (20% of the grid range) and $t_{n_T} = 0.4991$.

Figure 2.5 shows the theoretical conditional survival function and the probability of default for this model under the above conditions.

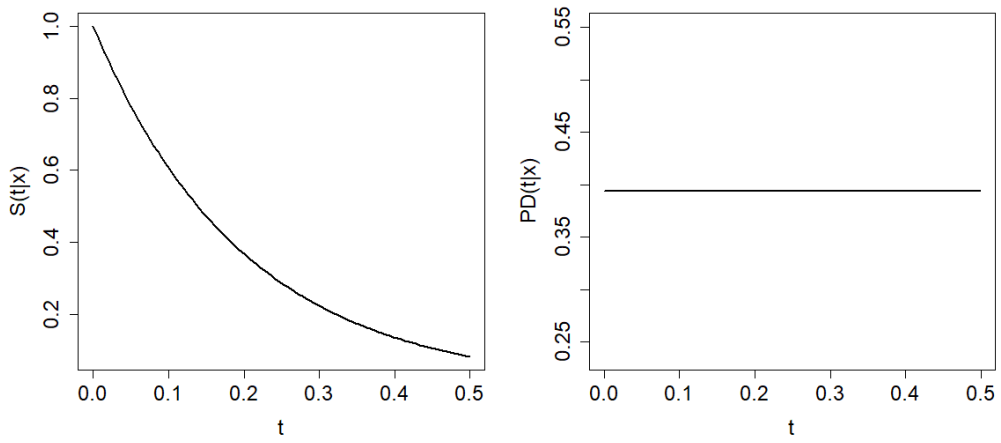

Figure 2.5: Theoretical conditional survival function $S(t|x)$ (left) and probability of default $PD(t|x)$ (right) in Model 1.

Model 2

Model 2 considers a $U(0,1)$ distribution for X. The time to default conditional to the credit scoring, $T|_{X=x}$, follows a Weibull distribution with parameters d and $\Gamma(x)^{-1/d}$, with $\Gamma(x) = c_0 + c_1 x$,

$$T|_{X=x} \sim \mathcal{W}(d, \Gamma(x)^{-1/d}),$$

and the censoring time conditional to the credit scoring follows a Weibull distribution with parameters d and $\Delta(x)^{-1/d}$, with $\Delta(x) = d_0 + d_1 x + d_2 x^2$,

$$C|_{X=x} \sim \mathcal{W}(d, (\Delta(x))^{-1/d}).$$

In this case, the conditional survival function and the probability of default are given by:

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

The censoring conditional probability is obtained as follows:

$$\begin{aligned} P(\delta = 0|X = x) &= H_0(\infty|X = x) = \int_0^\infty \Big(1 - F(u|x)\Big) dG(u|x) = \\ &= \int_0^\infty e^{-\Gamma(x)t^d} d\Big(1 - e^{-\Delta(x)t^d}\Big) = \int_0^\infty d\Delta(x) t^{d-1} e^{-\Gamma(x)t^d} e^{-\Delta(x)t^d} dt \\ &= -\frac{\Delta(x)}{\Gamma(x) + \Delta(x)} e^{-(\Gamma(x)+\Delta(x))t^d} \Big]_0^\infty = \frac{\Delta(x)}{\Gamma(x) + \Delta(x)} \end{aligned}$$

and the unconditional probability of censoring is given by

$$P(\delta = 0) = \int_{-\infty}^{+\infty} P(\delta = 0|X = x) m(x) dx.$$

The polynomials Γ and Δ used in this model are $\Gamma(x) = 1 + 5x$ and $\Delta(x) = 10 + d_1 x + 20x^2$. Having set the value of the credit scoring, $x = 0.6$ the value of d_1 is chosen so that the censoring conditional probability is 0.2, 0.5 and 0.8. The resulting values of d_1 and the probabilities associated with them are shown in Table

d_1	$P(\delta = 0\|X = 0.6)$	$P(\delta = 0)$
-27	0.2	0.43
-22	0.5	0.51
-2	0.8	0.82

Table 2.2: Values of d_1 and the associated censoring probabilities for Model 2.

In addition, Figures 2.6, 2.7 and 2.8 show the plots of the resulting polynomials $C(x)$ and $D(x)$ along with the plot of $P(\delta = 0|X = x)$ in each case. It can be seen that $\Gamma(x), \Delta(x) \geq 0 \ \forall x \in [0, 1]$.

The conditional survival function and the probability of default of Model 2 are estimated in a time grid of size n_T, $0 < t_1 < \cdots < t_{n_T}$, where $t_{n_T} + b = F^{-1}(0.95|x)$ for the value of the covariable $x = 0.6$. For the previously set parameters, one has $b = 0.15$ (20% of the time grid) and $t_{n_T} = 0.7154$.

Figure 2.9 show the theoretical conditional survival function and probability of default for this model under the above conditions.

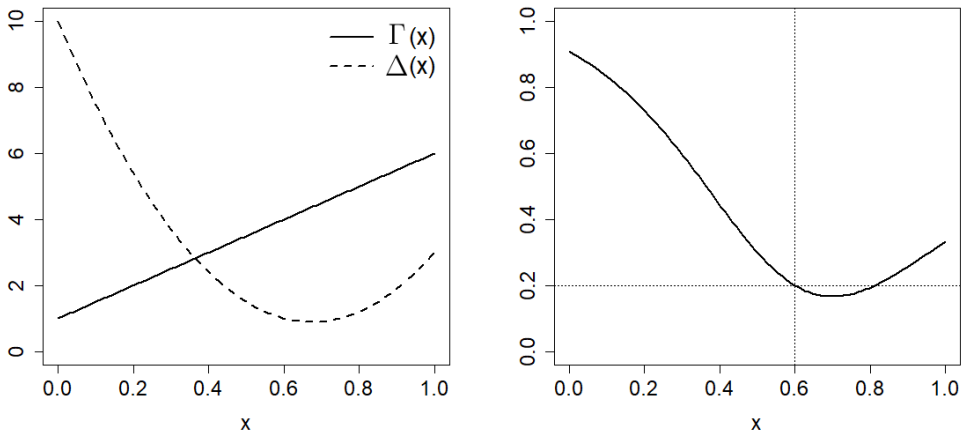

Figure 2.6: $\Gamma(x)$ (solid line) and $\Delta(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.6) = 0.2$ in Model 2.

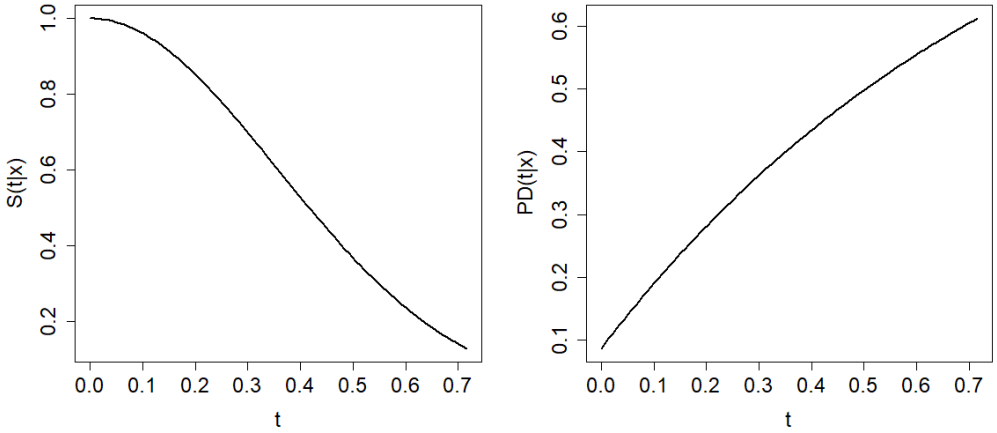

Figure 2.9: Theoretical conditional survival function $S(t|x)$ (left) and probability of default $PD(t|x)$ (right) in Model 2.

Model 1 and Model 2 are close to Cox models. By definition, the hazard function can be expressed in terms of the conditional survival function as follows:

$$\Lambda(t|x) = \int_0^t \lambda(s|x)ds = \int_0^t \frac{f(s|x)}{S(s|x)}ds = -\ln\Big(S(s|x)\Big)\Big]_{s=0}^{s=t} = -\ln\Big(S(t|x)\Big).$$

Then, $S(t|x) = \exp\Big(-\Lambda(t|x)\Big)$.

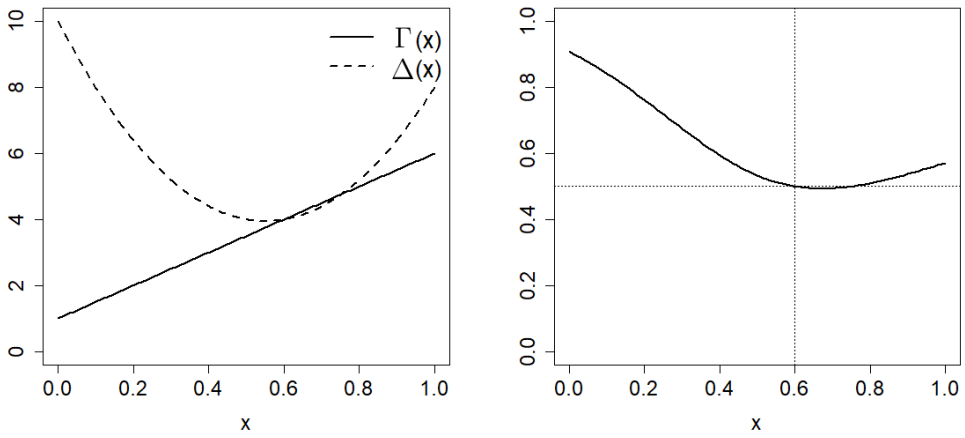

Figure 2.7: $\Gamma(x)$ (solid line) and $\Delta(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.6) = 0.5$ in Model 2.

Assuming a proportional hazards model, we have $\Lambda(t|x) = \Lambda_0(t)\exp(x^t\beta)$, thus,

$$S(t|x) = \exp\Big(-\Lambda_0(t)\exp(x^t\beta)\Big) = \Big(\exp\Big(-\Lambda_0(t)\Big)\Big)^{\exp(x^t\beta)} = S_0(t)^{\exp(x^t\beta)}.$$

Therefore, on a Cox model, the conditional survival function can be factorized such that $S(t|x) = S_0(t)^{\exp(x^t\beta)}$, where $S_0(t)$ is a survival function that does not depend on x.

Since the conditional survival function of Model 1 can be expressed as $S(t|x) = S_0(t)^{P(x)}$ by defining $S_0(t) = e^{-t}$ which is the survival function of the distribution $Exp(1)$, Model 1 is as close to a Cox model as the polynomial $P(x)$ is to a function of the form $\exp(\beta_0 + \beta_1 x)$.

If we consider $\beta_0 = 0$ and $\beta_1 = \ln 6$, then the polynomial $P(x) = 1 + 5x$ and its derivative coincide respectively with the expression $\exp(\beta_0 + \beta_1 x)$ and its derivative in both $x = 0$ and $x = 1$. Hence, the survival function of Model 1, $S(t|x) = e^{-P(x)t}$, is very close to the survival function of a Cox model.

The survival function of Model 2 is $S(t|x) = e^{-\Gamma(x)t^d}$ and it is also very close to the survival function of a Cox model, since it can be expressed as $S(t|x) = S_0(t)^{\Gamma(x)}$, by just defining $S_0(t) = e^{-t^d}$.

The polynomial $\Gamma(x)$ is identical to $P(x)$, so the above argument is valid for it,

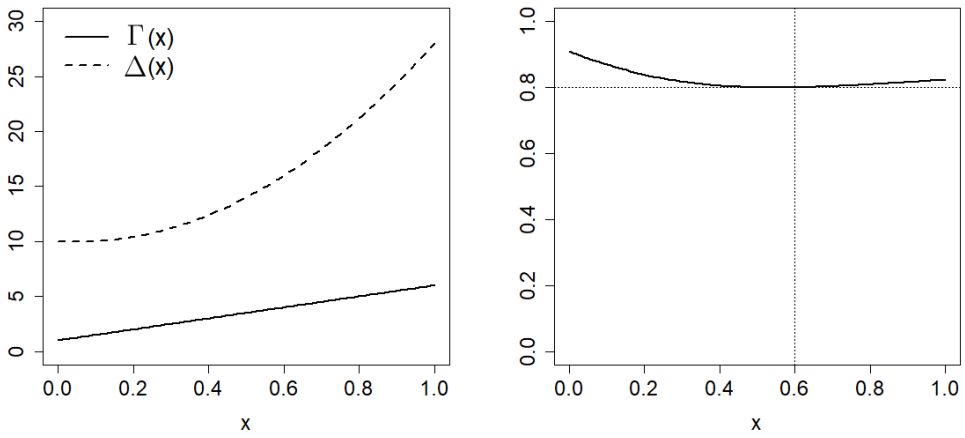

Figure 2.8: $\Gamma(x)$ (solid line) and $\Delta(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.6) = 0.8$ in Model 2.

concluding that $\Gamma(x)$ is very similar to a function of the form $\exp(\beta_0+\beta_1 x)$ choosing $\beta_0 = 0$ and $\beta_1 = \ln 6$. On the other hand, it is clear that $S_0(t) = e^{-t^d}$ is a survival function since $S_0(t) \geq 0$ for all $t \geq 0$ and $e^{-t^d} > e^{-v^d}$ if $0 \leq t < v$. Thus, Model 2 is also close to a Cox model.

With the intention that the simulations carried out to study the behaviour of the estimators cover widely clear scenarios, a third model which is far from a Cox model is introduced.

Model 3

For Model 3, a $U(0, 1)$ distribution is considered for the credit scoring, X. The time to default conditional to the credit scoring, $T|_{X=x}$, follows an exponential distribution of parameter $R(x)$,

$$T|_{X=x} \sim Exp\Big(R(x)\Big),$$

and the censoring time conditional to the credit scoring, $C|_{X=x}$, follows an exponential distribution with parameter $Q(x) = b_0 + b_1 x + b_2 x^2$,

$$C|_{X=x} \sim Exp\Big(Q(x)\Big).$$

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

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

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

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

Model 3 is as far from a Cox model as the polynomial $R(x)$ is from a function of the form $\exp(\beta_0 + \beta_1 x)$. Then, chosing $R(x) = 2 + 58x - 160x^2 + 107x^3$, Model 3 is far from a Cox model.

The distribution of the censoring time is defined by the polynomial $Q(x) = 10 + b_1 x + 20x^2$. Having set the value of the credit scoring, $x = 0.8$, the value of b_1 is chosen so that the censoring conditional probability is 0.2, 0.5 and 0.8. The resulting values for b_1 are shown in Table 2.3.

b_1	$P(\delta = 0\|X = 0.8)$	$P(\delta = 0)$
-113/4	0.2	0.268
-55/2	0.5	0.374
-123/5	0.8	0.534

Table 2.3: Values of b_1 and the associated censoring probabilities for Model 3.

The resulting polynomials $R(x)$ and $Q(x)$ along with the plot of $P(\delta = 0|X = x)$ are shown in Figures 2.10, 2.11 and 2.12. It can be seen that $R(x), Q(x) \geq 0$ $\forall x \in [0, 1]$.

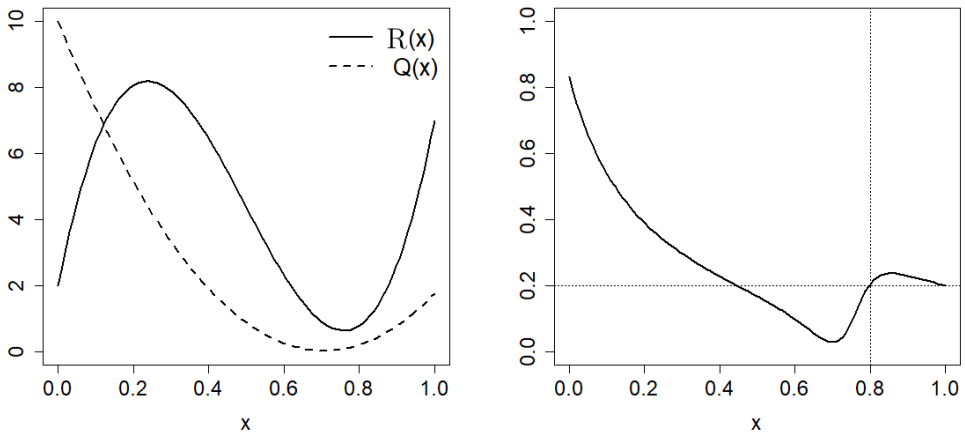

Figure 2.10: $R(x)$ (solid line) and $Q(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.8) = 0.2$ in Model 3.

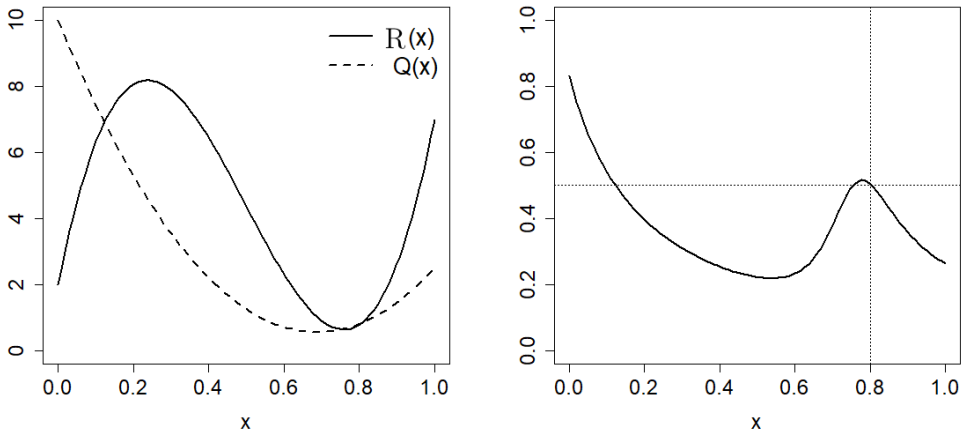

Figure 2.11: $R(x)$ (solid line) and $Q(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.8) = 0.5$ in Model 3.

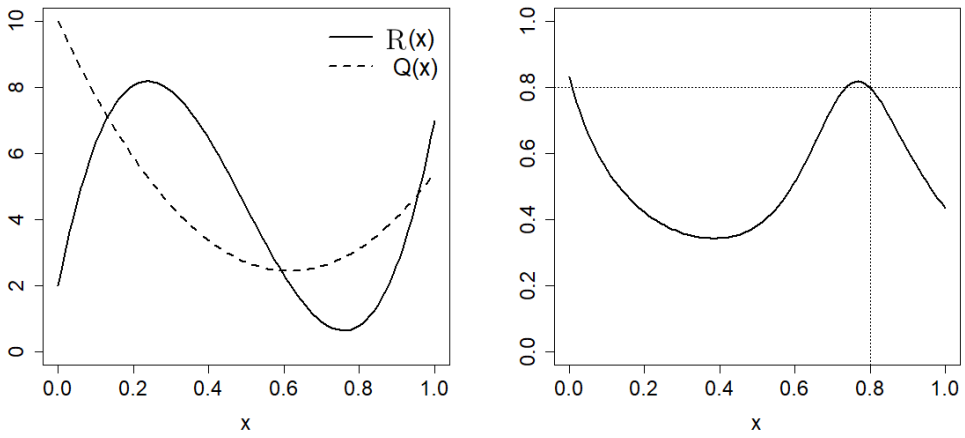

Figure 2.12: $R(x)$ (solid line) and $Q(x)$ (dashed line) in the left panel and $P(\delta = 0|X = x)$ in the right panel when $P(\delta = 0|X = 0.8) = 0.8$ in Model 3.

The conditional survival function and the probability of default in Model 3 are estimated in a time grid of size n_T, $0 < t_1 < \cdots < t_{n_T}$, where $t_{n_T} + b = F^{-1}(0.95|x)$ for the value of the covariable $x = 0.8$. For the previously set parameters, one has $b = 0.7$ (20% of the grid range) and $t_{n_T} = 3.1211$.

Figure 2.13 shows the theoretical conditional survival function and probability of default for this model under the above conditions.

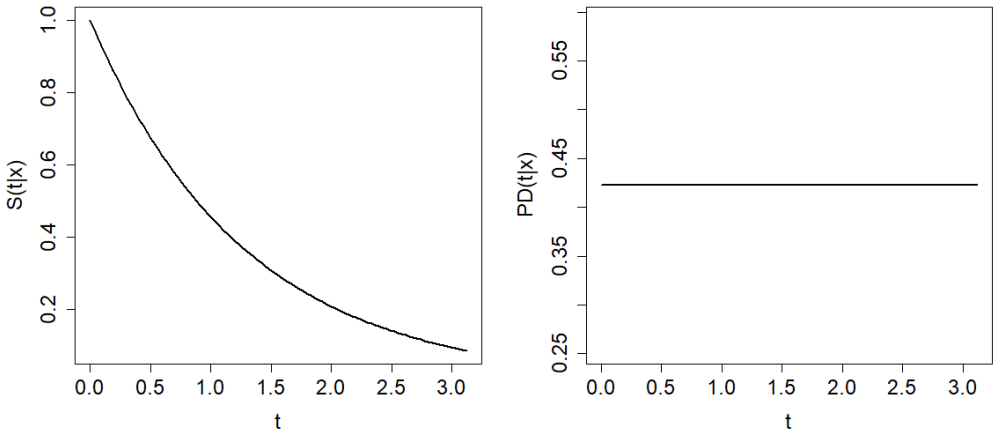

Figure 2.13: Theoretical conditional survival function $S(t|x)$ (left) and probability of default $PD(t|x)$ (right) in Model 2.

In the comparative study a parametric method is introduced as a benchmark method. The chosen technique is the Cox proportional hazards method which assumes that $S(t|x) = \exp(-\Lambda(t|x))$ and its estimation is obtained as follows

$$\widehat{S}^{PH}(t|x) = \exp(-\widehat{\Lambda}(t|x))$$

with

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{I_{\{Z_i \leq t, \delta_i = 1\}}}{\sum_{i=1}^{n} I_{\{Z_j \geq Z_i\}}}$$

and

$$\widehat{\beta} = \arg\max L(\beta)$$

where

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(x_i^t \beta)}{\sum_{i=1}^{n} I_{\{Z_j > Z_i\}} \exp(x_j^t \beta)}$$

is the likelihood function. This idea for estimating $S(t|x)$ in this context was introduced by Naraim (1992) and here it is applied to obtain a probability of default estimator by replacing $\widehat{S}_h(t|x)$ in (2.2) by $\widehat{S}^{PH}(t|x)$. The R package *survival* is used to obtain the PD estimations by this method (see Therneau (2015) for more details).

Note that Model 1 is close to a proportional hazards model, while Models 2 and 3 move away from this parametric model. For this reason, Cox method is supposed to have a reasonable behaviour in Model 1 but worse in Models 2 and 3.

The truncated Gaussian kernel with a truncation range $(-50, 50)$ is used, the sample size is $n = 400$, and the size of the lifetime grid is $n_T = 100$. The WLL estimator is corrected so that the estimations of the PD that it provides are contained in $[0, 1]$, simply setting the value 1 for $\widehat{PD}_h^{WLL}(t|x)$ if it is greater than 1 or the value 0 if it is negative. In addition, the boundary effect is corrected for all the estimators as explained below.

The boundary effect is corrected at $t = 0$ for the variable Z, since the time variable must be positive. The boundary effect causes an overestimation of $F(t|x)$ (and consequently an underestimation of $S(t|x)$) at values of $t > 0$ close to zero. In Silverman (1986) the reflection principle is proposed to correct the boundary effect in the estimation of a density function with compact support. This method consists

of obtaining $\widehat{f}(t)$ for positive t and setting the value of $\widehat{f}(t)$ to zero for all negative t. Once this is done, the estimator has to be corrected so that the estimated density integrates one. Then, positive values of $\hat{f}(t)$ for $t < 0$ are used to increase the estimation of $f(-t)$. This idea is carried over to the distribution function to obtain an estimator, $\widehat{F}^c(t|x)$, which corrects the boundary effect at zero as follows:

$$\widehat{F}^c(t|x) = \begin{cases} 0 & \text{si } t < 0, \\ \widehat{F}(t|x) - \widehat{F}(-t|x) & \text{si } t \geq 0, \end{cases}$$

where $\widehat{F}(t|x)$ is a nonparametric uncorrected estimator of the distribution function. Consequently, an estimator of the survival function, $\widehat{S}^c(t|x)$, which avoids the underestimation of $S(t|x)$ around $t = 0$, for $t > 0$ is given by:

$$\widehat{S}^c(t|x) = \begin{cases} 1 & \text{si } t < 0, \\ 1 + \widehat{S}(t|x) - \widehat{S}(-t|x) & \text{si } t \geq 0, \end{cases}$$

where $\widehat{S}(t|x) = 1 - \widehat{F}(t|x)$.

For every estimator, the optimal smoothing parameter h_{MISE} is selected as the value which minimises, in a grid of 50 bandwidth values, a Monte Carlo approximation of the MISE:

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

based on $N = 50$ simulated samples, with the integral approximated in a grid of the interval $[0, 1]$. Of course, this bandwidth cannot be used in practice, but this choice produces a fair comparison since the four estimators are constructed using their best possible bandwidths. The smoothing parameter chosen for the Van Keilegom-Akritas estimator is the optimal one for estimating the conditional distribution function by means of Beran's estimator, which is a previous step in the estimation of the PD with this technique. The value of $MISE$ using this smoothing parameter is approximated from $N = 1000$ simulated samples for every estimator and used, along with its square root ($RMISE$), as a measure of the estimation error. Tables 2.4, 2.5 and 2.6 show the estimation errors for Model 1, 2 and 3.

In some of the scenarios analysed, $MISE_x(h)$ turned out to be a decreasing function of h. For this reason, the MISE bandwidth selected was a high but reasonable value, considering that the variable X moves in the interval $[0, 1]$. This is the case for the MISE function for the WLL and WNW estimators in Model 1 and in Model 2 when the conditional probability of censoring is 0.8.

In Models 1 and 2, Beran's estimator provides a smaller error than Cox method in most of the cases, even though these models are close to a Cox model. The VKA estimator is competitive with Beran's estimator and Cox model in these scenarios. They are followed by the WNW estimator, which works significantly better than WLL. In Model 3, the smallest estimator error is the one coming from Beran's estimator in all cases. In this model, the VKA estimator is not competitive with Beran's not even when the censoring conditional probability is 0.8.

The higher the censoring probability, the greater the error is for any of the estimators. However, Beran's estimator behaves reasonably well in all scenarios. The WLL and WNW estimators present a much larger error than the rest of the estimators when the censoring probability increases in all models. This is more evident for the WLL estimator.

			Beran	WLL	WNW	VKA	Cox
$P(\delta = 0\|x = 0.8)$	0.2	h_{MISE}	0.24286	1.00000	1.00000	0.14285	—
		$MISE_x(h_{MISE})$	0.00398	0.00590	0.00439	0.00978	0.00850
		$RMISE_x(h_{MISE})$	0.06311	0.07681	0.06624	0.09888	0.09218
	0.5	h_{MISE}	0.39592	1.00000	1.00000	0.15000	—
		$MISE_x(h_{MISE})$	0.01129	0.03493	0.02291	0.01835	0.01312
		$RMISE_x(h_{MISE})$	0.10626	0.18689	0.15137	0.13546	0.11454
	0.8	h_{MISE}	0.42857	1.00000	1.00000	0.22143	—
		$MISE_x(h_{MISE})$	0.04379	0.10352	0.07567	0.04290	0.04054
		$RMISE(h_{MISE})$	0.20925	0.32175	0.27508	0.20711	0.20134

Table 2.4: Optimal bandwidth, $MISE$ and $RMISE$ of the PD estimators for each level of censoring conditional probability in Model 1.

			Beran	WLL	WNW	VKA	Cox
$P(\delta = 0\|x = 0.6)$	0.2	h_{MISE}	0.30204	0.43780	0.39898	0.24385	—
		$MISE_x(h_{MISE})$	0.00296	0.00493	0.00493	0.00543	0.00491
		$RMISE_x(h_{MISE})$	0.05441	0.07021	0.07021	0.07369	0.07006
	0.5	h_{MISE}	0.34082	0.51529	0.98064	0.22449	—
		$MISE_x(h_{MISE})$	0.01254	0.02871	0.02808	0.01731	0.01271
		$RMISE_x(h_{MISE})$	0.11198	0.16944	0.16757	0.13157	0.11274
	0.8	h_{MISE}	0.39898	1.00000	1.00000	0.22449	—
		$MISE_x(h_{MISE})$	0.06623	0.12551	0.11111	0.06424	0.06298
		$RMISE_x(h_{MISE})$	0.25735	0.35427	0.33333	0.25346	0.25097

Table 2.5: Optimal bandwidth, $MISE$ and $RMISE$ of the PD estimator for each level of censoring conditional probability and each estimator for Model 2.

			Beran	WLL	WNW	VKA	Cox
$P(\delta = 0\|x = 0.8)$	0.2	h_{MISE}	0.09898	0.09082	0.08571	0.04551	—
		$MISE_x(h_{MISE})$	0.07201	0.14970	0.14490	0.09989	0.19534
		$RMISE_x(h_{MISE})$	0.26835	0.38691	0.38066	0.31605	0.44197
	0.5	h_{MISE}	0.13163	0.10612	0.11122	0.05735	—
		$MISE_x(h_{MISE})$	0.20260	0.46798	0.46215	0.26169	0.30251
		$RMISE_x(h_{MISE})$	0.45011	0.68409	0.67982	0.51155	0.55001
	0.8	h_{MISE}	0.15204	0.65100	0.15204	0.13429	—
		$MISE_x(h_{MISE})$	0.42281	0.66306	0.65951	0.67132	0.47346
		$RMISE_x(h_{MISE})$	0.65024	0.81429	0.81210	0.81934	0.68808

Table 2.6: Optimal bandwidth, $MISE$ and $RMISE$ of the PD estimator for each level of censoring conditional probability and each estimator for Model 3.

In a second study, the following curves are calculated for each estimator and each level of censoring conditional probability from $N = 1000$ simulated samples:

$$\Big\{(t_k, PD(t_k|x))\Big\}_{k=1}^{n_T},$$

$$\Big\{(t_k, \widehat{PD}^{(5)}_{h_{MISE}}(t_k|x))\Big\}_{k=1}^{n_T},$$

$$\left\{(t_k, \widehat{PD}^{(50)}_{h_{MISE}}(t_k|x))\right\}_{k=1}^{n_T},$$
$$\left\{(t_k, \widehat{PD}^{(95)}_{h_{MISE}}(t_k|x))\right\}_{k=1}^{n_T},$$

where $\widehat{PD}^{(j)}(t_k|x)$ denotes the j-th percentile of all the PD estimations obtained in time t_k.

In this section the WLL estimator is not taken into account since the results obtained for it are similar but worse than those obtained with the WNW estimator. On the other hand, a very high censoring probability and a not very high sample size, such as that handled here, lead to not very accurate PD estimations, so the third censoring scenario ($P(\delta = 0|x) = 0.8$) is excluded. The resulting curves for each model are shown in the following figures.

Figures 2.14, 2.15 and 2.16 show that the greater the censoring conditional probability, the worse the obtained estimations are. Also the greater the time value in which PD is estimated, the larger the error is for the three estimators for Models 1, 2 and 3. The 50th percentile of the estimations that best fits the true probability of default curve is the one obtained with Beran's estimator.

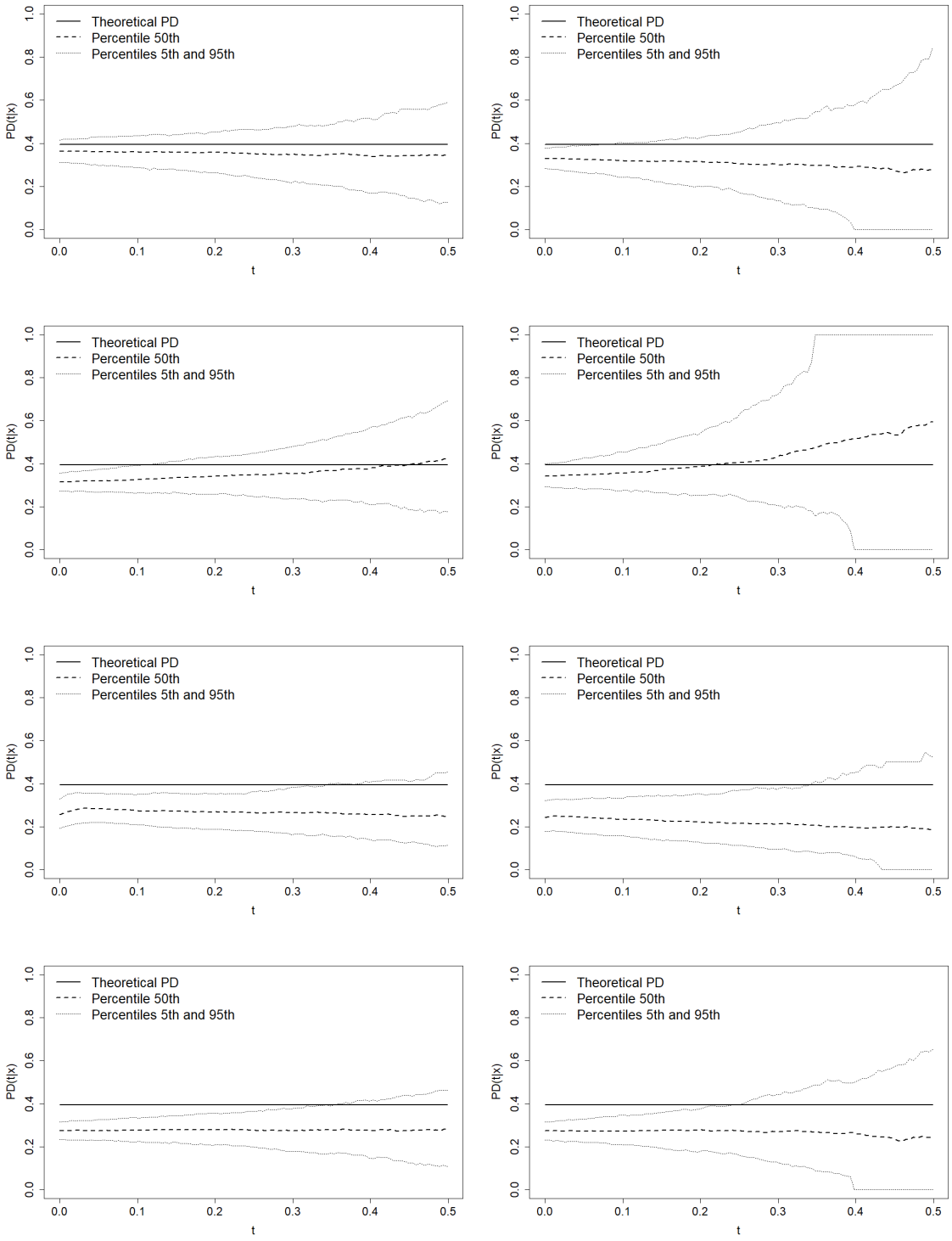

Figure 2.14: Theoretical $PD(t|x)$ (solid line), 50th percentile (dashed line) and 5th and 95th percentiles (dotted lines) obtained by means of Beran's (top), WNW (second), VKA (third) and Cox (bottom) for $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.5$ (right) for Model 1.

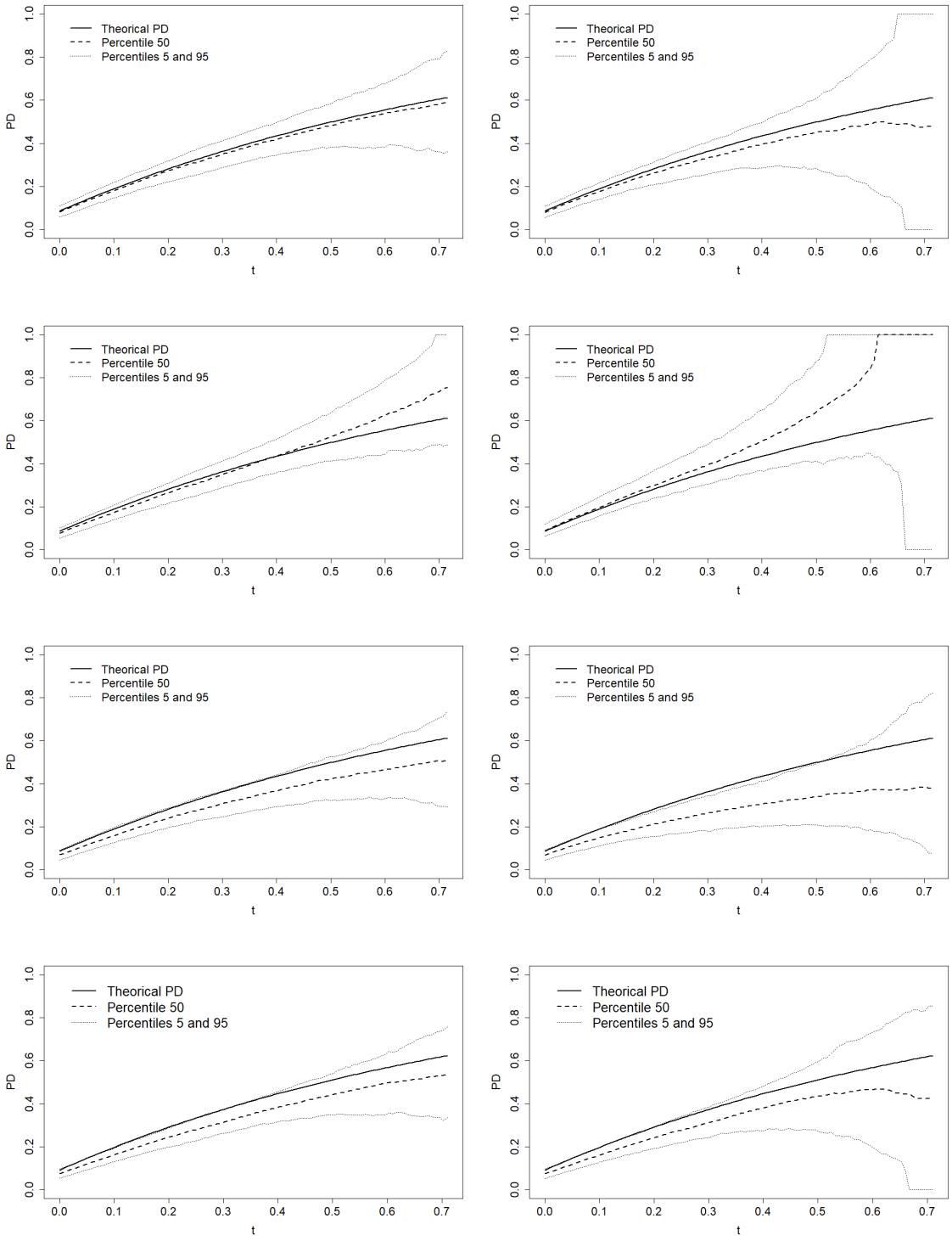

Figure 2.15: Theoretical $PD(t|x)$ (solid line), 50th percentile (dashed line) and 5th and 95th percentiles (dotted lines) obtained by means of Beran's (top), WNW (second), VKA (third) and Cox (bottom) for $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.5$ (right) for Model 2.

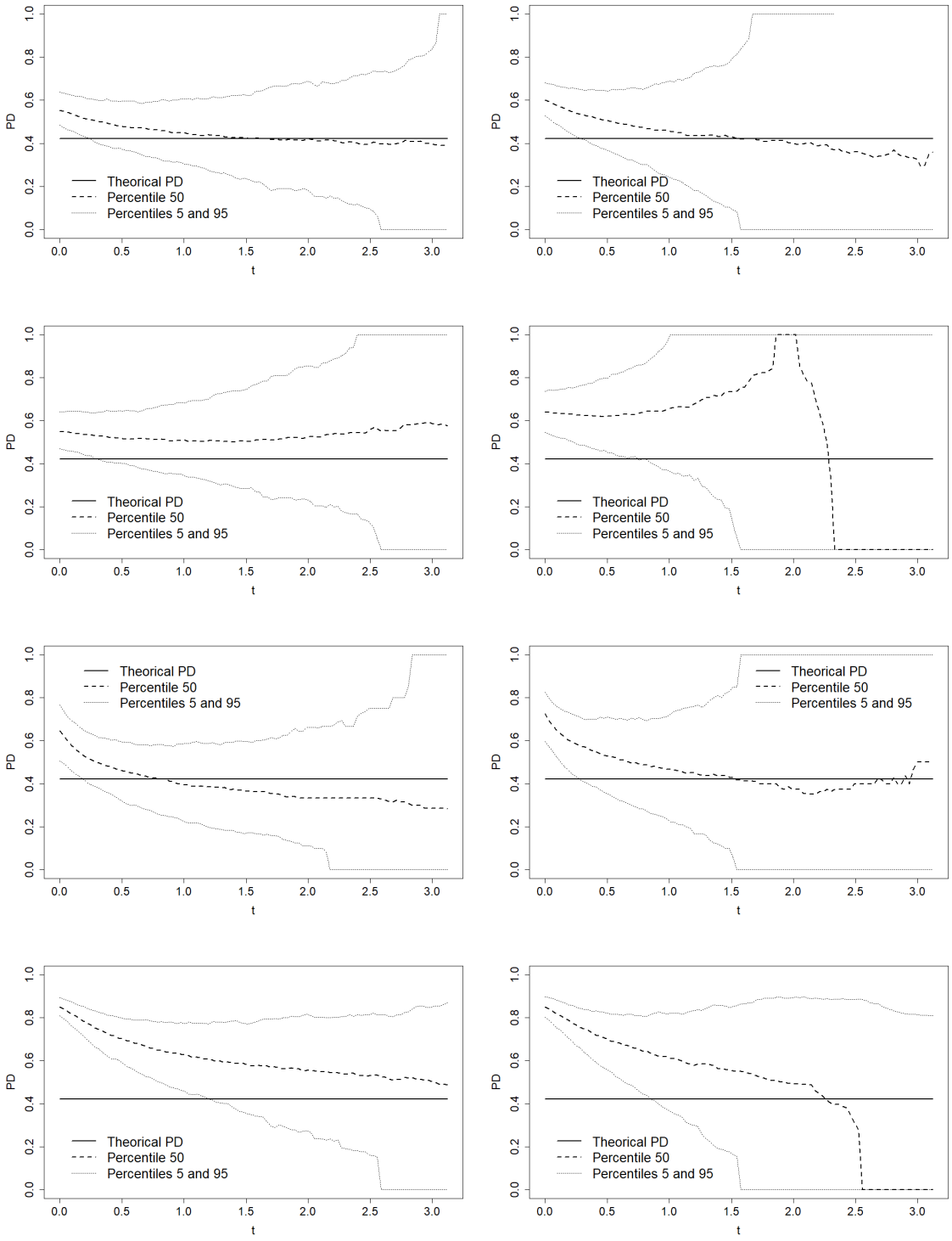

Figure 2.16: Theoretical $PD(t|x)$ (solid line), 50th percentile (dashed line) and 5th and 95th percentiles (dotted lines) obtained by means of Beran's (top), WNW (second), VKA (third) and Cox (bottom) for $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.5$ (right) for Model 3.

In Cai (2003) and Van Keilegom and Akritas (1999), it is proven that the con-

ditional survival function estimators given in (2.6) and (2.8) improve on Beran's estimator when the survival function is estimated in the right tail of the time distribution, most remarkably for heavy censoring. It is certainly of interest to check whether the PD estimators derived from these survival function estimators inherit this good feature.

The following paragraphs discuss the behaviour of these four estimators (along with the benchmark method) when estimating the probability of default in the right tail of the distribution through a simulation study similar to the previous one. The parameters of each model and simulation conditions remained, but the range of the time variable where the PD is estimated is changed. The aim is to obtain the optimal smoothing parameter, h_{MISE}, and approximate the value of $MISE(h_{MISE})$ when $PD(t|x)$ is estimated in a grid within the interval $[t_{0.7}, t_{0.95}]$, where t_α denotes the value of time that satisfies $F(t_\alpha + b|x) = \alpha$. For Model 1, these values are $t_{0.7} = 0.1408$ and $t_{0.95} = 0.4992$. For Model 2, they are $t_{0.7} = 0.3986$ and $t_{0.95} = 0.7154$. For Model 3, they are $t_{0.7} = 0.8357$ and $t_{0.95} = 3.1211$.

The results are shown in Tables 2.7, 2.8 and 2.9. In some of the scenarios analysed, $MISE(h)$ turned out to be a decreasing function of h. For this reason, the MISE bandwidth selected was a high but reasonable value, considering that the variable X moves in the interval $[0, 1]$. This is the case for the MISE function for the WLL and WNW estimators in Models 1 and 2 and for Beran's estimator in Model 3 when the conditional probability of censoring is 0.8.

In Model 1, WNW estimator provides the smallest estimation error when the censoring probability is low. When the censoring probability increases, Beran's estimator, VKA estimator and Cox method behave reasonably well. The VKA estimator provides the smallest error in that case.

When the censoring probability is low in Model 2, all the estimators have a similar behaviour and Beran's is the one which provides the smallest estimation error. In this case, Van Keilegom-Akritas estimator is not competitive with Cox method. When increasing the censoring probability, the estimation error of WLL

and WNW is higher than the rest and the best estimations are obtained by the VKA estimator.

In Model 3, for $P(\delta = 0|x = 0.8) = 0.2$, Beran's estimator provides the lowest estimation error by far. When increasing the censoring probability, the VKA estimator presents the highest mean integrated squared error. Although all estimators have a bad behaviour in this context, Beran's estimator is the most reasonable one, being competitive with Cox method.

			Beran	WLL	WNW	VKA	Cox
$P(\delta = 0\|x = 0.8)$	0.2	h_{MISE}	0.26327	1.50000	1.50000	0.13571	—
		$MISE_x(h_{MISE})$	0.00356	0.00508	0.00351	0.00775	0.00644
		$RMISE_x(h_{MISE})$	0.05969	0.07127	0.05924	0.08804	0.08024
	0.5	h_{MISE}	0.47959	1.50000	1.50000	0.15204	—
		$MISE_x(h_{MISE})$	0.01037	0.03376	0.02234	0.01474	0.01093
		$RMISE_x(h_{MISE})$	0.10181	0.18374	0.14947	0.12140	0.10457
	0.8	h_{MISE}	0.53469	1.50000	1.50000	0.21939	—
		$MISE_x(h_{MISE})$	0.04074	0.07773	0.06882	0.03619	0.03768
		$RMISE_x(h_{MISE})$	0.20184	0.27879	0.26234	0.19025	0.19412

Table 2.7: Optimal bandwidth, $MISE$ and $RMISE$ of the PD estimation in the right tail of the distribution for each level of censoring conditional probability and for each estimator in Model 1.

			Beran	WLL	WNW	VKA	Cox
$P(\delta=0\|x=0.6)$	0.2	h_{MISE}	0.3020	0.2827	0.2633	0.2245	—
		$MISE_x(h_{MISE})$	0.00259	0.00454	0.00454	0.00443	0.00381
		$RMISE_x(h_{MISE})$	0.05089	0.06738	0.06738	0.06656	0.06172
	0.5	h_{MISE}	0.3602	0.9418	1.000	0.2051	—
		$MISE_x(h_{MISE})$	0.01132	0.02678	0.02625	0.01454	0.01096
		$RMISE_x(h_{MISE})$	0.10640	0.16365	0.16202	0.12058	0.10469
	0.8	h_{MISE}	0.4959	1.0000	1.0000	0.2245	—
		$MISE_x(h_{MISE})$	0.06259	0.08916	0.08781	0.05511	0.05932
		$RMISE_x(h_{MISE})$	0.25018	0.29859	0.29633	0.23476	0.24359

Table 2.8: Optimal bandwidth, $MISE$ and $RMISE$ of the PD estimation in the right tail of the distribution for each level of censoring conditional probability and for each estimator in Model 2.

			Beran	WLL	WNW	VKA	Cox
$P(\delta=0\|x=0.8)$	0.2	h_{MISE}	0.1514	0.1312	0.1110	0.0504	—
		$MISE_x(h_{MISE})$	0.06131	0.13114	0.12891	0.09135	0.11269
		$RMISE_x(h_{MISE})$	0.24761	0.36213	0.35905	0.30224	0.33568
	0.5	h_{MISE}	0.2322	0.8384	0.1716	0.0706	—
		$MISE_x(h_{MISE})$	0.18619	0.41102	0.40943	0.23851	0.21622
		$RMISE_x(h_{MISE})$	0.43150	0.64111	0.63987	0.48837	0.46499
	0.8	h_{MISE}	1.0000	0.9592	0.2322	0.1857	—
		$MISE_x(h_{MISE})$	0.38754	0.42825	0.42824	0.55738	0.38660
		$RMISE_x(h_{MISE})$	0.62253	0.65441	0.65441	0.74658	0.62177

Table 2.9: Optimal bandwidth, $MISE$ and $RMISE$ of the PD estimation in the right tail of the distribution for each level of censoring conditional probability and for each estimator in Model 3.

Figures 2.17, 2.18 and 2.19 show in more detail the behaviour of the four estimators in the right tail of the distribution for the highest censoring conditional probability. The estimation of the PD is obtained in three fixed values of time, $t_{0.7}$, $t_{0.8}$ and $t_{0.95}$, from $N = 1000$ simulated samples and the boxplots of the esti-

mations are shown. It is easy to see that the performance of Beran's and the Van Keilegom-Akritas estimators is remarkably better than the performance of WLL and WNW estimators, taking into account that, in the right tail of the distribution with heavy censoring, the information provided by the data is very reduced. The VKA estimator is competitive in this context.

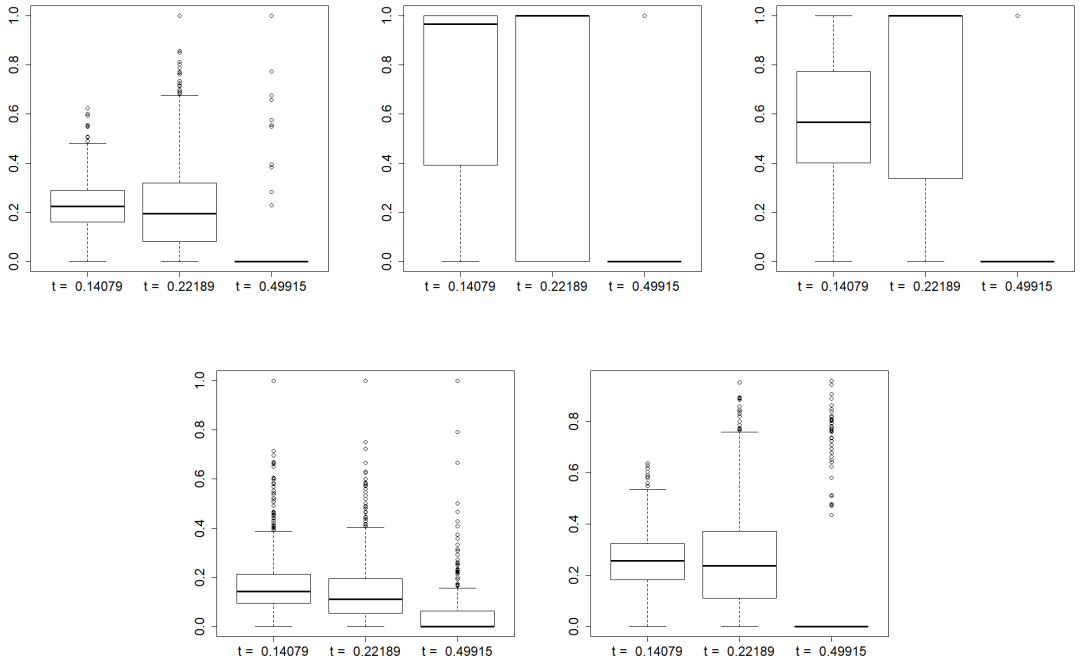

Figure 2.17: Boxplot of the estimations of $PD(t|x)$ obtained by Beran's, WLL, WNW, VKA and Cox estimator (from left to right) for $t = t_{0.7}, t_{0.8}, t_{0.95}$ when $P(\delta = 0|x) = 0.8$ for Model 1.

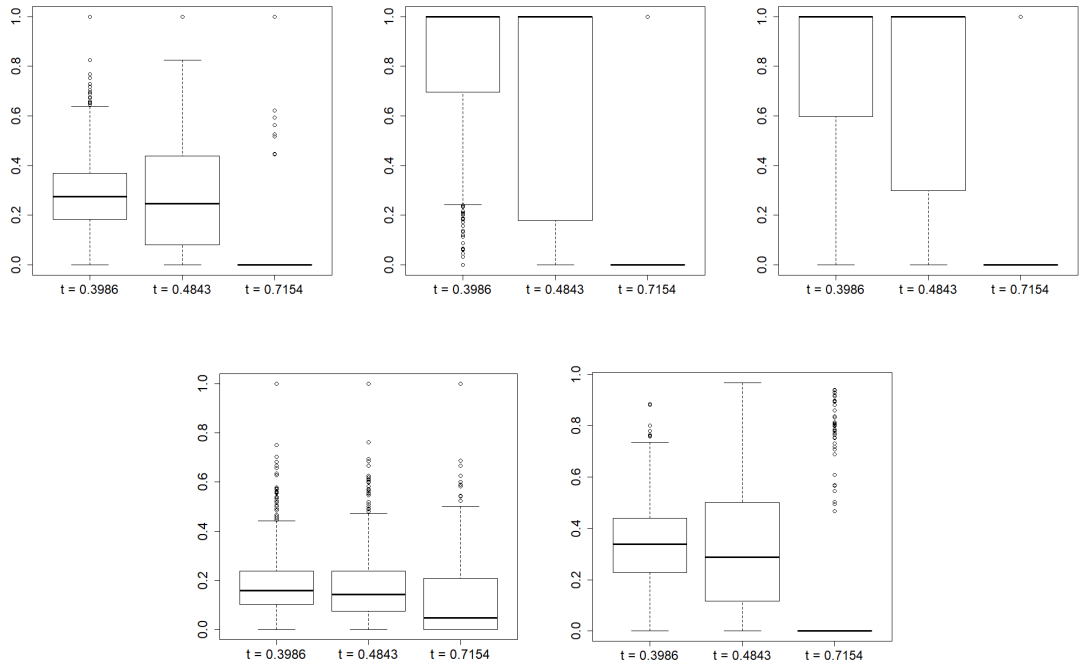

Figure 2.18: Boxplot of the estimations of $PD(t|x)$ obtained by Beran's, WLL, WNW, VKA and Cox estimator (from left to right) for $t = t_{0.7}, t_{0.8}, t_{0.95}$ when $P(\delta = 0|x) = 0.8$ for Model 2.

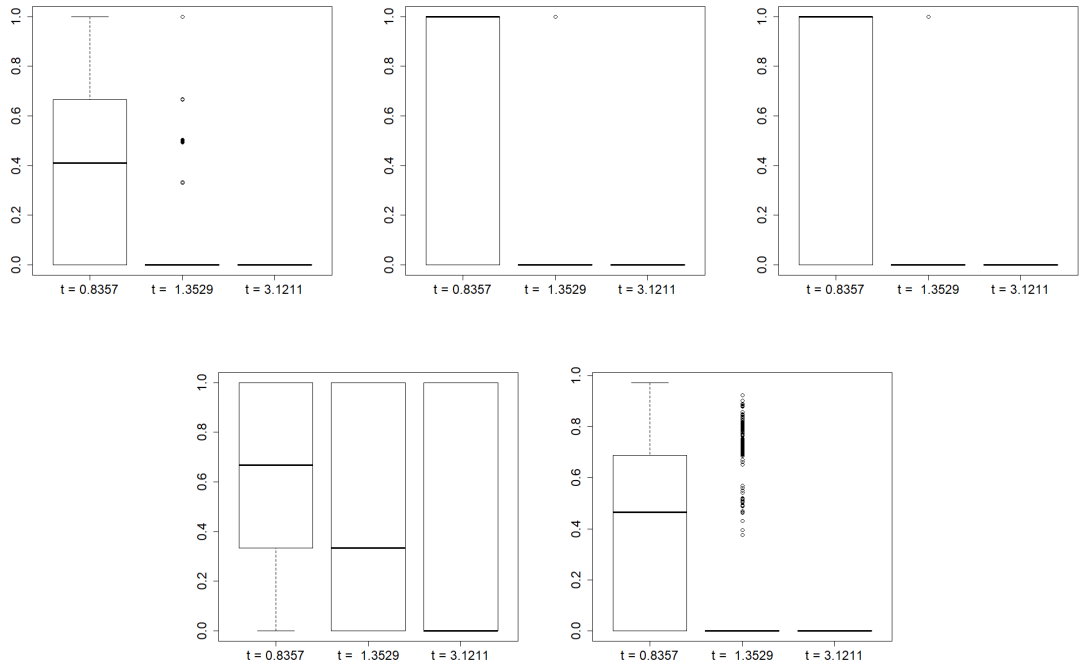

Figure 2.19: Boxplot of the estimations of $PD(t|x)$ obtained by Beran's, WLL, WNW, VKA and Cox estimator (from left to right) for $t = t_{0.7}, t_{0.8}, t_{0.95}$ when $P(\delta = 0|x) = 0.8$ for Model 3.

Another important aspect of the estimators which must be considered is their computation time. Table 2.10 shows the CPU times (in seconds) that each of the estimators spends in obtaining an estimation of the probability of default curve in a 100-point time's grid and a fixed value of x for different values of the sample size.

Beran's estimator is barely affected by the increase of the sample size and it is the fastest of the four studied estimators. Its CPU time is practically equal to the CPU time of the parametric method. The following one is the WNW estimator which CPU time is similar to that of the WLL estimator, but it is slightly faster. The slowest and most affected by the increase of the sample size is the VKA estimator.

n	Beran	WLL	WNW	VKA	Cox
50	0.02	0.05	0.05	0.27	0.01
100	0.02	0.07	0.06	1.04	0.01
200	0.02	0.09	0.08	5.55	0.01
400	0.02	0.19	0.17	28.50	0.02
1200	0.03	0.85	0.77	526.43	0.03

Table 2.10: CPU time for the estimation of $PD(t|x)$ in a time grid of size 100 for every estimator and different sample sizes.

2.5 Application to real data

The estimation methods given in previous sections are now applied to a real data set. The data consists of a sample of 10,000 consumer credits from a Spanish bank registered between July 2004 and November 2006. They were previously used in Cao et al. (2009). To obtain each client'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 of credits for which the default is observed is 7.2%. An intentionally biased subsample was obtained from the original sample, so as not to show the true solvency situation of the bank and thus preserve confidentiality. The variables considered are the following ones:

- X: it is the credit scoring observed for each borrower; its range lies inside the interval $[0, 1]$ and the higher its value, the greater solvency the debtor has.
- Z: it is the observed lifetime of the credit; it is measured in months and it takes values between 0 and 30,
- δ: it is the uncensoring indicator; it is equal to one when the default is observed.

Table 2.11 shows some summary statistics of the data. Figures 2.20 and 2.21 show the histograms of the observed lifetime and credit scoring variables. In all of them, data from censored and uncensored (and therefore defaulted) credits are distinguished.

Note that censored credits, which have not fallen into default during the study, have higher lifetimes and higher credit scorings. This is reasonable since a client with greater solvency will continue paying his or her credit longer and it will be more difficult to observe the default. See also that the credit scoring values, although they are higher in the censored group of credits, are generally high. This may be due to the fact that they correspond to credits actually granted by the financial company.

		min.	$1^{st}Q.$	median	mean	$3^{th}Q.$	max.
Censored group	Z	0.00	6.73	11.23	13.37	19.86	29.50
	X	0.56	0.89	0.95	0.92	0.98	0.99
Uncensored group	Z	0.03	2.97	5.35	7.54	11.44	24.77
	X	0.22	0.58	0.70	0.68	0.79	0.91

Table 2.11: Summary statistics for lifetime (Z) and credit scoring (X) for the uncensored group (defaulted credits) and the censored group.

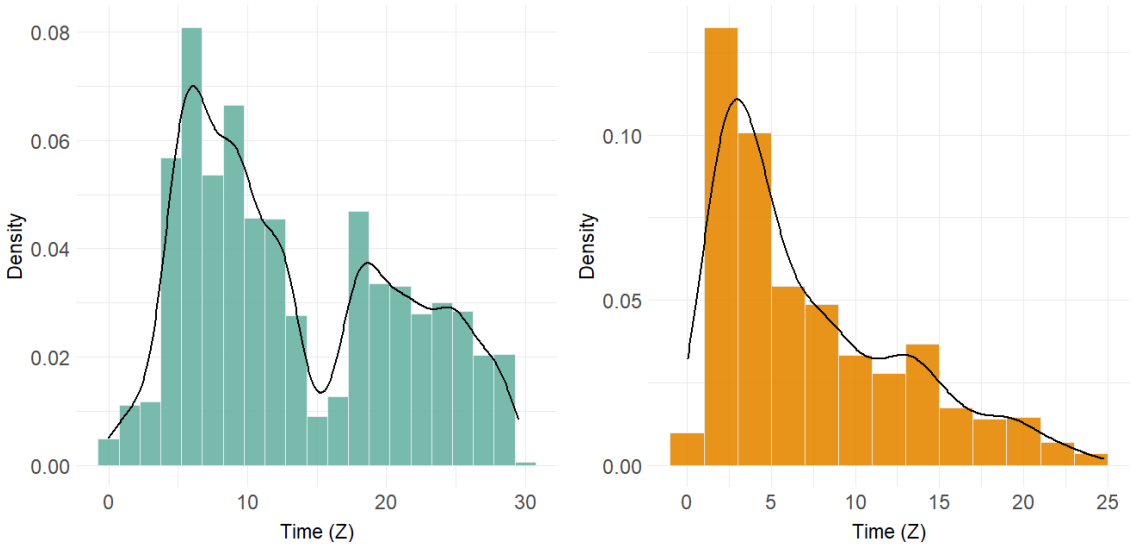

Figure 2.20: Histogram and kernel density estimation of the observed lifetime for the censored sample (left) and the uncensored sample (right).

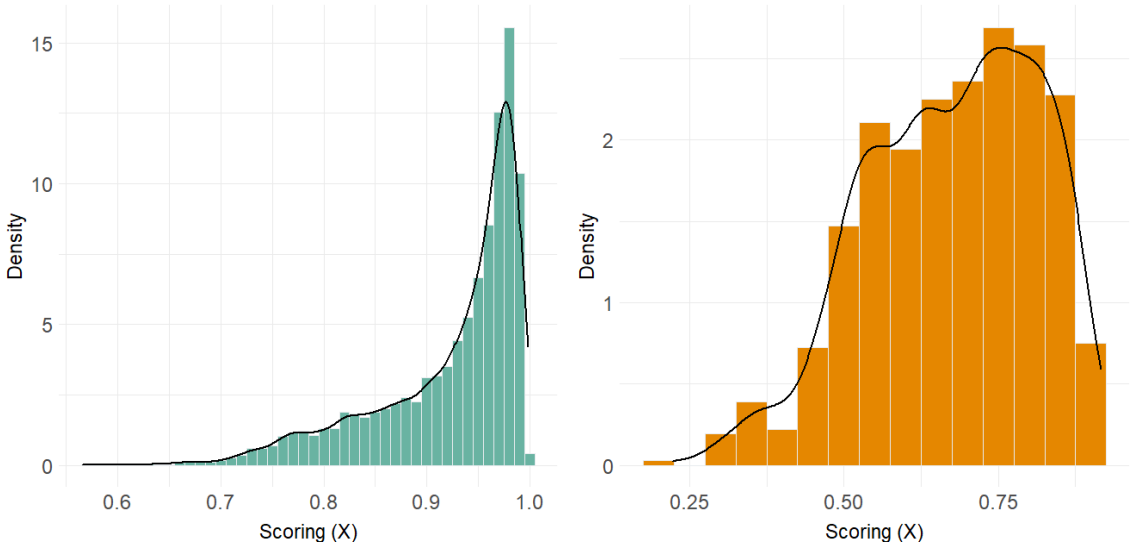

Figure 2.21: Histogram and kernel density estimation of the credit scoring for the censored sample (left) and the uncensored sample (right).

Next, the estimation of the probability of default for $x = 0.8$ at horizon $b = 5$ months is obtained, in a time grid along the interval $[0, 25]$, using the four estimators presented in Section 2.2 and the benchmark method used in Section 2.4.

The probability of default was estimated by each method with some different

possible values of the smoothing parameter in order to evaluate their influence in the estimation, which turned out to be very slight, and choose a reasonable one. The chosen bandwidths are $h = 0.05$ for Beran's estimator, $h = 0.4$ for the WLL estimator, $h = 0.01$ for WNW and $h = 0.2$ for the VKA estimator. The estimations obtained are shown in Figure 2.22.

It could be thought that Beran's PD estimation hardly presents variability or jumps, unlike the rest of the estimators. However, it is simply a scale factor. Figure 2.23 shows the estimated PD obtained using Beran's estimator in the personal credit dataset.

Since in this case the censoring is heavy (92.8%), the VKA estimator should be the most reliable of all of them, although the simulations showed that Beran's estimator was also accurate. According to Beran's estimation (Figure 2.23), the probability of default has a decreasing tendency and it is close to zero at all points. It follows from the first fact that the probability of falling into default is reduced while the debt maturity is increasing. The second fact is reasonable, given that the probability of default is being calculated for a considerably higher value of the covariate, which indicates a greater solvency of the borrower.

In practice, the financial institution measures different features of its clients, such as age, amount of money in the bank account, salary, years of employment, etc. They summarize, usually by logistic regression, these covariates into the single variable credit scoring. Subsequently, techniques such as those shown in this work allow the calculation of the probability of default at horizon b for all of them. The curve $PD(t|x)$ provides the probability that the client will default after a certain period of time b.

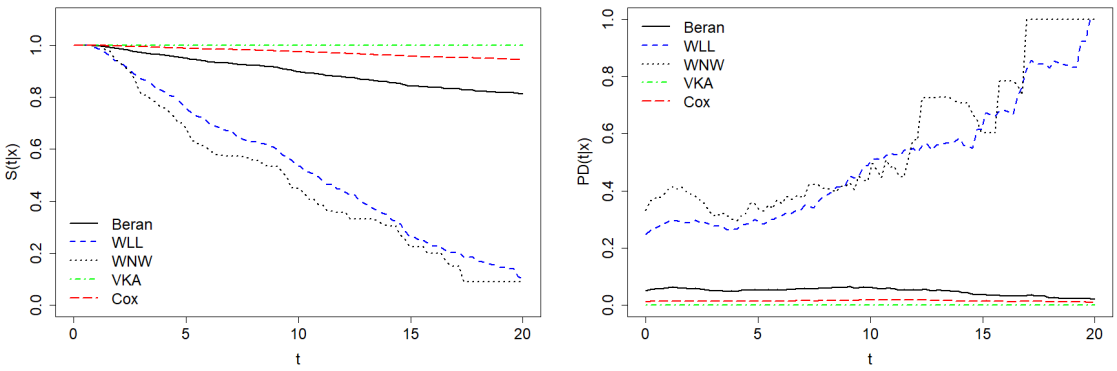

Figure 2.22: Estimation of $S(t|x)$ (left) and $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.8$ by means of Beran's (solid line), WLL (dashed line), WNW (dotted line), VKA (dash-dotted line) and Cox (long dashed line) estimators for the consumer credits dataset.

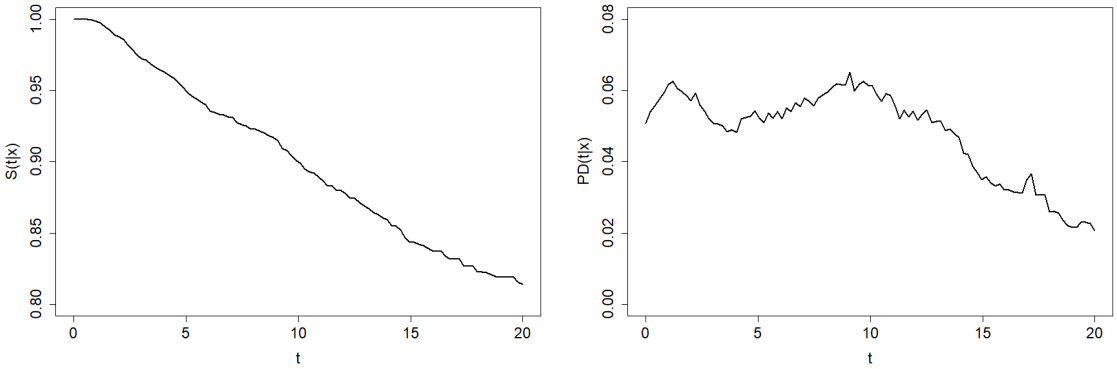

Figure 2.23: Estimation of $S(t|x)$ (left) and $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.8$ by means of Beran's estimator on the consumer credits dataset.

2.6 Proofs

Proof of Theorem 2.1

Denote $PD(t|x) = 1 - \dfrac{P}{Q}$ with $P = S(t+b|x)$ and $Q = S(t|x)$ and $\widehat{PD}(t|x) = 1 - \dfrac{\widehat{P}}{\widehat{Q}}$ with $\widehat{P} = \widehat{S}(t+b|x)$ and $\widehat{Q} = \widehat{S}(t|x)$. The following equation will be useful at some points along the proof:

$$\frac{1}{z} = 1 - (z-1) + \cdots + (-1)^p (z-1)^p + (-1)^{(p+1)} \frac{(z-1)^{(p+1)}}{z} . \tag{2.11}$$

First, an asymptotic expression of the bias of $\widehat{PD}(t|x)$ will be obtained. For $p=1$ and $z=\dfrac{\widehat{Q}}{E(\widehat{Q})}$, Equation (2.11) gives:

$$\begin{aligned}\frac{\widehat{P}}{\widehat{Q}} &= \frac{\widehat{P}}{E(\widehat{Q})}\frac{E(\widehat{Q})}{\widehat{Q}} = \frac{\widehat{P}}{E(\widehat{Q})}\left(1-\left(\frac{\widehat{Q}}{E(\widehat{Q})}-1\right)+\frac{E(\widehat{Q})}{\widehat{Q}}\left(\frac{\widehat{Q}}{E(\widehat{Q})}-1\right)^2\right) = \\ &= \frac{\widehat{P}}{E(\widehat{Q})}-\frac{\widehat{P}\big(\widehat{Q}-E(\widehat{Q})\big)}{E(\widehat{Q})^2}+\frac{\widehat{P}}{\widehat{Q}}\frac{\big(\widehat{Q}-E(\widehat{Q})\big)^2}{E(\widehat{Q})^2}.\end{aligned}$$

Taking expectations,

$$\begin{aligned}E\left(\frac{\widehat{P}}{\widehat{Q}}\right) &= \frac{E(\widehat{P})}{E(\widehat{Q})}-\frac{E\big[\widehat{P}\big(\widehat{Q}-E(\widehat{Q})\big)\big]}{E(\widehat{Q})^2}+\frac{E\left[\frac{\widehat{P}}{\widehat{Q}}\big(\widehat{Q}-E(\widehat{Q})\big)^2\right]}{E(\widehat{Q})^2} = \\ &= \frac{E(\widehat{P})}{E(\widehat{Q})}-\frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2}+\frac{E\left[\frac{\widehat{P}}{\widehat{Q}}\big(\widehat{Q}-E(\widehat{Q})\big)^2\right]}{E(\widehat{Q})^2}.\end{aligned} \tag{2.12}$$

On the other hand,

$$Bias\big(\widehat{PD}(t|x)\big) = E\left(1-\frac{\widehat{P}}{\widehat{Q}}\right)-\left(1-\frac{P}{Q}\right) = \frac{P}{Q}-E\left(\frac{\widehat{P}}{\widehat{Q}}\right).$$

Consequently,

$$Bias\big(\widehat{PD}(t|x)\big) = \alpha_1+\alpha_2+\alpha_3, \tag{2.13}$$

where $\alpha_1 = \dfrac{P}{Q}-\dfrac{E(\widehat{P})}{E(\widehat{Q})}$, $\alpha_2 = \dfrac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2}$ and $\alpha_3 = -\dfrac{E\left[\frac{\widehat{P}}{\widehat{Q}}\big(\widehat{Q}-E(\widehat{Q})\big)^2\right]}{E(\widehat{Q})^2}$.

Using standard algebra and Condition C.2 gives

$$\begin{aligned}\alpha_1 &= \frac{P}{Q}-\frac{P+B_0(t+b|x)h^2+o(h^2)}{Q+B_0(t|x)h^2+o(h^2)} \\ &= \frac{PB_0(t|x)h^2-QB_0(t+b|x)h^2+o(h^2)}{Q(Q+B_0(t|x)h^2+o(h^2))} \\ &= \frac{P/QB_0(t|x)h^2-B_0(t+b|x)h^2+o(h^2)}{Q+B_0(t|x)h^2+o(h^2)} \\ &= \frac{(1-PD(t|x))B_0(t|x)-B_0(t+b|x)}{S(t|x)}h^2+o(h^2).\end{aligned} \tag{2.14}$$

$$\alpha_2 = \frac{C_0(t,t+b|x)\dfrac{1}{nh}+o\left(\dfrac{1}{nh}\right)}{\big(Q+B_0(t|x)h^2+o(h^2)\big)^2} = O\left(\frac{1}{nh}\right)+o(h^2), \tag{2.15}$$

$$\begin{aligned}
\alpha_3 &= -\frac{E\Big[\frac{\widehat{P}}{\widehat{Q}}\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2} = -\frac{E\Big[\frac{P}{Q}\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2} - \frac{E\Big[\Big(\frac{\widehat{P}}{\widehat{Q}}-\frac{P}{Q}\Big)\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2} \\
&= -\frac{\Big(1-PD(t|x)\Big)Var(\widehat{Q})}{E(\widehat{Q})^2} - \frac{E\Big[\Big(\widehat{PD}(t|x)-PD(t|x)\Big)\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2} \\
&= \alpha_{31}+\alpha_{32},
\end{aligned}$$

where

$$\begin{aligned}
\alpha_{31} &= -\frac{\Big(1-PD(t|x)\Big)Var(\widehat{Q})}{E(\widehat{Q})^2} = -\frac{\Big(1-PD(t|x)\Big)V_0(t|x)\frac{1}{nh}+o\Big(\frac{1}{nh}\Big)}{\Big(Q^2+B_0(t|x)h^2+o(h^2)\Big)} \\
&= O\Big(\frac{1}{nh}\Big)
\end{aligned}$$

and

$$\begin{aligned}
|\alpha_{32}| &= \left|\frac{E\Big[\Big(\widehat{PD}(t|x)-PD(t|x)\Big)\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2}\right| \\
&\leq \frac{E\Big[\Big|\widehat{PD}(t|x)-PD(t|x)\Big|\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2} \leq \frac{E\Big[\Big(\widehat{Q}-E(\widehat{Q})\Big)^2\Big]}{E(\widehat{Q})^2} = \frac{Var(\widehat{Q})}{E(\widehat{Q})^2} \\
&= \frac{V_0(t|x)\frac{1}{nh}+o\Big(\frac{1}{nh}\Big)}{\Big(Q^2+B_0(t|x)h^2+o(h^2)\Big)} = O\Big(\frac{1}{nh}\Big).
\end{aligned}$$

Therefore,

$$\alpha_3 = O\left(\frac{1}{nh}\right). \tag{2.16}$$

Finally plugging (2.14), (2.15) and (2.16) into (2.13) the bias part in Theorem 2.1 is proven.

Next, an asymptotic expression for the variance will be found. To do this, equation (2.11) is used with $p=3$ and $z=\dfrac{\widehat{Q}^2}{E(\widehat{Q})^2}$:

$$\begin{aligned}
\frac{E(\widehat{Q})^2}{\widehat{Q}^2} &= 1+\sum_{i=1}^{3}(-1)^i\left(\frac{\widehat{Q}^2}{E(\widehat{Q})^2}-1\right)^i + \frac{(\widehat{Q}^2/E(\widehat{Q})^2-1)^4}{\widehat{Q}^2/E(\widehat{Q})^2} \\
&= 1+\sum_{i=1}^{3}(-1)^i\left(\frac{\widehat{Q}^2-E(\widehat{Q})^2}{E(\widehat{Q})^2}\right)^i + \left(\frac{\widehat{Q}^2-E(\widehat{Q})^2}{E(\widehat{Q})^2}\right)^4\frac{E(\widehat{Q})^2}{\widehat{Q}^2}.
\end{aligned} \tag{2.17}$$

Note that,

$$\begin{aligned}(\widehat{Q}-E(\widehat{Q}))^2 &= \widehat{Q}^2-2\widehat{Q}E(\widehat{Q})+E(\widehat{Q})^2+E(\widehat{Q})^2-E(\widehat{Q})^2\\ &= \widehat{Q}^2-2\widehat{Q}E(\widehat{Q})-E(\widehat{Q})^2+2E(\widehat{Q})^2\\ &= \widehat{Q}^2-E(\widehat{Q})^2-2E(\widehat{Q})(\widehat{Q}-E(\widehat{Q})),\end{aligned}$$

implies $\widehat{Q}^2-E(\widehat{Q})^2=(\widehat{Q}-E(\widehat{Q}))^2+2E(\widehat{Q})(\widehat{Q}-E(\widehat{Q}))$. Using this equation along with Newton's binomial formula, it is possible to obtain:

$$\begin{aligned}\left(\frac{\widehat{Q}^2-E(\widehat{Q})^2}{E(\widehat{Q})^2}\right)^i &= \left(\frac{(\widehat{Q}-E(\widehat{Q}))^2+2E(\widehat{Q})(\widehat{Q}-E(\widehat{Q}))}{E(\widehat{Q})^2}\right)^i\\ &= \left(\frac{(\widehat{Q}-E(\widehat{Q}))^2}{E(\widehat{Q})^2}+\frac{2E(\widehat{Q})(\widehat{Q}-E(\widehat{Q}))}{E(\widehat{Q})^2}\right)^i\\ &= \sum_{j=0}^{i}\binom{i}{j}\left(\frac{(\widehat{Q}-E(\widehat{Q}))^2}{E(\widehat{Q})^2}\right)^j\left(\frac{2E(\widehat{Q})(\widehat{Q}-E(\widehat{Q}))}{E(\widehat{Q})^2}\right)^{i-j}\\ &= \sum_{j=0}^{i}\binom{i}{j}\frac{(\widehat{Q}-E(\widehat{Q}))^{2j}}{E(\widehat{Q})^{2j}}\cdot\frac{2^{i-j}(\widehat{Q}-E(\widehat{Q}))^{i-j}}{E(\widehat{Q})^{i-j}}\\ &= \sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}(\widehat{Q}-E(\widehat{Q}))^{i+j}}{E(\widehat{Q})^{i+j}}.\end{aligned}$$

which is used to replace in expression (2.17), obtaining:

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

Thus, it is possible to calculate

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

$$
\begin{aligned}
&= \frac{E(\widehat{P}^2) - E(\widehat{P})^2 + E(\widehat{P})^2}{E(\widehat{Q})^2} + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j} E(\widehat{P}^2(\widehat{Q} - E(\widehat{Q})^{i+j}))}{E(\widehat{Q})^{i+j+2}} \\
&\quad + \sum_{j=0}^{4} \binom{4}{j} \frac{2^{4-j} E\Big(\frac{\widehat{P}^2}{\widehat{Q}^2}\big(\widehat{Q} - E(\widehat{Q})\big)^{4+j}\Big)}{E(\widehat{Q})^{4+j}} \\
&= \frac{E\big(\widehat{P}^2 - E(\widehat{P})^2\big)}{E(\widehat{Q})^2} + \frac{E(\widehat{P})^2}{E(\widehat{Q})^2} + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i} \binom{i}{j} \frac{2^{i-j} E(\widehat{P}^2(\widehat{Q} - E(\widehat{Q})^{i+j}))}{E(\widehat{Q})^{i+j+2}} \\
&\quad + \sum_{j=0}^{4} \binom{4}{j} \frac{2^{4-j} E\Big(\frac{\widehat{P}^2}{\widehat{Q}^2}\big(\widehat{Q} - E(\widehat{Q})\big)^{4+j}\Big)}{E(\widehat{Q})^{4+j}}.
\end{aligned}
\tag{2.18}
$$

Let us now define:

$$
\begin{aligned}
A_{ij} &= E\Big[\big(\widehat{P} - E(\widehat{P})\big)^i \big(\widehat{Q} - E(\widehat{Q})\big)^j\Big], \\
B_{ij} &= E\Big[\widehat{P}^i \big(\widehat{Q} - E(\widehat{Q})\big)^j\Big], \\
C_i &= E(\widehat{Q})^i, \\
D_{ij} &= E\Big[\Big(1 - \frac{\widehat{P}}{\widehat{Q}}\Big)^i \big(\widehat{Q} - E(\widehat{Q})\big)^j\Big],
\end{aligned}
$$

for $i, j = 0, 1, \ldots$. It is easy to verify that $A_{0j} = B_{0j}$, $\forall\ j = 0, 1, \ldots$ and $B_{2j} = A_{2j} + 2B_{10}A_{1j} - B_{10}^2 A_{0j}$.

Replacing these equations in expression (2.18) it is obtained:

$$
\begin{aligned}
E\Big(\frac{\widehat{P}^2}{\widehat{Q}^2}\Big) &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{B_{2,i+j}}{C_{i+j+2}} + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2,4+j}}{C_{4+j}} \\
&= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3}(-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{A_{2,i+j} + 2B_{10}A_{1,i+j} - B_{10}^2 A_{0,i+j}}{C_{i+j+2}} \\
&\quad + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2,4+j}}{C_{4+j}}.
\end{aligned}
\tag{2.19}
$$

Using Condition C.3 it is possible to prove that, for $i \geq 3$,

$$
A_{i0} = E\Big[\big(\widehat{P} - E(\widehat{P})\big)^i\Big] = o\Big(\frac{1}{nh}\Big), \quad A_{0i} = B_{0i} = E\Big[\big(\widehat{Q} - E(\widehat{Q})\big)^i\Big] = o\Big(\frac{1}{nh}\Big)
$$

for $i + j \geq 3$, $A_{ij} = o\Big(\frac{1}{nh}\Big)$ and for $j \geq 3$, $B_{ij} = o\Big(\frac{1}{nh}\Big)$ and $D_{ij} = o\Big(\frac{1}{nh}\Big)$.

Moreover, $A_{01} = 0 = A_{10}$. Therefore,

$$\begin{aligned} E\left(\frac{\widehat{P}^2}{\widehat{Q}^2}\right) &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} - \frac{4B_{10}A_{11}}{C_3} - \frac{3B_{10}^2A_{02}}{C_4} + o\left(\frac{1}{nh}\right) \\ &= \frac{Var(\widehat{P})}{E(\widehat{Q})^2} + \frac{E(\widehat{P})^2}{E(\widehat{Q})^2} - \frac{4E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3} + \frac{3E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4} + o\left(\frac{1}{nh}\right). \end{aligned}$$

On the other hand, using (2.12),

$$\begin{aligned} E\left(\frac{\widehat{P}}{\widehat{Q}}\right) &= \frac{B_{10}}{C_1} - \frac{A_{11}}{C_2} + \frac{A_{12}+B_{10}A_{02}}{C_3} - \frac{A_{13}+B_{10}A_{03}}{C_4} + \frac{D_{14}}{C_4} \\ &= \frac{E(\widehat{P})}{E(\widehat{Q})} - \frac{Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^2} + \frac{E(\widehat{P})Var(\widehat{Q})}{E(\widehat{Q})^3} + o\left(\frac{1}{nh}\right). \end{aligned}$$

Then, an expression of $Var\big(\widehat{PD}(t|x)\big)$ is as follows:

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

Then, using Condition C.2,

$$Var\big(\widehat{PD}(t|x)\big) = \beta_1 + \beta_2 + \beta_3 + o\left(\frac{1}{nh}\right) \tag{2.20}$$

where $\beta_1 = \dfrac{Var(\widehat{P})}{E(\widehat{Q})^2}$, $\beta_2 = -2\dfrac{E(\widehat{P})Cov(\widehat{P},\widehat{Q})}{E(\widehat{Q})^3}$ and $\beta_3 = \dfrac{E(\widehat{P})^2Var(\widehat{Q})}{E(\widehat{Q})^4}$.

Straightforward but tedious calculations and Condition C.2 give

$$\beta_1 = \frac{V_0(t+b|x)}{S(t|x)^2}\frac{1}{nh} + o\left(\frac{1}{nh}\right), \tag{2.21}$$

$$\beta_2 = -\frac{2S(t+b|x)C_0(t,t+b|x)}{S(t|x)^3}\frac{1}{nh} + o\left(\frac{1}{nh}\right), \tag{2.22}$$

$$\beta_3 = \frac{S(t+b|x)^2V_0(t|x)}{S(t|x)^4}\frac{1}{nh} + o\left(\frac{1}{nh}\right). \tag{2.23}$$

Equations (2.21), (2.22) and (2.23) can be plugged into (2.20) to prove the variance part in Theorem 2.1.

□

Chapter 3

Doubly smoothed conditional survival estimation

3.1 Introduction

The probability of default estimations obtained by means of the estimators presented in Chapter 2 are very reasonable, but they have excessive variability and are very rough curves (see Figures 2.22 and 2.23). For the conditional survival function estimators which were analysed there, smoothing is only performed with respect to the covariate. They are step functions with respect to t, each jump ocurring at uncensored observed lifetimes. This fact, along with the survival ratio structure of the PD estimator (see Equation (2.2)), is the reason why the obtained curves are so unstable.

This is a phenomenon that can be observed even in a simpler scenario. Let us consider a simple random sample $T_1, ..., T_n$ of an uncensored time variable, T, for which no covariate is taken into account. In such a case, an estimation of the survival function of T, $S(t)$, could be obtained from the empirical distribution function, $F_n(t)$, as follows:

$$S_n(t) = 1 - F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i > t). \tag{3.1}$$

Then, a straightforward estimator of the probability of default would be given by:

$$PD_n(t) = 1 - \frac{S_n(t+b)}{S_n(t)}.$$

Despite being under a simpler scenario (no censoring and no covariate) and achieving a decent estimation of the survival function by means of the empirical distribution, the ratio of survivals at times t and $t+b$ causes jumps in the PD estimation such as those shown in Figure 3.1.

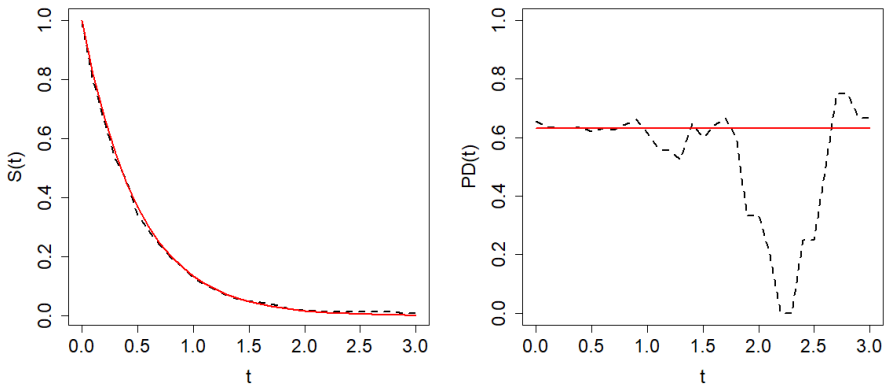

Figure 3.1: Theoretical curves (solid lines) and estimations (dashed lines) by means of the empirical distribution function of the survival function (left) and the probability of default (right) for a sample of size 400 from an $Exp(2)$.

The most commonly used nonparametric estimator of $F(t|x)$ under censoring was introduced by Beran (1981). This estimator turns out to be the Kaplan-Meier estimator (see Kaplan and Meier (1958)) in absence of covariates. Asymptotic properties of this estimator have been widely studied in the literature by Dabrowska (1989), González-Manteiga and Cadarso-Suárez (1994), Van Keilegom and Veraverbeke (1996) and Iglesias-Pérez and González-Manteiga (1999), among others. Another nonparametric estimator of the conditional distribution function with censored data was proposed by Van Keilegom and Akritas (1999) and Van Keilegom et al. (2001). It presents a better behaviour than Beran's estimator when estimating the distribution function in the right tail with heavy censoring. In Gannoun et al. (2005) and Gannoun et al. (2007) an alternative estimator based on the local linear method proposed in Cai (2003) was studied. All these nonparametric distribution estimators

are based on just covariate smoothing and will present the issue shown in Figure 3.1 for the estimation of the probability of default.

Our idea is to smooth somewhat the jumps that characterise the survival function estimator, so that they are not magnified in the PD estimation. The proposal, to be detailed in the following paragraphs, consists of obtaining a weighted average of the jumps that the survival estimator takes. In the uncensored and unconditional case above this would come to a weighted average of the jumps $1/n$ that the empirical survival estimator in (3.3) takes. Figure 3.2 shows the survival and PD estimated in this smoothed way. The large improvement over the empirical estimator shown in Figure 3.1 is evident.

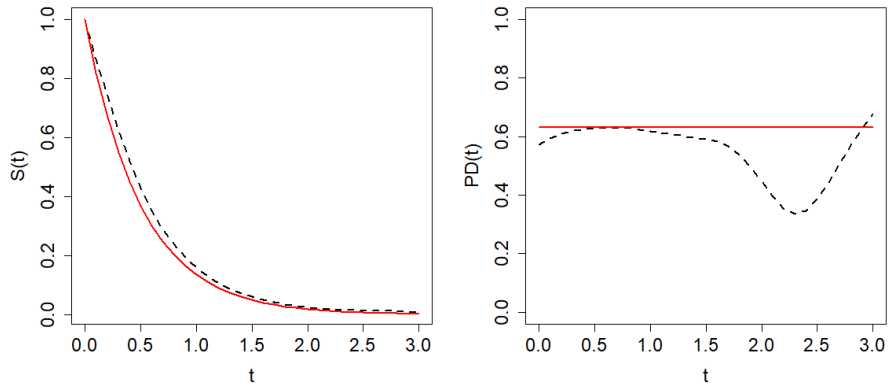

Figure 3.2: Theoretical curves (solid lines) and estimations (dashed lines) by means of the smoothed distribution estimator of the survival function (left) and the probability of default (right) for a sample of size 400 from an $Exp(2)$.

Time variable smoothing of the conditional survival function could be useful for the graphical representation, as well as to reduce the estimation error. In addition, as discussed above, it could be potentially useful in estimating the probability of default. In biomedical studies, predicting a patient's survival time, T, given a covariate X, is certainly a problem of interest. This fact also motivates the in-depth analysis of the doubly smoothed estimator of the survival function that is proposed in this chapter. The idea of a time variable smoothing was firstly used in Földes et al. (1981) to propose a smoothed Kaplan-Meier estimator. The work of Giné and Nickl

(2008) presents a smoothed empirical standard measure without covariates. This smoothing was also used in Portier and Segers (2018) to obtain a smoothed quantile function from a cumulative distribution function which is smoothed in the time variable. The local polynomial smoothing of the Kaplan–Meier estimator for fixed designs is explored and analysed in Bagkavos and Ioannides (2021). In Leconte et al. (2002) the smoothed Beran's estimator was studied by simulation, but the derivation of the asymptotic properties was not addressed.

In this chapter, a nonparametric estimator of the conditional survival function with double smoothing both in the covariate and in the time variable is defined. Asymptotic properties of the nonparametric estimator with double smoothing associated with Beran's estimator (Beran (1981)) are presented and a simulation study shows the improvement obtained by using the smoothed Beran's estimator of the conditional survival function for censored data. An illustration with real data is included.

The content of this chapter has been published in Peláez et al. (2022b).

3.2 Doubly smoothed conditional survival estimator

Consider a random sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ of (X, Z, δ). Let $\widehat{S}_h(t|x)$ be a nonparametric estimator of the conditional survival function with $h = h_n$ being the smoothing parameter for the covariate. The distribution function of T is denoted by $F(t)$ and the conditional distribution function of T evaluated at t given $X = x$ by $F(t|x)$. Functions $S(t)$ and $S(t|x)$ are the corresponding survival functions. The conditional distribution function of Z is denoted by $H(t|x)$, and the conditional distribution function of C is denoted by $G(t|x)$. The expression of the proposed doubly smoothed survival estimator is as follows:

$$\widetilde{S}_{h,g}(t|x) \;=\; 1 - \sum_{i=1}^{n} s_{(i)} \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right) \tag{3.2}$$

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 cumulative distribution function of the kernel K, $\mathbb{K}(t) = \int_{-\infty}^{t} K(u)du$, and $g = g_n$ is the smoothing parameter for the time variable.

This survival estimator is not only smoothed in the covariate but also in the time variable. It is based on estimating the survival function in a point t conditional to x by means of a weighted mean of the values that the estimator $\widehat{S}_h(t|x)$ takes in points near t so that a smoothed estimation is obtained. Its construction is fairly intuitive. Consider the kernel density estimator proposed by Parzen (1962) and Rosenblatt (1956) defined by

$$\widehat{f}_g(t) = \frac{1}{ng} \sum_{i=1}^{n} K\left(\frac{t - T_i}{g}\right)$$

The smoothed empirical distribution estimator derived from this density estimator was proposed in Rao (1983) and it is given by

$$\widehat{F}_g(t) \;=\; \int_{-\infty}^{t} \widehat{f}_g(u)du = \frac{1}{ng} \sum_{i=1}^{n} \int_{-\infty}^{t} K\left(\frac{u - T_i}{g}\right)du = \frac{1}{n} \sum_{i=1}^{n} \mathbb{K}\left(\frac{t - T_i}{g}\right).$$

Consequently, a smoothed empirical survival estimator could be defined by

$$\widehat{S}_g(t) = 1 - \widehat{F}_g(t) = 1 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{K}\left(\frac{t - T_i}{g}\right). \tag{3.3}$$

Note that the empirical ditribution function defined by $\widehat{F}(t) = \dfrac{1}{n} \sum_{i=1}^{n} I(T_i \leq t)$ is a step function that jumps up by $1/n$ at each point T_i. Following the idea of Rao (1983), a smoothed estimator of the conditional survival function is obtained by replacing the jumps $\dfrac{1}{n}$ of the empirical distribution in (3.3) by the jumps $s_{(i)}$ previously defined, obtaining the smoothed conditional survival estimator in (3.2).

This doubly smoothed estimator, $\widetilde{S}_{h,g}(t|x)$, could be obtained from any conditional survival estimator, $\widehat{S}_h(t|x)$, using Equation (3.2). In this chapter, the study focuses on the classic Beran's estimator of the conditional survival function, $\widehat{S}_h^B(t|x)$, defined in (2.3). The smoothed survival function estimator based on Beran's estimator, $\widetilde{S}_{h,g}^B(t|x)$, is obtained by replacing $\widehat{S}_h(t|x)$ in (3.2) with $\widehat{S}_h^B(t|x)$ in (2.3) as

follows:

$$\widetilde{S}^B_{h,g}(t|x) \;=\; 1-\sum_{i=1}^{n} s_{(i)}\mathbb{K}\left(\frac{t-Z_{(i)}}{g}\right) \tag{3.4}$$

where $s_{(i)} = \widehat{S}^B_h(Z_{(i-1)}|x) - \widehat{S}^B_h(Z_{(i)}|x)$ and $s_{(1)} = 1 - \widehat{S}^B_h(Z_{(1)}|x)$.

Other conditional survival function estimators could also be considered, for instance, the Weighted Nadaraya-Watson estimator and the Van Keilegom-Akritas estimator. Their expressions are shown in Section 2.2 and the corresponding doubly smoothed versions could be built using Equation (3.2).

3.3 Asymptotic results for the smoothed Beran's estimator

The derivation of the asymptotic properties of the smoothed Beran's estimator, $\widetilde{S}^B_{h,g}(t|x)$, is addressed in this section. First, the necessary assumptions are set out, secondly the theoretical results are established, and finally a discussion on the optimal bandwidths is given.

3.3.1 Assumptions and notation

The assumptions required to establish and prove the asymptotic properties and the notation used are presented below.

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 and C are conditionally independent given $X = x$.

(c) Denoting $l_x = \inf\{t : H(t|x) > 0\}$ and $u_x = \inf\{t : H(t|x) = 1\}$, for any $x \in I_c$, $0 \leq l_x$, $0 \leq u_x < \infty$.

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

A.4. The first and second derivatives of m, $m'(x)$ and $m''(x)$, respectively, exist and are continuous on 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 on $[l, u]$ for some existing l and the value u considered in Assumption A.3d.

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

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

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 $nh \to \infty$ when $n \to \infty$.

These assumptions are standard in the literature and affordable in this context. They were previously required in Dabrowska (1989) and Iglesias-Pérez and González-Manteiga (1999). Assumptions A.1 and A.2 are about characteristics and independence of the variables involved. Assumptions A.3-A.7 are needed to bound

some population functions. Conditions A.2, A.3a, A.3b and A.4 are assumed in Dabrowska (1989) to obtain exponential bounds for the tails of the distribution of $\widehat{S}_h^B(t|x)$ and, from them, to obtain the weak and strong convergence of this estimator. Assumptions A.3c and A.3d are necessary to estimate the tails of the distribution functions involved. Conditions A.5, A.6 and A.7 along with those imposed on the kernel function in Assumption A.8 ensure asymptotic unbiasedness of $\widehat{S}_h^B(t|x)$. Bandwidth requirements are covered by Assumption A.9. The imposed conditions ensure that the point (t, x), where the theoretical results are established, is not a boundary point.

The following notation will be used. Let $R : \mathbb{R} \longrightarrow \mathbb{R}$ be any function, the constants c_R and d_R are defined as follows

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

In particular, one can consider the kernel K and its distribution function $\mathbb{K}$ to define these constants. In this case, Assumption A.8 guarantees that c_K and d_K are finite. Being that,

$$c_K = \int K(t)^2 dt \leq \int_{-1}^{1} \|K\|_\infty^2 dt \leq \|K\|_\infty^2 \int_{-1}^{1} dt = 2\|K\|_\infty^2 < \infty,$$

$$d_K = \int t^2 K(t) dt \leq \int_{-1}^{1} t^2 \|K\|_\infty dt \leq \|K\|_\infty \int_{-1}^{1} t^2 dt = \frac{2}{3}\|K\|_\infty < \infty.$$

From A.8, $K(u) = K(-u)$ for all $u \in \mathbb{R}$. Then, it is satisfied that $\int_{-\infty}^{u} K(t)dt = \int_{-u}^{\infty} K(t)dt$ for $u < 0$, which implies $\mathbb{K}(u) = 1 - \mathbb{K}(-u)$. As a consequence,

$$\begin{aligned} c_{\mathbb{K}} &= \int_{-\infty}^{+\infty} \mathbb{K}^2(u)du = \int_{-1}^{1} \mathbb{K}^2(u)du = \int_{-1}^{0} \mathbb{K}^2(u)du + \int_{0}^{1} \mathbb{K}^2(u)du \\ &= \int_{-1}^{0} \Big(1 - \mathbb{K}(-u)\Big)^2 du + \int_{0}^{1} \mathbb{K}^2(u)du = \int_{0}^{1} \Big(1 - \mathbb{K}(v)\Big)^2 dv + \int_{0}^{1} \mathbb{K}^2(u)du \\ &= \int_{0}^{1} \Big(1 - 2\mathbb{K}(u) + \mathbb{K}^2(u) + \mathbb{K}^2(u)\Big)du = \int_{0}^{1} \Big(1 - 2\mathbb{K}(u)\Big(1 - \mathbb{K}(u)\Big)\Big)du, \end{aligned}$$

Given that $2\mathbb{K}(u)\Big(1 - \mathbb{K}(u)\Big) \geq 0$,

$$c_{\mathbb{K}} = \int_{0}^{1} \Big(1 - 2\mathbb{K}(u)\Big(1 - \mathbb{K}(u)\Big)\Big)du \leq \int_{0}^{1} 1du = 1.$$

The following functions are also defined,

$$K_l(u) = u^l K(u), \quad \mathbb{K}_l(u) = \int_{-\infty}^{u} K_l(t)dt. \tag{3.5}$$

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 $\ddot{f}(x_1, ..., x_k)$. Finally, let $f * g$ be the convolution of any two functions f and g defined as $f * g = \int f(t-u)g(u)du$.

The following functions are required to state the asymptotic results:

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

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

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

$$\Phi_\eta(u, t, x) = \int K(v)\Big(1 - F(t - gv|x)\Big)\Phi_\xi(u, t - gv, x)dv,$$

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

An additional assumption related to the differentiability of the above functions is then 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 at $u = x$. 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 $u = x$.

3.3.2 Asymptotic results

An almost sure representation for the smoothed Beran's estimator of the conditional survival function is presented here. Asymptotic expressions for the bias and variance of the estimator are found and the asymptotic normality is stated.

In Iglesias-Pérez and González-Manteiga (1999), an almost sure representation is found for a generalized Beran's estimator of the conditional survival function when the data are subject to random left truncation and right censoring. We do not consider truncation but only right censoring. Then, an almost sure representation of Beran's estimator can be obtained as a direct consequence of Theorem 2(c) in Iglesias-Pérez and González-Manteiga (1999) by just assuming a degenerated in zero distribution for the left truncation time variable.

Theorem 3.1 (Almost sure representation for Beran's estimator of the conditional survival function)**.** *Under assumptions A.1-A.10, if $l < l_x$ for any $x \in I$, then*

$$\widehat{S}_h^B(t|x) - S(t|x) = (1 - F(t|x)) \sum_{i=1}^{n} w_{h,i}(x)\xi(Z_i, \delta_i, t, x) + R_n(t|x)$$

for $t \in [l, u]$, $x \in I$, where

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

A similar result is obtained below for the smoothed Beran's estimator.

Theorem 3.2 (Almost sure representation for the smoothed Beran's estimator of the conditional survival function)**.** *Under assumptions A.1-A.10, if $l < l_x$ for any $x \in I$, then*

$$\widetilde{S}_{h,g}^B(t|x) - S(t|x) = \sum_{i=1}^{n} w_{h,i}(x)\eta(Z_i, \delta_i, t, x) - \frac{1}{2} d_K F''(t|x) g^2 + R_n^1(t|x) + R_n^2(t|x)$$

for $t \in [a', b']$, $x \in I$, where $a' = l + \varepsilon$, $b' = u - \varepsilon$ for $\varepsilon > 0$,

$$\sup_{(t,x)\in[a',b']\times I} \left| R_n^1(t|x) \right| = O\left(\frac{\ln n}{nh}\right)^{3/4} \quad a.s.,$$

and

$$\sup_{(t,x)\in[a',b']\times I} \left| R_n^2(t|x) \right| = o(g^2).$$

Applying Theorem 3.2, the asymptotic bias and covariance of the smoothed Beran's estimator of the conditional survival function are obtained. Firstly, the smoothed Beran's estimator $\widetilde{S}^B_{h,g}(t|x)$ is written as the sum of two terms: one dominant term and a negligible one. This is shown in Lemma 3.1.

Lemma 3.1. *Under the assumptions of Theorem 3.2, the smoothed Beran's estimator $\widetilde{S}^B_{h,g}(t|x)$ can be written as follows*

$$\widetilde{S}^B_{h,g}(t|x) = \widetilde{S}^{AB}_{h,g}(t|x) + \widetilde{R}_n(t|x)$$

where

$$\widetilde{S}^{AB}_{h,g}(t|x) = S(t|x) + \sum_{i=1}^{n} w^A_{h,i}(x)\eta(Z_i, \delta_i, t, x) - \frac{1}{2}d_K F''(t|x)g^2,$$

with

$$w^A_{h,i}(x) = \frac{1}{nh}\frac{K\Big((x - X_i)/h\Big)}{m(x)}$$

for all $i = 1, ..., n$, and

$$\sup_{[l,u]\times I} |\widetilde{R}_n(t|x)| = O\left(\frac{\ln n}{nh}\right)^{3/4} + o(g^2) + O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)^2.$$

Theorem 3.3 (Bias and covariance of $\widetilde{S}^{AB}_{h,g}(t|x)$)**.** *Under the assumptions of Theorem 3.2, the asymptotic expressions for the bias and the variance of $\widetilde{S}^B_{h,g}(t|x)$ are the following:*

$$\begin{aligned} Bias\Big(\widetilde{S}^{AB}_{h,g}(t|x)\Big) &= \frac{d_K\Big(1 - F(t|x)\Big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)m(x)\Big)h^2 \\ &\quad -\frac{1}{2}d_K F''(t|x)g^2 + o(h^2), \end{aligned}$$

$$\begin{aligned} Var\Big(\widetilde{S}^{AB}_{h,g}(t|x)\Big) &= \frac{c_K}{m(x)}\Big(1 - F(t|x)\Big)^2 L(t|x)\frac{1}{nh} \\ &\quad +\frac{c_K(c_{\mathbb{K}} - 1)}{m(x)}\Big(1 - F(t|x)\Big)^2 L'(t|x)\frac{g}{nh} + O\left(\frac{h^2 + g^2}{nh}\right). \end{aligned}$$

Finally, the asymptotic distribution of the smoothed Beran's estimator of the conditional survival function is obtained.

Theorem 3.4 (Limit distribution of $\widetilde{S}^B_{h,g}(t|x)$)**.** *Under the assumptions of Theorem 3.2 and assuming* $\dfrac{(\ln n)^3}{nh} \to 0$, $C_h := \lim_{n\to\infty} n^{1/5}h > 0$ *and* $C_g := \lim_{n\to\infty} n^{1/5}g > 0$, *the limit distribution of* $\widetilde{S}^B_{h,g}(t|x)$ *is given by*

$$\sqrt{nh}\Big(\widetilde{S}^B_{h,g}(t|x) - S(t|x)\Big) \xrightarrow{d} N(\mu, \sigma),$$

where

$$\begin{aligned}\mu &= C_h^{5/2}\frac{d_K\big(1-F(t|x)\big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)m(x)\Big)\\ &\quad -C_h^{1/2}C_g^{4/2}\frac{1}{2}d_K F''(t|x)\end{aligned}$$

and

$$\sigma^2 = \frac{c_K}{m(x)}\big(1-F(t|x)\big)L(t|x).$$

Remark 3.1. *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 estimator is*

$$\sqrt{nh}\Big(\widetilde{S}^B_{h,g}(t|x) - S(t|x)\Big) \xrightarrow{d} N(\mu_1, \sigma),$$

with

$$\mu_1 = C_h^{5/2}\frac{d_K\big(1-F(t|x)\big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)m(x)\Big).$$

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

The asymptotic properties of Beran's estimator for the conditional survival function were proven in both works Dabrowska (1989) and Iglesias-Pérez and González-Manteiga (1999). It is worth noting that the asymptotic bias of Beran's estimator and the smoothed Beran's estimator have the same order as long as g is negligible with respect to h, i.e., $g = o(h)$. On the other hand, assuming $h \to 0$ and $g \to 0$, the asymptotic variance of Beran's estimator and the smoothed Beran's estimator have the same order since the terms g/nh and h/n are negligible compared to $1/nh$.

Proofs of these results can be found in Section 3.6.

3.3.3 Asymptotically optimal bandwidths

In this section, a discussion about the smoothing parameters of the smoothed Beran's survival estimator takes place in order to find the asymptotic optimal bandwidths defined as those that minimise the asymptotic mean squared error (MSE).

Considering only the dominant terms of the bias and the variance of the asymptotic estimator $\widetilde{S}^{AB}_{h,g}(t|x)$, from the expressions given in Theorems 3.3 and 3.4, it follows that

$$\mathrm{Var}\Big(\widetilde{S}^{AB}_{h,g}(t|x)\Big) = c_1\frac{1}{nh} - c_2\frac{g}{nh} + o\bigg(\frac{g}{nh}\bigg),$$
$$\mathrm{Bias}\Big(\widetilde{S}^{AB}_{h,g}(t|x)\Big) = c_3h^2 + c_4g^2 + o(h^2),$$

where the constants c_1, c_2, c_3 and c_4 are defined by

$$\begin{aligned}
c_1 &= c_K\frac{\Big(1-F(t|x)\Big)^2L(t|x)}{m(x)} > 0,\\
c_2 &= c_K(1-c_{\mathbb{K}})\frac{\Big(1-F(t|x)\Big)^2L'(t|x)}{m(x)} > 0,\\
c_3 &= \frac{d_K\Big(1-F(t|x)\Big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)m(x)\Big),\\
c_4 &= -\frac{1}{2}d_KF''(t|x).
\end{aligned}$$

Then, the asymptotic bandwidths that minimise the dominant terms of the MSE can be obtained by minising the function:

$$\Psi(h,g) = c_1\frac{1}{nh} - c_2\frac{g}{nh} + c_3^2h^4 + c_4^2g^4 + 2c_3c_4h^2g^2. \tag{3.6}$$

Then, it is necessary to consider the partial derivatives of Ψ with respect to both h and g, equal them to zero and distinguish three different cases depending on the relative asymptotic behaviour of h and g. The partial derivative of Ψ with respect to h is

$$\frac{\partial\Psi}{\partial h} = -c_1\frac{1}{nh^2} + c_2\frac{g}{nh^2} + 4c_3^2h^3 + 4c_3c_4hg^2,$$

but, since $g\to 0$, the term $c_2\dfrac{g}{nh^2}$ is negligible with respect to the term $\dfrac{c_1}{nh^2}$. Similarly,

$$\frac{\partial\Psi}{\partial g} = -c_2\frac{1}{nh} + 4c_4^2g^3 + 4c_3c_4h^2g.$$

Therefore, the equations to be taken into account are the following ones

$$-c_1\frac{1}{nh^2}+4c_3^2h^3+4c_3c_4hg^2=0, \tag{3.7}$$

$$-c_2\frac{1}{nh}+4c_4^2g^3+4c_3c_4h^2g=0. \tag{3.8}$$

There are three possible cases for the asymptotic behaviour of $\frac{g}{h}$.

Case 1 $g=o(h)$

Equations asymptotically equivalent to (3.7) and (3.8) in this case are

$$-c_1\frac{1}{nh^2}+4c_3^2h^3=0,$$
$$-c_2\frac{1}{nh}+4c_3c_4h^2g=0.$$

Then, the optimal bandwidths are $h_{opt}=c_0n^{-1/5}$ and $g_{opt}=d_0n^{-2/5}$ with $c_0=\left(\frac{c_1}{4c_3^2}\right)^{1/5}$ and $d_0=\frac{c_2c_3^{1/5}}{4^{2/5}c_1^{3/5}c_4}$. In this case,

$$\Psi(h_{opt},g_{opt})=\left(\frac{c_1}{c_0}+c_3^2c_0^4\right)n^{-4/5}+\left(c_3c_0+2c_3c_4c_0^2d_0^2-\frac{c_2d_0}{c_0}\right)n^{-6/5}+c_4^2d_0^4n^{-8/5}.$$

Case 2 $h=o(g)$

When $h=o(g)$, asymptotically equivalent versions of Equations (3.7) and (3.8) are

$$-\frac{c_1}{nh^2}+4c_3c_4hg^2=0,$$
$$-\frac{c_2}{nh}+4c_4^2g^3=0.$$

and the solution of this system is $h_{opt}=e_0n^{-1/7}$ and $g_{opt}=\left(\frac{c_2}{4c_4^2e_0}\right)^{1/3}n^{-2/7}$ with $e_0=\left(\frac{c_1(4c_4^2)^{2/3}}{4c_3c_4c_2^{2/3}}\right)^{3/7}$. So, $g_{opt}=o(h_{opt})$ which contradicts the initial hypothesis. Case 2 is discarded.

Case 3 $\lim_{h\to\infty} \frac{h}{g} = \alpha$ for some $\alpha > 0$.

In this case, $\frac{h}{g} = \alpha$ asymptotically and the asymptotic expression for Ψ becomes

$$\Psi(h, g) \;=\; \frac{c_1}{n\alpha g} - \frac{c_2}{n\alpha} + c_5 g^4$$

with $c_5 = c_3^2\alpha^4 + c_4^2 + 2c_3c_4\alpha^2 = (c_3\alpha^2 + c_4)^2$.

The option $c_5 = 0$ is discarded because it leads to an optimal bandwidth g which does not tend to zero. Therefore, $c_5 = (c_3\alpha^2 + c_4)^2 > 0$ and the minimun is reached at $h_{opt} = \alpha l_0 n^{-1/5}$ and $g_{opt} = l_0 n^{-1/5}$ with $l_0 = \left(\frac{c_1}{4c_5\alpha}\right)^{1/5}$. Straightforward calculations lead to

$$\Psi(h_{opt}, g_{opt}) \;=\; \frac{5c_1^{4/5}c_5^{1/5}}{4^{4/5}\alpha^{4/5}} n^{-4/5} - \frac{c_2}{n\alpha}.$$

This means that the minimal value of Ψ is attained at $\alpha = \infty$ which contradicts Case 3.

From the arguments above it follows that $g = o(h)$ is the only feasible case, obtaining the corresponding optimal bandwidths for the estimator $\widetilde{S}^{AB}_{h,g}(t|x)$.

3.4 Simulation study

A simulation study was conducted in order to compare the performance of the smoothed Beran's estimator of the conditional survival function with Beran's estimator. The models considered in the study are those presented in Section 2.4.

The conditional survival function is estimated in a time grid of size n_T, $0 < t_1 < \cdots < t_{n_T}$. For Model 1, $t_{n_T} = 0.4992$ is about the 90th percentile of the corresponding time variable distribution for $x = 0.8$. Model 2 considers $t_{n_T} = 0.7154$ is about the 90th percentile of the corresponding time variable distribution for $x = 0.6$. For Model 3, $x = 0.8$ and $t_{n_T} = 3.1211$ is about the 90th percentile of the corresponding time variable distribution.

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

The optimal bandwidth for $\widehat{S}_h^B(t|x)$, h_1, is taken (from a meshgrid of 50 values of h) as the value which minimises a Monte Carlo approximation of the MISE:

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

based on $N = 100$ simulated samples. The value of $MISE$ 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 of $\widehat{S}_h^B(t|x)$.

The smoothed survival estimator $\widetilde{S}_{h,g}^B(t|x)$ depends on two bandwidths. Three strategies are be used in order to obtain these smoothing parameters.

Strategy 1 It consists of fixing the covariate smoothing parameter to the optimal one, h_1, for Beran's estimator and approximating the time variable smoothing parameter. The error to minimise is

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

considered as a function of the bandwidth g. It is approximated from $N = 100$ simulated samples in a grid of 50 values of g 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 $\widetilde{S}_{h,g}^B(t|x)$. The main advantage of using this strategy is its relatively low computational cost.

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

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

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

Strategy 3 The optimal bandwidth (h_3, g_3) is obtained by minimising some Monte Carlo approximations of the function $MISE_x(h, g)$ based on $N = 100$ simulated samples using a limited-memory algorithm. An optimisation method for solving large nonlinear optimisation problems allowing box contraints (L-BFGS-B) is chosen. It was proposed by Byrd et al. (1995) for solving optimization problems subject to simple bounds on the variables in which information on the Hessian matrix is difficult to obtain. This uses a limited-memory modification of the BFGS quasi-Newton method published simultaneously by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970). Results of numerical studies about this method are shown in Byrd et al. (1995). It is available at the stats package from the Comprehensive R Archive Network (CRAN) using Fortran 77 subroutines (see Zhu et al. (1997)). The value of the $MISE$ made by $\widetilde{S}^B_{h_3,g_3}(t|x)$ is approximated from $N = 1000$ simulated samples.

Neither the bandwidth for Beran's estimator nor the bandwidths for the smoothed Beran's estimator with any of these strategies can be used in practice but their choice produces a fair comparison since the estimators are built using the best possible choice for the smoothing parameters.

Figure 3.3 shows the function $MISE_x(h)$ over a grid of 50 values of h for Models 1, 2 and 3 when the conditional censoring probability is 0.5. These graphs show the function to minimise in order to obtain the optimal bandwidth for Beran's estimator and the region where this minimum is attained. The results for other levels of censoring probability, which are not shown here, are quite similar.

Figure 3.4 shows the function $MISE_x(h_1, g)$ for each level of censoring conditional probability and each model. These graphs show the error curve to minimise in order to obtain the optimal time smoothing parameter. It follows from this that the optimal bandwidth g is easily approximated by Strategy 1.

Figure 3.5 shows the function $MISE_x(h, g)$ over a meshgrid of 50 values of h and 50 values of g for Models 1, 2 and 3 when the censoring conditional probability is 0.5. These graphs show the two-dimensional functions to be minimised in Strategies 2

and 3 in order to obtain the optimal bandwidths for the smoothed Beran's estimator. The red zone is where this minimum is reached and the coordinates of the point at which the minimum is attained provide the optimal smoothing bandwidths. The results for other levels of censoring probability, which are not shown here, are quite similar. These graphs show that the smoothing parameters for the smoothed Beran's estimator can be well approximated in all scenarios.

It is clear that the magnitude of the estimation error is notably affected by the choice of the time smoothing bandwidth (g). However, for a fixed value of h, the value of g for which the smallest error is made does not seem to vary too much depending on the value of the covariate smoothing bandwidth (h). This can be seen in Figure 3.6. There, $MISE_x(h, g)$ is shown as a function of g for some fixed values of h within the interval where the optimum is reached. The obtained curves have similar shape and they are close for all the values of h in Model 2. They are not that close in Model 1 or 3, but the minimum of $MISE_x(h, g)$ is reached for similar values of g in all of them.

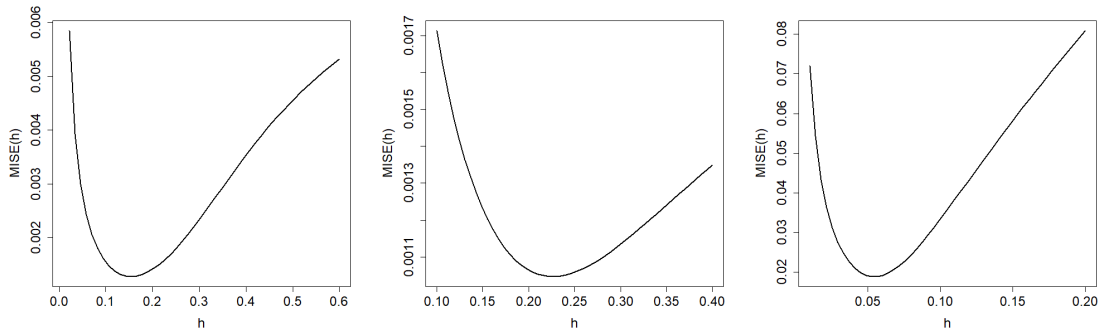

Figure 3.3: $MISE_x(h)$ function approximated via Monte Carlo for Beran's estimator using $N = 100$ simulated samples from Model 1 (left), Model 2 (center) and Model 3 (right) when $P(\delta = 0|x) = 0.5$.

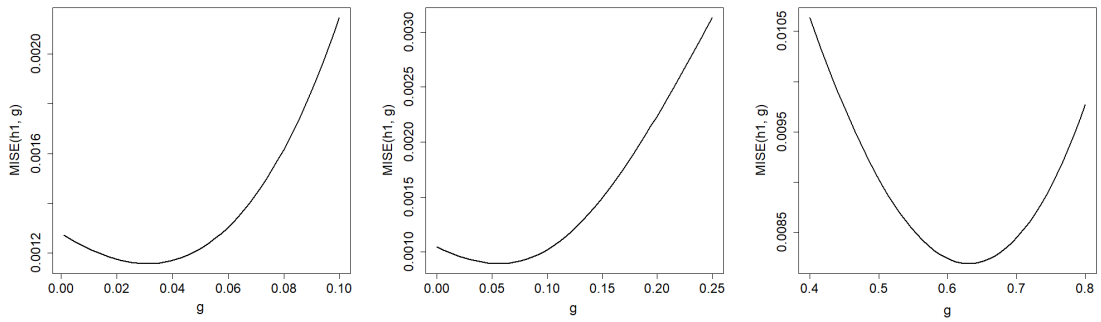

Figure 3.4: $MISE_x(h_1, g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (left), Model 2 (center) and Model 3 (right) when $P(\delta = 0|x) = 0.5$.

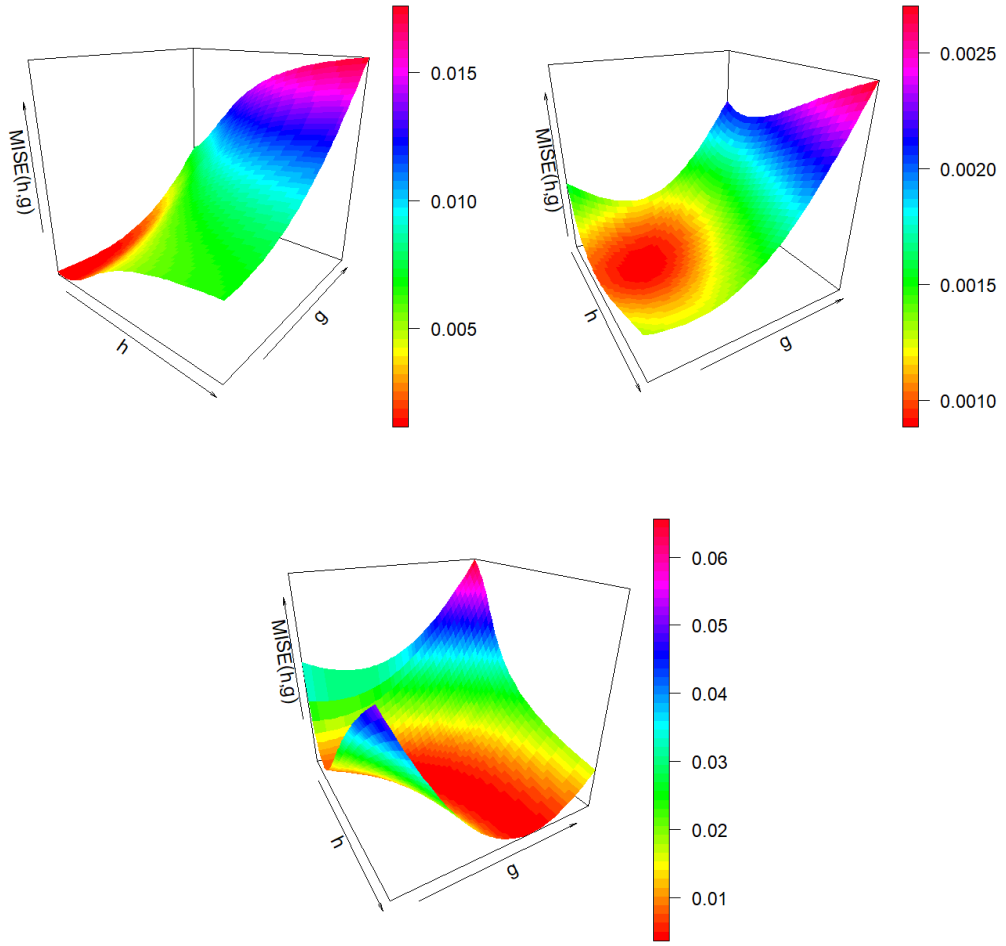

Figure 3.5: $MISE_x(h, g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (topleft), Model 2 (topright) and Model 3 (bottom) when $P(\delta = 0|x) = 0.5$.

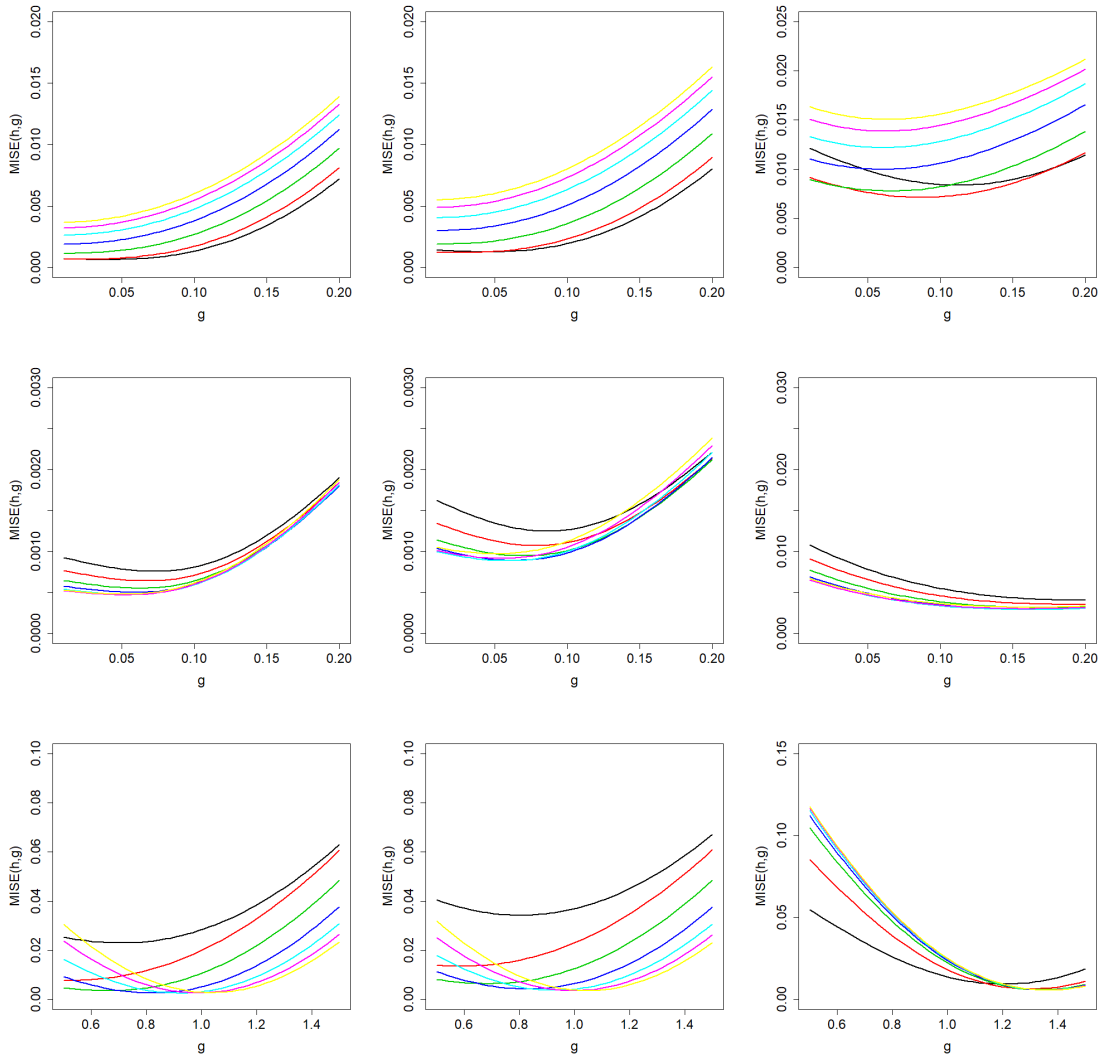

Figure 3.6: $MISE_x(h,g)$ as a function of g approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples for some fixed equispaced values of $h \in [0.1, 0.8]$ for Model 1 (top), $h \in [0.1, 0.4]$ for Model 2 (middle) and $h \in [0.01, 0.18]$ for Model 3 (bottom) with $P(\delta = 0|x) = 0.2$ (left), $P(\delta = 0|x) = 0.5$ (center) and $P(\delta = 0|x) = 0.8$ (right).

Tables 3.1, 3.2 and 3.3 show the MISE bandwidths and the estimation errors of Beran's estimator and the smoothed Beran's estimator for each model obtained by means of each strategy. In order to compare the behaviour of the estimators and quantify the improvement of the smoothing over the original estimator, the ratio $R_i(x)$ is defined

$$R_i(x) \quad = \quad \frac{RMISE_x\Big(\widetilde{S}^B_{h_i,g_i}(\cdot|x)\Big)}{RMISE_x\Big(\widehat{S}^B_{h_1}(\cdot|x)\Big)}$$

for $i = 1, 2, 3$ depending on the chosen smoothing strategy. The closer to 0 the value of $R_i(x)$, the greater the improvement of the smoothed Beran's estimator with respect to Beran's estimator. The relation between R_1, R_2 and R_3 also informs which of the three strategies reduces the error most.

In all cases, $RMISE$ values are lower for the smoothed Beran's estimator than for Beran's estimator and this difference becomes bigger when increasing the censoring conditional probability. This is confirmed by looking at the values of $R_i(x)$ for all $i = 1, 2, 3$. When the censoring conditional probability is 0.2 or 0.5, the time smoothing reduces the error by about 5% in Model 1 and 8% in Model 2. This improvement is about 22% in Model 1 and 40% in Model 2 when the probability of conditional censoring is 0.8. The error reduction in Model 3 with respect to the nonsmoothed survival estimator is more significant, reaching 50% and 70% when censoring is moderate or heavy, respectively.

The approximations of the optimal bandwidths obtained by Strategy 3 are similar to those obtained by Strategy 2. The corresponding estimation errors made by the smoothed Beran's estimator are also very close. Both Strategies 2 and 3 improve on Strategy 1. Therefore, the results are equally compelling for both Strategies 2 and 3. The advantage of using Strategy 3 lies in the computing times required to obtain an approximation of the optimal bandwidths (h, g). In later paragraphs it will be seen that the minimisation method based on the iterative method turns out to be significantly faster.

A brief study not included here shows that the results of these simulations hold even if the distribution of X is not uniform.

Model 1				
$P(\delta = 0\|x)$		0.2	0.5	0.8
$\widehat{S}^B_{h_1}$	h_1	0.14245	0.15449	0.21469
	$RMISE_x(h_1)$	0.02577	0.03549	0.10122
$\widetilde{S}^B_{h_1,g_1}$	h_1	0.14245	0.15449	0.21469
	g_1	0.02727	0.03131	0.07576
	$RMISE_x(h_1, g_1)$	0.02466	0.03364	0.08010
	R_1	0.95693	0.94787	0.79135
$\widetilde{S}^B_{h_2,g_2}$	h_2	0.13673	0.13673	0.19184
	g_2	0.02939	0.03714	0.07980
	$RMISE_x(h_2, g_2)$	0.02460	0.03349	0.07952
	R_2	0.95460	0.94365	0.78562
$\widetilde{S}^B_{h_3,g_3}$	h_3	0.13108	0.13936	0.19323
	g_3	0.03036	0.03480	0.08129
	$RMISE_x(h_3, g_3)$	0.02459	0.03348	0.07951
	R_3	0.95421	0.94336	0.78547

Table 3.1: Optimal bandwidths, $RMISE$, R_1, R_2 and R_3 of the survival estimation for Beran's estimator, the smoothed Beran's estimator with Strategy 1, Strategy 2 and Strategy 3 in each level of conditional censoring probability for Model 1.

Model 2				
$P(\delta = 0\|x)$		0.2	0.5	0.8
$\widehat{S}^B_{h_1}$	h_1	0.25918	0.22857	0.23469
	$RMISE_x(h_1)$	0.02304	0.03186	0.08641
$\widetilde{S}^B_{h_1,g_1}$	h_1	0.25918	0.22857	0.23469
	g_1	0.05110	0.05620	0.16330
	$RMISE_x(h_1, g_1)$	0.02144	0.02943	0.05185
	R_1	0.93055	0.92373	0.60005
$\widetilde{S}^B_{h_2,g_2}$	h_2	0.24082	0.20408	0.20408
	g_2	0.05265	0.06041	0.16510
	$RMISE_x(h_2, g_2)$	0.02129	0.02907	0.05067
	R_2	0.92405	0.91243	0.58639
$\widetilde{S}^B_{h_3,g_3}$	h_3	0.24189	0.20475	0.20511
	g_3	0.05265	0.06040	0.16324
	$RMISE_x(h_3, g_3)$	0.02130	0.02908	0.05067
	R_3	0.92448	0.91274	0.58639

Table 3.2: Optimal bandwidths, $RMISE$, R_1, R_2 and R_3 of the survival estimation for Beran's estimator, the smoothed Beran's estimator with Strategy 1, Strategy 2 and Strategy 3 in each level of conditional censoring probability for Model 2.

Model 3				
$P(\delta = 0\|x)$		0.2	0.5	0.8
$\widehat{S}^B_{h_1}$	h_1	0.04490	0.05265	0.12837
	$RMISE_x(h_1)$	0.11112	0.14644	0.28914
$\widetilde{S}^B_{h_1,g_1}$	h_1	0.04490	0.05265	0.12837
	g_1	0.54082	0.62857	1.23469
	$RMISE_x(h_1, g_1)$	0.07817	0.10091	0.08886
	R_1	0.70347	0.68909	0.30733
$\widetilde{S}^B_{h_2,g_2}$	h_2	0.11061	0.16980	1.00000
	g_2	0.88776	1.03061	1.35714
	$RMISE_x(h_2, g_2)$	0.05248	0.06379	0.07550
	R_2	0.47228	0.43561	0.26112
$\widetilde{S}^B_{h_3,g_3}$	h_3	0.11391	0.17195	1.00000
	g_3	0.89506	1.03011	1.35999
	$RMISE_x(h_3, g_3)$	0.05241	0.06375	0.07565
	R_3	0.47165	0.43533	0.26164

Table 3.3: Optimal bandwidths, $RMISE$, R_1, R_2 and R_3 of the survival estimation for Beran's estimator, the smoothed Beran's estimator with Strategy 1, Strategy 2 and Strategy 3 in each level of v conditional censoring probability for Model 3.

Figure 3.7 shows a cloud of estimated survival curves (50 out of 1000), the theoretical survival function, the mean curve and the 5th and 95th percentiles of the total estimated curves for the smoothed Beran's estimator in Model 1, Model 2 and Model 3. These figures show clearly how the estimated curves are distributed and the variability they present, as well as to notice the worsening of the estimations as the probability of censoring increases.

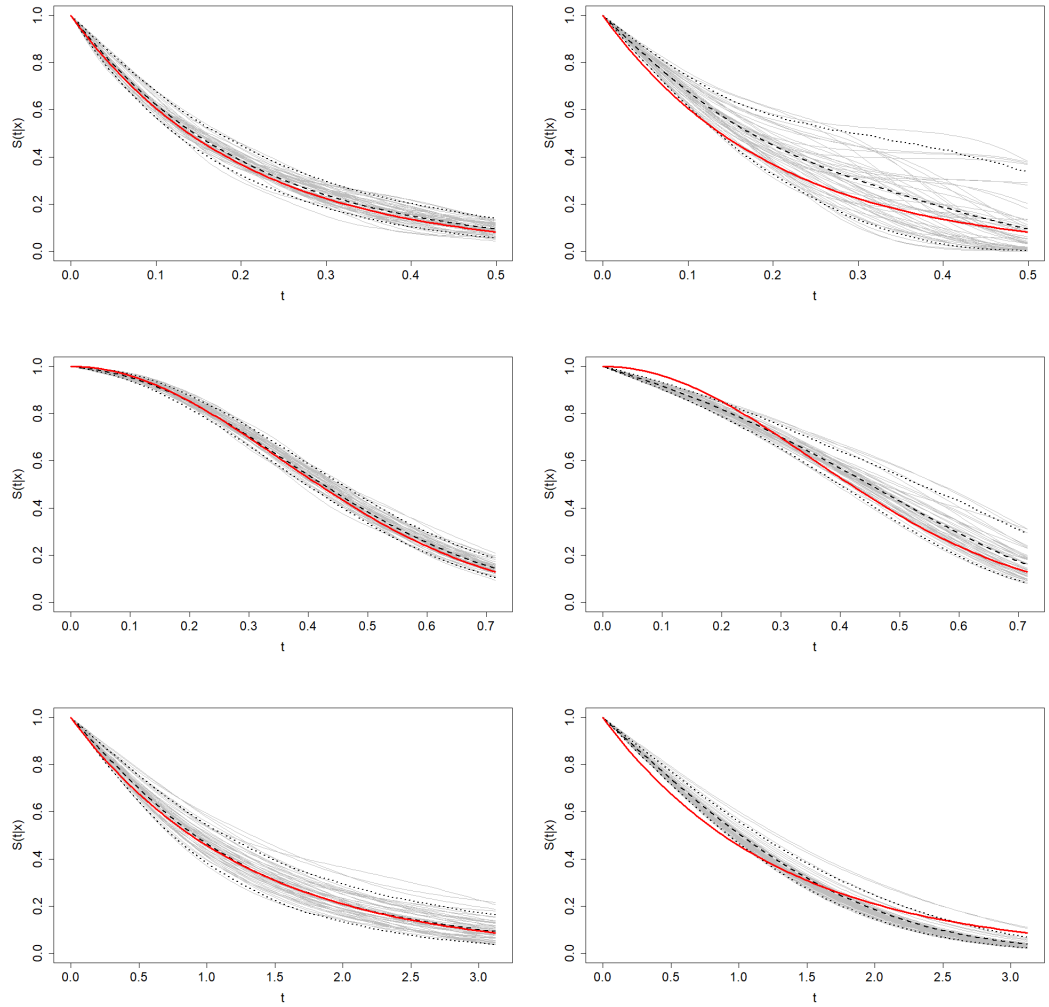

Figure 3.7: Theoretical $S(t|x)$ (solid line), mean curve (dashed line) and 5th and 95th percentiles (dotted line) obtained by means of the smoothed Beran's estimator when $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.8$ (right) in Model 1 (top), Model 2 (middle) and Model 3 (bottom).

Table 3.4 shows the computation time (in seconds) of Beran's estimator and smoothed Beran's estimator when estimating the conditional survival curve in a 100-point time grid and a fixed value of x for different values of the sample size in Model 1. The smoothing parameters are fixed to the optimal values. Time variable smoothing results in an increase of the CPU time. The smoothed Beran's estimator is the most affected by the increase of the sample size and its CPU times is higher than the CPU of Beran's estimator.

n	50	100	200	400	1200
Beran	0.01	0.01	0.01	0.02	0.02
SBeran	0.01	0.01	0.02	0.03	0.07

Table 3.4: CPU time (in seconds) for estimating $S(t|x)$ in a time grid of size 100 for each estimator and different sample sizes (n).

It is also interesting to compare the computational efficiency of the strategies used to find the optimal bandwidths, since Strategy 1 seems to be faster but Strategies 2 and 3 provides smaller estimation errors. Table 3.5 shows the CPU time (in minutes) for each strategy and several number of trials.

In all strategies the sample size is $n = 400$ and the conditional survival function is estimated in a time grid of size $n_T = 100$. The number of simulated samples (N) used to approximate the $MISE$ by Monte Carlo is the parameter that varies to compare the time each strategy takes to obtain the optimal bandwidths. The results clearly show the computational advantage of using Strategy 3, since Strategy 1 and 2 are significantly slower.

N	50	100	150	200
Strategy 1	1.12	1.95	2.99	3.86
Strategy 2	37.58	79.83	117.79	159.99
Strategy 3	0.55	1.13	2.04	2.71

Table 3.5: CPU time (in minutes) for approximating the optimal bandwidth (h, g) for $\widetilde{S}^B_{h,g}(t|x)$ with Strategies 1, 2 and 3 and different numbers of trials (N).

3.5 Application to real data

A brief illustration of the use of the smoothed Beran's estimator is provided here. The survival function of the time that COVID-19 patients remain hospitalised in the Intensive Care Unit (ICU) is estimated by means of Beran's and the smoothed

Beran's estimators. A dataset from SERGAS (Galician health service) with dates of admission and discharge (if applicable) to ICU and age of 288 COVID-19 patients in Galicia (Spain) during the first weeks of the outbreak (February-April, 2020) is used. The censoring rate of this dataset is 14.60%. Figure 3.8 shows the conditional survival estimation at $x = 66.04$ years, which is the sample mean of age in the dataset, along the time interval $[0, 40]$. The bandwidths were empirically chosen through visual inspection and considering the ranges in which the variables lie: $h = 7.03$ for Beran's estimator and $(h, g) = (7.03, 10.80)$ for the smoothed Beran's estimator. The differences between the two estimations are remarkable. Although the tendency of the survival curve is similar in both estimations, Beran's estimation has the classic roughness of a jump function in the time variable, while the smoothed Beran's estimation is a much smoother curve and presumably with lower estimation error.

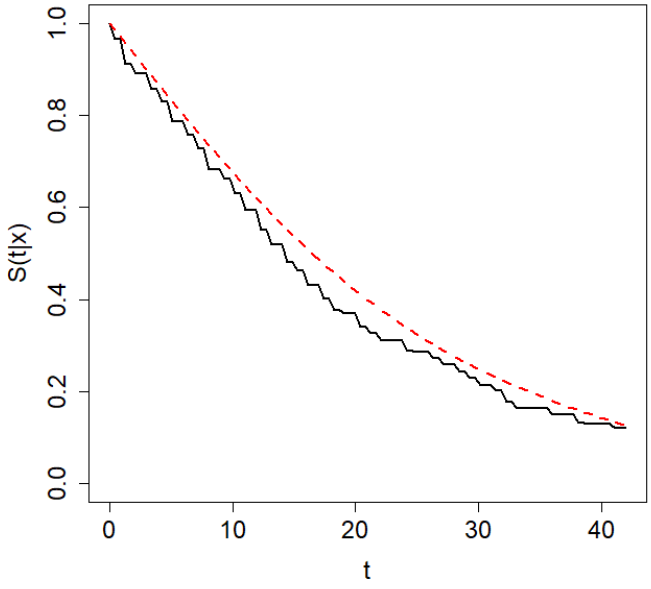

Figure 3.8: Conditional survival function estimated with Beran's (solid line) and the smoothed Beran's estimator (dashed line) for the COVID-19 SERGAS dataset.

3.6 Proofs

The following lemmas will be used in the proofs.

Lemma 3.2 (Integration by parts formula for Riemann-Stieltjes integral with a piecewise-defined function)**.** *Let $u : [0, L] \longrightarrow \mathbb{R}$ be a differentiable function in $[0, L]$ and let $v : [0, L] \longrightarrow \mathbb{R}$ be a nondecreasing piecewise function, i.e.,*

$$v(x) = \sum_{j=1}^{k-1} b_j 1_{[a_{j-1},a_j)}(x) + b_k 1_{[a_{k-1},a_k]}(x)$$

where $0 = a_0 < a_1 < \cdots < a_k = L$ and $b_i \in \mathbb{R}$ for all $i = 1, ..., k$, $b_1 < b_2 < \cdots < b_k$. Then,

$$\int_0^L u(x)v(dx) = \Big[u(x)v(x)\Big]_{x=0}^{x=L} - \int_0^L u'(x)v(x)dx.$$

Proof of Lemma 3.2.

On the one hand,

$$\begin{array}{rcl} \int_0^L u(x)v(dx) & = & \sum_{i=1}^{k-1} u(a_i)\Big(v(a_i) - v(a_i^-)\Big) = \sum_{i=1}^{k-1} u(a_i)\Big(v(a_i) - v(a_{i-1})\Big) \\ & = & \sum_{i=1}^{k-1} u(a_i)\Big(b_{i+1} - b_i\Big) \end{array} \tag{3.9}$$

On the other hand,

$$\Big[u(x)v(x)\Big]_{x=0}^{x=L} - \int_0^L u'(x)v(x)dx =$$

$$\begin{aligned} &= u(L)v(L) - u(0)v(0) - \sum_{j=1}^{k-1} b_j \int_0^L u'(x)1_{[a_{j-1},a_j)}(x)dx - b_k \int_0^L u'(x)1_{[a_{k-1},a_k]}(x)dx \\ &= u(L)v(L) - u(0)v(0) - \sum_{j=1}^{k-1} b_j \int_{a_{j-1}}^{a_j} u'(x)dx - b_k \int_{a_{k-1}}^{a_k} u'(x)dx \\ &= u(L)v(L) - u(0)v(0) - \sum_{j=1}^{k-1} b_j\Big(u(a_j) - u(a_{j-1})\Big) - b_k\Big(u(a_k) - u(a_{k-1})\Big) \\ &= u(L)v(L) - u(0)v(0) - \sum_{j=1}^{k-1} b_j u(a_j) + \sum_{j=1}^{k-1} b_j u(a_{j-1}) - b_k u(a_k) + b_k u(a_{k-1}) \end{aligned}$$

$$
\begin{aligned}
&= u(L)v(L) - u(0)v(0) - \sum_{j=1}^{k-1} b_j u(a_j) + \sum_{j=0}^{k-2} b_{j+1} u(a_j) - b_k u(a_k) + b_k u(a_{k-1}) \\
&= u(L)v(L) - u(0)v(0) - \sum_{j=1}^{k-2} b_j u(a_j) - b_{k-1} u(a_{k-1}) + \sum_{j=1}^{k-2} b_{j+1} u(a_j) + b_1 u(a_0) \\
&\quad -b_k u(a_k) + b_k u(a_{k-1}) \\
&= u(L)v(L) - u(0)v(0) + b_1 u(a_0) - b_k u(a_k) + \sum_{j=1}^{k-1} (b_{j+1} - b_j) u(a_j)
\end{aligned}
$$

Since $a_0 = 0$, $a_k = L$, $v(a_0) = b_1$ and $v(a_k) = b_k$, we have

$$
\Big[u(x)v(x)\Big]_{x=0}^{x=L} - \int_0^L u'(x)v(x)dx = \sum_{j=1}^{k-1} (b_{j+1} - b_j) u(a_j). \tag{3.10}
$$

Now, using (3.9) and (3.10), the lemma is proved.

□

Lemma 3.3. *Under Assumptions A.8 and A.10, then*

$$
\begin{aligned}
&E\left[K\left(\frac{x - X_1}{h}\right)\eta(Z_1, \delta_1, t, x)\right] \\
&\quad = \frac{d_K}{2}\Big(1 - F(t|x)\Big)\Big(2\Phi'_\xi(x, t, x)m'(x) + \Phi''_\xi(x, t, x)m(x)\Big)h^3 + o(h^3).
\end{aligned}
$$

Proof of Lemma 3.3.

First,

$$
\begin{aligned}
&E\left[K\left(\frac{x - X_1}{h}\right)\eta(Z_1, \delta_1, t, x)\right] \\
&= E\left[K\left(\frac{x - X_1}{h}\right)E\Big[\eta(Z_1, \delta_1, t, x)|X_1\Big]\right] = \int_{-\infty}^{+\infty} K\left(\frac{x - u}{h}\right)\Phi_\eta(u, t, x)m(u)du \\
&= \int_{-\infty}^{+\infty} hK(v)\Phi_\eta(x - hv, t, x)m(x - hv)dv \\
&= \int_{-\infty}^{+\infty} hK(v)\left(\Phi_\eta(x, t, x)m(x) - hv\frac{\partial\Phi_\eta(u, t, x)m(u)}{\partial u}\Big|_{u=x}\right. \\
&\quad \left. + \frac{h^2v^2}{2}\frac{\partial^2\Phi_\eta(u, t, x)m(u)}{\partial u^2}\Big|_{u=x} + o(h^2)\right)dv.
\end{aligned}
$$

Then,

$$E\Big[K\Big(\frac{x-X_1}{h}\Big)\eta(Z_1,\delta_1,t,x)\Big]$$

$$= \Phi_\eta(x,t,x)m(x)h + \frac{d_K}{2}\Big(\Phi_\eta(x,t,x)m''(x) + 2\Phi'_\eta(x,t,x)m'(x) + \Phi''_\eta(x,t,x)m(x)\Big)h^3 + o(h^3).$$

Next, an explicit expression for Φ_η is obtained,

$$\Phi_\eta(u,t,x) = \int K(v)\Big(1-F(t-gv|x)\Big)\Phi_\xi(u,t-gv,x)dv,$$

where $\Phi_\xi(u,t,x) = E\Big[\xi(Z_1,\delta_1,t,x)|X_1=u\Big]$ can be written as follows:

$$\Phi_\xi(u,t,x) = \int_0^t \frac{dH_1(z|u)}{1-H(z|x)} - \int_0^t \frac{1-H(v|u)}{\Big(1-H(v|x)\Big)^2}dH_1(v|x).$$

Then, $\Phi_\xi(x,t,x) = 0$ for any x and t and, consequently, $\Phi_\eta(x,t,x) = 0$ for any x and t. Furthermore, using Taylor's formula, expressions for $\Phi'_\eta(x,t,x)$ and $\Phi''_\eta(x,t,x)$ are as follows

$$\Phi'_\eta(x,t,x) = \Big(1-F(t|x)\Big)\Phi'_\xi(x,t,x) + o(g),$$
$$\Phi''_\eta(x,t,x) = \Big(1-F(t|x)\Big)\Phi''_\xi(x,t,x) + o(g).$$

Hence,

$$E\Big[K\Big(\frac{x-X_1}{h}\Big)\eta(Z_1,\delta_1,t,x)\Big]$$

$$= \frac{d_K}{2}\Big(1-F(t|x)\Big)\Big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)m(x)\Big)h^3 + o(h^3).$$

□

Lemma 3.4. *Under Assumptions A.8 and A.10, then*

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

$$= c_K m(x)V_g^1(t_1,t_2,x)h + c_K m(x)V_g^2(t_1,t_2,x)hg + c_K m(x)V_g^3(t_1,t_2,x)hg^2 + d_{K^2}V_g^4(t_1,t_2,x)h^3 + O(h^4) + O(h^3g) + O(hg^3).$$

where

$$
\begin{aligned}
V_g^1(t_1,t_2,x) &= 2J(t_1|x)\Big(1-F(t_2|x)\Big)\mathbb{K}*K\left(\frac{t_2-t_1}{g}\right),\\
V_g^2(t_1,t_2,x) &= 2J(t_1|x)f(t_2|x)\mathbb{K}*K_1\left(\frac{t_1-t_2}{g}\right)\\
&\quad +2J'(t_1|x)\Big(1-F(t_2|x)\Big)\mathbb{K}*K_1\left(\frac{t_2-t_1}{g}\right),\\
V_g^3(t_1,t_2,x) &= J''(t_1|x)\Big(1-F(t_2|x)\Big)\mathbb{K}*K_2\left(\frac{t_2-t_1}{g}\right)\\
&\quad -J(t_1|x)f'(t_2|x)\left(d_K-\mathbb{K}*K_2\left(\frac{t_1-t_2}{g}\right)\right)\\
&\quad +2J'(t_1|x)f(t_2|x)\mathbb{K}_1*K_1\left(\frac{t_2-t_1}{g}\right),\\
V_g^4(t_1,t_2,x) &= m(x)\Big(1-F(t_1|x)\Big)\Big(1-F(t_2|x)\Big)\Phi'_\xi(x,t_1,x)\Phi'_\xi(x,t_2,x)\\
&\quad +\frac{1}{2}D''_g(x,t_1,t_2,x),\\
J(t|x) &= \Big(1-F(t|x)\Big)L(t|x),\\
D_g(u,t_1,t_2,x) &= Cov\Big[\eta(Z_1,\delta_1,t_1,x),\eta(Z_1,\delta_1,t_2,x)|X_1=u\Big]m(u).
\end{aligned}
$$

Proof of Lemma 3.4.

Using the Law of Total Covariance,

$$
Cov\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t_2,x)\right] = C_{11}-C_{12}+C_2, \tag{3.11}
$$

where

$$
C_{11}=E\left[K^2\left(\frac{x-X_1}{h}\right)\Phi_\eta(X_1,t_1,x)\Phi_\eta(X_1,t_2,x)\right],
$$

$$
C_{12}=E\left[K\left(\frac{x-X_1}{h}\right)\Phi_\eta(X_1,t_1,x)\right]E\left[K\left(\frac{x-X_1}{h}\right)\Phi_\eta(X_1,t_2,x)\right]
$$

and

$$
C_2=E\left[\mathrm{Cov}\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t_2,x)\Big|X_1\right]\right].
$$

Asymptotic expressions for the terms involved in (3.11) are found. The first one

becomes

$$\begin{aligned} C_{11} &= \int K^2\Big(\frac{x-u}{h}\Big)\Phi_\eta(u,t_1,x)\Phi_\eta(u,t_2,x)m(u)du \\ &= \int K^2\Big(\frac{x-u}{h}\Big)\Big(\int K(v_1)\big(1-F(t_1-gv_1|x)\big)\Phi_\xi(u,t_1-gv_1,x)dv_1\Big) \\ &\quad \cdot\Big(\int K(v_2)\big(1-F(t_2-gv_2|x)\big)\Phi_\xi(u,t_2-gv_2,x)dv_2\Big)\, m(u)du \\ &= \int\int\int hK^2(w)K(v_1)K(v_2)\big(1-F(t_1-gv_1|x)\big)\big(1-F(t_2-gv_2|x)\big) \\ &\quad \cdot B(x-hw,t_1-gv_1,t_2-gv_2,x)dv_1dv_2dw, \end{aligned}$$

where $B(u,z_1,z_2,x) := \Phi_\xi(u,z_1,x)\Phi_\xi(u,z_2,x)m(u)$.

Since $\Phi_\xi(x,z_1,x) = 0 = \Phi_\xi(x,z_2,x)$ for any x, z_1 and z_2, we have

$$B(x,z_1,z_2,x) = 0,$$

$$\frac{\partial B(u,z_1,z_2,x)}{\partial u}\bigg|_{u=x} = 0$$

and

$$\frac{\partial^2 B(u,z_1,z_2,x)}{\partial u^2}\bigg|_{u=x} = 2\Phi'_\xi(x,z_1,x)\Phi'_\xi(x,z_2,x)m(x).$$

Now, by means of a Taylor expansion of $B(u,t_1-gv_1,t_2-gv_2,x)$ when $u=x-hw$ around $u=x$,

$$B(x-hw,t_1-gv_1,t_2-gv_2,x) = h^2w^2\Phi'_\xi(x,t_1-gv_1,x)\Phi'_\xi(x,t_2-gv_2,x)m(x)+O(h^3).$$

Thus,

$$\begin{aligned} C_{11} &= \int\int\int h^3w^2K^2(w)K(v_1)K(v_2)\big(1-F(t_1-gv_1|x)\big)\big(1-F(t_2-gv_2|x)\big) \\ &\quad \cdot\Phi'_\xi(x,t_1-gv_1,x)\Phi'_\xi(x,t_2-gv_2,x)m(x)dv_1dv_2dw + O(h^4), \end{aligned}$$

and using Taylor expansions of the functions involved when $z_1 = t_2 - gv_1$ and $z_2 = t_2 - gv_2$ around $z_1 = t_1$ and $z_2 = t_2$, respectively, leads to

$$\begin{aligned} C_{11} &= d_{K^2}m(x)\big(1-F(t_1|x)\big)\big(1-F(t_2|x)\big)\Phi'_\xi(x,t_1,x)\Phi'_\xi(x,t_2,x)h^3 \\ &\quad +O(h^4)+O(h^3g^2). \end{aligned} \tag{3.12}$$

From Lemma 3.3, $E\left[K\left(\frac{x-X_1}{h}\right)\Phi_\eta(X_1,t,x)\right] = O(h^3)$. Hence,

$$C_{12} \;=\; O(h^6). \tag{3.13}$$

Now,

$$\begin{aligned} C_2 &= \int K^2\left(\frac{x-z}{h}\right)\mathrm{Cov}\Big[\eta(Z_1,\delta_1,t_1,x),\eta(Z_1,\delta_1,t_2,x)|X_1=z\Big]m(z)dz \\ &= \int hK^2(v)\mathrm{Cov}\Big[\eta(Z_1,\delta_1,t_1,x),\eta(Z_1,\delta_1,t_2,x)|X_1=x-hv\Big]m(x-hv)dv \\ &= c_K D_g(x,t_1,t_2,x)h + d_{K^2}D_g''(x,t_1,t_2,x)h^3 + O(h^4), \end{aligned} \tag{3.14}$$

where $D_g(u,t_1,t_2,x)$ is defined in the statement of Lemma 3.4. An expression for $D_g(x,t_1,t_2,x)$ is calculated. Since

$$E\Big[\eta(Z_1,\delta_1,t,x)|X_1=x\Big] = \int\int K(v)\Big(1-F(t-gv|x)\Big)\Phi_\xi(x,t,x) = 0$$

and

$$\begin{aligned} &E\Big[\eta(Z_1,\delta_1,t_1,x)\eta(Z_1,\delta_1,t_2,x)|X_1=x\Big] \\ &\quad= \int\int K(v_1)K(v_2)\Big(1-F(t_1-gv_1|x)\Big)\Big(1-F(t_2-gv_2|x)\Big) \\ &\qquad E\Big[\xi(Z_1,\delta_1,t_1-gv_1,x)\xi(Z_1,\delta_1,t_2-gv_2,x)|X_1=x\Big]dv_1dv_2, \end{aligned}$$

it follows that

$$\begin{aligned} D_g(x,t_1,t_2,x) &= E\Big[\eta(Z_1,\delta_1,t_1,x)\eta(Z_1,\delta_1,t_2,x)|X_1=x\Big]m(x) \\ &= m(x)\int\int K(v_1)K(v_2)\Big(1-F(t_1-gv_1|x)\Big)\Big(1-F(t_2-gv_2|x)\Big) \\ &\qquad E\Big[\xi(Z_1,\delta_1,t_1-gv_1,x)\xi(Z_1,\delta_1,t_2-gv_2,x)|X_1=x\Big]dv_1dv_2. \end{aligned}$$

Long calculations lead to the following expression for $D_g(x,t_1,t_2,x)$:

$$\begin{aligned} D_g(x,t_1,t_2,x) &= m(x)V_g^1(t_1,t_2,x) + m(x)V_g^2(t_1,t_2,x)g \\ &\quad + m(x)V_g^3(t_1,t_2,x)g^2 + O(g^3), \end{aligned} \tag{3.15}$$

where $V_g^1(t_1,t_2,x)$, $V_g^2(t_1,t_2,x)$ and $V_g^3(t_1,t_2,x)$ are defined in the statement of Lemma 3.4.

By means of similar but more tedious calculations, omitted here, a general expression for $D_g(u,t_1,t_2,x)$ could be obtained. Thus, using expression (3.15) in (3.14), gives:

$$\begin{aligned} C_2 &= c_K m(x) V_g^1(t_1,t_2,x)h + c_K m(x) V_g^2(t_1,t_2,x)hg \\ &\quad + c_K m(x) V_g^3(t_1,t_2,x)h^2 g + d_{K^2} D_g''(x,t_1,t_2,x) m(x) V_g^1(t_1,t_2,x)h^3 \\ &\quad + O(h^4) + O(h^3 g) + O(hg^3). \end{aligned} \tag{3.16}$$

Now, plugging (3.12), (3.13) and (3.16) in (3.11) gives

$$\begin{aligned} & Cov\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t_2,x)\right] \\ &= c_K m(x) V_g^1(t_1,t_2,x)h + c_K m(x) V_g^2(t_1,t_2,x)hg + c_K m(x) V_g^3(t_1,t_2,x)hg^2 \\ &\quad + d_{K^2} V_g^4(t_1,t_2,x)h^3 + O(h^4) + O(h^3 g) + O(hg^3). \end{aligned}$$

□

Lemma 3.5. *Under Assumptions A.8 and A.10, then*

$$\begin{aligned} & Var\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right] \\ &= c_K m(x)\Big(1-F(t|x)\Big)^2 L(t|x)h \\ &\quad + c_K m(x)(c_{\mathbb{K}}-1)\Big(1-F(t|x)\Big)^2 L'(t|x)hg + c_K m(x) V(t,x) hg^2 \\ &\quad + \frac{d_{K^2}}{m^2(x)}\left(m(x)\Big(1-F(t|x)\Big)^2\Big(\Phi'_\xi(x,t,x)\Big)^2 + \frac{1}{2}D(t,x)\right)\frac{h}{n} \\ &\quad + o\left(\frac{h^2+g^2}{nh}\right). \end{aligned}$$

where

$$\begin{aligned} V(t,x) &= -d_K\Big(1-F(t|x)\Big)L(t|x)f'(t|x) + \left(\frac{1}{2}-\mu_1(\mathbb{K}^2)\right)\Big(1-F(t|x)\Big)^2 L''(t|x) \\ &\quad + \Big(2\mu_1(\mathbb{K}^2)-1\Big)\Big(1-F(t|x)\Big)L'(t|x)f(t|x), \end{aligned}$$

$$
\begin{aligned}
D(t,x) &= \Big(1-F(t|x)\Big)^2\Big(m''(x)N(x,t,t,x)+m(x)N''(x,t,t,x)\\
&\quad +2m'(x)N'(x,t,t,x)-2c_K m(x)\Phi'_\xi(x,t,x)\Phi'_\xi(x,t,x)\Big),\\
N(u,t_1,t_2,x) &= E\Big[\xi(Z_1,\delta_1,t_1,x)\xi(Z_1,\delta_1,t_2,x)\Big|X_1=u\Big].
\end{aligned}
$$

Proof of Lemma 3.5.

Using Lemma 3.4 for $t_1=t_2=t$,

$$
\begin{aligned}
&Var\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right]\\
&\quad = Cov\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x),K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right]\\
&\quad = c_K m(x)V_g^1(t,t,x)h+c_K m(x)V_g^2(t,t,x)hg+c_K m(x)V_g^3(t,t,x)hg^2\\
&\qquad +d_{K^2}V_g^4(t,t,x)h^3+O(h^4)+O(h^3g)+O(hg^3),
\end{aligned}
$$

where

$$
\begin{aligned}
V_g^1(t,t,x) &= 2\Big(1-F(t|x)\Big)^2L(t|x)\mathbb{K}*K(0),\\
V_g^2(t,t,x) &= 2\Big(1-F(t|x)\Big)^2L'(t|x)\mathbb{K}*K_1(0),\\
V_g^3(t,t,x) &= \Big(-f'(t|x)L(t|x)\Big(1-F(t|x)\Big)\\
&\qquad -2f(t|x)L'(t|x)\Big(1-F(t|x)\Big)\Big)\mathbb{K}*K_2(0)\\
&\qquad -\Big(1-F(t|x)\Big)L(t|x)f'(t|x)\Big(d_K-\mathbb{K}*K_2(0)\Big)\\
&\qquad \Big(-2f^2(t|x)L(t|x)+\Big(1-F(t|x)\Big)L'(t|x)f(t|x)\Big)\mathbb{K}_1*K_1(0),\\
V_g^4(t,t,x) &= m(x)\Big(1-F(t|x)\Big)^2\Big(\Phi'_\xi(x,t,x)\Big)^2+\frac{1}{2}D''_g(x,t,t,x).
\end{aligned}
$$

Definitions of $K_l(u)$ and $\mathbb{K}_l(u)$ in (3.5) and assumption A.8 give:

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

$$
\begin{aligned}
\mathbb{K} * K_1(0) &= \int \mathbb{K}(u)K_1(-u)du = \int \mathbb{K}(u)K_1(-u)du = -\int uK(u)\mathbb{K}(u)du \\
&= -\frac{1}{2}\Big[u\mathbb{K}^2(u)\Big]_{-1}^{+1} + \frac{1}{2}\int \mathbb{K}^2(u)du = -\frac{1}{2} + \frac{1}{2}c_{\mathbb{K}}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{K} * K_2(0) &= \int \mathbb{K}(u)K_2(-u)du = \int \mathbb{K}(u)u^2K(-u)du = \int u^2K(u)\mathbb{K}(u)du \\
&= \frac{1}{2}\Big[u^2\mathbb{K}(u)\Big]_{-1}^{+1} - \int u\mathbb{K}^2(u)du = \frac{1}{2} - \mu_1(\mathbb{K}^2)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{K}_1 * K_1(0) &= \int \mathbb{K}_1(u)K_1(-u)du = \int \mathbb{K}_1(u)(-u)K(-u)du \\
&= -\int uK(u)\mathbb{K}_1(u)du = -\int uK(u)\bigg(\int_{-\infty}^{u} K_1(v)dv\bigg)du \\
&= -\int uK(u)\bigg(\int_{-\infty}^{u} vK(v)dv\bigg)du = -\int\int_{\{v\le u\}} uvK(u)K(v)dudv \\
&= -\frac{1}{2}\bigg(\int\int_{\{v\le u\}} uvK(u)K(v)dudv + \int\int_{\{u\le v\}} vuK(v)K(u)dvdu\bigg) \\
&= -\frac{1}{2}\bigg(\int\int_{\mathbb{R}^2} uvK(u)K(v)dudv + 0\bigg) = 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
V_g^1(t,t,x) &= \Big(1 - F(t|x)\Big)^2 L(t|x), \\
V_g^2(t,t,x) &= (c_{\mathbb{K}} - 1)\Big(1 - F(t|x)\Big)^2 L'(t|x), \\
V_g^3(t,t,x) &= -d_K\Big(1 - F(t|x)\Big)L(t|x)f'(t|x) \\
&\quad + \bigg(\frac{1}{2} - \mu_1(\mathbb{K}^2)\bigg)\Big(1 - F(t|x)\Big)^2 L''(t|x) \\
&\quad + \Big(2\mu_1(\mathbb{K}^2) - 1\Big)\Big(1 - F(t|x)\Big)L'(t|x)f(t|x).
\end{aligned}
$$

Now, an expression for $D_g(u,t_1,t_2,x)$ is found:

$$\begin{aligned}
D_g(u,t_1,t_2,x) &= E\Big[\eta(Z_1,\delta_1,t_1,x)\eta(Z_1,\delta_1,t_1,x)\Big|X_1=u\Big]m(u)\\
&\quad -E\Big[\eta(Z_1,\delta_1,t_1,x)\Big|X_1=u\Big]E\Big[\eta(Z_1,\delta_1,t_1,x)\Big|X_1=u\Big]m(u)\\
&= \int\int K(v_1)K(v_2)\Big(1-F(t_1-gv_1|x)\Big)\Big(1-F(t_2-gv_2|x)\Big)\\
&\quad \cdot N(u,t_1-gv_1,t_2-gv_2,x)m(u)dv_1dv_2\\
&\quad -\int K(v)^2\Big(1-F(t_1-gv|x)\Big)\Big(1-F(t_2-gv|x)\Big)\\
&\quad \cdot\Phi_\xi(u,t_1-gv,x)\Phi_\xi(u,t_2-gv,x)m(u)dv.
\end{aligned}$$

Differentiating $D_g(u,t_1,t_2,x)$ twice with respect to u and using Taylor's formula when g tends to 0, an expression for $D_g''(u,t_1,t_2,x)$ derives

$$\begin{aligned}
&D_g''(u,t_1,t_2,x)\\
&= \int\int K(v_1)K(v_2)dv_1dv_2\Big(1-F(t_1|x)\Big)\Big(1-F(t_2|x)\Big)\\
&\quad \cdot\Big(m''(u)N(u,t_2,t_2,x)+m(u)N''(u,t_1,t_2,x)+2m'(u)N'(u,t_1,t_2,x)\Big)\\
&\quad -\int K^2(v)dv\Big(1-F(t_1|x)\Big)\Big(1-F(t_2|x)\Big)\Big(\Phi_\xi''(u,t_1,x)\Phi_\xi(u,t_2,x)m(u)\\
&\quad +\Phi_\xi(u,t_1,x)\Phi_\xi''(u,t_2,x)m(u)+\Phi_\xi(u,t_1,x)\Phi_\xi(u,t_2,x)m''(u)\\
&\quad +2\Phi_\xi'(u,t_1,x)\Phi_\xi'(u,t_2,x)m(u)+2\Phi_\xi(u,t_1,x)\Phi_\xi'(u,t_2,x)m'(u)\\
&\quad +2\Phi_\xi'(u,t_1,x)\Phi_\xi(u,t_2,x)m'(u)\Big)+o(1).
\end{aligned}$$

Note that $\Phi_\xi(x,t,x)=0$, $\int\int K(v_1)K(v_2)dv_1dv_2=1$ and $c_K=\int K^2(v)dv$. Hence,

$$\begin{aligned}
D_g''(x,t,t,x) &= \Big(1-F(t|x)\Big)^2\Big(m''(x)N(x,t,t,x)+m(x)N''(x,t,t,x)\\
&\quad +2m'(x)N'(x,t,t,x)-2c_Km(x)\Phi_\xi'(x,t,x)\Phi_\xi'(x,t,x)\Big)+o(1)\\
&= D(t,x)+o(1).
\end{aligned}$$

Therefore,

$$V_g^4(t,t,x)=m(x)\Big(1-F(t|x)\Big)^2\Big(\Phi_\xi'(x,t,x)\Big)^2+\frac{1}{2}D(t,x)+o(1),$$

and, consequently,

$$Var\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right]$$

$$\begin{aligned}
&= c_K m(x) V_g^1(t,t,x)h + c_K m(x) V_g^2(t,t,x)hg + c_K m(x) V_g^3(t,t,x)hg^2 \\
&\quad + \frac{d_{K^2}}{m^2(x)}\Big(m(x)\big(1-F(t|x)\big)^2\big(\Phi'_\xi(x,t,x)\big)^2 + \frac{1}{2}D(t,x)\Big)\frac{h}{n} + o\left(\frac{g^2}{nh}+\frac{h}{n}\right).
\end{aligned}$$

□

Proof of Theorem 3.2.

Denoting $\widetilde{F}^B_{h,g}(t|x) = 1-\widetilde{S}^B_{h,g}(t|x)$ and $\widehat{F}^B_h(dt|x) = 1-\widehat{S}^B_h(dt|x)$, standard algebra gives

$$\widetilde{F}^B_{h,g}(t|x) - F(t|x) = \int \mathbb{K}\left(\frac{t-u}{g}\right)\widehat{F}^B_h(du|x) - F(t|x) = A_1 + A_2, \tag{3.17}$$

where

$$A_1 = \int \mathbb{K}\left(\frac{t-u}{g}\right)\big(\widehat{F}^B_h(du|x) - F(du|x)\big)$$

and

$$A_2 = \int \mathbb{K}\left(\frac{t-u}{g}\right)F(du|x) - F(t|x).$$

Using Lemma 3.2 and Theorem 3.1, it is obtained

$$\begin{aligned}
A_1 &= \int \mathbb{K}\left(\frac{t-y}{g}\right)\big(\widehat{F}^B_h(dy|x) - F(dy|x)\big) \\
&= \left[\mathbb{K}\left(\frac{t-y}{g}\right)\big(\widehat{F}^B_h(y|x) - F(y|x)\big)\right]_{y=-\infty}^{y=+\infty} \\
&\quad + \int \frac{1}{g}K\left(\frac{t-y}{g}\right)\big(\widehat{F}^B_h(y|x) - F(y|x)\big)dy \\
&= \int \frac{1}{g}K\left(\frac{t-y}{g}\right)\big(\widehat{F}^B_h(y|x) - F(y|x)\big)dy \\
&= \int K(u)\big(\widehat{F}^B_h(t-gu|x) - F(t-gu|x)\big)du \\
&= \int K(u)\Big(\big(F(t-gu|x)-1\big)\sum_{i=1}^n w_{h,i}(x)\xi(Z_i,\delta_i,t-gu,x) + R_n(t-gu|x)\Big)du
\end{aligned}$$

Then,

$$A_1 = A_{11} + A_{12} \tag{3.18}$$

where

$$A_{11} = -\int K(u)\Big(1 - F(t - gu|x)\Big)\sum_{i=1}^{n} w_{h,i}(x)\xi(Z_i, \delta_i, t - gu, x)du$$

and

$$A_{12} = \int K(u)R_n(t - gu|x)du.$$

First considering A_{11} in (3.18),

$$\begin{aligned} A_{11} &= -\int K(u)\Big(1 - F(t - gu|x)\Big)\sum_{i=1}^{n} w_{h,i}(x)\xi(Z_i, \delta_i, t - gu, x)du \\ &= -\sum_{i=1}^{n} w_{h,i}(x)\int K(u)\Big(1 - F(t - gu|x)\Big)\xi(Z_i, \delta_i, t - gu, x)du, \end{aligned}$$

and considering $\eta(Z, \delta, t, x)$ defined in Section 3.3.1, it is obtained

$$A_{11} = -\sum_{i=1}^{n} w_{h,i}(x)\eta(Z_i, \delta_i, t, x). \tag{3.19}$$

Considering A_{12} in (3.18) and using A.8, it follows that

$$|A_{12}| = \left|\int K(u)R_n'(t - gu|x)du\right| \leq \int_{-1}^{1} K(u)\Big|R_n'(t - gu|x)\Big|du \leq \sup_{z\in[t-g,t+g]}\Big|R_n'(z|x)\Big|.$$

Fix $\varepsilon > 0$ and define $a' = l + \varepsilon$, $b' = u - \varepsilon$. Then,

$$\sup_{(t,x)\in[a',b']\times I}|A_{12}| \leq \sup_{(t,x)\in[a',b']\times I}\left\{\sup_{z\in[t-g,t+g]}\Big|R_n(z|x)\Big|\right\} \tag{3.20}$$

On the one hand, there exists $n_0 \in \mathbb{N}$ such that $g = g_n \leq \varepsilon$ for all $n \geq n_0$. So, $z \in [t - g, t + g]$ implies that $|z - t| \leq g \leq \varepsilon$ and equivalently, $t - \varepsilon \leq z \leq t + \varepsilon$.

On the other hand, $t \in [a', b']$ implies that $l + \varepsilon = a' \leq t \leq b' = u - \varepsilon$.

Therefore,

$$z \leq t + \varepsilon \leq (u - \varepsilon) + \varepsilon = u \Rightarrow z \leq u$$

and also,

$$z \geq t - \varepsilon \geq (l + \varepsilon) - \varepsilon = l \Rightarrow z \geq l.$$

Hence, $z \in [l, u]$ and $x \in I$. So, for $t \in [a', b']$,

$$\sup_{z\in[t-g,t+g]}\Big|R_n(z|x)\Big| \leq \sup_{(t',x')\in[l,u]\times I}\Big|R_n(t'|x')\Big|. \tag{3.21}$$

Recalling the inequality obtained in (3.20) and applying the inequality in (3.21), one has

$$\sup_{(t,x)\in[a',b']\times I}|A_{12}| \leq \sup_{(t,x)\in[a',b']\times I}\left\{\sup_{(t',x')\in[l,u]\times I}\left|R_n(t'|x')\right|\right\}$$

and from Theorem 3.1,

$$\sup_{(t',x')\in[l,u]\times I}\left|R_n(t'|x')\right| = O\left(\frac{\ln n}{nh}\right)^{3/4} \quad \text{a. s.}$$

Finally, defining $R_n^1(t|x) = A_{12}$, the following is obtained

$$\sup_{(t,x)\in[a',b']\times I}\left|R_n^1(t|x)\right| = \sup_{(t,x)\in[a',b']\times I}|A_{12}| = O\left(\frac{\ln n}{nh}\right)^{3/4} \quad \text{a. s.} \tag{3.22}$$

Now, considering A_2 in (3.17) and using Lemma 3.2, it follows that

$$\begin{aligned}
A_2 &= \int \mathbb{K}\left(\frac{t-y}{g}\right)F(dy|x) - F(t|x) \\
&= \left[\mathbb{K}\left(\frac{t-y}{g}\right)F(y|x)\right]_{y=-\infty}^{y=+\infty} + \int_{-\infty}^{+\infty}\frac{1}{g}K\left(\frac{t-y}{g}\right)F(y|x)dy - F(t|x) \\
&= \int_{-\infty}^{+\infty}\frac{1}{g}K\left(\frac{t-y}{g}\right)F(y|x)dy - F(t|x) = \int_{-\infty}^{+\infty}K(u)F(t-gu|x)du - F(t|x)
\end{aligned}$$

Assuming that $g = g_n$ tends to zero when n tends to infinity, Taylor's formula for $F(t-gu|x)$ gives:

$$F(t-gu|x) = F(t|x) - guF'(t|x) + \frac{1}{2}(gu)^2F''(t|x) + o(g^2)$$

Then,

$$\begin{aligned}
A_2 &= \int_{-\infty}^{+\infty}K(u)\Big[F(t|x) - guF'(t|x) + \frac{1}{2}(gu)^2F''(t|x) + o(g^2)\Big]du - F(t|x) \\
&= F(t|x)\int_{-\infty}^{+\infty}K(u)du - F'(t|x)g\int_{-\infty}^{+\infty}uK(u)du + \frac{1}{2}g^2F''(t|x)\int_{-\infty}^{+\infty}u^2K(u)du \\
&\quad + o(g^2)\int_{-\infty}^{+\infty}K(u)du - F(t|x)
\end{aligned}$$

Using assumption A.8,

$$A_2 = \frac{1}{2}d_KF''(t|x)g^2 + R_n^2(t|x) \tag{3.23}$$

with $\sup\left\{\left|R_n^2(t|x)\right| : (t,x)\in[l,u]\times I\right\} = o(g^2)$.

Finally, Equations (3.17) - (3.23) give

$$\widetilde{F}^B_{h,g}(t|x) - F(t|x) = -\sum_{i=1}^{n} w_{h,i}(x)\eta(Z_i,\delta_i,t,x) + \frac{1}{2}d_K F''(t|x)g^2 + R^1_n(t|x) + R^2_n(t|x)$$

where

$$\sup_{(t,x)\in[a',b']\times I}\Big|R^1_n(t|x)\Big| = O\left(\frac{\ln n}{nh}\right)^{3/4} \quad \text{a. s.}$$

which proves Theorem 3.2 since $\widetilde{S}^B_{h,g}(t|x) - S(t|x) = F(t|x) - \widetilde{F}^B_{h,g}(t|x)$.

□

Proof of Lemma 3.1.

Theorem 3.2 gives

$$\widetilde{S}^B_{h,g}(t|x) - S(t|x) = \sum_{i=1}^{n} w_{h,i}(x)\eta(Z_i,\delta_i,t,x) - \frac{1}{2}d_K F''(t|x)g^2 + R^1_n(t|x) + R^2_n(t|x) \quad a.s.,$$

where

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

Note that

$$\widehat{m}_h(x) := \frac{1}{nh}\sum_{j=1}^{n} K\Big((x-X_j)/h\Big)$$

is the Parzen-Rosenblatt estimator of the density function of X, $m(x)$. Then,

$$\begin{aligned}
\sum_{i=1}^{n} w_{h,i}(x)\eta(Z_i,\delta_i,t,x) &= \sum_{i=1}^{n}\frac{1}{nh}\frac{K\Big((x-X_i)/h\Big)}{\widehat{m}_h(x)}\eta(Z_i,\delta_i,t,x) \\
&= \sum_{i=1}^{n}\frac{1}{nh}\frac{K\Big((x-X_i)/h\Big)}{m(x)}\eta(Z_i,\delta_i,t,x) \\
&\quad + \frac{m(x)-\widehat{m}_h(x)}{\widehat{m}_h(x)}\sum_{i=1}^{n}\frac{1}{nh}\frac{K\Big((x-X_i)/h\Big)}{m(x)}\eta(Z_i,\delta_i,t,x) \\
&= \sum_{i=1}^{n} w^A_{h,i}(x)\eta(Z_i,\delta_i,t,x) + R^3_n(t|x)
\end{aligned}$$

where

$$w^A_{h,i}(x) = \frac{1}{nh}\frac{K\Big((x-X_i)/h\Big)}{m(x)}$$

for all $i = 1, ..., n$ and

$$R^3_n(t|x) = \frac{m(x)-\widehat{m}_h(x)}{\widehat{m}_h(x)}\sum_{i=1}^{n} w^A_{h,i}(x)\eta(Z_i,\delta_i,t,x)$$

Since $\widehat{m}_h(x)$ is a consistent estimator of $m(x)$ and its bias and variance convergence rates are $O(h^2)$ and $O(1/nh)$, respectively (see Silverman (1986)),

$$R_n^3(t|x) = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)\sum_{i=1}^{n} w_{h,i}^A(x)\eta(Z_i,\delta_i,t,x).$$

From Lemma 3.3,

$$E\left[\sum_{i=1}^{n} w_{h,i}^A(x)\eta(Z_i,\delta_i,t,x)\right] = nE\left[\frac{1}{nh}\frac{1}{m(x)}K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right] = O(h^2).$$

From Lemma 3.5,

$$Var\left[\sum_{i=1}^{n} w_{h,i}^A(x)\eta(Z_i,\delta_i,t,x)\right] = nVar\left[\frac{1}{nh}\frac{1}{m(x)}K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right] = O\left(\frac{1}{nh}\right).$$

Therefore,

$$\sum_{i=1}^{n} w_{h,i}^A(x)\eta(Z_i,\delta_i,t,x) = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)$$

and

$$\sup_{[l,u]\times I} |R_n^3(t|x)| = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)^2.$$

Finally,

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

□

Proof of Theorem 3.3.

Lemma 3.1 gives

$$\begin{aligned}\widetilde{S}_{h,g}^{AB}(t|x) - S(t|x) &= \sum_{i=1}^{n} w_{h,i}^A(x)\eta(Z_i,\delta_i,t,x) - \frac{1}{2}d_K F''(t|x)g^2 \\ &= \sum_{i=1}^{n} \varphi_{n,i}(t,x) - \frac{1}{2}d_K F''(t|x)g^2,\end{aligned}$$

where $\varphi_{n,i}(t,x) = \dfrac{1}{nh}\dfrac{1}{m(x)}K\left(\dfrac{x-X_i}{h}\right)\eta(Z_i,\delta_i,t,x)$ are independent and identically distributed random variables for all $i = 1,...,n$. Consequently,

$$\text{Bias}\Big(\widetilde{S}_{h,g}^{AB}(t|x)\Big) = nE\Big(\varphi_{n,1}(t,x)\Big) - \frac{1}{2}d_K F''(t|x)g^2 \tag{3.24}$$

and

$$\mathrm{Var}\Big(\widetilde{S}^{AB}_{h,g}(t|x)\Big) \;=\; n\mathrm{Var}\Big(\varphi_{n,1}(t,x)\Big). \tag{3.25}$$

First, an expression for $E\Big(\varphi_{n,1}(t,x)\Big)$ is found. From Lemma 3.3,

$$\begin{aligned} E\Big(\varphi_{n,1}(t,x)\Big) &= \frac{1}{nh}\frac{1}{m(x)}E\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right] \\ &= \frac{d_K\Big(1-F(t|x)\Big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x)+\Phi''_\xi(x,t,x)m(x)\Big)\frac{h^2}{n}+o\left(\frac{h^2}{n}\right) \end{aligned}$$

and replacing it in (3.24) the bias of $\widetilde{S}^{AB}_{h,g}(t|x)$ is obtained.

To derive the variance of the estimator, an asymptotic expression for $\mathrm{Var}\Big(\varphi_{n,1}(t,x)\Big)$ is found. Using Lemma 3.5,

$$\begin{aligned} \mathrm{Var}\Big(\varphi_{n,1}(t,x)\Big) &= \frac{1}{n^2h^2m^2(x)}Var\left[K\left(\frac{x-X_1}{h}\right)\eta(Z_1,\delta_1,t,x)\right] \\ &= \frac{c_K}{m(x)}V_1(t,x)\frac{1}{n^2h}+\frac{c_K}{m(x)}V_2(t,x)\frac{g}{n^2h}+\frac{c_K}{m(x)}V_3(t,x)\frac{g^2}{n^2h} \\ &\quad +\frac{d_{K^2}}{m^2(x)}\Big(m(x)\Big(1-F(t|x)\Big)^2\Big(\Phi'_\xi(x,t,x)\Big)^2+\frac{1}{2}D(t,x)\Big)\frac{h}{n^2} \\ &\quad +o\left(\frac{g}{n^2h}+\frac{h}{n^2}\right). \end{aligned} \tag{3.26}$$

Plugging (3.26) in (3.25), the variance part of the theorem is proved.

□

Proof of Theorem 3.4.

Lemma 3.1 gives:

$$\begin{aligned} \sqrt{nh}\Big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\Big) &= \sqrt{nh}\sum_{i=1}^{n}\varphi_{n,i}(t,x)-\sqrt{nh}\frac{1}{2}d_KF''(t|x)g^2 \\ &\quad +\sqrt{nh}\widetilde{R}_n(t|x) \quad a.s. \end{aligned} \tag{3.27}$$

The variables $\varphi_{n,i}(t,x) = \dfrac{1}{nh}\dfrac{1}{m(x)}K\left(\dfrac{x-X_i}{h}\right)\eta(Z_i,\delta_i,t,x)$ are independent and identically distributed random variables for all $i=1,...,n$. Assuming $\dfrac{(\ln n)^3}{nh}\to 0$,

the remainder term $\sqrt{nh}\widetilde{R}_n(t|x)$ is negligible with respect to the dominant terms in (3.27). According to Theorem 3.3, the variance of the dominant terms in (3.27) is given by

$$Var\left[\sqrt{nh}\sum_{i=1}^{n}\varphi_{n,i}(t,x)\right] \;=\; nh\frac{c_K}{m(x)}\Big(1-F(t|x)\Big)L(t|x)\frac{1}{nh}+O\left(nh\left(\frac{h}{n}+\frac{g}{nh}\right)\right).$$

Since the supports of the functions K and m are compact and $F(t|x)$ is bounded, this variance is finite.

Therefore, the asymptotic distribution of $\sqrt{nh}\Big(\widetilde{S}^B_{h,g}(t|x)-S(t|x)\Big)$ is the same as the asymptotic distribution of $\sqrt{nh}\sum_{i=1}^{n}\varphi_{n,i}(t,x)$.

If Lindeberg's condition for triangular arrays (Theorem 7.2 in Billingsley (1968)) is satisfied, then

$$\sum_{i=1}^{n}\left(\sqrt{nh}\varphi_{n,i}(t,x)-E\Big[\sqrt{nh}\varphi_{n,i}(t,x)\Big]\right)\xrightarrow{d}N(0,\sigma),$$

where

$$\sigma^2=\frac{c_K}{m(x)}\Big(1-F(t|x)\Big)L(t|x).$$

Defining the following indicator function

$$\mathbb{1}_{n,i} \;=\; \mathbb{1}\Big(\Big|\sqrt{nh}\varphi_{n,i}(t,x)-E[\sqrt{nh}\varphi_{n,i}(t,x)]\Big|>\varepsilon\sigma\Big),$$

Lindeberg's condition requires that

$$\lim_{n\to\infty}\frac{1}{\sigma^2}E\left[\sum_{i=1}^{n}\left(\sqrt{nh}\varphi_{n,i}(t,x)-E\Big[\sqrt{nh}\varphi_{n,i}(t,x)\Big]\right)^2\mathbb{1}_{n,i}\right]=0, \tag{3.28}$$

for every $\varepsilon>0$.

By applying assumption A.3d, it is easy to prove that $\xi(Z,\delta,t,x)$ is bounded:

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

and, consequently, η is also bounded:

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

Since η is bounded, K and $m(x)$ have compact support and $nh\to\infty$, $\Big\{\varphi_{n,i}(t,x)-E\Big[\varphi_{n,i}(t,x)\Big], i=1,...,n,\ n\in\mathbb{N}\Big\}$ is a sequence of random variables which is bounded by a convergent to zero nonrandom sequence, $\dfrac{\varepsilon\sigma}{\sqrt{nh}}$. Hence, there exists $n_0\in\mathbb{N}$ such that for all $i=1,...,n$, $\mathbb{1}_{n,i}=0$ for all $n\geq n_0$ and

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

which proves Lindeberg's condition in (3.28). Then,

$$\sum_{i=1}^{n}\left(\sqrt{nh}\varphi_{n,i}(t,x)-E\Big[\sqrt{nh}\varphi_{n,i}(t,x)\Big]\right)\xrightarrow{d}N(0,\sigma)$$

and, therefore, using (3.27) and Slutsky lemma, the asymptotic normality of the estimator holds:

$$\sqrt{nh}\Big(\widetilde{S}^B_{h,g}(t|x)-S(t|x)\Big)\xrightarrow{d}N(\mu,\sigma).$$

Using Theorem 3.3 under assumptions of Theorem 3.4, $C_h:=\lim_{n\to\infty}n^{1/5}h>0$ and $C_g:=\lim_{n\to\infty}n^{1/5}g>0$,

$$\begin{aligned}\mu &= C_h^{5/2}\frac{d_K\Big(1-F(t|x)\Big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x)+\Phi''_\xi(x,t,x)m(x)\Big) \\ &\quad -C_h^{1/2}C_g^{4/2}\frac{1}{2}d_KF''(t|x).\end{aligned}$$

□

Chapter 4

Bootstrap bandwidth selection for the smoothed Beran's survival estimator

4.1 Introduction

Beran's estimator and the smoothed Beran's estimator of the conditional survival functions were presented and compared in Chapter 3. Their asymptotic properties have been deeply analysed in Iglesias-Pérez and González-Manteiga (1999) and Chapter 3, respectively. Simulation studies carried out in previous chapter show a good performance of the smoothed Beran's estimator. However, these preceding studies were carried out using the smoothing parameters that minimised the mean integrated squared error obtained from the theoretical survival curve. Since the asymptotic bias and variance expressions are complex and depend on several population parameters, they are not useful in practice to obtain plug-in estimations of the theoretical bandwidths. The goal of this chapter is to propose resampling techniques to approximate them.

Bootstrap has become a strong tool in many statistical applications since it was

first introduced by Efron (1979). Bootstrap for right censored data was first proposed by Efron (1981) and the bootstrap method and its applications were studied in Efron and Tibshirani (1993). Asymptotic theory for the bootstrap for right censored data was stablished by Reid (1981), Akritas (1986) and Lo and Singh (1986). In Van Keilegom and Veraverbeke (1997), bootstrap for nonparametric regression with right censored observations at fixed covariate values was studied. A bootstrap approach for the nonparametric censored regression setup was studied in Li and Datta (2001). In Geerdens et al. (2017) a local cross-validation bandwidth selector was proposed.

Our approach follows the ideas of Li and Datta (2001), and it is based on the obvious bootstrap. Both Beran's and the smoothed Beran's estimators are bootstrapped in order to approximate their corresponding optimal bandwidths. The existing theoretical results only allow to obtain pointwise and theoretical confidence intervals, which are not computable in practice, since the variance of the estimator again depends on unknown population quantities. Therefore, the bootstrap is also useful to compute confidence regions.

Bootstrap selectors for the bandwidths of Beran's and the smoothed Beran's estimators are proposed. A simulation study shows the behaviour of the survival estimators with bootstrap bandwidths. The issue of obtaining confidence regions for the conditional survival function, $S(t|x)$, for a fixed value of $x \in I \subseteq \mathbb{R}$ and t covering the interval $I_T \subseteq \mathbb{R}^+$, is also addressed using Beran and the smoothed Beran's estimators. This work is motivated by a real data application based on studying the survival times of COVID-19 patients in Galicia, Spain, during the first weeks of the breakdown.

4.2 Bandwidth selection for Beran's and the smoothed Beran's survival estimators

In this section, methods for the automatic selection of the bandwidths for Beran's estimator in (2.3) and the smoothed Beran's estimator in (3.4) of the conditional survival function are proposed. These estimators are based on the right censored random sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^{n}$ of (X, Z, δ).

4.2.1 Beran's estimator

There are two classic methods for bootstrap resampling in a censoring context: the obvious bootstrap and the simple bootstrap. The equivalence between both methods in an unconditional setup is proved in Efron (1981). In Li and Datta (2001), this result is extended to the case where a covariate is involved, assuming there is no ties in the sample values of the covariate. This was done by proving the equivalence of the two resampling methods, the obvious bootstrap and the simple weighted bootstrap. In this chapter, the following obvious bootstrap method combined with a smoothed bootstrap for the covariate is proposed for the automatic selection of the covariate bandwidth h of Beran's estimator, $\widehat{S}_h^B(t|x)$, defined in (2.3). Here, this estimator is simply denoted by $\widehat{S}_h(t|x)$.

Algorithm for bootstrap resampling based on Beran's estimator

Let $I_1 \subseteq \mathbb{R}$ be an interval containing appropriate bandwidth values and let $r \in I_1$ be pilot bandwidth for the bootstrap resampling:

1. Obtain $U_1, \ldots, U_n$ iid with $U_i \sim U(0,1)$ and $V_1, \ldots, V_n$ iid with common density K for all $i = 1, \ldots, n$.

2. For each $i = 1, \ldots, n$, define

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

Generate T_i^* from Beran's estimator of the conditional distribution of T using the sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ and bandwidth r, denoted by $\widehat{F}_r(t|X_i^*)$, and C_i^* from the Beran's estimator of the conditional distribution of C using the sample $\{(X_i, Z_i, 1-\delta_i)\}_{i=1}^n$ and bandwidth r, denoted by $\widehat{G}_r(t|X_i^*)$.

The estimators $\widehat{F}_r(t|X_i^*)$ and $\widehat{G}_r(t|X_i^*)$ 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

$$Z_i^* = \min\{T_i^*, C_i^*\},$$

$$\delta_i^* = I\Big(T_i^* \leq C_i^*\Big).$$

4. Consider the bootstrap resample $\Big\{(X_i^*, Z_i^*, \delta_i^*)\Big\}_{i=1}^n$.

In this chapter, in order to estimate the survival function, $S(t|x)$, for a fixed $x \in I$ and t covering the interval $I_T \subset \mathbb{R}$, our benchmark is the bandwidth $h_{MISE} \in I_1$, that minimizes the mean integrated squared error given by

$$MISE_x(h) = E\left(\int_{I_T} \Big(\widehat{S}_h(t|x) - S(t|x)\Big)^2 dt\right) \tag{4.1}$$

whose bootstrap approximation is

$$MISE_x^*(h) = E^*\left(\int_{I_T} \Big(\widehat{S}_h^*(t|x) - \widehat{S}_r(t|x)\Big)^2 dt\right)$$

where $\widehat{S}_r(t|x)$ is the estimation of the theoretical survival function with pilot bandwidth, r, using the sample $\Big\{(X_i, Z_i, \delta_i)\Big\}_{i=1}^n$ and $\widehat{S}_h^*(t|x)$ is the bootstrap estimation of $S(t|x)$ with bandwidth h, using the bootstrap resample $\Big\{(X_i^*, Z_i^*, \delta_i^*)\Big\}_{i=1}^n$.

The resampling distribution of $\widehat{S}_h^*(t|x)$ cannot be computed in a close form, so the Monte Carlo method is used. It is based on obtaining B bootstrap resamples and estimating $\widehat{S}_h^*(t|x)$ for each of them. Thus, the distribution of $\widehat{S}_h^*(t|x)$ is approximated by the empirical one of $\widehat{S}_h^{*,1}(t|x), \ldots, \widehat{S}_h^{*,B}(t|x)$, obtained from B bootstrap resamples and the bootstrap version of the estimation error of Beran's estimator for any smoothing parameter h is given by

$$MISE_x^*(h) \simeq \frac{1}{B}\sum_{k=1}^{B}\left(\int_{I_T} \Big(\widehat{S}_h^{*,k}(t|x) - \widehat{S}_r(t|x)\Big)^2 dt\right). \tag{4.2}$$

Likewise, the integral is approximated by a Riemann sum.

Algorithm for bootstrap bandwidth selector for Beran's estimator

Let $x \in I$ be a fixed value of the covariate, $t \in I_T$ and $r \in I_1$:

1. Compute $\widehat{S}_r(t|x)$ from the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$.
2. Obtain B bootstrap resamples of the form $\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\}_{i=1}^n$ with $k = 1, ..., B$ using the bootstrap based on Beran's estimator with pilot bandwidth $r \in I_1$ and calculate $\widehat{S}_h^{*,k}(t|x)$ for each of them.
3. Approximate $MISE_x^*(h)$ according to (4.2).
4. Repeat Steps 1–3 for values of h in a grid of I_1.
5. Select the value of h that provides the smallest $MISE_x^*(h)$ as the bootstrap bandwidth h^*.

Concerning the auxiliary bandwidth $r \in I_1$, a preliminary analysis not shown here suggests the following choice for the pilot bandwidth:

$$r = c\frac{\Big(Q_X(0.975) - Q_X(0.025)\Big)}{2}\left(\sum_{i=1}^{n}\delta_i\right)^{-1/3}, \tag{4.3}$$

where $Q_X(u)$ is the u quantile of the sample $\Big\{X_i\Big\}_{i=1}^n$, as a suitable pilot bandwidth in this context. Equation (4.3) considers the variability of the covariate, $Q_X(0.975) - Q_X(0.025)$, and the uncensored sample size, $\sum_{i=1}^n \delta_i$. The exponent of this sample size, $-1/3$, is typically appropriate in selection of the optimal bandwidth for estimating the distribution function (Azzalini (1981), Jones (1990)). This expression was derived after several attempts in the simulation studies. These analyses show that choosing $c < 1$ increases the estimation error of Beran's estimator since the bootstrap method provides excessively small bandwidths. In general, $c \geq 1$ is considered, with the choice $c = 3/2$ being appropriate. In cases where the function

$E(T|X = x)$ is found to be highly variable with respect to x, smaller bandwidths may be considered and our proposal there is $c = 1$.

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

$$MSE_{t,x}(h) = E\left(\left(\widehat{S}_h(t|x) - S(t|x)\right)^2\right).$$

4.2.2 The smoothed Beran's estimator

Given the good performance that the doubly smoothed survival estimator showed in previous simulation studies, it is interesting to propose a method for automatic selection of the two-dimensional bandwidth on which it depends. Then, consider the smoothed Beran's estimator of the conditional survival function, $\widetilde{S}^B_{h,g}(t|x)$, defined in (3.4). For simplicity of notation, the smoothed Beran's estimator of the survival function is denoted by $\widetilde{S}_{h,g}(t|x)$ in this chapter. A bootstrap method is proposed for the automatic selection of the bivariate bandwidth (h, g).

Algorithm for bootstrap resampling based on the smoothed Beran's estimator

Let $I_1 \subseteq \mathbb{R}$ and $I_2 \subseteq \mathbb{R}$ be intervals containing appropriate bandwidth values and let $r \in I_1$ and $s \in I_2$ be pilot bandwidths for the smoothed resample of X, T and C:

1. Obtain $U_1, \ldots, U_n$ iid with $U_i \sim U(0, 1)$ and $V_1, \ldots, V_n$ iid with common density K, $W^1_1, \ldots, W^1_n$ iid with common density K and $W^2_1, \ldots, W^2_n$ iid with common density K for all $i = 1, \ldots, n$.

2. For each $i = 1, \ldots, n$, obtain

$$X^*_i = X_{[nU_i]+1} + rV_i,$$

$$T_i^* = T_{0,i}^* + sW_i^1$$

$$C_i^* = C_{0,i}^* + sW_i^2$$

where $T_{0,i}^*$ is resampled from $\widehat{F}_r(t|X_i^*)$, constructed using Beran's estimator with the sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$, and $C_{0,i}^*$ is resampled from $\widehat{G}_r(t|X_i^*)$, constructed using Beran's estimator with the sample $\{(X_i, Z_i, 1-\delta_i)\}_{i=1}^n$.

3. For each $i = 1, \ldots, n$, obtain

$$Z_i^* = \min\{T_i^*, C_i^*\},$$

$$\delta_i^* = I\Big(T_i^* \leq C_i^*\Big).$$

4. Consider the bootstrap resample $\Big\{(X_i^*, Z_i^*, \delta_i^*)\Big\}_{i=1}^n$.

The conditional distribution functions of $T^*|X^*$ and $C^*|X^*$ are, respectively, the smoothed Beran's estimators $\widetilde{F}_{r,s}(t|X_i^*)$ and $\widetilde{G}_{r,s}(t|X_i^*)$.

The optimal bivariate bandwidth, $(h_{MISE}, g_{MISE}) \in I_1 \times I_2$ is defined as the pair of bandwidths that minimizes the mean integrated squared error given by

$$MISE_x(h,g) = E\left(\int_{I_T}\Big(\widetilde{S}_{h,g}(t|x) - S(t|x)\Big)^2 dt\right). \tag{4.4}$$

The bootstrap version of $MISE_x(h,g)$ is given by

$$MISE_x^*(h,g) = E^*\left(\int_{I_T}\Big(\widetilde{S}_{h,g}^*(t|x) - \widetilde{S}_{r,s}(t|x)\Big)^2 dt\right),$$

where $\widetilde{S}_{r,s}(t|x)$ is the smoothed Beran's survival estimation with pilot bandwidths $(r,s) \in I_1 \times I_2$ using the sample $\Big\{(X_i, Z_i, \delta_i)\Big\}_{i=1}^n$ and $\widetilde{S}_{h,g}^*(t|x)$ is the bootstrap estimation of $S(t|x)$ with bandwidths (h,g), using the bootstrap resample $\Big\{(X_i^*, Z_i^*, \delta_i^*)\Big\}_{i=1}^n$. Since the sampling distribution of $\widetilde{S}_{h,g}^*(t|x)$ is unknown, the Monte Carlo method gives the following approximation

$$MISE_x^*(h,g) \simeq \frac{1}{B}\sum_{k=1}^{B}\left(\int_{I_T}\Big(\widetilde{S}_{h,g}^{*,k}(t|x) - \widetilde{S}_{r,s}(t|x)\Big)^2 dt\right), \tag{4.5}$$

based on the empirical distribution of $\widetilde{S}_{h,g}^*(t|x)$ obtained from B bootstrap resamples. The integral is approximated by a Riemann sum.

Algorithm for bootstrap bandwidth selector for the smoothed Beran's estimator

Let x be a fixed value of the covariate, $t \in I_T$ and $(r, s) \in I_1 \times I_2$:

1. Compute $\widetilde{S}_{r,s}(t|x)$ from the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$.
2. Obtain B bootstrap resamples of the form $\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\}_{i=1}^n$ with $k = 1, ..., B$ using the bootstrap based on the smoothed Beran's estimator and calculate $\widetilde{S}_{h,g}^{*,k}(t|x)$ for each of them.
3. Approximate $MISE_x^*(h)$ according to (4.5).
4. Repeat Steps 1–3 for pairs of values (h, g) in a grid of $I_1 \times I_2$.
5. Obtain the pair (h, g) that provides the smallest $MISE_x^*(h, g)$ as the bootstrap bandwidth (h^*, g^*).

The auxiliary bandwidth $r \in I_1$ was defined in (4.3). The pilot bandwidth $s \in I_2$ for the time variable smoothing is chosen using the following formula

$$s = \frac{3}{4}\Big(Q_Z(0.975) - Q_Z(0.025)\Big)\left(\sum_{i=1}^n \delta_i\right)^{-1/7}, \tag{4.6}$$

where $Q_Z(u)$ is the u quantile of the sample $\left\{Z_i\right\}_{i=1}^n$. This expression was derived after several attempts in the simulation studies. It takes into account the variability of the observed time variable, $Q_Z(0.975) - Q_Z(0.025)$, and the sample size of the uncensored population, $\sum_{i=1}^n \delta_i$. The exponent of this sample size, $-1/7$, is heuristically deduced from the asymptotic expression of the MISE of the survival estimators (see Chapter 3).

4.3 Simulation study for bandwidth selection

A simulation study is conducted in order to show the behaviour of bootstrap bandwidth selectors for Beran's and smoothed Beran's estimators proposed in Section 4.2.

Due to the computational cost of the resampling methods, only Models 2 and 3, presented in Section 2.4, and the low and medium censoring scenarios ($P(\delta = 0|x) = 0.2$ and $P(\delta = 0|x) = 0.5$) will be considered in this chapter.

Model 2 considers a uniform distribution for the credit scoring and Weibull life and censoring times. The conditional survival function for this model is estimated at $x = 0.6$ in a time grid over the interval $I_T = (0, 0.8654)$.

Model 3 also considers a uniform distribution for the credit scoring and exponential distributions for life time and censoring time. The conditional survival function for this model is estimated at $x = 0.8$ in a time grid over the interval $I_T = (0, 3.8211)$.

Regarding the pilot bandwidth defined in (4.3) Model 2 considers $c = 3/2$, while Model 3 considers $c = 1$. The reason for this choice is that the conditional distribution of $T|X = x$ for Model 3 changes quite a lot with x, thus requiring smaller bandwidths. This is shown in Figure 4.1 where the theoretical regression function, $r(x) = E(T|X = x)$, for both models is shown.

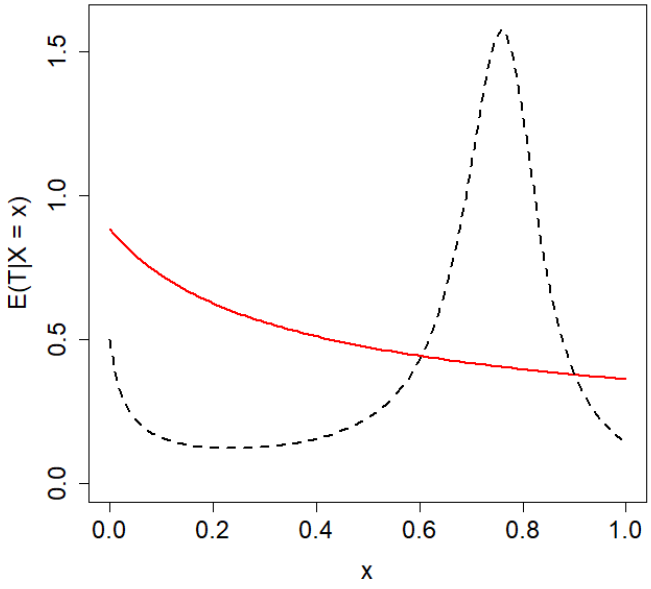

Figure 4.1: Theoretical regression function $r(x) = E(T|X = x)$ for Model 2 (solid line) and Model 3 (dashed line).

For more details about these simulation models, see Section 2.4. The simulation setup is similar to the one introduced in Section 3.4. Strategy 3 presented there is considered to obtain the bandwidths that minimised the bootstrap version of the MISE in this section.

4.3.1 Simulation study for Beran's estimator

In this subsection, the behaviour of the bootstrap bandwidth selector for Beran's estimator is analysed. For each model, the estimation error function $MISE_x(h)$ is approximated via Monte Carlo using 300 simulated samples. The bandwidth that minimises $MISE_x(h)$ is obtained and denoted by h_{MISE}. The values of h_{MISE} and $MISE_x(h_{MISE})$ are used as a benchmark.

In the simulation study, $N = 300$ simulated samples are used. For each sample, $B = 500$ bootstrap resamples are generated to approximate the bootstrap MISE function, $MISE_x^*(h)$, and obtain the bootstrap bandwidth associated to each simulated sample, h_j^*, $j = 1, 2, \ldots, N$. The mean value of the N bootstrap bandwidths and the standard deviation are defined as follows

$$\overline{h^*} = \frac{1}{N}\sum_{j=1}^{N} h_j^*, \quad sd\big(h^*\big) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\Big(h_j^* - \overline{h^*}\Big)^2}.$$

As a relative measure of the difference between the bootstrap bandwidth and the optimal one, we compute

$$H_j^* = \frac{h_j^* - h_{MISE}}{h_{MISE}},$$

with $j = 1, \ldots, N$. The mean of the absolute value of these relative deviations, $\overline{H^*} = \frac{1}{N}\sum_{j=1}^{N}|H_j^*|$, is a good measure of how close the bootstrap bandwidth is to the optimal one.

For each sample, the estimation error of Beran's estimator with the corresponding bootstrap bandwidth,

$$MISE_x(h_j^*) = E\bigg(\int_{I_T}\Big(\widehat{S}_{h_j^*}(t|x) - S(t|x)\Big)^2 dt\bigg),$$

and its square root, $RMISE_x(h_j^*)$, are approximated via Monte Carlo using 300 simulated samples. The mean of these estimation errors given by

$$\overline{RMISE_x(h^*)} = \frac{1}{N}\sum_{j=1}^{N} RMISE_x(h_j^*)$$

is used as a measure of the estimation error made by the bootstrap bandwidth, when compared with the estimation error made by the MISE bandwidth.

As a relative measure of the difference between the estimation errors using the bootstrap and the MISE bandwidths, the following ratios are defined:

$$R_j^* = \frac{RMISE_x(h_j^*) - RMISE_x(h_{MISE})}{RMISE_x(h_{MISE})}$$

satisfying $R_j^* \geq 0$ for all $j = 1, \ldots, N$. The mean of the R_j^* values with $j = 1, \ldots, N$ is denoted by $\overline{R^*} = \frac{1}{N}\sum_{j=1}^{N} R_j^*$.

Small values (close to zero) of $\overline{H^*}$ and $\overline{R^*}$ indicate good behavior of the bootstrap bandwidth. Values of the bootstrap bandwidths, estimation errors and relative measures for Models 2 and 3 are included in Table 4.1. Figures 4.2 and 4.3 show the MISE function and the bootstrap MISE approximation for Models 2 and 3 and the boxplots of the relative error measures. The results show a good performance of the proposed bootstrap selector.

	Model 2		Model 3	
$P(\delta = 0 \mid X = x)$	0.2	0.5	0.2	0.5
h_{MISE}	0.23939	0.21212	0.04515	0.05687
$RMISE_x(h_{MISE})$	0.02411	0.03652	0.11612	0.15576
$\overline{h^*}$ (sd)	0.23815 (0.093)	0.21897 (0.082)	0.06718 (0.007)	0.08082 (0.011)
$\overline{H^*}$	0.29033	0.26199	0.48794	0.42119
$\overline{RMISE_x(h^*)}$	0.02548	0.03809	0.13391	0.17242
$\overline{R^*}$	0.05762	0.04373	0.15316	0.10698

Table 4.1: MISE and average bootstrap bandwidths and estimation errors of Beran's survival estimator in each level of conditional censoring probability for Models 2 and 3. Numbers within brackets are standar deviations.

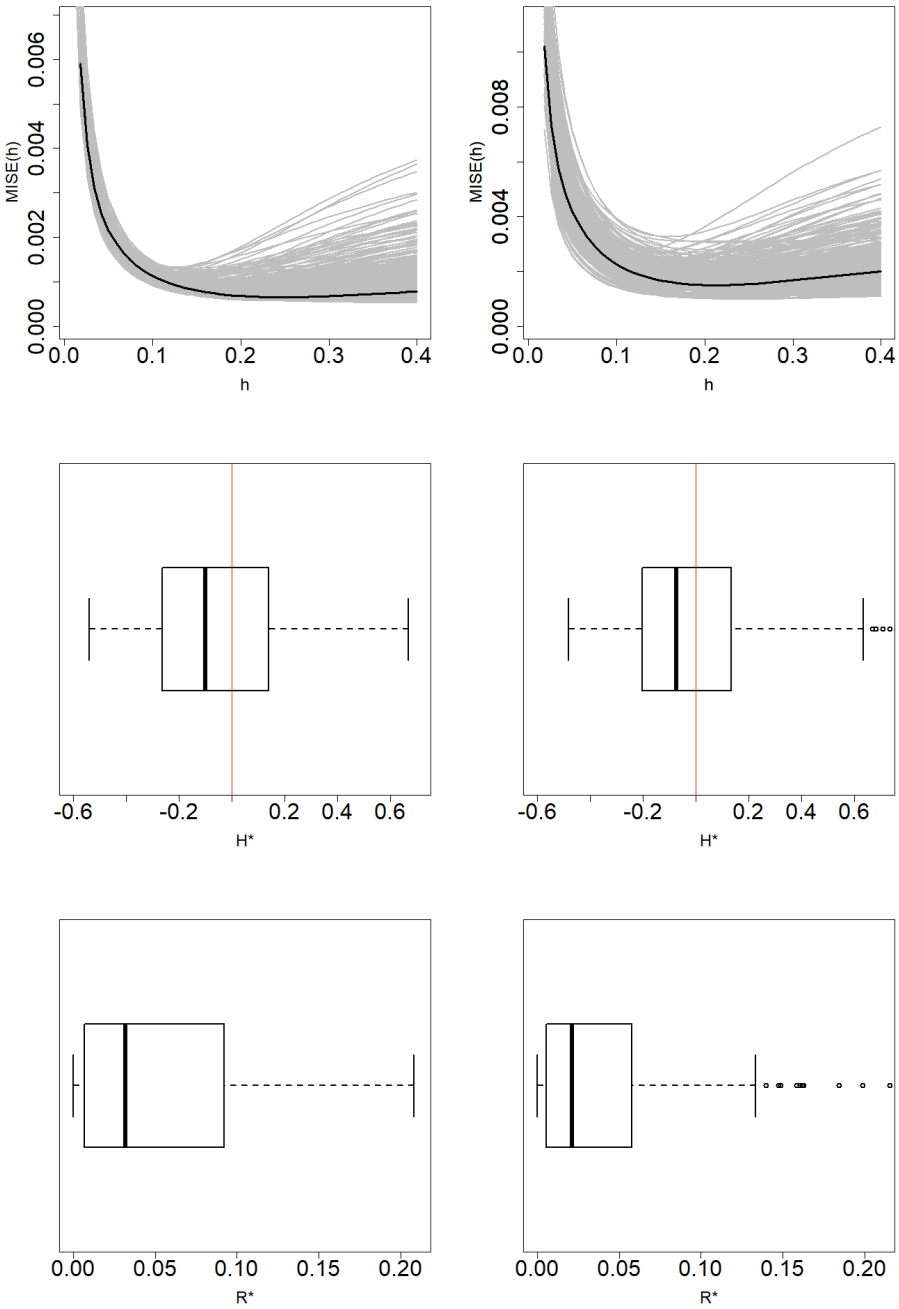

Figure 4.2: $MISE_x(h)$ function (black line) approximated via Monte Carlo and $MISE_x^*(h)$ functions (gray lines) for $N = 300$ samples (top), boxplot of $H_1^*, ..., H_N^*$ values (middle) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 2.

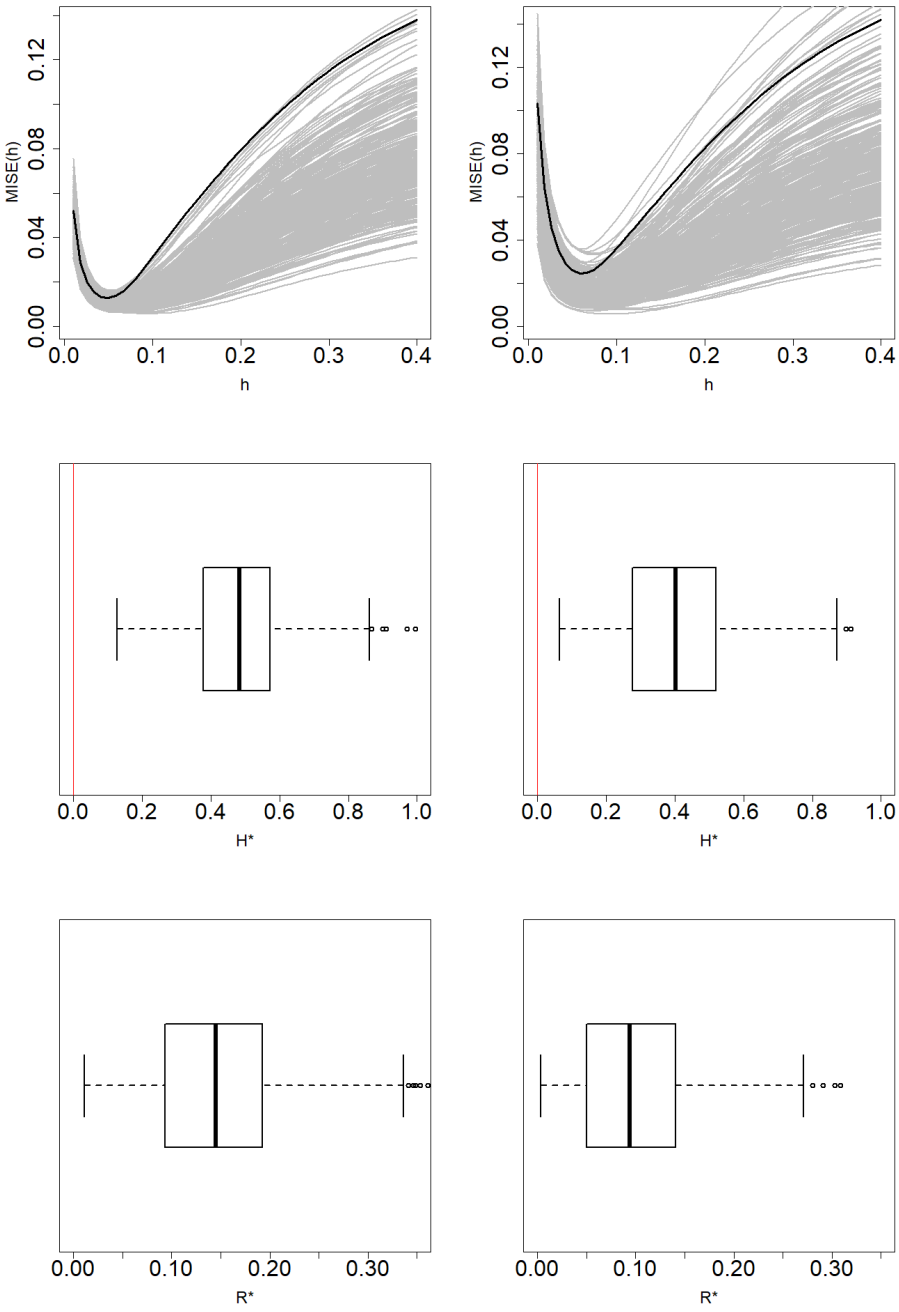

Figure 4.3: $MISE_x(h)$ function (black line) approximated via Monte Carlo and $MISE_x^*(h)$ functions (gray lines) for $N = 300$ samples (top), boxplot of $H_1^*, ..., H_N^*$ values (middle) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 3.

Figure 4.4 shows the theoretical survival function and Beran's estimation with optimal and bootstrap bandwidths for one sample from Models 2 and 3 when the

conditional probability of censoring is 0.5.

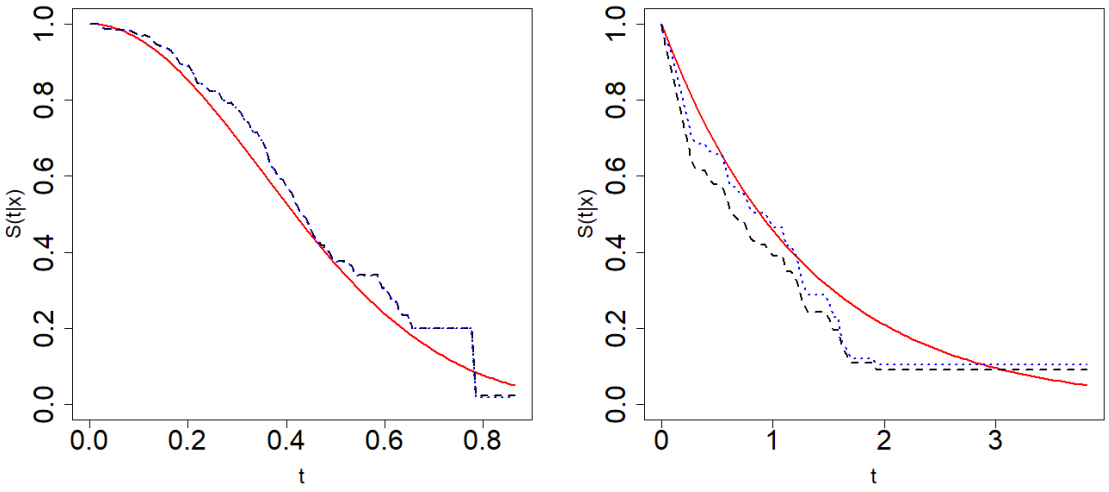

Figure 4.4: Theoretical conditional survival function $S(t|x)$ (solid line), Beran's estimation with MISE bandwidth (dotted line) and Beran's estimation with bootstrap bandwidth (dashed line) for one sample from Model 2 (left) and Model 3 (right) with $P(\delta = 0|x) = 0.5$.

4.3.2 Simulation study for the smoothed Beran's estimator

In this section, a simulation study on the bootstrap bandwidth selector for the smoothed Beran's estimator in (3.4) is carried out. The resampling technique and Monte Carlo approximation of the MISE presented in Section 4.2.2 are used.

For each model, the error function $MISE_x(h, g)$ is approximated via Monte Carlo from 300 simulated samples and the bivariate bandwidth that minimises $MISE_x(h, g)$ is obtained and denoted by (h_{MISE}, g_{MISE}). The values of (h_{MISE}, g_{MISE}) and $MISE_x(h_{MISE}, g_{MISE})$ are used as a benchmark.

In the study, $N = 300$ 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 value of the N bootstrap bandwidths and the standard deviation are the following:

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

$$sd\big(h^*\big) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\Big(h_j^* - \overline{h^*}\Big)^2}, \quad sd\big(g^*\big) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\Big(g_j^* - \overline{g^*}\Big)^2}.$$

In order to measure the distance of the bootstrap two-dimensional bandwidth of the j-th sample, (h_j^*, g_j^*), from the corresponding MISE bandwidth, (h_{MISE}, g_{MISE}), consider the vector

$$D_j^* = \left(\frac{h_j^* - h_{MISE}}{h_{MISE}}, \frac{g_j^* - g_{MISE}}{g_{MISE}}\right) \in \mathbb{R}^2.$$

and its Euclidean norm denoted by $H_j^* = \|D_j^*\|_2$ with $j = 1, \ldots, N$. The mean value, $\overline{H^*} = \frac{1}{N}\sum_{j=1}^{N} H_j^*$ is a measure of how close the bootstrap bandwidths are to the MISE one.

For each sample, the estimation error of the smoothed Beran's estimator with the corresponding bootstrap bandwidth,

$$MISE_x(h_j^*, g_j^*) = E\bigg(\int_{I_T}\Big(\widetilde{S}_{h_j^*, g_j^*}(t|x) - S(t|x)\Big)^2 dt\bigg),$$

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

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

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

The ratio

$$R_j^* = \frac{RMISE_x(h_j^*, g_j^*) - RMISE_x(h_{MISE}, g_{MISE})}{RMISE_x(h_{MISE}, g_{MISE})}$$

is defined as a relative measure of the difference between the error of the estimator with bootstrap bandwidth and MISE bandwidth. The mean of the positive values R_j^* with $j = 1, \ldots, N$ is denoted by $\overline{R^*} = \frac{1}{N}\sum_{j=1}^{N} R_j^*$. Values of the bootstrap bivariate bandwidths, estimation errors and relative measures for Models 2 and 3 are included in Table 4.2.

	Model 2		Model 3	
$P(\delta = 0\|X = x)$	0.2	0.5	0.2	0.5
h_{MISE}	0.23469	0.20408	0.11122	0.23979
g_{MISE}	0.05143	0.08102	0.91530	1.13878
$RMISE_x(h_{MISE}, g_{MISE})$	0.02158	0.03024	0.05594	0.06938
$\overline{h^*}$ (sd)	0.24172 (0.086)	0.22600 (0.085)	0.31289 (0.132)	0.36221 (0.142)
$\overline{g^*}$ (sd)	0.09701 (0.017)	0.11474 (0.022)	0.85749 (0.117)	0.86538 (0.112)
$\overline{H^*}$	0.94544	0.54256	1.82913	0.73757
$\overline{RMISE_x(h^*, g^*)}$	0.02769	0.03851	0.11035	0.12633
$\overline{R^*}$	0.20027	0.17301	1.01971	0.84987

Table 4.2: MISE and average bootstrap bandwidths and estimation errors of the smoothed Beran's survival estimator in each level of censoring conditional probability for Models 2 and 3. Numbers within brackets are standar deviations.

Figures 4.5 and 4.6 show the $MISE_x(h, g)$ function of the smoothed Beran's estimator and its bootstrap approximation along with the corresponding contour plot for one sample of both Models 2 and 3 when the conditional probability of censoring is 0.5. It is approximated over a meshgrid of 50×50 values of (h, g). Note that both $MISE_x(h, g)$ and $MISE_x^*(h, g)$ curves for each fixed h value are quite similar in the region close to the minimum value of $MISE_x^*(h, g)$. Thus, the influence of the covariate smoothing parameter, h, is weak when estimating the survival function using values of bandwidth g close to the optimal one.

Figures 4.7 and 4.8 show the boxplots of H_j^* and R_j^* with $j = 1, \ldots, N$. In general, the selector tends to overestimate the value of the covariate bandwidths. Due to the behaviour of the $MISE_x(h, g)$ curves, mentioned above, this does not lead to a significant increase in the estimation error.

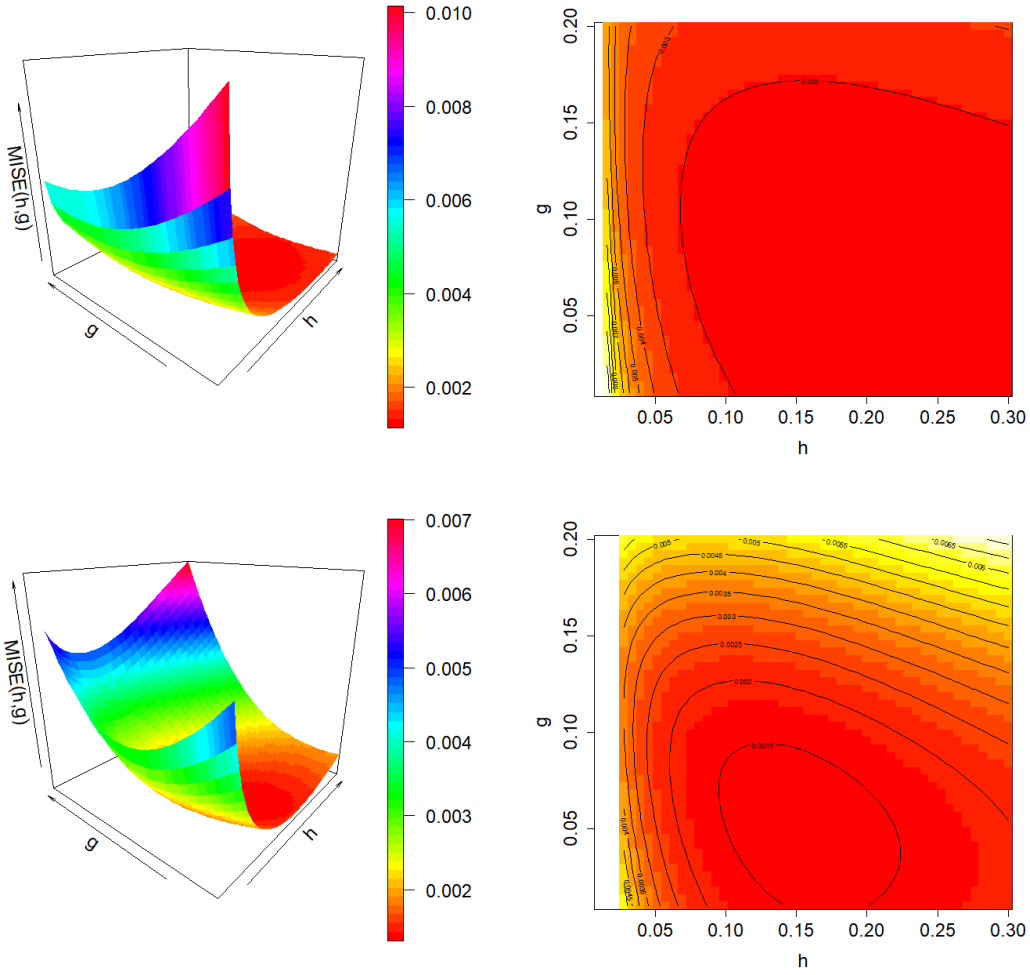

Figure 4.5: $MISE_x(h, g)$ function (top left) and contour plot of $MISE_x(h, g)$ (top right) and $MISE_x^*(h, g)$ function (bottom left) and contour plot of $MISE_x^*(h, g)$ (bottom right) for one sample from Model 2 when $P(\delta = 0|x) = 0.5$.

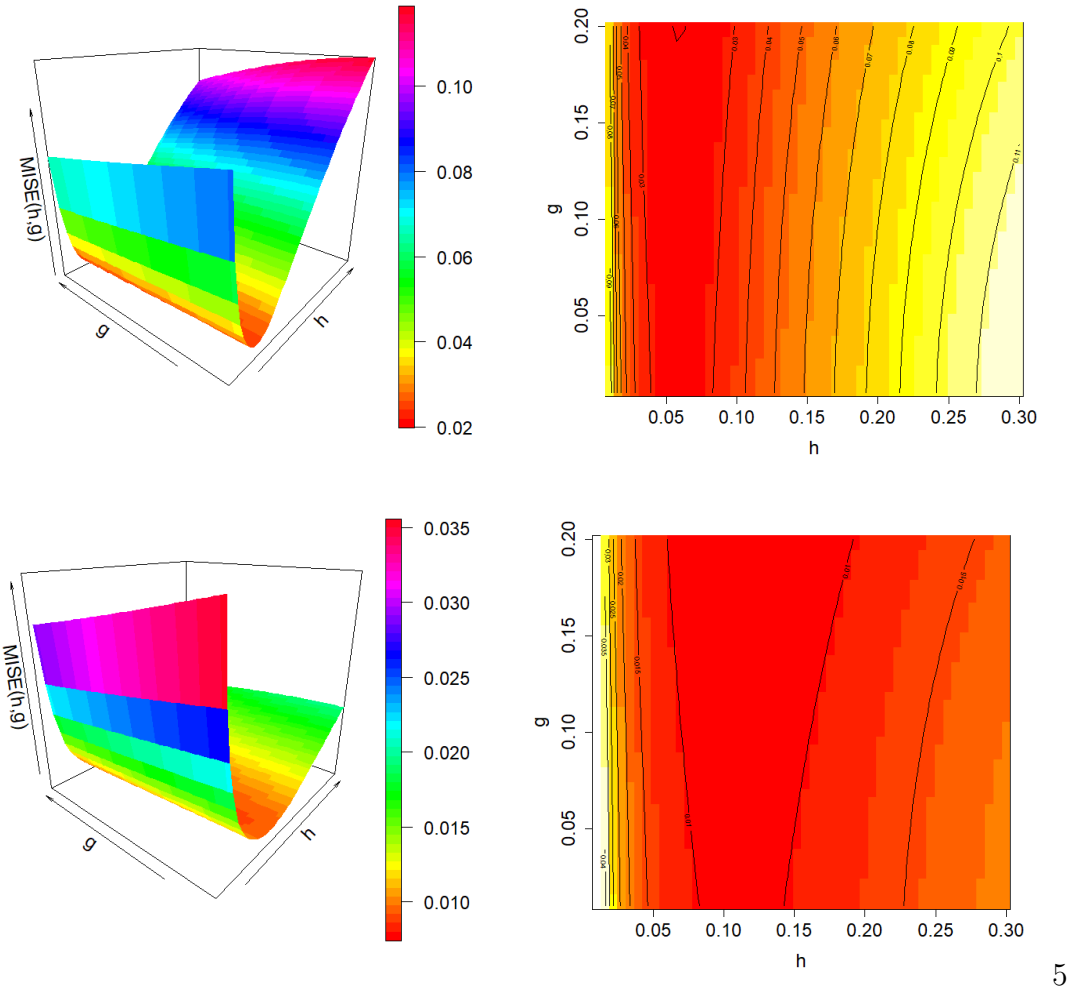

Figure 4.6: $MISE_x(h, g)$ function (top left) and contour plot of $MISE_x(h, g)$ (top right) and $MISE_x^*(h, g)$ function (bottom left) and contour plot of $MISE_x^*(h, g)$ (bottom right) for one sample from Model 3 when $P(\delta = 0|x) = 0.5$.

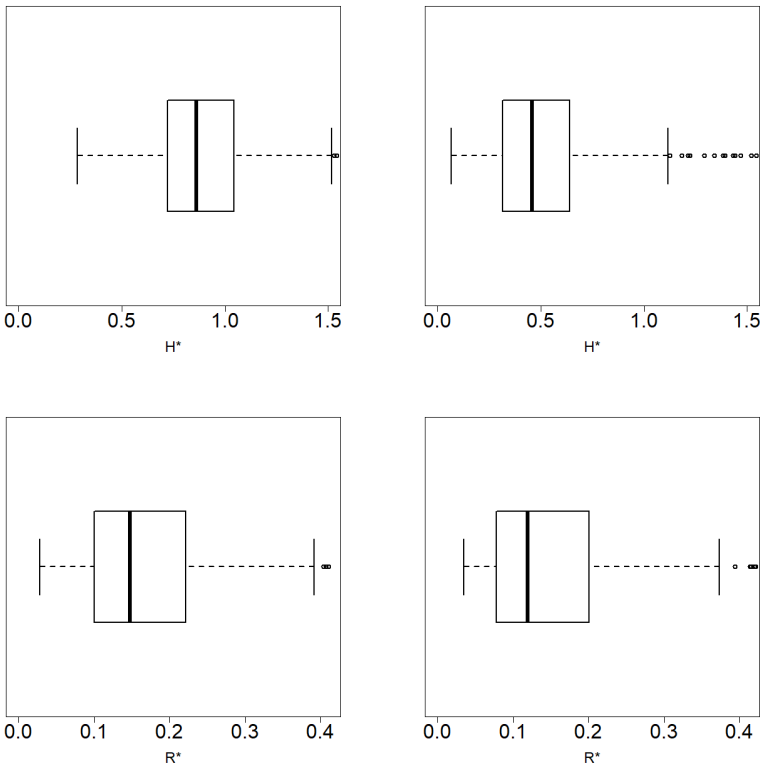

Figure 4.7: Boxplot of $H_1^*, ..., H_N^*$ values (top) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 2.

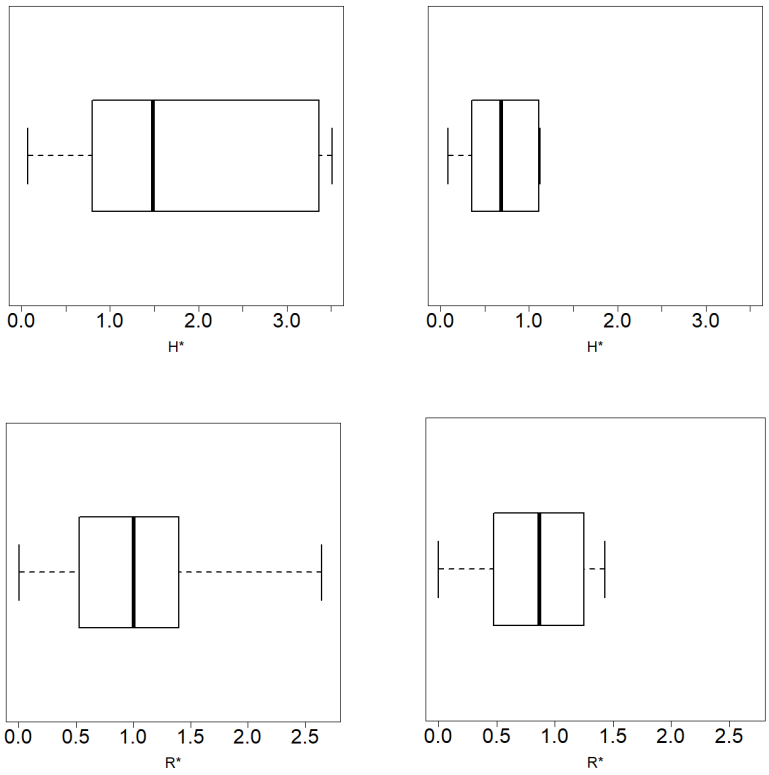

Figure 4.8: Boxplot of $H_1^*, ..., H_N^*$ values (top) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 3.

Figure 4.9 shows the theoretical survival function and the smoothed Beran's estimation with optimal and bootstrap bandwidths for one sample of each model when the conditional probability of censoring is 0.5.

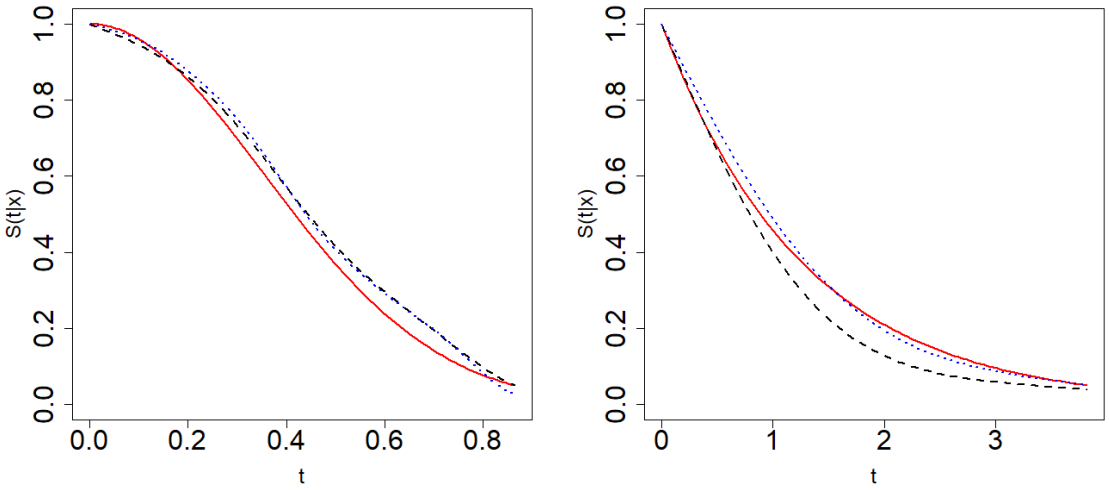

Figure 4.9: Theoretical conditional survival function, $S(t|x)$, (solid line), smoothed Beran's estimation with MISE bandwidth (dotted line) and smoothed Beran's estimation with bootstrap bandwidth (dashed line) for one sample from Model 2 (left), Model 3 (right) with $P(\delta = 0|x) = 0.5$.

The results showed in Tables 4.1 and 4.2 are summarised in Table 4.3 to compare the behaviour of Beran's and the smoothed Beran's estimators and to evaluate whether the improvement that smoothing in the time variable provides for survival estimation is preserved when approximating the smoothing parameters by resampling techniques. The behaviour of both estimators with bootstrap bandwidths is very similar in Model 2, while there is a significant decrease in estimation error due to double smoothing in Model 3.

		Model 2		Model 3	
$P(\delta = 0\|X = x)$		0.2	0.5	0.2	0.5
$\widehat{S}_h(t\|x)$	$\overline{h^*}$	0.23815	0.21897	0.06718	0.08082
	$\overline{RMISE_x(h^*)}$	0.02548	0.03809	0.13391	0.17242
$\widetilde{S}_{h,g}(t\|x)$	$\overline{h^*}$	0.24172	0.22600	0.31289	0.36220
	$\overline{g^*}$	0.09701	0.11474	0.85749	0.86537
	$\overline{RMISE(h^*, g^*)}$	0.02770	0.03851	0.11035	0.12633

Table 4.3: Comparative table of the bootstrap bandwidths and RMISE with Beran's estimator and the smoothed Beran's estimator in Model 2 and 3.

4.4 Confidence regions using Beran's and the smoothed Beran's estimators

Let $x \in I$ be a fixed value of the covariate and consider $S(t|x)$ the conditional survival curve with $t \in I_T$. The curve $S(t|x)$ belongs to the function space $\mathcal{F}(I_T)$ whose elements are real-valued functions with domain I_T. From the sample $\{(X_i, Z_i, \delta_i), i = 1, ..., n\}$, Beran's estimation of $S(t|x)$, $\widehat{S}_h(t|x)$, is obtained and a confidence region of $S(t|x)$ at $1-\alpha$ confidence level associated to Beran's estimator can be constructed. A similar construction is done for the smoothed Beran's estimator. This confidence region of $S(t|x)$ is a random subset of $I_T \times \mathcal{F}(I_T)$ denoted by R_α that satisfies

$$P\Big((t, S(t|x)) \in R_\alpha, \forall t \in I_T\Big) = 1 - \alpha.$$

In this section we propose two different methods to obtain confidence regions of the $S(t|x)$ curve based on resampling techniques. Both Beran's estimator and the smoothed Beran's estimator can be used with these two methods.

Method 1 for confidence regions

First, Beran's estimator of the conditional survival function, $\widehat{S}_h(t|x)$, given in (2.3) is used. This method follows the ideas of Cao et al. (2010) to obtain prediction regions. It is based on finding the value of $\lambda_\alpha \in \mathbb{R}^+$ such that

$$P\Big(|\widehat{S}_h(t|x) - S(t|x)| < \lambda_\alpha \sigma(t), \forall t \in I_T\Big) = 1 - \alpha$$

with $\sigma^2(t) = Var\Big(\widehat{S}_h(t|x)\Big)$. Thus, the theoretical confidence region is defined by

$$R^1_\alpha = \Big\{(t, y) : t \in I_T, y \in \Big(\widehat{S}_h(t|x) - \lambda_\alpha \sigma(t), \widehat{S}_h(t|x) + \lambda_\alpha \sigma(t)\Big)\Big\}.$$

Since λ_α and $\sigma(t)$ are unknown, they are approximated by means of a bootstrap technique. The bootstrap confidence region is defined as follows:

$$R^{1*}_\alpha = \Big\{(t, y) : t \in I_T, y \in \Big(\widehat{S}^*_h(t|x) - \lambda^*_\alpha \sigma^*(t), \widehat{S}^*_h(t|x) + \lambda^*_\alpha \sigma^*(t)\Big)\Big\}.$$

where $\widehat{S}_h^*(t|x)$ is the bootstrap estimation of $S(t|x)$ with bandwidth h and λ_α^* and $\sigma^*(t)$ are the bootstrap analogue of λ_α and $\sigma(t)$. The confidence region R_α^{1*} satisfies

$$p(\lambda_\alpha^*) = P^*\Big((t, \widehat{S}_r(t|x)) \in R_\alpha^{1*}, \forall t \in I_T\Big) = 1 - \alpha. \tag{4.7}$$

From the original sample $\Big\{(X_i, Z_i, \delta_i)\Big\}_{i=1}^n$, Beran's estimator of $S(t|x)$ is obtained with appropriate bandwidth h, $\widehat{S}_h(t|x)$. The algorithm to obtain the bootstrap confidence region for $S(t|x)$ at confidence level $1-\alpha$ associated to $\widehat{S}_h(t|x)$ is explained below. The Monte Carlo method is used to approximate $\sigma^*(t)$, and an iterative method is used to approximate the value of λ_α^* so that the confidence region has a confidence level approximately equal to $1 - \alpha$.

1. Compute Beran's estimator $\widehat{S}_r(t|x)$ from the original sample $\Big\{(X_i, Z_i, \delta_i)\Big\}_{i=1}^n$ and pilot bandwidth $r \in I_1$.

2. Generate B bootstrap resamples of the form $\Big\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\Big\}_{i=1}^n$ by means of the resampling algorithm for Beran's estimator presented in Subsection 4.2.1 and pilot bandwidth r.

3. For $k = 1, \ldots, B$, compute $\widehat{S}_h^{*,k}(t|x)$ with the k-th bootstrap resample and bandwidth h, obtaining $\Big\{\widehat{S}_h^{*,k}(t|x)\Big\}_{k=1}^B$.

4. Approximate the standard deviation of $\widehat{S}_h^*(t|x)$ by

$$\sigma^*(t) \simeq \left(\frac{1}{B}\sum_{k=1}^{B}\left(\widehat{S}_h^{*,k}(t|x) - \frac{1}{B}\sum_{l=1}^{B}\widehat{S}_h^{*,l}(t|x)\right)^2\right)^{1/2}, \quad t \in I_T.$$

5. Use an iterative method to obtain an approximation of the value λ_α^* defined in (4.7).

6. The confidence region is given by

$$\widehat{R}_\alpha^1 = \Big\{(t, y) : t \in I_T, y \in \Big(\widehat{S}_h(t|x) - \lambda_\alpha^*\sigma^*(t), \widehat{S}_h(t|x) + \lambda_\alpha^*\sigma^*(t)\Big)\Big\}.$$

Iterative method to approximate λ_α^*

The iterative method to approximate the value of $\lambda_\alpha^* \in \mathbb{R}^+$ so that the confidence region R_α^* has a confidence level approximately equal to $1-\alpha$ is explained below. This algorithm allows to quickly and efficiently approximate the parameter λ_α^*.

Let $\left\{\widehat{S}_h^{*,k}(t|x)\right\}_{k=1}^B$ be the Beran's estimations of the survival function with bandwidth h over a set of B bootstrap resamples of $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$. Define the Monte Carlo approximation of $p(\lambda)$ in (4.7), for any $\lambda \in \mathbb{R}^+$, as follows:

$$p(\lambda) \simeq \frac{1}{B}\sum_{k=1}^{B} I\Big(\widehat{S}_r(t|x) \in \Big(\widehat{S}_h^{*,k}(t|x) - \lambda\sigma^*(t), \widehat{S}_h^{*,k}(t|x) + \lambda\sigma^*(t)\Big), \forall t \in I_T\Big). \quad (4.8)$$

Let $\lambda_L, \lambda_H \in \mathbb{R}^+$ be such that $p(\lambda_L) \leq 1-\alpha \leq p(\lambda_H)$ and let $\zeta > 0$ be a tolerance, for example, $\zeta = 10^{-4}$.

1. Obtain $\lambda_M = \dfrac{\lambda_L + \lambda_H}{2}$ and compute Monte Carlo approximations of $p(\lambda_L)$, $p(\lambda_M)$ and $p(\lambda_H)$ according to (4.8).
2. If $p(\lambda_M) = 1-\alpha$ or $p(\lambda_H) - p(\lambda_L) < \zeta$, then $\lambda_\alpha^* = \lambda_M$. Otherwise,
   (a) If $1-\alpha < p(\lambda_M)$, then $\lambda_H = \lambda_M$ and return to Step 1.
   (b) If $p(\lambda_M) < 1-\alpha$, then $\lambda_L = \lambda_M$ and return to Step 1.

This method to obtain confidence regions for the curve $S(t|x)$ for fixed $x \in I$ and t covering I_T based on Beran's estimator can be adapted to obtain confidence regions using the smoothed Beran's estimator. Simply replace Beran's estimator $\widehat{S}_h(t|x)$ by the smoothed Beran's estimator $\widetilde{S}_{h,g}(t|x)$ given in (3.4) where necessary, and obtain the analogous bootstrap approximations of λ_α and $\sigma(t)$:

1. Compute the smoothed Beran's estimator $\widetilde{S}_{r,s}(t|x)$ from the original sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$ and pilot bandwidths $r \in I_1$ and $s \in I_2$.
2. Generate B bootstrap resamples of the form $\left\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\right\}_{i=1}^n$ by means of the resampling algorithm for the smoothed Beran's estimator presented in Subsection 4.2.1 and pilot bandwidths r and s.

3. For $k = 1, \ldots, B$, compute $\widetilde{S}^{*,k}_{h,g}(t|x)$ with the k-th bootstrap resample and bandwidth h, obtaining $\left\{\widetilde{S}^{*,k}_{h,g}(t|x)\right\}_{k=1}^{B}$.

4. Approximate the standard deviation of $\widetilde{S}^{*}_{h,g}(t|x)$ by

$$\sigma^*(t) \simeq \left(\frac{1}{B}\sum_{k=1}^{B}\left(\widetilde{S}^{*,k}_{h,g}(t|x) - \frac{1}{B}\sum_{l=1}^{B}\widetilde{S}^{*,l}_{h,g}(t|x)\right)^2\right)^{1/2}, \quad t \in I_T.$$

5. Use an iterative method to obtain an approximation of the value λ^*_α defined in (4.7).

6. The confidence region is given by

$$\widetilde{R}^1_\alpha = \left\{(t,y) : t \in I_T, y \in \left(\widetilde{S}_{h,g}(t|x) - \lambda^*_\alpha\sigma^*(t), \widetilde{S}_{h,g}(t|x) + \lambda^*_\alpha\sigma^*(t)\right)\right\}.$$

The pilot bandwidths defined in (4.3) and (4.6) are used for the confidence region algorithm based on both Beran's and the smoothed Beran's estimators.

Method 2 for confidence regions

An alternative procedure to obtain a confidence region for $S(t|x)$, with fixed $x \in I$ and t covering the interval I_T, is based on considering that the curve $S(t|x)$ belongs to the functional space $L_p(I_T)$ defined using a natural generalization of the p-norm for finite-dimensional vector spaces, $\|\cdot\|_p$. The confidence region for $S(t|x)$ computed at the $1-\alpha$ confidence level is a ball around $\widehat{S}_h(t|x)$ of radius ρ_α, where ρ_α is such that

$$P\left(\|\widehat{S}_h(t|x) - S(t|x)\|_p < \rho_\alpha\right) = 1-\alpha.$$

This idea was presented in Zhun and Politis (2017) to obtain prediction regions in functional autoregression models. Since $S(t|x)$ is unknown, the distribution of $R = \|\widehat{S}_h(t|x) - S(t|x)\|_p$ is not available and the value of ρ_α can not be calculated. Therefore, a bootstrap approximation is given by $R^* = \|\widehat{S}^*_h(t|x) - \widehat{S}_r(t|x)\|_p$. The bootstrap confidence region is a ball in $L_p(I_T)$ around $\widehat{S}_h(t|x)$ of radius ρ^*_α, where ρ^*_α is such that

$$P^*\left(\|\widehat{S}^*_h(t|x) - \widehat{S}_r(t|x)\|_p < \rho^*_\alpha\right) = 1-\alpha.$$

and $r \in I_h$ is an auxiliary bandwidth.

From the original sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^{n}$, Beran's estimator of $S(t|x)$ is obtained with appropriate bandwidth h, $\widehat{S}_h(t|x)$. The algorithm to obtain the bootstrap confidence region for $S(t|x)$ at confidence level $1-\alpha$ associated to $\widehat{S}_h(t|x)$ is explained below. The Monte Carlo method is used to approximate the radius ρ_α, so that the confidence region has a confidence level approximately equal to $1-\alpha$.

1. Compute Beran's estimator $\widehat{S}_r(t|x)$ with the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^{n}$ and pilot bandwidth $r \in I_h$.

2. Generate B bootstrap resamples $\left\{(X_i^{*k}, Z_i^{*k}, \delta_i^{*k})\right\}_{i=1}^{n}$, for $k = 1, ..., B$, by means of the resampling algorithm for Beran's estimator presented in Subsection 4.2.1 and pilot bandwidth r.

3. For $k = 1, ..., B$, compute $\widehat{S}_h^{*k}(t|x)$ with the k-th bootstrap resample and bandwidth r and obtain
$$R_k^* = \|\widehat{S}_h^{*k}(t|x) - \widehat{S}_r(t|x)\|_p$$

4. Sort the values $R_1^*, ..., R_B^*$ by obtaining $R_{(1)}^*, ..., R_{(B)}^*$ and select $\rho_\alpha^* = R_{([B(1-\alpha)])}^*$.

5. The confidence region is the ball in $L_p(I_T)$ around $\widehat{S}_h(t|x)$ with radius ρ_α^*.

Regarding the norm to be used, the usual norms of the function spaces L_1 and L_2 allow us to mathematically define the confidence region and to check whether or not a given curve belongs to this region. The disadvantage of these function spaces is that they do not allow a graphical representation of the confidence region.

Choosing the function space L_∞ and its associated norm, $\|\cdot\|_\infty$, then the statistic used to obtain the confidence region is defined as follows

$$R = \|\widehat{S}_h(t|x) - S(t|x)\|_\infty = \sup_{t\in[l,u]} |\widehat{S}_h(t|x) - S(t|x)|$$

and the confidence region is

$$R_\alpha^2 = \left\{(t, y) : t \in I_T, y \in \left(\widehat{S}_h(t|x) - \rho_\alpha, \widehat{S}_h(t|x) + \rho_\alpha\right)\right\}.$$

whose graphical representation may be useful. The disadvantage of this choice of the space is that the confidence region R_α has the same radius ρ_α at all points $t \in I_T$, so it does not capture the variability of the estimator, $\sigma^2(t) = Var(\widehat{S}_h(t|x))$.

This method can be adapted to obtain confidence regions using the smoothed Beran's estimator. Simply replace Beran's survival estimator $\widehat{S}_h(t|x)$ by the smoothed Beran's estimator $\widetilde{S}_{h,g}(t|x)$ given in (3.4) where necessary. The confidence region for $S(t|x)$ based on the smoothed Beran's estimator at $1-\alpha$ confidence level is a ball in $L_p(I_T)$ around $\widetilde{S}_{h,g}(t|x)$ of radius ρ_α, where ρ_α is such that

$$P(R < \rho_{1-\alpha}) = 1 - \alpha$$

with

$$R = \|\widetilde{S}_{h,g}(t|x) - S(t|x)\|_p.$$

A similar procedure to the one shown in the previous paragraphs for Beran's estimator allows us to obtain the bootstrap approximation of ρ_α.

The pilot bandwidths defined in (4.3) and (4.6) are used for the confidence region algorithm based on both Beran's and the smoothed Beran's estimators.

If the aim were the point estimation of $S(t|x)$, the algorithms proposed in this section can be easily adapted to obtain confidence intervals for $S(t|x)$ for fixed $x \in I$ and $t \in I_T$.

4.5 Simulation study for confidence regions

A simulation study is carried out to analyse the performance of the bootstrap confidence regions obtained by means of the two methods proposed in Section 4.4 and based on both Beran's and the smoothed Beran's estimator.

Models 2 and 3 are considered and the simulation setup is the one introduced in Section 4.3. Two conditional probabilities of censoring are considered for each model: $P(\delta = 0|x) = 0.2$ and $P(\delta = 0|x) = 0.5$. The number of simulated samples

of each model is $N = 300$ and $B = 500$ bootstrap resamples are obtained for each sample. The sample size is $n = 400$. The confidence level is $1 - \alpha$ with $\alpha = 0.05$. When Beran's estimator is considered, the optimal bandwidth that minimises the mean integrated squared error is used ($h = h_1$ from Tables 3.2 and 3.3). Similarly, the two-dimensional bandwidth that minimises the MISE is considered when using the smoothed Beran's estimator ($(h, g) = (h_3, g_3)$ from Tables 3.2 and 3.3). These bandwidths are unknown in practice, but they allow a fair comparison of the methods in the simulation study. Regarding the pilot bandwidth defined in (4.3), Model 2 considers $c = 3/2$, while Model 3 considers $c = 1$, as explained in Section 4.3.

Denote the lower and upper bounds of the confidence region by $l(t, x)$ and $u(t, x)$, respectively. It may happen that the lower bound of the confidence region is less than 0 or the upper bound is greater than one for some points (t_0, x_0). When this happens, we set $l(t_0, x_0) = 0$ or $u(t_0, x_0) = 1$, as appropriate.

It is clear that $S(t|x) = 1$ when $t = 0$ and $S(t|x)$ is not necessarily 1 when $t = 0 + \varepsilon$ with any $\varepsilon > 0$. However, due to the lack of information provided by the data at times close to zero, it is the case that the estimation of $S(t|x)$ is 1 for the smallest values of the time grid in most of the samples of the study. As a consequence, using Method 1, $l(t, x) = 1 = u(t, x)$ for such small values of t and the confidence region does not contain the true survival curve, so the coverage decreases. The proposed solution is to artificially increase the width of the confidence region at the first points of the grid: for those values of t such that $l(t, x) = 1 = u(t, x)$, we make $l(t, x) = l(t', x)$ where $t' \in \{t_1, \ldots, t_n\}$ is the first grid point such that $l(t_0, x) < 1$. This is a problem that Method 2 does not present, since the variability that the conditional survival estimations have in the right tail of the time ditribution is inherited by the width of the confidence region at all points of the time grid.

A confidence region performs well if its coverage is close to the nominal one, in this case $1 - \alpha = 0.95$, and has a small area or average width. The following values measure the performance of the confidence region and allow for the comparison of results.

Coverage is the percentage of bootstrap regions that contain the whole theoretical survival curve and it is defined as follows

$$\frac{1}{N}\sum_{j=1}^{N} I\Big\{S(t_k|x) \in \Big(l(t_k,x), u(t_k,x)\Big), \forall k = 1, ..., n_T\Big\}.$$

The mean pointwise coverage is the mean of the proportion of time grid values for which the confidence region contains the theoretical conditional survival curve. It is given by

$$\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_T}\sum_{k=1}^{n_T} I\Big\{S(t_k|x) \in \Big(l(t_k,x), u(t_k,x)\Big)\Big\}\right).$$

Average width of the bootstrap confidence region is defined by

$$\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_T}\sum_{k=1}^{n_T}\Big(u(t_k,x) - l(t_k,x)\Big)\right).$$

Winkler score (see Winkler (1972)) is also used to compare the behaviour of the methods. For classical confidence or prediction intervals, it is defined as the length of the interval plus a penalty if the theoretical value is outside the interval. Thus, it combines width and coverage. For values that fall within the interval, the Winkler score is simply the length of the interval. So low scores are associated with narrow intervals. When the theoretical value falls outside the interval, the penalty is proportional to how far the observation is from the interval. The formula of the Winkler score (WS) as a function of the time and covariate variables is as follows:

$$\begin{aligned}\mathrm{WS}(t,x) &= u(t,x) - l(t,x) + \frac{2}{\alpha}(l(t,x) - S(t|x))I\Big(S(t|x) < l(t,x)\Big) \\ &\quad + \frac{2}{\alpha}(S(t|x) - u(t,x))I\Big(S(t|x) > u(t,x)\Big).\end{aligned}$$

Since we are working with confidence regions for fixed $x \in I$ and t varying over the interval I_T, the integrated Winkle score is proposed as a criteria for the comparison of the confidence regions. It is defined by

$$IWS(x) = \int_{I_T} WS(t,x)dt.$$

and the lower the value of IWS, the better the performance of the confidence region.

The results obtained are shown in Tables 4.4 and 4.5. The high values of pointwise coverage in all scenarios are remarkable. Furthermore, these coverage percentages are preserved when using double smoothing, while the average width of the confidence regions is decreased. This is reflected in the IWS, which presents much larger values in the Beran's estimator-based confidence regions.

Method 1 has lower mean coverage, but equal pointise coverage and smaller width than Method 2 in Model 2. In Model 3, the coverages of the two methods are similar, with Method 1 providing confidence regions of smaller width. The coverage indicates the percentage of times the theoretical curve is completely contained in the confidence band. This coverage decreases as soon as the curve goes outside the region at a single point on the time grid. This, that only a few points go out of the region, is what mainly happens here.

In some of the scenarios, the mean coverage of Method 1 is remarkably low. For example, the average coverage of the confidence region based on the Beran's estimator for Model 2 is 40%. This value indicates that only in 60 out of 100 trials does the confidence region obtained by the proposed method entirely contain the theoretical curve. However, in the same scenario, the average point coverage is 96%, so the survival curve is within the confidence region at 96 out of 100 grid points, which is a considerably high value of the pointwise coverage.

In conclusion, the two proposed methods for the confidence regions have reasonable behaviours, both presenting very high pointwise coverages. Method 1 provides confidence regions of variable width at the cost of slightly decreasing the average coverage. Method 2 has higher coverage percentages but also a larger width, which is also constant everywhere. The results obtained using the smoothed Beran's estimator in either method are promising.

Model 2	Beran				SBeran			
$P(\delta = 0 \mid X = 0.6)$	0.2		0.5		0.2		0.5	
Method	Met 1	Met 2	Met 1	Met 2	Met 1	Met 2	Met 1	Met 2
Width	0.16264	0.21677	0.21664	0.35643	0.15759	0.16426	0.17985	0.21985
Coverage (%)	39.33	97.67	40.33	97.00	97.67	97.33	58.67	98.67
Pointwise coverage(%)	96.45	99.93	95.85	99.82	98.71	99.44	96.57	99.67
IWS	0.15076	0.17167	0.21480	0.26943	0.13759	0.13550	0.16372	0.17469

Table 4.4: Coverage, average width and IWS of the 95% confidence regions by means Methods 1 and 2 and Beran's and the smoothed Beran's estimators using $N = 300$ simulated samples from Model 2.

Model 3	Beran				SBeran			
$P(\delta = 0 \mid X = 0.8)$	0.2		0.5		0.2		0.5	
Method	Met 1	Met 2	Met 1	Met 2	Met 1	Met 2	Met 1	Met 2
Width	0.34203	0.34511	0.42486	0.41146	0.24070	0.19981	0.37740	0.27440
Coverage (%)	85.33	89.00	66.67	83.33	88.67	93.67	96.00	99.67
Pointwise coverage(%)	97.56	99.32	92.90	98.91	98.24	98.94	98.67	99.94
IWS	1.37220	1.18335	2.06192	1.38213	0.93535	0.77965	1.45099	0.92742

Table 4.5: Coverage, average width and IWS of the 95% confidence regions by means Methods 1 and 2 and Beran's and the smoothed Beran's estimators using $N = 300$ simulated samples from Model 3.

This analysis is also illustrated in following figures. Figure 4.10 shows the theoretical survival curve, the estimation with MISE bandwidths and the bootstrap estimations of $S(t|x)$ from $B = 500$ resamples using both Beran's and the smoothed Beran's estimator. These graphs show the higher variability of the Beran's estimations in the resamples with respect to the smoothed Beran's estimations.

Figure 4.11 shows the confidence regions for the conditional survival function obtained by Method 1 for one sample from Models 2 and 3. The confidence regions obtained by Method 2 are shown in Figure 4.12. The higher variability of the Beran's estimations in the resamples with respect to the smoothed Beran's estimations leads to much wider confidence regions. When using Method 1, this only affects the width

of the confidence region at the right tail of the time distribution. When using Method 2, this variability causes the confidence region to have a larger width for all points on the time grid.

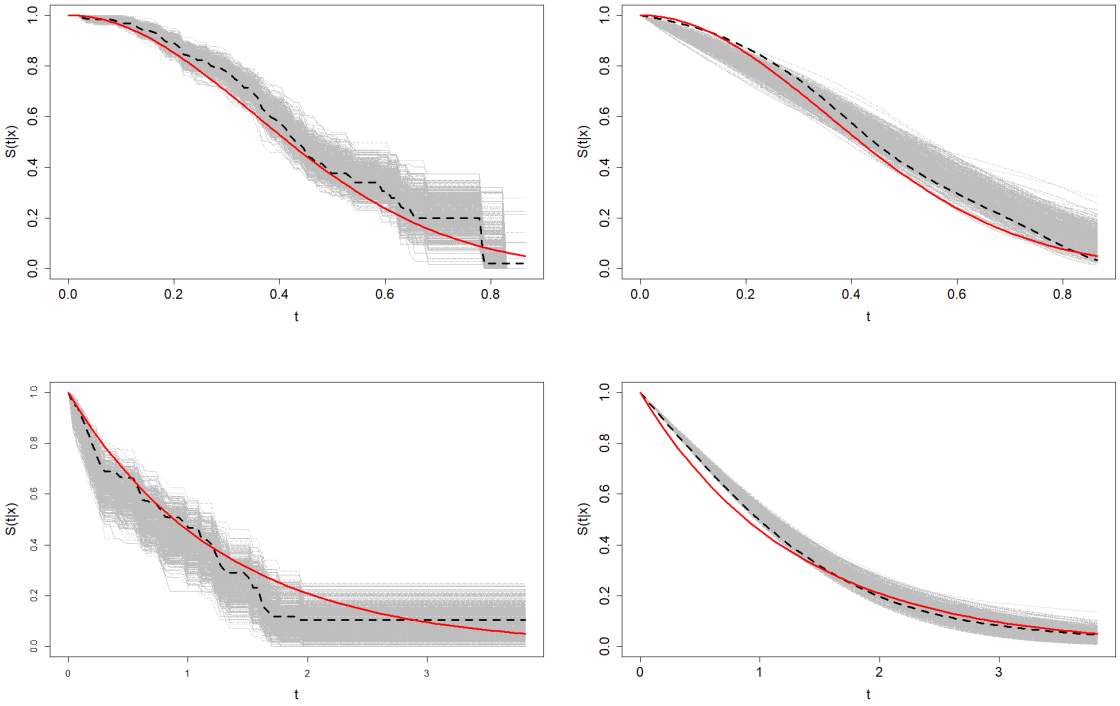

Figure 4.10: Theoretical $S(t|x)$ (red solid line) and estimation with MISE bandwidths (black dashed line) along with the bootstrap estimations of $S(t|x)$ from $B = 500$ resamples (gray dashed lines) by means of Beran's estimator (left) and the smoothed Beran's estimator (right) for one sample from Model 2 (top) and Model 3 (bottom) when $P(\delta = 0|x) = 0.5$.

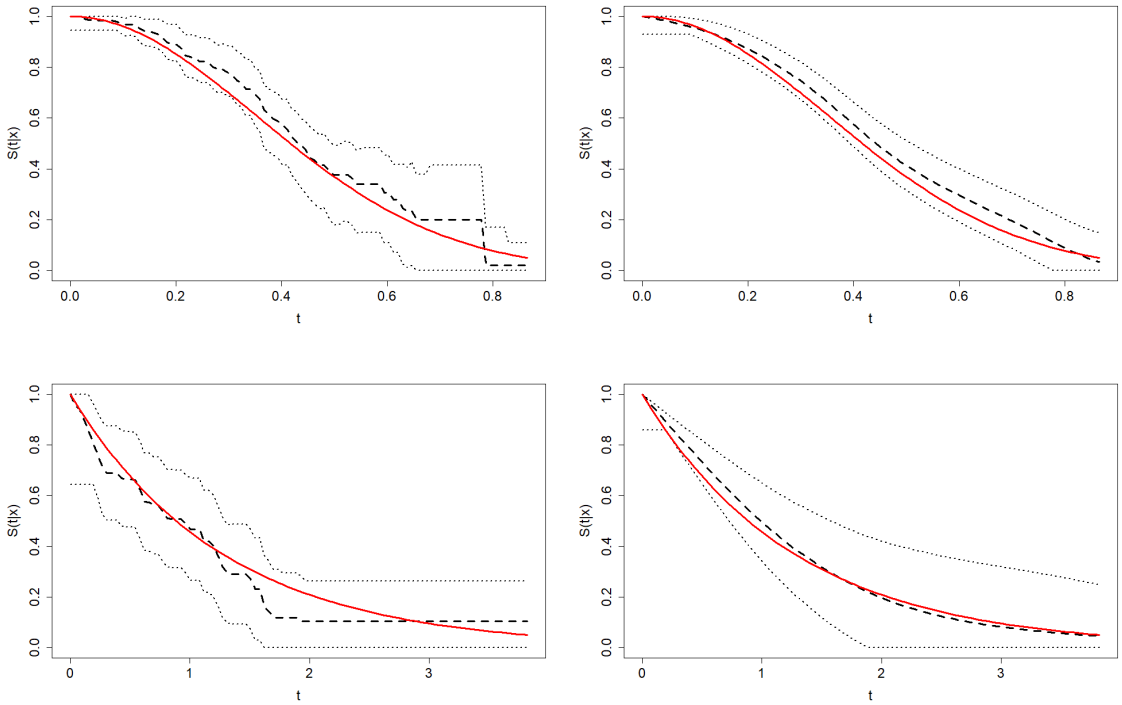

Figure 4.11: Theoretical $S(t|x)$ (red solid line), estimation with MISE bandwidths (black dashed line) and 95% confidence region (black dotted lines) by means of Beran's estimator (left) and the smoothed Beran's estimator (right) for one sample from Model 2 (top) and Model 3 (bottom) when $P(\delta = 0|x) = 0.5$ using Method 1.

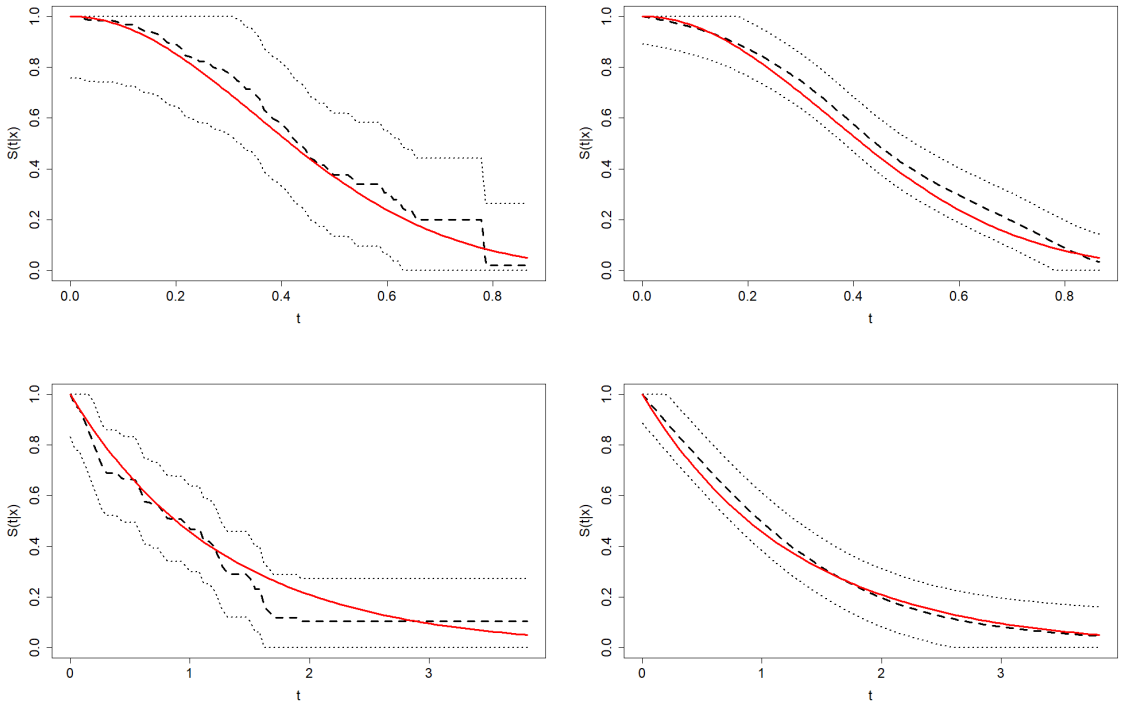

Figure 4.12: Theoretical $S(t|x)$ (red solid line), estimation with MISE bandwidths (black dashed line) and 95% confidence region (black dotted lines) by means of Beran's estimator (left) and the smoothed Beran's estimator (right) for one sample from Model 2 (top) and Model 3 (bottom) when $P(\delta = 0|x) = 0.5$ using Method 2.

4.6 Analysis of the computational times

This section includes a brief analysis of the computation times required in the processes detailed above.

Note that the CPU time of the resampling is the same both when using the Beran's estimator and when using the smoothed Beran's estimator. This is due to the fact that the resampling is identical in both cases, except for the perturbation on the lifetime and censoring variables, which is insignificant in terms of computation time.

Regarding the selection of a bootstrap bandwidth, the difference between the two methods lies in the function to be minimised, since it will be unidimensional in

the case of the Beran's estimator, but two-dimensional in the case of the smoothed Beran's estimator. Since the optimization of the error function was conducted using an external R package, its computational efficiency is not analysed here.

In the first scenario, the number of samples $N = 1$ and the number of resamples $B = 100$ are set and different sample sizes are considered. Table 4.6 shows the resulting times.

n	100	200	400	800	1600	3200	6400
Time	0h 0' 12"	0h 0' 49"	0h 4' 32"	0h 29' 41"	3h 31' 48'	27h 36' 40"	227h 58' 3"

Table 4.6: CPU times of the resampling method for $N = 1$ sample of size n and $B = 100$ bootstrap resamples.

In the second scenario the sample size, $n = 100$, is fixed and either the number of samples, N, or the number of resamples, B, are varied as a verification that the processes are linear over these parameters. Tables 4.7 and 4.8 show the results.

Approximating the bootstrap bandwidth or computing a confidence region by 500 resamples from a sample of size 100 requires one minute and a half, while a sample of size 500 requires 25 minutes to obtain the result. These times seem to increase quadratically as the sample size grows, which may lead to prohibitive times for very large sample sizes.

N	50	100	500	1000
Time	0h 14' 49"	0h 31' 17"	1h 29' 24"	2h 59' 7"

Table 4.7: CPU times of the resampling method for N samples of size $n = 100$ and $B = 100$ bootstrap resamples.

B	50	100	500	1000	5000
Time	0h 0' 8"	0h 0' 16"	0h 1' 26"	0h 3' 2"	0h 15' 9"

Table 4.8: CPU times of the resampling method for $N = 1$ sample of size $n = 100$ and B bootstrap resamples.

4.7 Application to real data

The usefulness of the automatic bootstrap selector of the bandwidths of Beran's and the smoothed Beran's estimator is illustrated in this section. The survival function of the time that COVID-19 patients remain hospitalised in ward or the Intensive Care Unit (ICU) is estimated by means of Beran's and the smoothed Beran's estimators. A dataset from SERGAS (Galician health service) with dates of admission and discharge (if applicable), age, gender and previous diseases of COVID-19 patients in Galicia (Spain) is available.

The event of interest is the patient leaving ward or ICU, so the time variable which is subject to right random censoring is the time until the patients leave the ward or the ICU. An informative covariate of the survival time is the age of the patient. Other factors like sex or previous diaseases are used to disaggregate interesting subpopulations. There are certain risk factors for COVID-19 that could affect hospitalisation and recovery times. Two of these are obesity and COPD. COPD (chronic obstructive pulmonary disease) is a chronic inflammatory lung disease that causes obstructed airflow from the lungs. The following paragraphs take into consideration whether or not patients have obesity or COPD in order to analyse their influence on the hospitalisation times.

4.7.1 Time until leaving ward

The time until a COVID-19 positive patient leaves the ward is first considered. A patient leaves the ward because he/she is discharged, admitted to the ICU or dies. When none of these three circumstances is observed for a patient before the end of the study, the censoring time is what is observed. The total number of hospitalised patients followed up is 2453 and the censoring rate of this dataset is 8.85%.

Table 4.9 shows summary statistics of the hospitalisation time in ward and the age of COVID-19 patients disaggregating the censored and uncensored groups. Figure 4.13 shows the histogram of the time in ward and the age for all patients.

Figure 4.14 informs about the proportion of men and women suffering from each of the above pathologies considered in the study.

		min.	$1^{st}Q.$	median	mean	$3^{th}Q.$	max.
Censored data	Time	1.00	5.00	15.00	18.22	28.00	105.00
	Age	4.00	69.00	80.00	76.04	87.00	100.00
Uncensored data	Time	1.00	6.00	10.00	13.02	16.00	75.00
	Age	0.00	60.00	72.00	69.61	82.00	106.00

Table 4.9: Summary statistics for time of the stay in ward (Z) and age (X) for the uncensored group (patients who left ward) and the censored group (patients in ward).

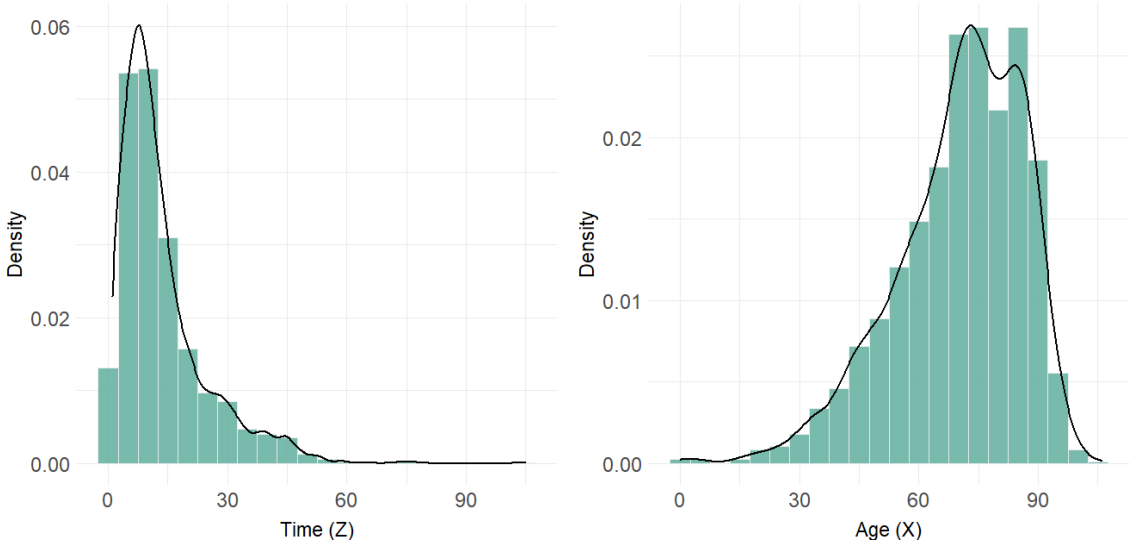

Figure 4.13: Histogram and kernel density estimation for the time in ward (left) and age (right).

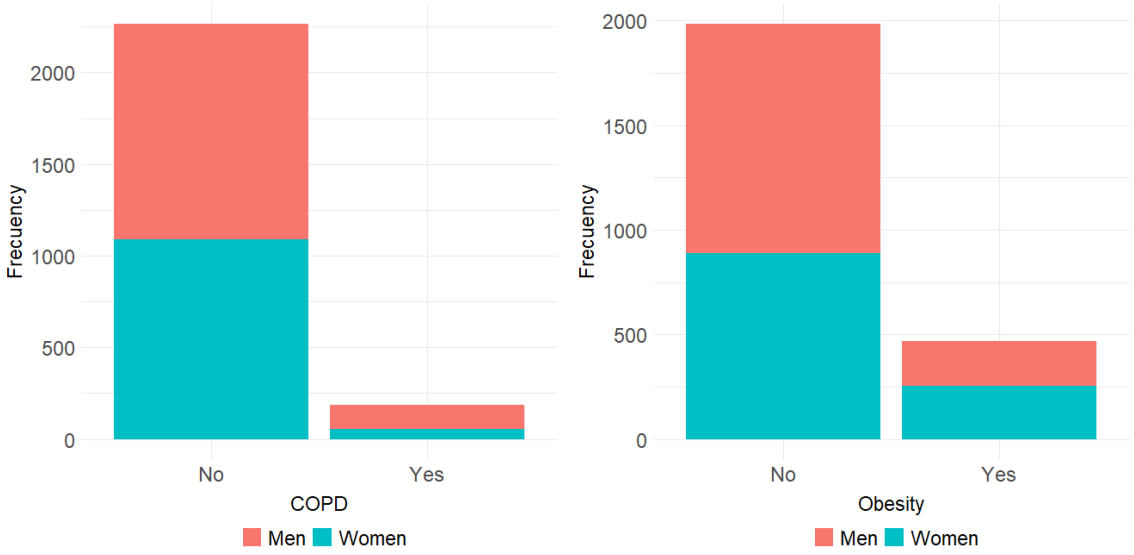

Figure 4.14: Bar chart of the COPD variable (left) and the obesity variable (right) by gender of the patients in ward.

The bootstrap algorithm shown in the previous section is used here to compute the bootstrap bandwidths for estimating the survival function of the time in ward of the Galician COVID-19 patients. Due to the good results that the smoothed Beran's estimator showed in the previous simulations, this is the estimator mainly used in this section. Some interesting confidence regions based on the smoothed Beran's estimator are also obtained. Method 1, which was proposed in Section 4.4 and provides confidence regions of varying width, is used.

The bootstrap estimation is obtained in a grid of time $t_1 < \cdots < t_{n_T}$ with $t_{n_T} = \widehat{Q}(0.95)$ and $n_T = 100$. The pilot bandwidth for the covariate used in the bootstrap algorithm is the one defined in (4.3) with $c = 3/2$. Figure 4.15 shows the reason for the choice of the constant c, as it can be seen that the regression function of T on X, estimated using the Nadaraya-Watson estimator for censored data, is not very variable with respect to x. The pilot bandwidth for the time variable is the one defined in (4.6).

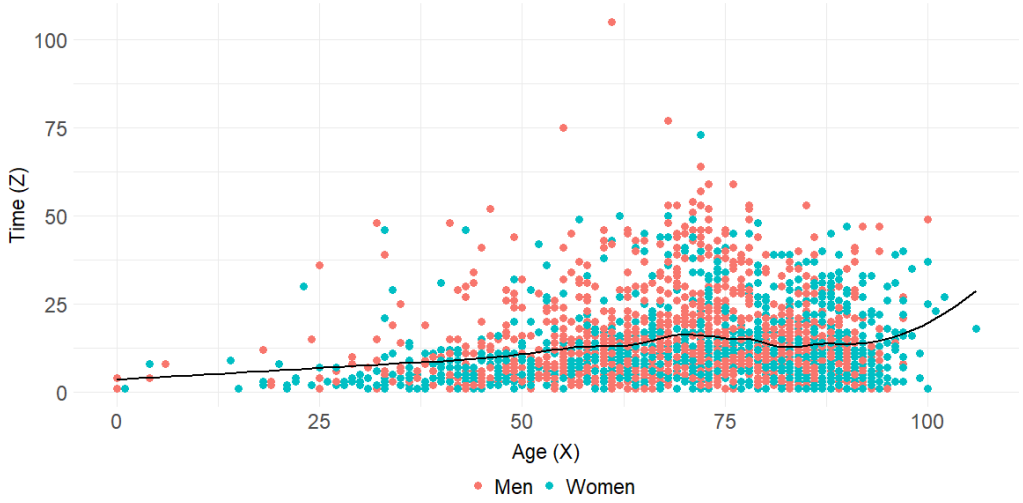

Figure 4.15: Scatter plot and estimated regression function of age versus time in ward.

Firstly, three age profiles are considered: 40, 60 and 80 years old. In some cases, due to the differences found, a brief analysis will be included for the 30 year old profile. In other cases, because of sample limitations, only 60 and 80 year old profiles will be considered.

The bandwidth that minimises the Monte Carlo approximation of the bootstrap MISE, $MISE_x^*(h^*)$ for Beran's estimator along with the square root of MISE, $RMISE_x^*(h^*)$ and the two-dimensional bandwidth that minimises the Monte Carlo approximation of the bootstrap MISE, $MISE_x^*(h^*, g^*)$, for the smoothed Beran's estimator along with the square root of MISE, $RMISE_x^*(h^*, g^*)$, are shown in Table 4.10. For $x = 80$, the $RMISE_x^*(h, g)$ function is decreasing in h, so the bandwidth selector for the smoothed Beran's estimator proposes as the bootstrap bandwidth the upper end of the interval considered for this smoothing parameter.

	Beran		SBeran		
x	h^*	$RMISE_x^*(h_1^*)$	h_2^*	g_2^*	$RMISE_x^*(h_2^*, g_2^*)$
40	4.765306	0.024209	5.507370	1.266695	0.160623
60	4.571429	0.016797	5.548651	0.784956	0.115867
80	13.387760	0.012351	30.000000	2.348188	0.055815

Table 4.10: Bootstrap bandwidth and bootstrap RMISE for Beran's estimation and the smoothed Beran's estimation of the conditional survival function of the time in ward for some different values of age.

Figure 4.16 shows the boostrap estimations of the survival function by means of Beran's and the smoothed Beran's estimator. The differences between the two estimations are not remarkable, except for the reduction of the roughness of the smoothed Beran's estimation. Only 20% of the 40 year old patients spend more than 15 days in ward. Meanwhile, 40% of COVID-19 positive patient of 60 or 80 years old spend more than 15 days in ward and only 20% of these patients spend more than 25 days in ward.

Figure 4.17 shows the estimation of the conditional survival function of the time in the ward of a 60-year-old patient and the bootstrap confidence region at the 95% confidence level obtained by Method 1. The average width of the confidence region is 0.1227.

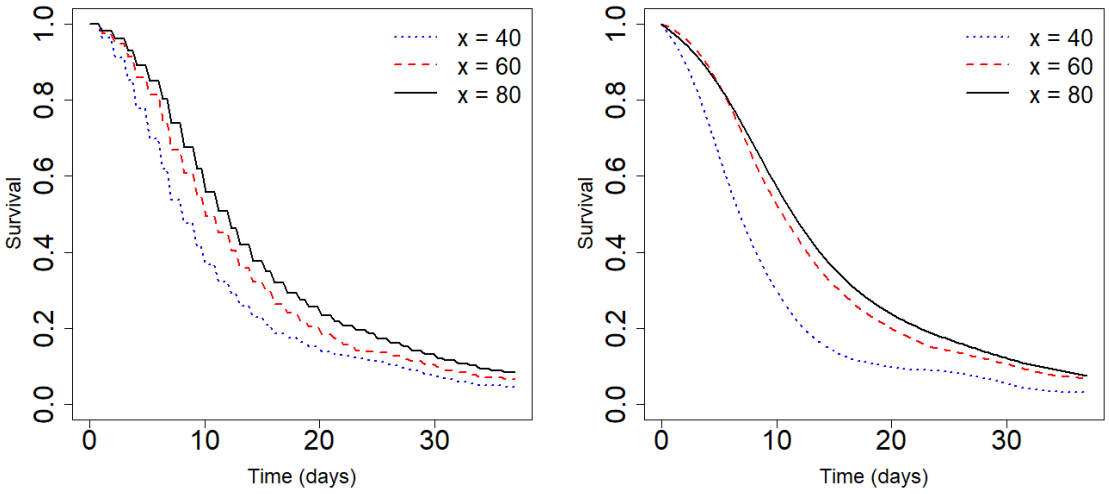

Figure 4.16: Estimation of $S(t|x)$ for time in ward with Beran's estimator (left) and the smoothed Beran's estimator (right) using the bootstrap bandwidths for $x = 40$ (dotted line), $x = 60$ (dashed line) and $x = 80$ (solid line).

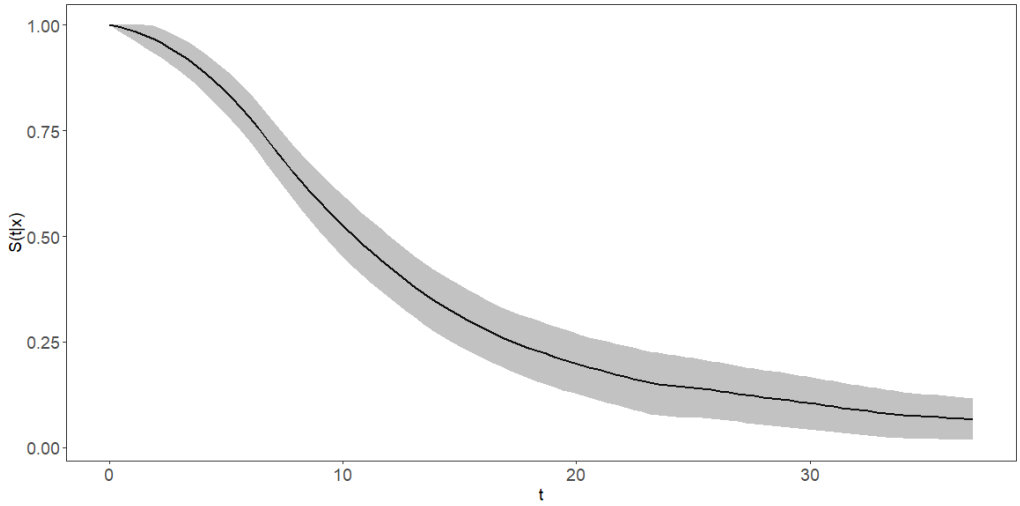

Figure 4.17: Estimation of $S(t|x)$ with bootstrap bandwidths for time in ward and bootstrap confidence region by means of Method 1 based on the smoothed Beran's estimator for $x = 60$.

Recovery times were also analysed by classifying individuals into two gender populations. The main conclusions are shown in following paragraphs.

Figure 4.18 shows that there is no remarkable differences between ages when restricting to the men subpopulation. On the contrary, the distribution of the time

in ward seems to be slightly different for women of different ages. About 20% of women aged 60-80 spend more than 20 days in ward. Meanwhile, only 10% of 40-year-old women spend more than 20 days in ward. Furthermore, Figure 4.19 shows that young women have shorter recovery times than young men.

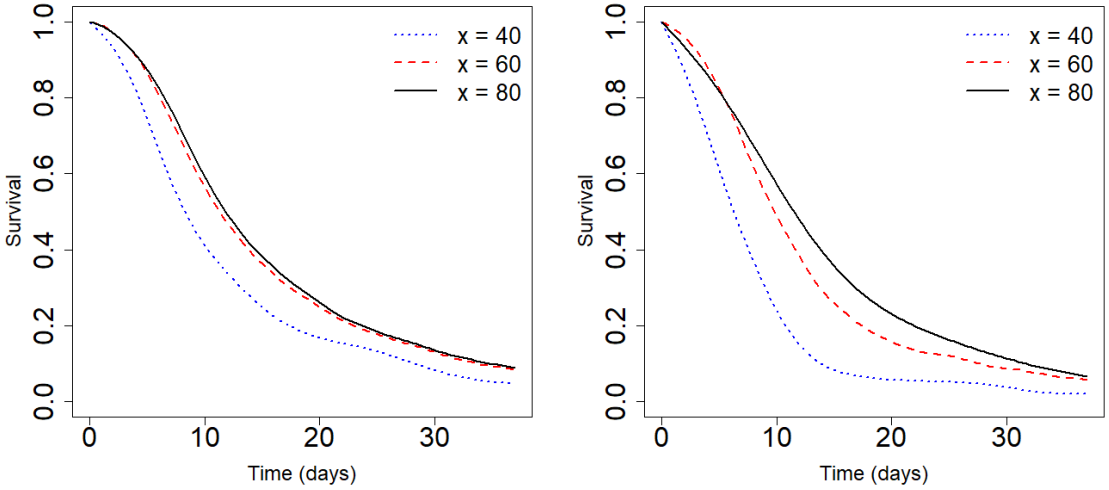

Figure 4.18: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidths in the men subpopulation (left) and in the women subpopulation (right) for $x = 40$ (dotted line), $x = 60$ (dashed line) and $x = 80$ (solid line).

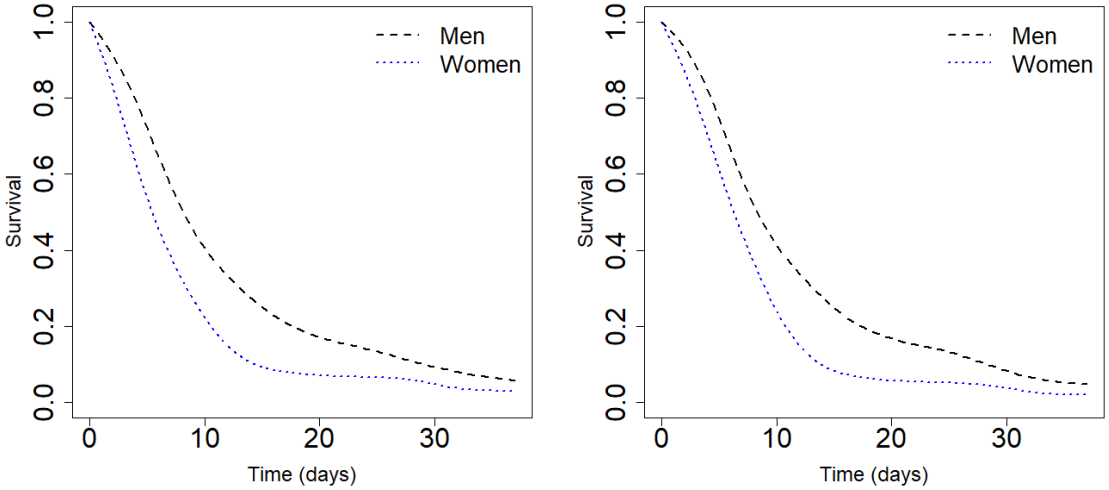

Figure 4.19: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidths with $x = 30$ (left) and $x = 40$ (right) in the men (dashed lines) and women (dotted lines) subpopulations.

Figure 4.20 shows the estimation of the conditional survival function of the time in ward of a 60-year-old man and a 60-year-old woman and the corresponding bootstrap confidence regions at the 95% confidence level. The average width of the confidence region for the men subpopulation is 0.1224 and for the women subpopulation is 0.1272.

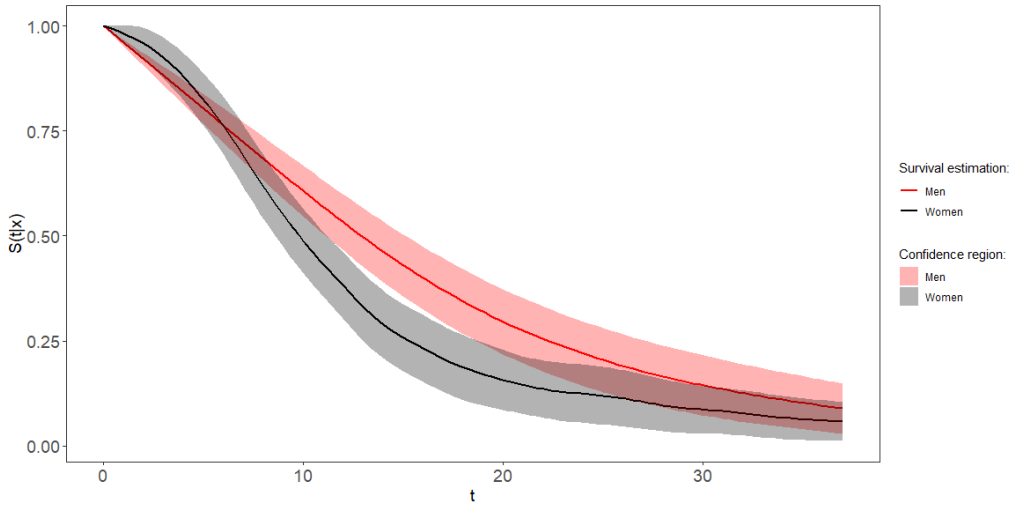

Figure 4.20: Estimation of $S(t|x)$ with bootstrap bandwidths for time in ward and bootstrap confidence region by means of Method 1 based on the smoothed Beran's estimator for $x = 60$ in the men (red lines) and the women (black lines) subpopulations.

Now, it is considered whether or not patients have COPD. The possible effect of this risk factor on recovery times is discussed below. The age profiles considered here are 60 and 80 years because the proportion of young patients in the sample diagnosed with COPD is low.

Figure 4.21 shows that there is no significant difference in the recovery time of patients with and without COPD for these ages. Although the recovery times of COPD patients in their 60s might be very slightly higher than in non-COPD patients, no differences are observed for patients in their 80s (Figure 4.22).

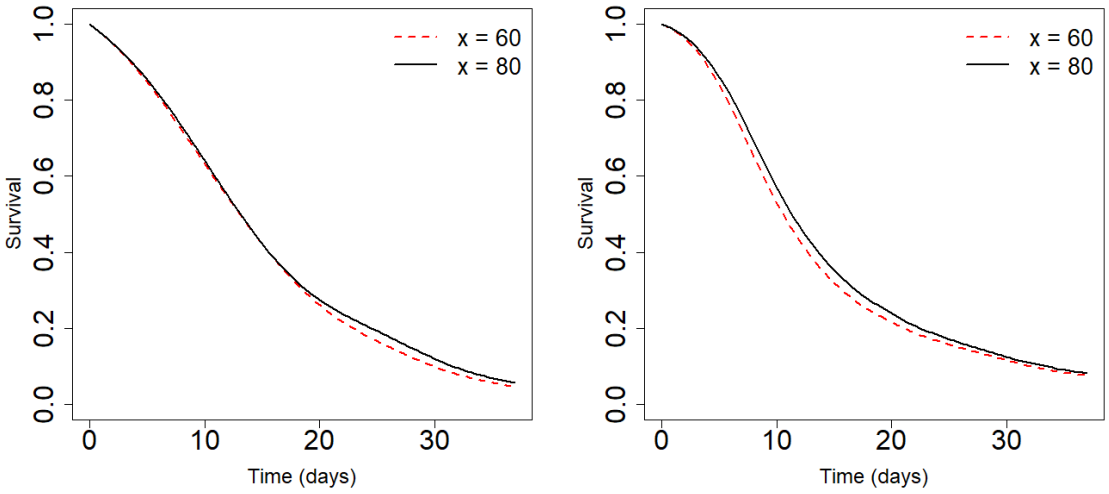

Figure 4.21: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidth in the COPD patients subpopulation (left) and non-COPD patients subpopulation (right) for $x = 60$ (dashed line) and $x = 80$ (solid line).

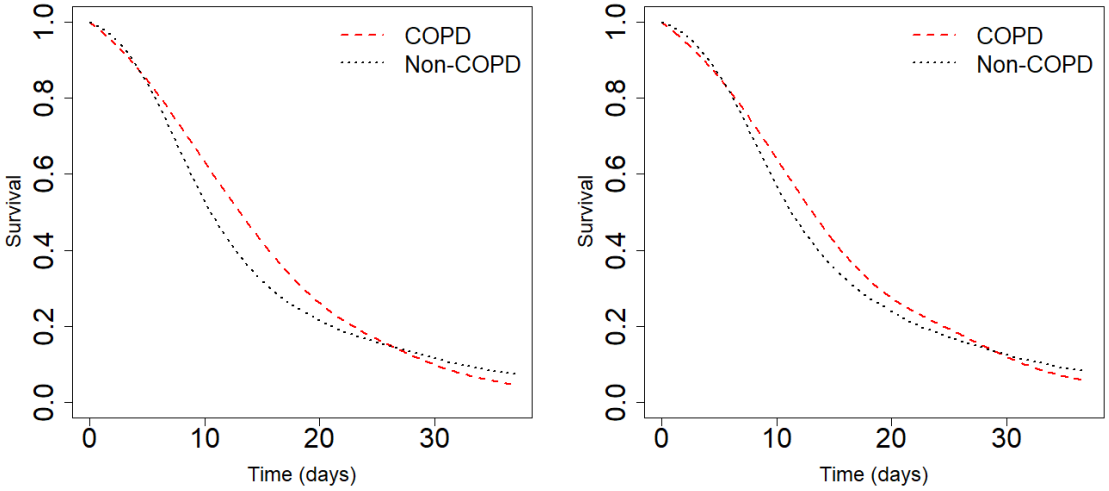

Figure 4.22: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidth with $x = 60$ (left) and $x = 80$ (right) in the COPD (dashed lines) and non-COPD (dotted lines) subpopulations.

Another risk factor for COVID-19 disease is obesity, so its possible effect on the time of hospitalisation (in ward) is studied.

Figure 4.23 shows that the effect of age on recovery time is greatly attenuated

by obesity. That is, in the case of obesity, the hospitalisation time is similar for all considered ages. Figure 4.24 shows that hospitalisation times are somewhat longer in 40-year-old patients with obesity than in 40-year-old patients without obesity. In contrast, at older ages, the effect of this risk factor is not appreciable: hospitalisation times in ward do not differ between patients with and without obesity in their 80s.

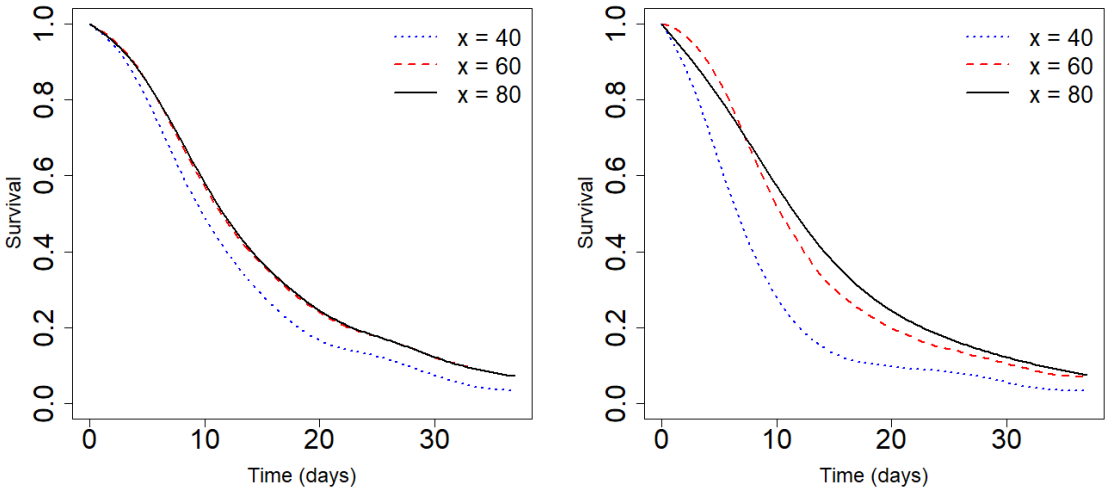

Figure 4.23: Estimation of $S(t|x)$ for time in ward with Beran's estimator using the bootstrap bandwidth in the obesity patients subpopulation (left) and non-obesity patients subpopulation (right) for $x = 40$ (dotted line), $x = 60$ (dashed line) and $x = 80$ (solid line).

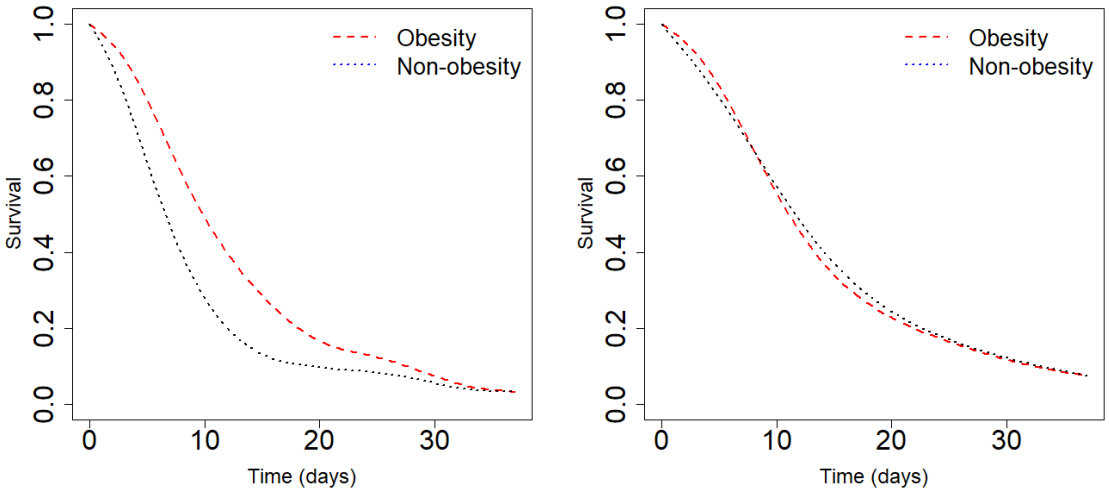

Figure 4.24: Estimation of $S(t|x)$ for time in ward with Beran's estimator using the bootstrap bandwidth with $x = 40$ (left) and $x = 80$ (right) in the obesity (dashed lines) and non-obesity (dotted lines) subpopulations.

4.7.2 Time until leaving ICU

The time until leaving the Intensive Care Unit is considered in this section. A COVID-19 positive patient leaves the ICU because he/she is discharged (from the ICU) or dies and his/her time until the event of interest is knwon. In other case, the censoring time is what is observed. The total number of patients in the ICU who were followed up is 288 and the censoring rate of this dataset is 14.58%. Table 4.11 shows summary statistics of the hospitalisation time in the ICU and the age of COVID-19 patients disaggregating the censored and uncensored groups. Figure 4.25 shows the histogram of the time in the ICU and the age for all patients. Figure 4.26 informs about the proportion of men and women suffering from each of the above pathologies considered in the study.

		min.	$1^{st}Q.$	median	mean	$3^{th}Q.$	max.
Censored data	Time	1.00	24.00	33.50	30.45	43.75	51.00
	Age	42.00	64.75	71.00	69.21	75.00	84.00
Uncensored data	Time	0.00	5.00	12.00	14.04	20.00	53.00
	Age	25.00	60.00	68.00	65.50	73.00	89.00

Table 4.11: Summary statistics for time of the stay in ICU (Z) and age (X) for the uncensored group (patients who left ICU) and the censored group (patients in ICU).

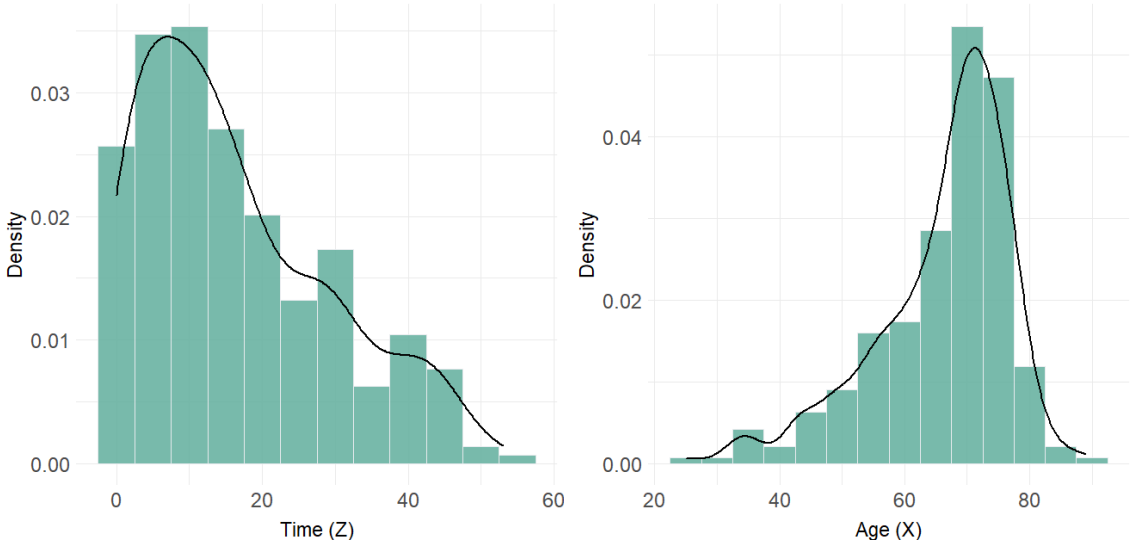

Figure 4.25: Histogram and kernel density estimation for the time in ICU (left) and age (right).

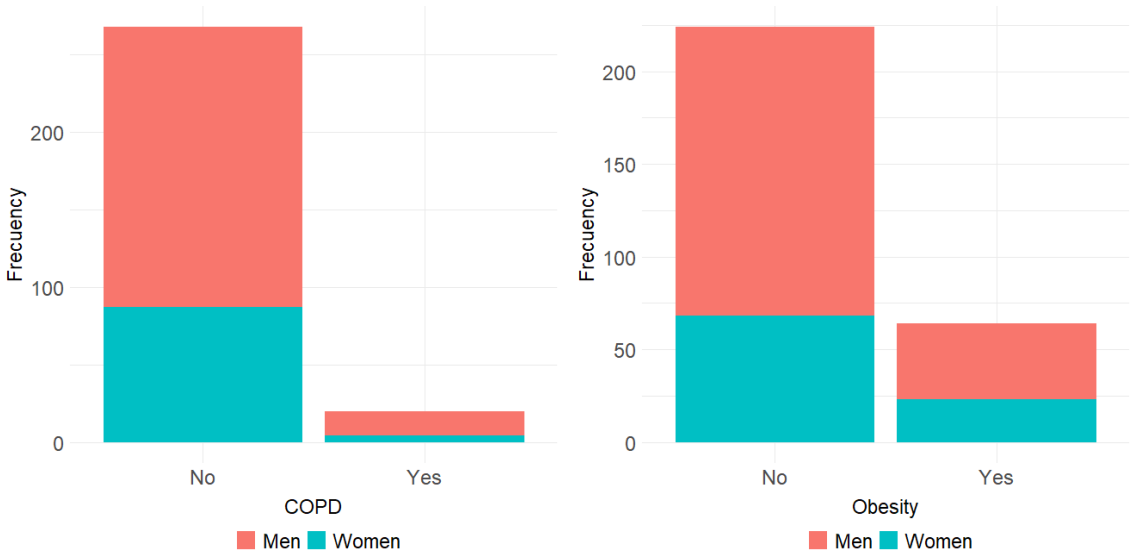

Figure 4.26: Bar chart of the COPD variable (left) and the obesity variable (right) by gender of the patients in the ICU.

As in the previous subsection, the smoothed Beran's estimator is used for estimating the survival function of the time in ICU of the Galician COVID-19 patients with bootstrap bandwidths obtained by the automatic selector proposed in Section 4.2.2. Some interesting confidence regions based on the smoothed Beran's estimator are also obtained. Method 1, which was proposed in Section 4.4 and provides confidence regions of varying width, is used. The bootstrap estimation is obtained in a grid of time $t_1 < \cdots < t_{n_T}$ with $t_{n_T} = \widehat{Q}(0.95)$ and $n_T = 100$. The pilot bandwidth for the covariate used in the bootstrap algorithm is the one defined in (4.3) with $c = 3/2$. The choice of c can be justified by the low variability of the regression function of the time in ICU as a function of age (see Figure 4.27). The pilot bandwidth for the time variable was defined in (4.6). Again, three age profiles are considered: 40, 60 and 80 years old.

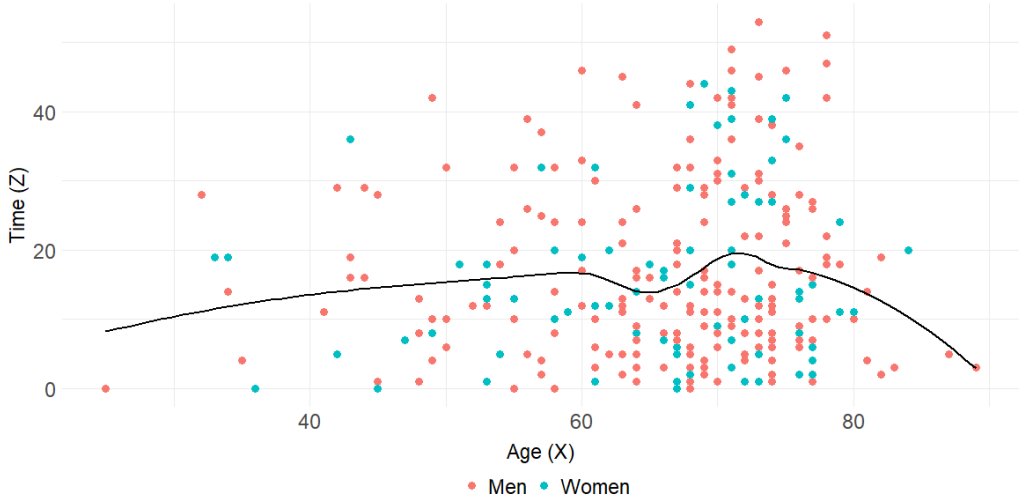

Figure 4.27: Scatter plot and estimated regression function of age versus time in ICU.

The bandwidth that minimises the Monte Carlo approximation of the bootstrap MISE, $MISE_x^*(h^*)$ for Beran's estimator along with the square root of MISE, $RMISE_x^*(h^*)$, and the two-dimensional bandwidth that minimises the Monte Carlo approximation of the bootstrap MISE, $MISE_x^*(h^*, g^*)$, for the smoothed Beran's estimator along with the square root of MISE, $RMISE_x^*(h^*, g^*)$, are shown in Table 4.12. For $x = 60$ and $x = 80$, the $RMISE_x^*(h, g)$ function is decreasing in h, so the bandwidth selector for the smoothed Beran's estimator proposes as the bootstrap bandwidth the upper end of the interval considered for this smoothing parameter.

	Beran		SBeran		
x	h^*	$RMISE_x^*(h_1^*)$	h_2^*	g_2^*	$RMISE_x^*(h_2^*, g_2^*)$
40	12.438780	0.055774	13.102690	2.949216	0.320261
60	11.244900	0.034801	30.000000	2.030116	0.218001
80	15.622450	0.031022	30.000000	5.739782	0.160175

Table 4.12: Bootstrap bandwidth and bootstrap RMISE for Beran' estimation and the smoothed Beran's estimation of the conditional survival function of the time in ICU for some different values of age.

Figure 4.28 shows the survival function of time in the ICU estimated for several

ages by means of Beran's and the smoothed Beran's estimator.

It can be seen, in contrast to the time in ward, that age has little effect on time in the ICU, except in hospitalisations of more than 20 days where slight differences can be seen with time in ICU being shorter in younger age groups.

Figure 4.29 shows the estimation of the coditional survival function of the time in the ICU of a 60-year-old patient and the bootstrap confidence region at the 95% confidence level obtained by Method 1 based on the smoothed Beran's estimator. The average width of the confidence region is 0.1529.

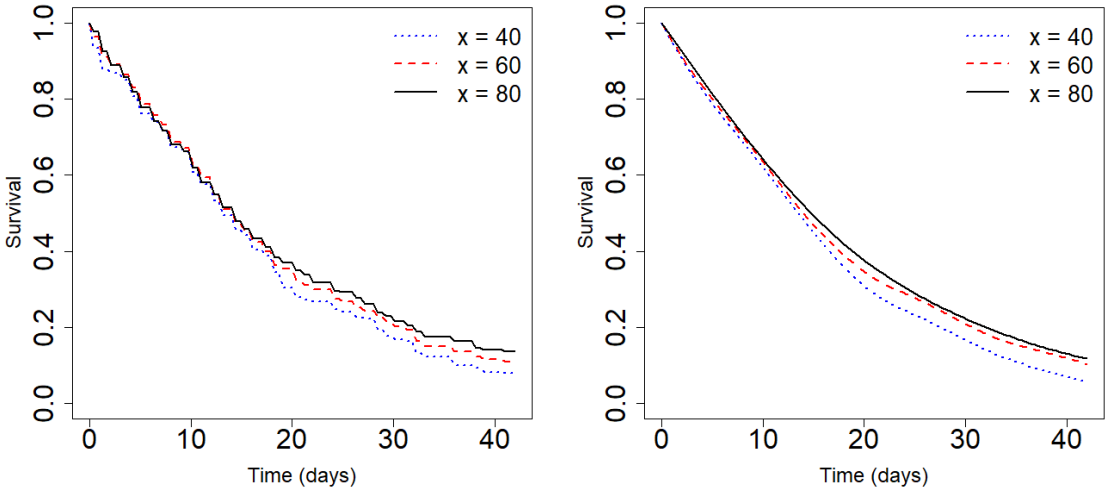

Figure 4.28: Estimation of $S(t|x)$ for time in ICU with Beran's estimator (left) and the smoothed Beran's estimator (right) using the bootstrap bandwidths for $x = 40$ (dotted line), $x = 60$ (dashed line) and $x = 80$ (solid line).

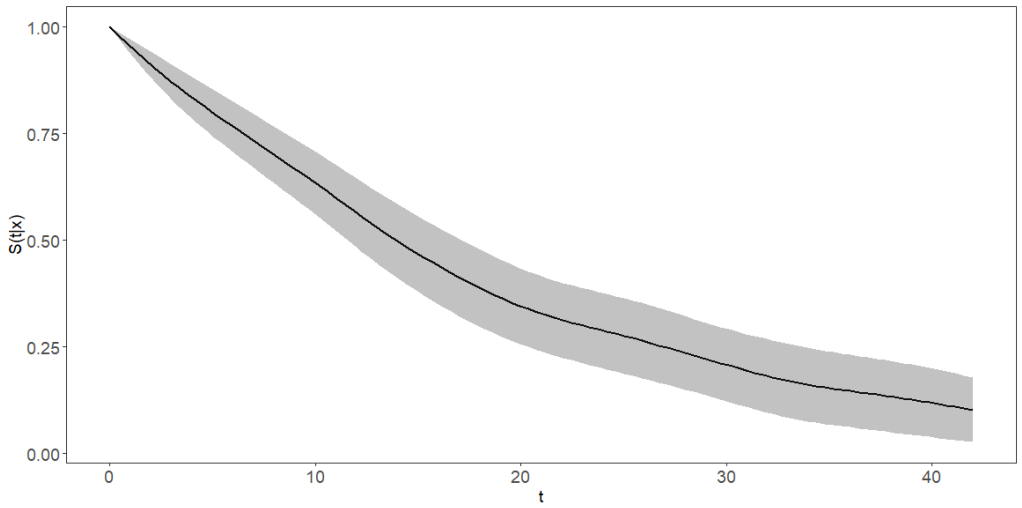

Figure 4.29: Estimation of $S(t|x)$ with bootstrap bandwidths for time in ICU and bootstrap confidence region by means of Method 1 based on the smoothed Beran's estimator for $x = 60$.

An analysis of the factors sex, diagnosis of COPD and obesity, parallel to the one carried out for time in ward, is included here for time in the ICU. The conclusions of this study for ICU time do not differ from those obtained for ward time.

The percentage of the patients spending at least 20 days in the ICU is nearly 40%. The estimated survival function shows that 20% of the COVID-19 patients spend more than 30 days in the ICU. According to Figure 4.30 there is no significant differences in the probability of survival with respect to age. Although age has no global impact on the ICU time, it does have an effect when we consider the male and female populations independently: young women have shorter ICU times than young men, but there are no differences between the older age groups (Figure 4.31).

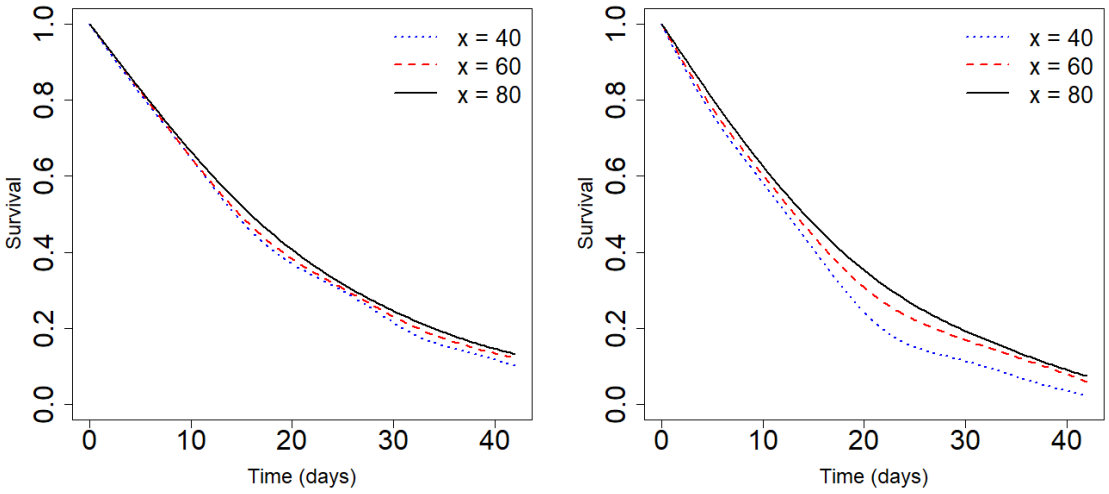

Figure 4.30: Estimation of $S(t|x)$ for time in the ICU with the smoothed Beran's estimator using the bootstrap bandwidth in the men subpopulation (left) and in the women subpopulation (right) for $x = 40$ (dotted line), $x = 60$ (dashed line) and $x = 80$ (solid line).

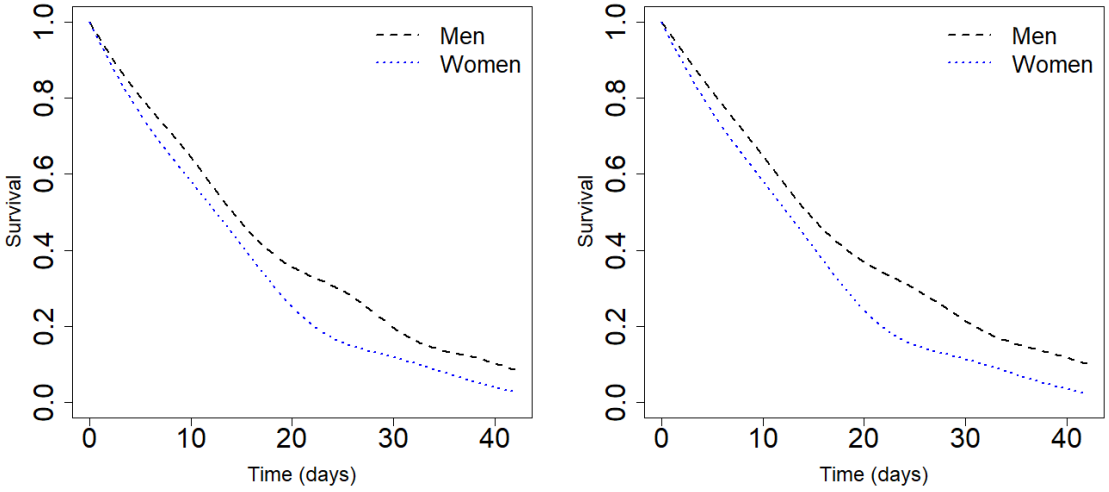

Figure 4.31: Estimation of $S(t|x)$ for time in the ICU with the smoothed Beran's estimator using the bootstrap bandwidth with $x = 30$ (left) and $x = 40$ (right) in the men (dashed lines) and women (dotted lines) subpopulations.

Figure 4.32 shows the estimation of the conditional survival function of the time in the ICU of a 60-year-old man and a 60-year-old woman and the corresponding bootstrap confidence regions at the 95% confidence level obtained by Method 1 based

on the smoothed Beran's estimator. The average width of the confidence region for the men subpopulation is 0.1752 and for the women subpopulation is 0.2306.

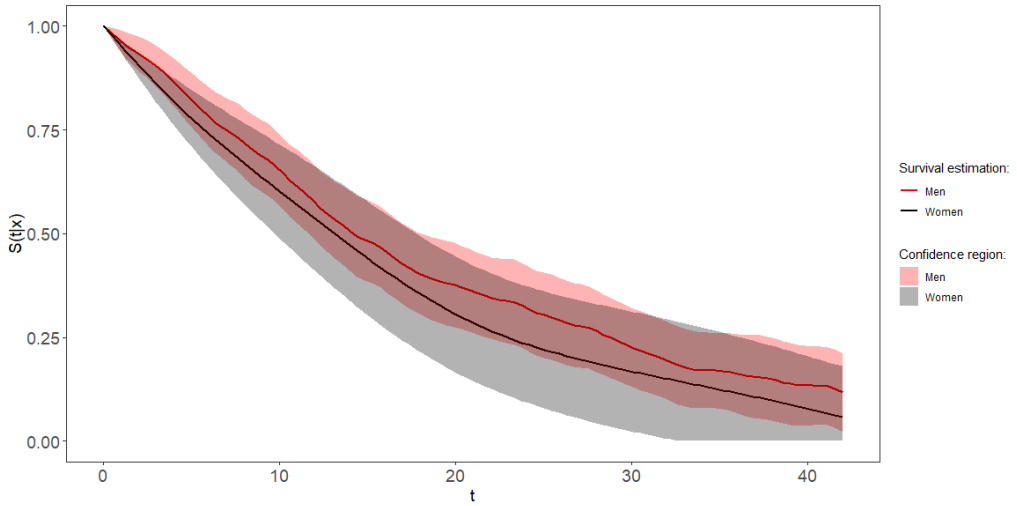

Figure 4.32: Estimation of $S(t|x)$ with bootstrap bandwidths for time in the ICU and bootstrap confidence region by means of Method 1 based on the smoothed Beran's estimator for $x = 60$ in the men (red lines) and the women (black lines) subpopulations.

Both risk factors, COPD and obesity, have a negative effect on patients' recovery times. In both cases this effect is attenuated by age. See Figures 4.33 and 4.34 for the COPD results and Figures 4.35 and 4.36 for obesity results.

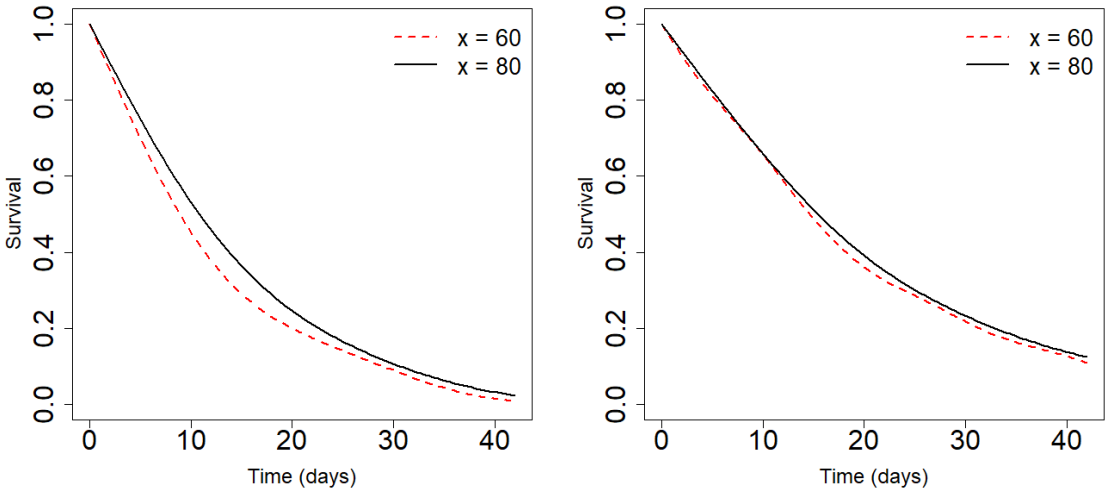

Figure 4.33: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidth in the COPD patients subpopulation (left) and non-COPD patients subpopulation (right) for$x = 60$ (dashed line) and $x = 80$ (solid line).

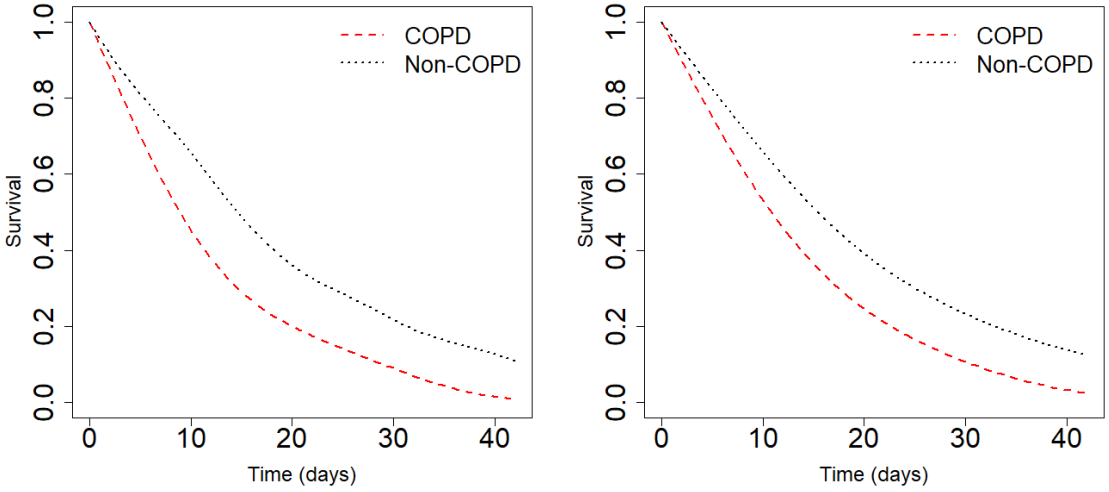

Figure 4.34: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidth with $x = 60$ (left) and $x = 80$ (right) in the COPD (dashed lines) and non-COPD (dotted lines) subpopulations.

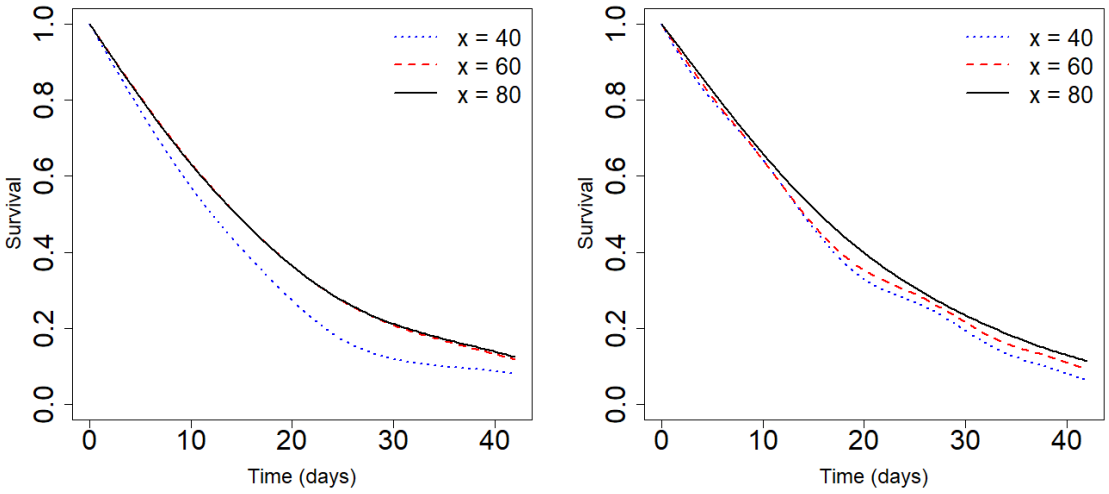

Figure 4.35: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidth in the obesity patients subpopulation (left) and non-obesity patients subpopulation (right) for $x = 40$ (dotted line), $x = 60$ (dashed line) and $x = 80$ (solid line).

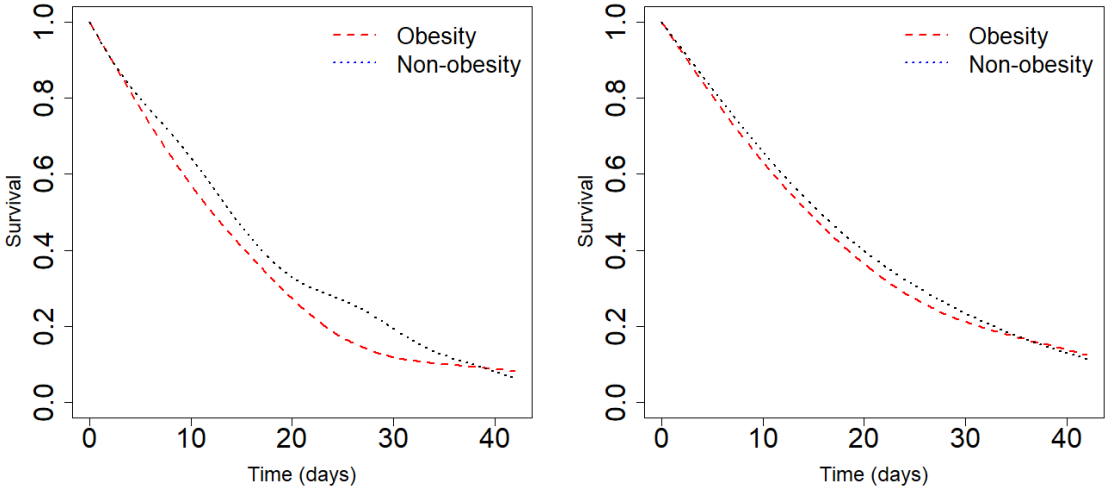

Figure 4.36: Estimation of $S(t|x)$ for time in ward with the smoothed Beran's estimator using the bootstrap bandwidth with $x = 40$ (left) and $x = 80$ (right) in the obesity (dashed lines) and non-obesity (dotted lines) subpopulations.

Chapter 5

Doubly smoothed estimator of the probability of default

5.1 Introduction

The works of Cao et al. (2009) or Peláez et al. (2021b) and also Chapter 2 consider probability of default estimators based on survival estimators which are smoothed with respect to the covariate, but step functions with respect to t, each jump ocurring at uncensored observed lifetimes. This fact along with the survival ratio required to compute the PD by Equation (2.1) are the cause of the roughness and variability observed in the probability of default estimations obtained in Chapter 2.

In the previous chapter 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. This doubly smoothed survival estimator can be used to obtain a doubly smoothed version of the PD estimator according to Equation (2.2). It is expected that the resulting estimator will no longer exhibit the roughness

problem observed so far. The aim is not only to improve the graphical representation of the estimated PD, but also to reduce the estimation error and obtain more realistic estimations in their application.

For this purpose, a general nonparametric estimator of the PD with double smoothing derived from the smoothed survival estimator is proposed in this chapter. The asymptotic properties of the doubly smoothed PD estimator based on Beran's estimator are studied. A simulation study shows the improvement obtained by using the double smoothing on a number of nonparametric estimators of the probability default, including Beran's estimator. Finally, the doubly smoothed PD estimator based on Beran's estimator is applied to a set of modified real data.

The content of this chapter is published in Peláez et al. (2021a).

5.2 Doubly smoothed PD estimator

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a 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. Consider the doubly smoothed survival estimator $\widetilde{S}_{h,g}(t|x)$, defined in (3.2). Replacing $S(t|x)$ in (2.1) by $\widetilde{S}_{h,g}(t|x)$, the doubly smoothed nonparametric estimator of the probability of default is as follows:

$$\widetilde{PD}_{h,g}(t|x) = 1 - \frac{\widetilde{S}_{h,g}(t+b|x)}{\widetilde{S}_{h,g}(t|x)}. \tag{5.1}$$

Since $\widehat{S}_h(t|x)$ in (3.2) is any arbitrary conditional survival estimator, and therefore so is $\widetilde{S}_{h,g}(t|x)$, the probability of default estimator $\widetilde{PD}_{h,g}(t|x)$ is very general. Nevertheless, this chapter mainly focuses on the smoothed Beran's estimator $\widetilde{S}^B_{h,g}(t|x)$ given in (3.4). Using $\widetilde{S}^B_{h,g}(t|x)$ in (5.1), the smoothed probability of default estimator based on Beran's survival estimator is obtained as follows:

$$\widetilde{PD}^B_{h,g}(t|x) = 1 - \frac{\widetilde{S}^B_{h,g}(t+b|x)}{\widetilde{S}^B_{h,g}(t|x)}. \tag{5.2}$$

5.3 Asymptotic results for the smoothed Beran's estimator

Asymptotic theory for the smoothed Beran's estimator of the PD, $\widetilde{PD}^{B}_{h,g}(t|x)$, is derived from the asymptotic properties of the smoothed Beran's survival estimator presented in Section 3.3.

Certain functions need to be defined in order to state these results. Some of them were already established in Section 3.3.1. Additional functions are defined below:

$$b_1(t,x) = \frac{d_K\Big(1-F(t|x)\Big)}{2m(x)}\Big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)m(x)\Big),$$

$$b_2(t,x) = -\frac{1}{2}d_K F''(t|x),$$

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

$$V_2(t,x) = \frac{c_K(c_{\mathbb{K}}-1)}{m(x)}\Big(1-F(t|x)\Big)^2 L'(t|x),$$

$$C(t_1,t_2,x) = \frac{c_K}{m(x)}2\Big(1-F(t_1|x)\Big)\Big(1-F(t_2|x)\Big)L(t_1|x),$$

The assumptions required to state the results are standard in the literature and not too restrictive in this context. They were previously assumed in the nonparametric conditional survival function estimation setup and listed in Section 3.3.1.

Theorem 5.1. *Let $(t,x) \in [l,u] \times I_c$ be such that $S(t|x) > 0$. Under assumptions A.1-A.10 and assuming $nh^3 \to \infty$ and $nhg^2 \to \infty$ when $n \to \infty$, expressions for the asymptotic bias and the asymptotic variance of $\widetilde{PD}^{B}_{h,g}(t|x)$ are the following:*

$$\begin{aligned} Bias\Big(\widetilde{PD}^{B}_{h,g}(t|x)\Big) &= \frac{\Big(1-PD(t|x)\Big)b_1(t,x) - b_1(t+b,x)}{S(t|x)}h^2 \\ &\quad + \frac{\Big(1-PD(t|x)\Big)b_2(t,x) - b_2(t+b,x)}{S(t|x)}g^2 \\ &\quad + o(h^2) + o(g^2) + O\left(\frac{1}{nh}\right), \end{aligned}$$

$$
\begin{aligned}
&Var\Big(\widetilde{PD}^{B}_{h,g}(t|x)\Big) \\
&= \left(\frac{V_1(t+b,x)}{S(t|x)^2} - 2\frac{S(t+b|x)C(t,t+b,x)}{S(t|x)^3} + \frac{S(t+b|x)^2V_1(t,x)}{S(t|x)^4}\right)\frac{1}{nh} \\
&\quad + \left(\frac{V_2(t+b,x)}{S(t|x)^2} + \frac{S(t+b|x)^2V_2(t,x)}{S(t|x)^4}\right)\frac{g}{nh} + o\left(\frac{1}{nh}\right) + O\left(\frac{h^2+g^2}{nh}\right).
\end{aligned}
$$

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 their expressions are complex and depend on too many population functions.

Theorem 5.2. *Under the assumptions of Theorem 5.1 and assuming*

$$
C_h := \lim_{n\to\infty} n^{1/5}h > 0, \quad C_g := \lim_{n\to\infty} n^{1/5}g > 0,
$$

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

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

Remark 5.1. *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* $\sqrt{nh}\Big(\widetilde{PD}^{B}_{h,g}(t|x) - PD(t|x)\Big) \xrightarrow{d} N(0, s_0)$.

Proofs of these results are included in Section 5.6.

5.4 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.

Models presented in Section 2.4 are again considered. This makes it possible to compare the results obtained in both studies. The simulation setup is similar to the one proposed in Section 3.4.

The probability of default curve is estimated in a time grid $0 < t_1 < \cdots < t_{n_T}$ of size n_T where $t_{n_T} + b = F^{-1}(0.95|x)$ and the horizon of default b is about 20% of the time range. For Model 1, $x = 0.8$, $b = 0.1$ and $t_{n_T} = 0.4991$. Model 2 considers $x = 0.6$, $b = 0.15$ and $t_{n_T} = 0.7154$. For Model 3, $x = 0.8$, $b = 0.7$ and $t_{n_T} = 3.1211$.

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).

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 MISE as explained in Section 2.4.

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.

The optimal bandwidth (h_2, g_2) is chosen as the pair which minimises some Monte Carlo approximations of

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

based on $N = 100$ simulated samples. Then, the value of the $MISE$ of $\widetilde{PD}_{h_2,g_2}^B(t|x)$

is approximated from $N = 1000$ simulated samples.

In order to minimise the function $MISE_x(h, g)$ without increasing CPU time more than necessary, the limited-memory algorithm based on the quasi-Newton method, explained in Section 3.4, is considered.

Neither the bandwidth for Beran's estimator nor the bandwidths for the smoothed Beran's estimator can be used in practice but their choice produces a fair comparison since the estimators are built using their best possible smoothing parameters.

Figure 5.1 shows the function $MISE_x(h, g)$ over a meshgrid of 50 values of h and 50 values of g for Models 1, 2 and 3 when the conditional censoring probability is 0.5. These graphs show the two-dimensional functions to minimise in order to obtain the optimal bandwidths for the smoothed Beran's PD estimator. The red zone is where this minimum is reached and the coordinates of the point at which the minimum is attained provide the optimal smoothing bandwidths. The results for other levels of censoring probability, which are not shown here, are quite similar.

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. However, for a fixed value of h, the value of g for which the smallest error is made does not seem to vary too much depending on the value of the covariate smoothing bandwidth (h). Figure 5.2 shows this analysis. There, $MISE_x(h, g)$ is shown as a function of g for some fixed values of h within the interval where the optimum is reached. The obtained curves have similar shape and they are close for all the values of h, mainly at the highest level of censoring conditional probability. The minimum of $MISE_x(h, g)$ is reached for similar values of g in all the scenarios.

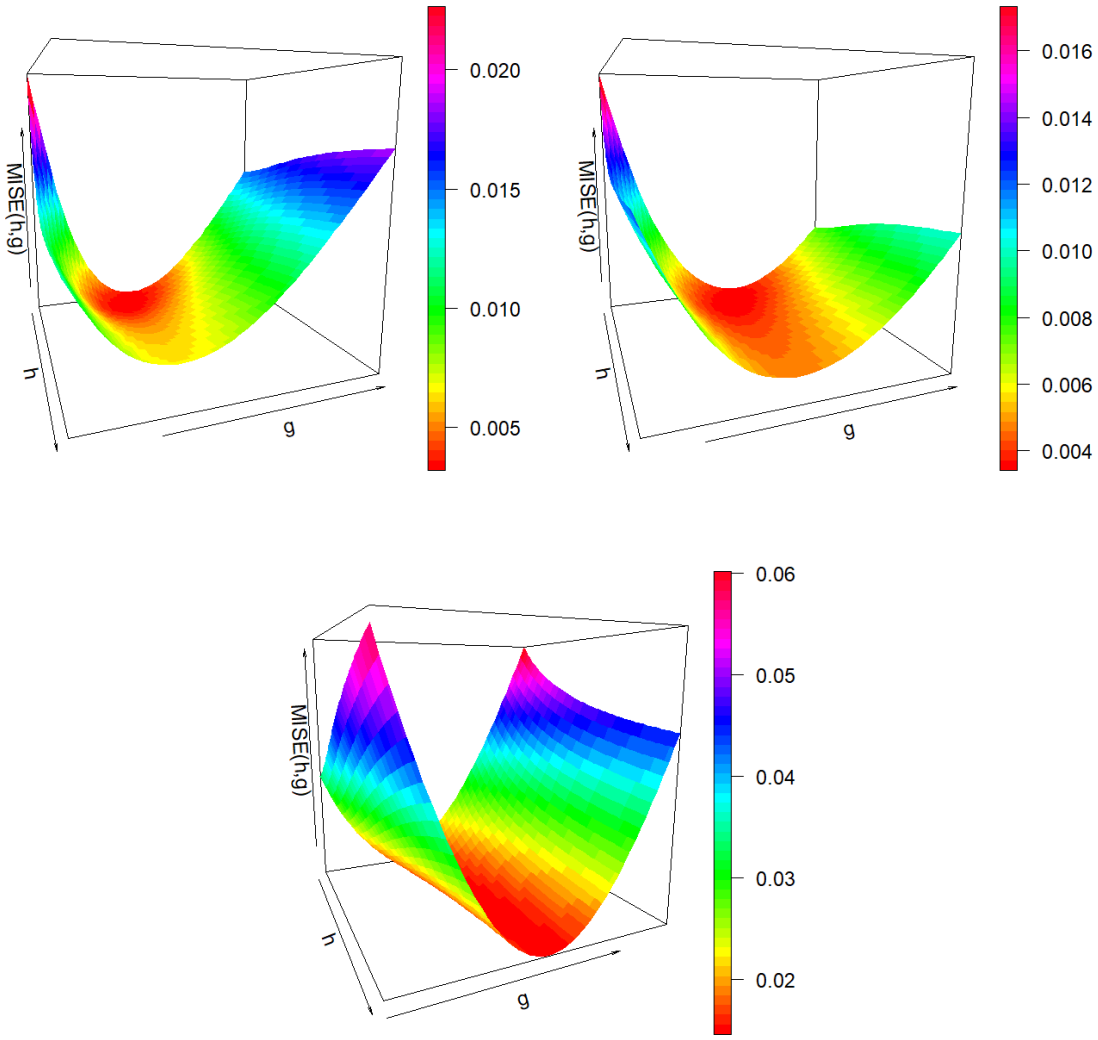

Figure 5.1: $MISE_x(h, g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (topleft), Model 2 (topright) and Model 3 (bottom) when $P(\delta = 0|x) = 0.5$.

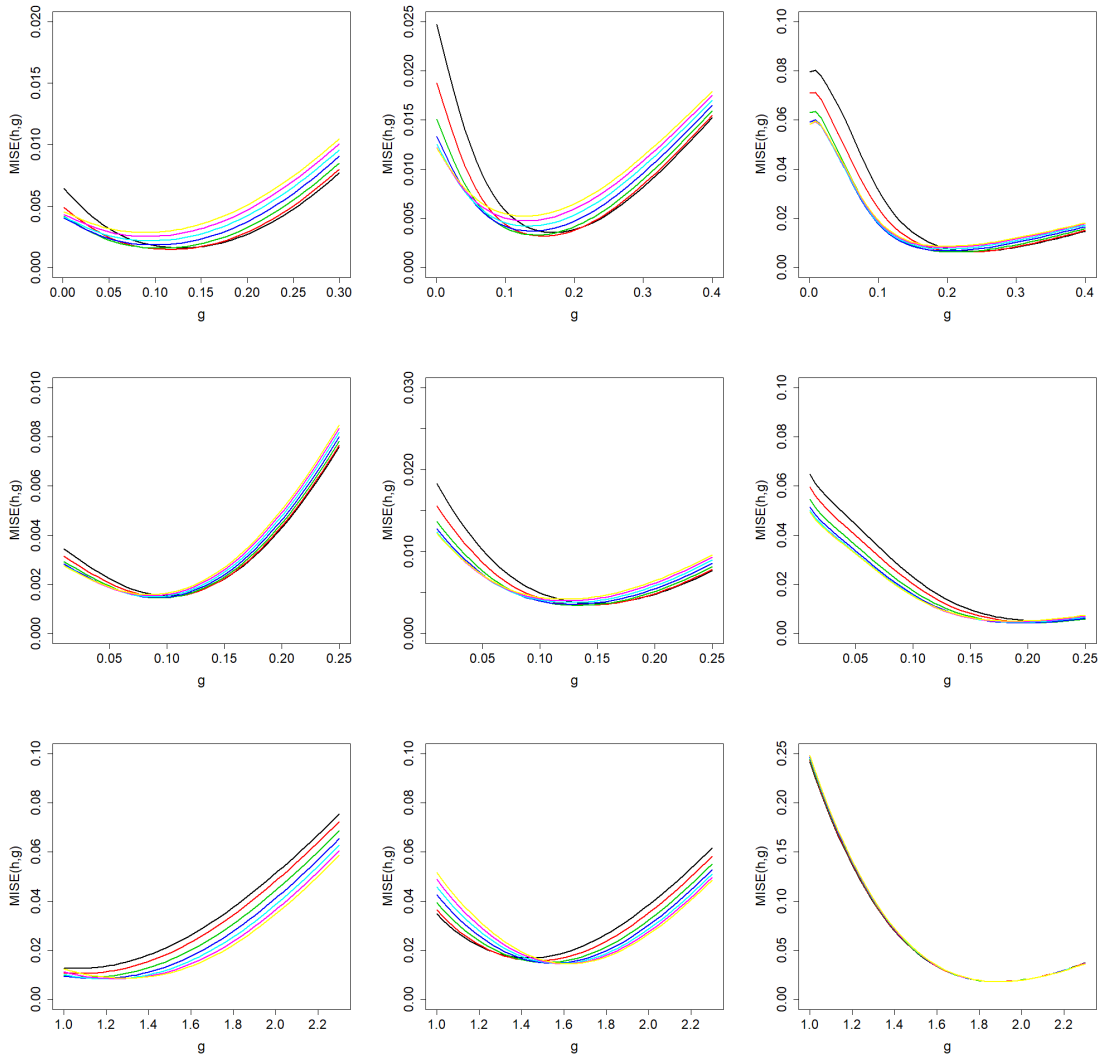

Figure 5.2: $MISE_x(h, g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top), Model 2 (middle) and Model 3 (bottom) for some fixed equispaced values of $h \in [0.1, 0.5]$ when $P(\delta = 0|x) = 0.5$.

Tables 5.1, 5.2 and 5.3 show the optimal bandwidths and the square root of the MISE (RMISE) of Beran's estimator and the smoothed Beran's estimator for each model. In order to compare the behaviour of the estimators and quantify the improvement of the smoothing over the original estimator, the ratio R_x is defined

$$R_x = \frac{RMISE_x\Big(\widetilde{PD}^B_{h_2,g_2}(\cdot|x)\Big)}{RMISE_x\Big(\widehat{PD}^B_{h_1}(\cdot|x)\Big)}$$

The closer to 0 the value of R_x, the greater the improvement of the smoothed Beran's estimator with respect to Beran's estimator.

In all cases, $RMISE$ values are lower for the smoothed Beran's estimator and this difference becomes bigger when increasing the censoring conditional probability. This is confirmed by looking at the values of R_x.

When the censoring conditional probability is 0.2 or 0.5 in Models 1 and 2, the time smoothing reduces the error by about $40-50\%$ and this improvement is more than 60% when the conditional probability of censoring is 0.8. The error reduction in Model 3 with respect to the nonsmoothed PD estimator is more significant, reaching 75% and 80% when censoring is moderate or heavy, respectively.

Model 1				
$P(\delta=0\|x)$		0.2	0.5	0.8
$\widehat{PD}^B_{h_1}$	h_1	0.24286	0.39592	0.42857
	$RMISE_x(h_1)$	0.06311	0.10626	0.20925
$\widetilde{PD}^B_{h_2,g_2}$	h_2	0.14438	0.15233	0.18917
	g_2	0.11510	0.15228	0.21839
	$RMISE_x(h_2,g_2)$	0.03687	0.05498	0.07647
	R_x	0.58422	0.51741	0.36545

Table 5.1: Optimal bandwidths, $RMISE$ and R_x of the PD estimation for Beran's estimator and the smoothed Beran's estimator with optimal bandwidths in each level of conditional censoring probability for Model 1.

Model 2				
$P(\delta = 0\|x)$		0.2	0.5	0.8
$\widehat{PD}^B_{h_1}$	h_1	0.30204	0.34082	0.39898
	$RMISE_x(h_1)$	0.05437	0.11195	0.25738
$\widetilde{PD}^B_{h_2,g_2}$	h_2	0.21687	0.15559	0.18962
	g_2	0.09320	0.13651	0.19811
	$RMISE_x(h_2, g_2)$	0.03846	0.05946	0.06198
	R_x	0.70738	0.53113	0.24083

Table 5.2: Optimal bandwidths, $RMISE$ and R_x of the PD estimation for Beran's estimator and the smoothed Beran's estimator with optimal bandwidths in each level of conditional censoring probability for Model 2.

Model 3				
$P(\delta = 0\|x)$		0.2	0.5	0.8
$\widehat{PD}^B_{h_1}$	h_1	0.09898	0.13163	0.15204
	$RMISE_x(h_1)$	0.27128	0.49813	0.67999
$\widetilde{PD}^B_{h_2,g_2}$	h_2	0.10722	0.26967	1.00000
	g_2	1.20340	1.61882	1.89462
	$RMISE_x(h_2, g_2)$	0.09208	0.12337	0.13431
	R_x	0.33944	0.24767	0.19751

Table 5.3: Optimal bandwidths, $RMISE$ and R_x of the PD estimation for Beran's estimator and the smoothed Beran's estimator with optimal bandwidths in each level of conditional censoring probability for Model 3.

Figure 5.3 shows a cloud of estimated survival curves (50 out of 1000), the theoretical survival function, the mean curve and the 5th and 95th percentiles of the total estimated curves for the smoothed Beran's estimator in Model 1, Model 2 and Model 3. These figures show clearly how the estimated curves are distributed and the variability they present. Note how the behaviour of the estimators become worse when the conditional probability of censoring increases, since lack of information leads to poor performance of the estimators, especially in Model 3.

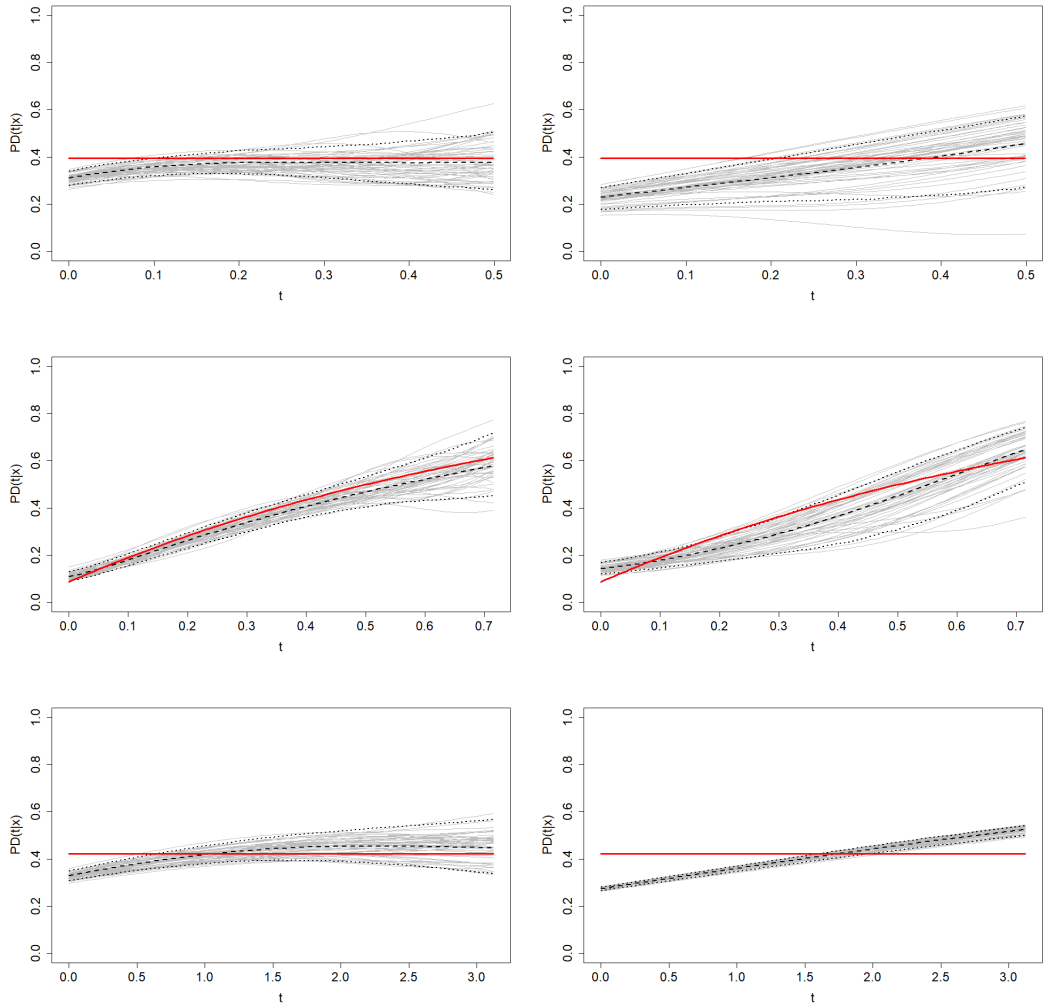

Figure 5.3: Theoretical $PD(t|x)$ (solid line), mean curve (dashed line) and 5th and 95th percentiles (dotted line) obtained by means of the smoothed Beran's estimator when $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.8$ (right) in Model 1 (top), Model 2 (middle) and Model 3 (bottom).

As explained in the introduction of this dissertation, credit scoring is usually obtained by means of a logistic regression using different characteristics of the client. Typical scoring values range from 0 to 1, and summarises the client's solvency. Clients with lower credit scoring are previously rejected for receiving the credit. Then, a negative skewed distribution of the credit scoring is expected in a real data set.

Based on the definition of credit scoring, assuming a uniform distribution of this variable is not realistic in this context. The reason for choosing the uniform

distribution for the simulation models is just following the simulations performed by Van Keilegom et al. (2001) for comparison purposes. A small analysis of Beran's estimator and the smoothed Beran's estimator with modified Models 2 and 3 is included below.

The beta distribution has a compact support, $[0, 1]$, and its parameters can be chosen to be negatively asymmetric. We chose $X \equiv Beta(10, 2)$. Its density function is shown in Figure 5.4. Models 2 and 3 consider the same distributions and parameters except for the beta distribution of the covariate. Strategy 1 is used to approximate the optimal bandwidths.

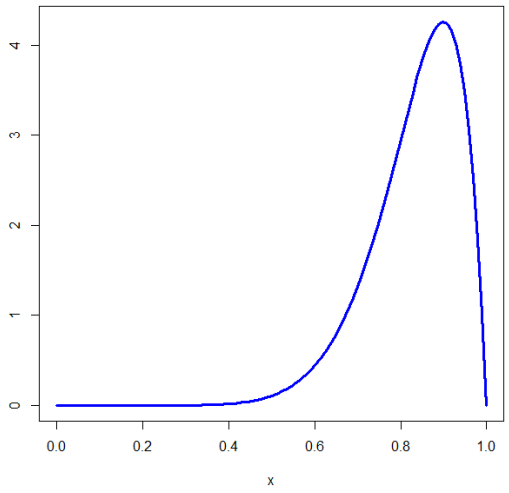

Figure 5.4: Probability density function of $Beta(\alpha = 10, \beta = 2)$.

Table 5.4 shows the optimal bandwidths, the estimation error and the ratio R_x for Beran's estimator and the smoothed Beran's estimator. When the censoring conditional probability is 0.2 or 0.5 in Model 2, the time smoothing reduces the error by about 60% and this improvement is about 80% when the conditional probability of censoring is 0.8. The error reduction in Model 3 with respect to the nonsmoothed PD estimator is also significant, reaching 80% when censoring is moderate or heavy. This brief analysis shows that the results of these simulations and the good appealing behaviour of the smoothed Beran's estimator hold when the distribution of X is not uniform but a more realistic asymmetric distribution for the credit scoring.

		Model 2 ($X \equiv Beta(10,2)$)			Model 3 ($X \equiv Beta(10,2)$)		
$P(\delta = 0\|x)$		0.2	0.5	0.8	0.2	0.5	0.8
$\widehat{PD}_h^B$	h_1	0.16122	0.18510	0.47551	0.06571	0.10551	0.10551
	$RMISE_x(h_1)$	0.20854	0.32296	0.38625	0.19733	0.40411	0.66071
$\widetilde{PD}_{h,g}^B$	h_2	0.16122	0.18510	0.47551	0.06571	0.10551	0.10551
	g_2	0.35347	0.35347	0.39388	1.01020	1.33674	1.77959
	$RMISE_x(h_2, g_2)$	0.07905	0.07415	0.06612	0.07600	0.11423	0.13945
	R_x	0.37906	0.22959	0.17118	0.38514	0.28267	0.21106

Table 5.4: Optimal bandwidths, $RMISE$ and R_1 of the PD estimation for Beran's estimator and the smoothed Beran's estimator with Strategy 1 in each level of censoring conditional probability for Model 1 and Model 2 with beta distribution for the covariate.

The computation time of both estimators should be considered in the comparison. Table 5.5 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 5.5 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.

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

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.

Any other estimator of the conditional survival function could be considered to obtain the corresponding smoothed estimator defined in (3.2) and then, to estimate the probability of default through the expression given in (5.1). In particular, two other survival estimators are considered in this work: the Weighted Nadaraya-Watson estimator (WNW) defined in (2.7) and the Van Keilegom-Akritas estimator (VKA) defined in (2.8). They are respectively denoted by $\widehat{S}_h^{WNW}(t|x)$ and $\widehat{S}_h^{VKA}(t|x)$. Their smoothed versions are built according to Equation (3.2), obtaining the following smoothed survival estimators: $\widetilde{S}_{h,g}^{WNW}(t|x)$ and $\widetilde{S}_{h,g}^{VKA}(t|x)$. Replacing $\widetilde{S}_{h,g}(t|x)$ with $\widetilde{S}_{h,g}^{WNW}(t|x)$ and $\widetilde{S}_{h,g}^{VKA}(t|x)$ in Equation (5.1) gives the nonparametric smoothed estimators of $PD(t|x)$ denoted by $\widetilde{PD}_{h,g}^{WNW}(t|x)$ and $\widetilde{PD}_{h,g}^{VKA}(t|x)$.

The optimal bandwidths for these estimators are obtained following the strategy used for the smoothed Beran's estimator of the PD introduced in this section. In some of the scenarios analysed for the smoothed WNW estimator, $MISE(h,g)$ turned out to be a decreasing function of h. For this reason, the bandwidth h_2 selected was a high but reasonable value, considering that the variable X moves in the interval $[0,1]$.

In order to quantify the improvement that the smoothing provides to the PD estimators and compare the performance of the three estimators, the ratios $R_S^\bullet(x)$ and $R_c(x)$ are defined for a fixed value of x as follows:

$$R_S^\bullet(x) = \frac{RMISE_x\Big(\widetilde{PD}_{h_2,g_2}^{\bullet}(\cdot|x)\Big)}{RMISE_x\Big(\widehat{PD}_{h_1}^{\bullet}(\cdot|x)\Big)}$$

$$R_c^\bullet(x) = \frac{RMISE_x\Big(\widetilde{PD}_{h_2,g_2}^{B}(\cdot|x)\Big)}{RMISE_x\Big(\widetilde{PD}_{h_2,g_2}^{\bullet}(\cdot|x)\Big)}$$

being $\bullet = B, WNW, VKA$ and they are included in Tables 5.6, 5.7 and 5.8 along with the approximation of the optimal smoothing parameters and the estimation error of each estimator.

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, the smoothing in the time variable is confirmed to be 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 smoothed WNW estimator is the estimator whose error is reduced the most, followed by Beran's estimator.

The values of $R_c^\bullet$ indicate how much better or worse the smoothed Beran's estimator is. The lower the value of $R_c^\bullet$, the smaller the estimation error of the smoothed Beran with comparison to the rest of the estimators. These values are, in most cases, less than one. This implies that, in general terms, the smoothed Beran's estimator is the one that provides the lowest estimation error. The smoothed WNW estimator is competitive with the smoothed Beran's estimator in some of the analysed scenarios.

	$P(\delta=0\|x)=0.2$			$P(\delta=0\|x)=0.5$			$P(\delta=0\|x)=0.8$		
	SBeran	SWNW	SVKA	SBeran	SWNW	SVKA	SBeran	SWNW	SVKA
h_2	0.14438	1.00000	0.14286	0.15233	1.00000	0.15000	0.18917	1.00000	0.22143
g_2	0.11510	0.09878	0.06327	0.15228	0.19347	0.13429	0.21839	0.27755	0.19939
$RMISE$	0.03687	0.04912	0.09343	0.05498	0.05740	0.11580	0.07647	0.05536	0.10859
$R_S^\bullet$	0.58422	0.74155	0.94488	0.51741	0.37920	0.85486	0.36545	0.20125	0.52431
$R_c^\bullet$	1.00000	0.75061	0.39463	1.00000	0.95784	0.47478	1.00000	1.38132	0.70421

Table 5.6: Optimal bandwidths, $RMISE$, $R_S^\bullet$ and $R_c^\bullet$ of the PD estimation for smoothed Beran's estimator, the smoothed WNW and the smoothed VKA for each level of conditional censoring probability in Model 1.

	$P(\delta=0\|x)=0.2$			$P(\delta=0\|x)=0.5$			$P(\delta=0\|x)=0.8$		
	SBeran	SWNW	SVKA	SBeran	SWNW	SVKA	SBeran	SWNW	SVKA
h_2	0.21687	0.38776	0.25918	0.15559	0.90102	0.22857	0.18962	1.00000	0.23469
g_2	0.09320	0.14020	0.06327	0.13651	0.20531	0.11653	0.19811	0.28367	0.19347
$RMISE$	0.03846	0.03513	0.06418	0.05946	0.03260	0.09957	0.06198	0.04705	0.09816
$R_S^\bullet$	0.70738	0.50036	0.88744	0.53113	0.19457	0.76112	0.24083	0.14115	0.38976
$R_c^\bullet$	1.00000	1.09479	0.59925	1.00000	1.82393	0.59717	1.00000	1.31732	0.63142

Table 5.7: Optimal bandwidths, $RMISE$, $R_S^\bullet$ and $R_c^\bullet$ of the PD estimation for smoothed Beran's estimator, the smoothed WNW and the smoothed VKA for each level of conditional censoring probability in Model 2.

	$P(\delta=0\|x)=0.2$			$P(\delta=0\|x)=0.5$			$P(\delta=0\|x)=0.8$		
	SBeran	SWNW	SVKA	SBeran	SWNW	SVKA	SBeran	SWNW	SVKA
h_2	0.10722	0.09143	0.04567	0.26967	0.10694	0.05380	1.00000	0.11857	0.12837
g_2	1.20340	1.55102	1.44286	1.61882	1.77551	1.45714	1.89462	1.92857	1.52857
$RMISE$	0.09208	0.12628	0.49730	0.12337	0.13406	0.37621	0.13431	0.13375	0.11410
$R_S^\bullet$	0.33944	0.33177	1.63226	0.24767	0.19828	0.88273	0.19751	0.16480	0.16868
$R_c^\bullet$	1.00000	0.72917	0.18516	1.00000	0.92026	0.32793	1.00000	1.00419	1.17713

Table 5.8: Optimal bandwidths, $RMISE$, $R_S^\bullet$ and $R_c^\bullet$ of the PD estimation for smoothed Beran's estimator, the smoothed WNW and the smoothed VKA for each level of conditional censoring probability in Model 3.

Analysing the differences between the computacional times os these techniques is also useful. Table 5.9 shows the CPU time (in seconds) that is needed by each estimator to obtain the estimated probability of default curve in a time grid of size 100 and a fixed value of x for different values of the sample size.

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.

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

Table 5.9: CPU time (in seconds) for estimating $PD(t|x)$ in a time grid of size 100 for every estimator and different sample sizes.

5.5 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 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 were previously used in Section 2.5, where an explanatory analysis was performed. The data set provides the credit scoring computed for each borrower, the observed lifetime of the credit in months and the uncensoring indicator.

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 5.5 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 5.6. Although the survival estimations are very similar with both estimators, it can be seen how the roughness of the curve estimation is reduced and the jumps are removed when using the smoothed Beran's estimator. This is even more remarkable when estimating the probability of default.

According to the smoothed Beran's estimation, the probability of default has an increasing tendency. It follows from it that the higher the debt maturity, the higher the probability of falling into default for an individual with this credit scoring.

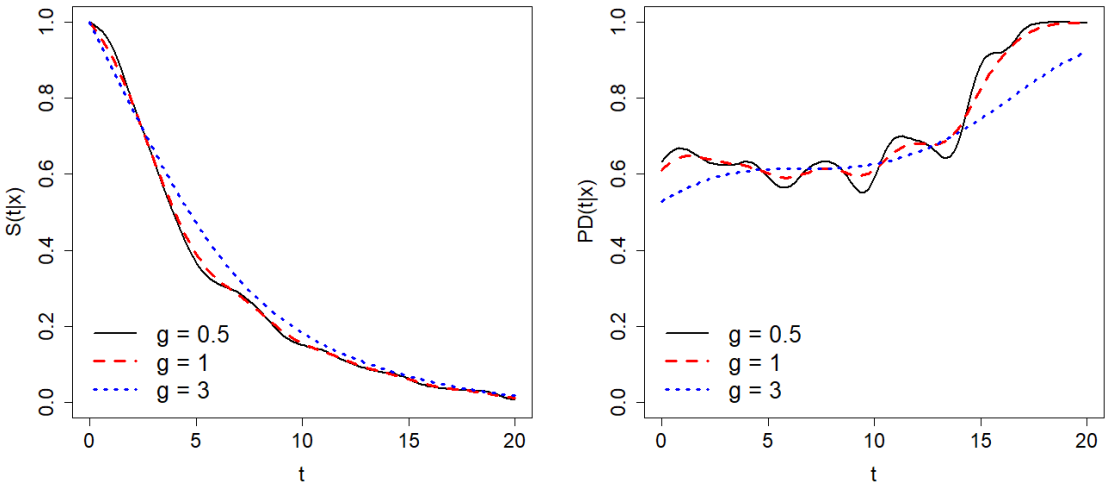

Figure 5.5: Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.5$ by means of the smoothed Beran's estimator on the consumer credits dataset for $h = 0.05$ and $g = 0.5$ (solid line), $g = 1$ (dashed line) and $g = 3$ (dotted line).

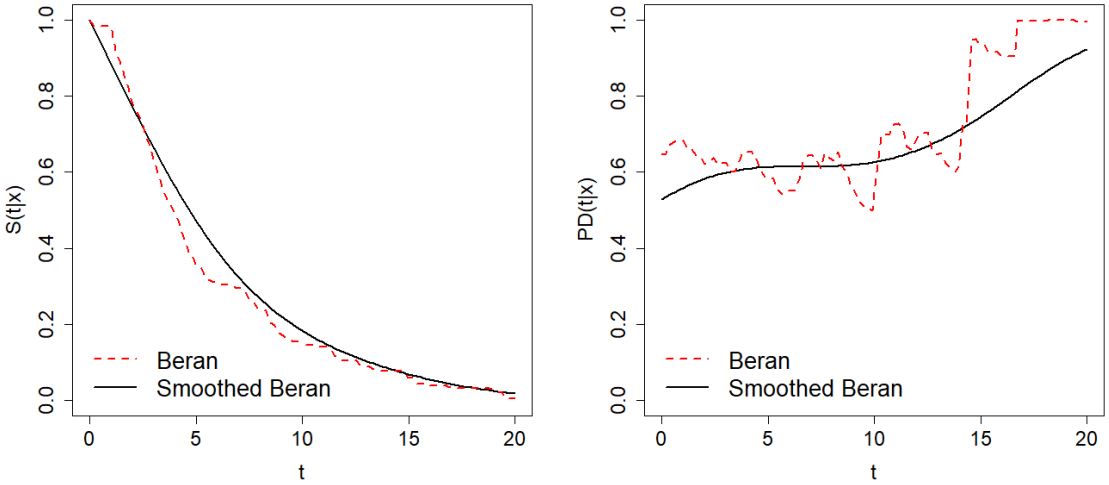

Figure 5.6: Estimation of $S(t|x)$ (left) and $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.5$ by means of Beran's estimator (dashed line) and smoothed Beran's estimator (solid line) using the bandwidths $h = 0.05$ and $g = 3$ on the consumer credits dataset.

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 5.7 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 $g = 3$.

Note how the PD estimations are closer to each other and closer to zero for all time points as the credit scoring value increases. The increasing spike of the estimation in the nondefaulted (censored) group in the right tail of the time distribution is probably due to the lack of information in that region.

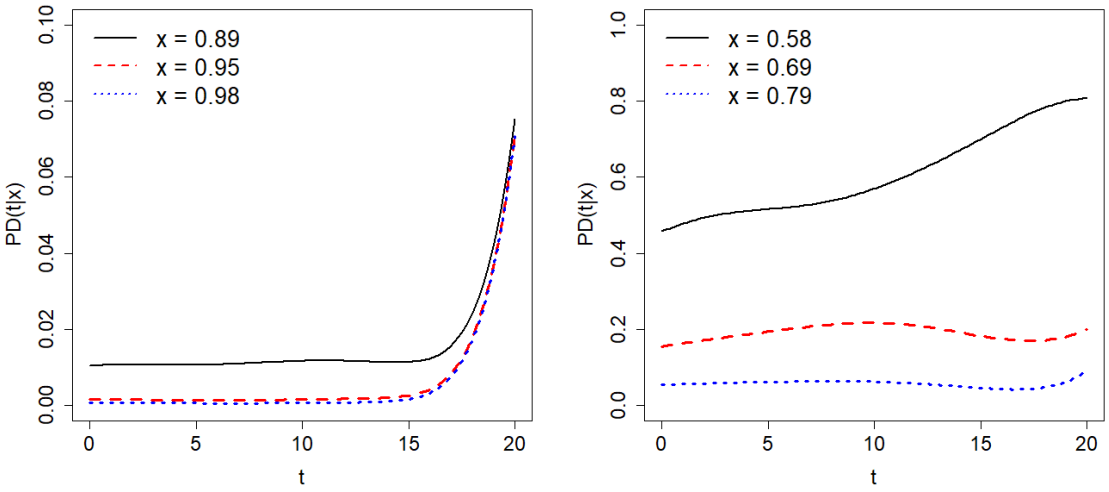

Figure 5.7: Smoothed Beran's estimation of $PD(t|x)$ at horizon $b = 5$, for large (left) and small (right) values of the score x, using bandwidths $h = 0.05$ and $g = 3$. The large values chosen are the three sample quartiles of the score for nondefaulted credits, while the small values are the three sample quartiles of the score for the defaulted credits.

In a real practical problem the true default probability curve is unknown. Therefore, unlike in simulations, in real data analysis it is not possible to evaluate the behaviour of the proposed estimator. Alternatively, it is possible to assess whether the PD curve obtained is reasonable and fits the real credit risk scenario. The results obtained by the doubly smoothed Beran's estimator seem to be more appropriate, since the roughness of the Beran's estimaton is not expected in this type of curve. Supporting the conclusion of our real data analysis, the estimation of the PD over time for the assessment of risk in portfolios and bond rating obtained in Barnard (2017) and dos Reis and Smith (2018) have shapes similar to those obtained here.

5.6 Proofs

Proof of Theorem 5.1.

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 2.1:

$$ABias\Big(\widetilde{PD}^B_{h,g}(t|x)\Big) = \alpha_1 + \alpha_2 + \alpha_3, \tag{5.3}$$

$$AVar\Big(\widetilde{PD}^B_{h,g}(t|x)\Big) = \beta_1 + \beta_2 + \beta_3, \tag{5.4}$$

where

$$\alpha_1 = \frac{P}{Q} - \frac{E(\widetilde{P})}{E(\widetilde{Q})}, \quad \alpha_2 = \frac{Cov(\widetilde{P},\widetilde{Q})}{E(\widetilde{Q})^2}, \quad \alpha_3 = -\frac{E\Big[\frac{\widetilde{P}}{\widetilde{Q}}\Big(\widetilde{Q} - E(\widetilde{Q})\Big)^2\Big]}{E(\widetilde{Q})^2} \tag{5.5}$$

and

$$\beta_1 = \frac{Var(\widetilde{P})}{E(\widetilde{Q})^2}, \quad \beta_2 = -2\frac{E(\widetilde{P})Cov(\widetilde{P},\widetilde{Q})}{E(\widetilde{Q})^3}, \quad \beta_3 = \frac{E(\widetilde{P})^2 Var(\widetilde{Q})}{E(\widetilde{Q})^4}. \tag{5.6}$$

The asymptotic expressions for the bias and the variance of the survival estimator $\widetilde{S}^B_{h,g}(t|x)$ are obtained from Theorem 3.3:

$$\text{Bias}\Big(\widetilde{S}^B_{h,g}(t|x)\Big) \;=\; b_1(t,x)h^2 + b_2(t,x)g^2 + o(h^2), \tag{5.7}$$

$$\text{Var}\Big(\widetilde{S}^B_{h,g}(t|x)\Big) \;=\; V_1(t,x)\frac{1}{nh} + V_2(t,x)\frac{g}{nh} + O\left(\frac{h^2+g^2}{nh}\right) \tag{5.8}$$

From Lemma 3.4, the covariance of the survival estimator $\widetilde{S}^B_{h,g}(t|x)$ is obtained:

$$\begin{aligned}\text{Cov}\Big(\widetilde{S}^B_{h,g}(t_1|x), \widetilde{S}^B_{h,g}(t_2|x)\Big) \;=\; & \frac{c_K}{m(x)}V_g^1(t_1,t_2,x)\frac{1}{nh} + \frac{c_K}{m(x)}V_g^2(t_1,t_2,x)\frac{g}{nh} \\ & + O\left(\frac{h^2+g^2}{nh}\right),\end{aligned}$$

where the functions $V_g^1(t_1,t_2,x)$ and $V_g^2(t_1,t_2,x)$ were defined in Lemma 3.4.

Considering $t_1 = t$, $t_2 = t+b$,

$$V_g^1(t,t+b,x) \;=\; 2J(t|x)\Big(1 - F(t+b|x)\Big)\mathbb{K} * K\left(\frac{b}{g}\right).$$

Given that $g = g_n \to 0$ when n tends to infinity, $b/g \to \infty$ and

$$\begin{aligned} \lim_{n\to\infty} \mathbb{K} * K\left(\frac{b}{g}\right) &= \lim_{u\to\infty} \mathbb{K} * K(u) = \lim_{u\to\infty} \int_{-\infty}^{+\infty} K(y)\mathbb{K}(u-y)dy \\ &= \int_{-\infty}^{+\infty} \lim_{u\to\infty} \mathbb{K}(u-y)K(y)dy = \int_{-\infty}^{+\infty} K(y)dy = 1. \end{aligned} \tag{5.9}$$

Then,

$$V_g^1(t, t+b, x) \xrightarrow[n\to\infty]{} 2\Big(1 - F(t|x)\Big)\Big(1 - F(t+b|x)\Big)L(t|x).$$

On the other hand,

$$\begin{aligned} V_g^2(t, t+b, x) &= 2J(t|x)f(t+b|x)\mathbb{K} * K_1\bigg(-\frac{b}{g}\bigg) \\ &\quad +2J'(t|x)\Big(1 - F(t+b|x)\Big)\mathbb{K} * K_1\left(\frac{b}{g}\right). \end{aligned}$$

Since

$$\begin{aligned} \lim_{n\to\infty} \mathbb{K} * K_1\bigg(-\frac{b}{g}\bigg) &= \lim_{u\to-\infty} \mathbb{K} * K(u) = \lim_{u\to-\infty} \int_{-\infty}^{+\infty} K_1(y)\mathbb{K}(u-y)dy \\ &= \int_{-\infty}^{+\infty} \lim_{u\to-\infty} \mathbb{K}(u-y)K_1(y)dy = 0. \end{aligned}$$

and

$$\begin{aligned} \lim_{n\to\infty} \mathbb{K} * K_1\left(\frac{b}{g}\right) &= \lim_{u\to\infty} \mathbb{K} * K_1(u) = \lim_{u\to\infty} \int_{-\infty}^{+\infty} K_1(y)\mathbb{K}(u-y)dy \\ &= \int_{-\infty}^{+\infty} \lim_{u\to\infty} \mathbb{K}(u-y)K_1(y)dy = \int_{-\infty}^{+\infty} K_1(y)dy \\ &= \textstyle\int_{-\infty}^{+\infty} yK(y)dy = 0, \end{aligned}$$

we have

$$V_g^2(t, t+b, x) \xrightarrow[n\to\infty]{} 0.$$

Therefore,

$$\mathrm{Cov}\Big(\widetilde{S}_{h,g}^B(t|x), \widetilde{S}_{h,g}^B(t+b|x)\Big) = C(t, t+b, x)\frac{1}{nh} + O\left(\frac{h^2+g^2}{nh}\right), \tag{5.10}$$

where $C(t_1, t_2, x)$ is defined in Section 5.3.

Considering Equations (5.5), (5.8) and (5.10), detailed expressions for α_1, α_2 and α_3 are obtained as follows:

$$\begin{aligned}
\alpha_1 &= \frac{P}{Q} - \frac{P + b_1(t+b,x)h^2 + b_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q + b_1(t,x)h^2 + b_2(t,x)g^2 + o(h^2) + o(g^2)} \\
&= \frac{PQ + Pb_1(t,x)h^2 + Pb_2(t,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + b_1(t,x)h^2 + b_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} + \\
&\quad \frac{-PQ - Qb_1(t+b,x)h^2 - Qb_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + b_1(t,x)h^2 + b_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} \\
&= \frac{Pb_1(t,x)h^2 - Qb_1(t+b,x)h^2 + o(h^2) + o(g^2)}{Q\Big(Q + b_1(t,x)h^2 + b_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} + \\
&\quad \frac{Pb_2(t,x)g^2 - Qb_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + b_1(t,x)h^2 + b_2(t,x)g^2 + o(h^2) + o(g^2)\Big)}.
\end{aligned}$$

Then,

$$\begin{aligned}
\alpha_1 &= \frac{\Big(1 - PD(t|x)\Big)b_1(t,x) - b_1(t+b,x)}{S(t|x)}h^2 \\
&\quad + \frac{\Big(1 - PD(t|x)\Big)b_2(t,x) - b_2(t+b,x)}{S(t|x)}g^2 + o(h^2) + o(g^2),
\end{aligned} \tag{5.11}$$

$$\alpha_2 = \frac{C(t,t+b,x)}{S(t|x)^2}\frac{1}{nh} + O\left(\frac{h^2+g^2}{nh}\right), \tag{5.12}$$

$$\begin{aligned}
\alpha_3 &= \frac{E\Big[\frac{\widetilde{P}}{\widetilde{Q}}\big(\widetilde{Q} - E(\widetilde{Q})\big)^2\Big]}{E(\widetilde{Q})^2} \leq \frac{\mathrm{Var}(\widetilde{Q})}{E(\widetilde{Q})^2} \\
&= \frac{V_1(t,x)}{S(t|x)^2}\frac{1}{nh} + \frac{V_2(t,x)}{S(t|x)^2}\frac{g}{nh} + O\left(\frac{h^2+g^2}{nh}\right).
\end{aligned} \tag{5.13}$$

By plugging (5.11), (5.12) and (5.13) into (5.3), the bias part in Theorem 5.1 is proved.

Now, expressions (5.6), (5.8) and (5.10) lead to

$$\beta_1 = \frac{V_1(t+b,x)}{S(t|x)^2}\frac{1}{nh} + \frac{V_2(t+b,x)}{S(t|x)^2}\frac{g}{nh} + O\left(\frac{h^2+g^2}{nh}\right), \tag{5.14}$$

$$\beta_2 = -2\frac{S(t+b,x)}{S(t|x)^3}C(t,t+b,x)\frac{1}{nh} + O\left(\frac{h^2+g^2}{nh}\right), \tag{5.15}$$

$$\beta_3 = \frac{S(t+b,x)^2 V_1(t,x)}{S(t|x)^4}\frac{1}{nh} + \frac{S(t+b,x)^2 V_2(t,x)}{S(t|x)^4}\frac{g}{nh} + O\left(\frac{h^2+g^2}{nh}\right). \quad (5.16)$$

Plugging Equations (5.14), (5.15) and (5.16) in (5.4) the variance part in Theorem 5.1 is proved.

□

Proof of Theorem 5.2.

From Equations (2.1) and (5.1) follows:

$$\frac{\widetilde{S}^B_{h,g}(t+b|x)}{\widetilde{S}^B_{h,g}(t|x)} - \frac{S(t+b|x)}{S(t|x)} = -\Big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\Big). \quad (5.17)$$

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

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

it holds

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

and considering (5.17):

$$\begin{aligned}PD(t|x) - \widetilde{PD}^B_{h,g}(t|x) &= a_1\Big(\widetilde{S}^B_{h,g}(t+b|x) - S(t+b|x)\Big) + a_2\Big(\widetilde{S}^B_{h,g}(t|x) - S(t|x)\Big) \\ &\quad + C\Big(\widetilde{S}^B_{h,g}(t|x)\Big)\left(1 - \frac{\widetilde{S}^B_{h,g}(t|x)}{S(t|x)}\right). \quad (5.18)\end{aligned}$$

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 3.1, $\widetilde{S}^B_{h,g}(t|x)$ can be written as a sum of the following terms

$$\widetilde{S}^B_{h,g}(t|x) \;=\; S(t|x) + \sum_{i=1}^{n} \varphi_{n,i}(t,x) + b_2(t,x)g^2 + R_n(t|x), \tag{5.19}$$

where $\varphi_{n,i}(t,x) = \dfrac{1}{nh}\dfrac{1}{m(x)} K\Big((x-X_i)/h\Big)\eta(Z_i,\delta_i,t,x)$ 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 (5.19),

$$\begin{aligned} & a_1\sqrt{nh}\Big(\widetilde{S}^B_{h,g}(t+b|x) - S(t+b|x)\Big) + a_2\sqrt{nh}\Big(\widetilde{S}^B_{h,g}(t|x) - S(t|x)\Big) \\ & \quad = \sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x) + a_1 b_2(t+b,x)g^2\sqrt{nh} + a_2 b_2(t,x)g^2\sqrt{nh} + \widetilde{R}_n(t,x), \end{aligned} \tag{5.20}$$

where

$$\widetilde{\varphi}_{n,i}(t,x) \;=\; \sqrt{nh}\Big(a_1\varphi_{n,i}(t+b,x) + a_2\varphi_{n,i}(t,x)\Big) \tag{5.21}$$

and

$$\begin{aligned} \widetilde{R}_n(t,x) \;&=\; \sqrt{nh}\Big(a_1 R_n(t+b,x) + a_2 R_n(t,x)\Big) \\ &=\; \sqrt{nh}(a_1+a_2)O_p\left(\frac{\ln n}{nh}\right)^{3/4} + \sqrt{nh}(a_1+a_2)o(g^2) \\ &\quad + O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right)\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x). \end{aligned} \tag{5.22}$$

Since $h \to 0$ and $nh \to \infty$, the term $O_p\left(h^2 + \dfrac{1}{\sqrt{nh}}\right)\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x)$ in (5.22) is negligible with respect to $\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x)$ in (5.20). Given that $g \to 0$, the term $\sqrt{nh}(a_1+a_2)o(g^2)$ in (5.22) is negligible with respect to $a_1 b_2(t+b,x)g^2\sqrt{nh} + a_2 b_2(t,x)g^2\sqrt{nh}$ in (5.20). Finally, the term $\sqrt{nh}(a_1+a_2)O_p\left(\dfrac{\ln n}{nh}\right)^{3/4}$ in (5.22) is negligible with respect to $\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x)$ in (5.20) because $\dfrac{nh}{(\ln n)^3} = \dfrac{C_h n^{4/5}}{(\ln n)^3} \to \infty$.

The variance of the dominant term in (5.20) is $O(1)$:

$$\begin{aligned}\mathrm{Var}\Big(\textstyle\sum_{i=1}^n \widetilde{\varphi}_{n,i}(t,x)\Big) &= n\mathrm{Var}\Big(\widetilde{\varphi}_{n,1}(t,x)\Big)\\ &= n^2h\Big(a_1^2\mathrm{Var}\Big(\varphi_{n,1}(t+b,x)\Big)+a_2^2\mathrm{Var}\Big(\varphi_{n,1}(t,x)\Big) \qquad (5.23)\\ &\quad +2a_1a_2\mathrm{Cov}\Big(\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\Big)\Big).\end{aligned}$$

From the proof of Lemma 3.4,

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

In particular, for $t_1=t$, $t_2=t+b$, $\mathbb{K}*K\left(\dfrac{t_2-t_1}{g}\right)=\mathbb{K}*K\left(\dfrac{b}{g}\right)$ and from Equation (5.9), $\lim_{n\to\infty}\mathbb{K}*K\left(\dfrac{b}{g}\right)=1$. Consequently,

$$\begin{aligned}&\mathrm{Cov}\Big(\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\Big)\\ &= \frac{2c_K}{m(x)n^2}\Big(1-F(t|x)\Big)\Big(1-F(t+b|x)\Big)L(t|x)\frac{1}{h}+O\left(\frac{g}{n^3h}\right)+o\left(\frac{1}{n^2h}\right). \qquad (5.24)\end{aligned}$$

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,

$$\mathrm{Var}\Big(\varphi_{n,1}(t,x)\Big) = \frac{c_K}{m(x)n^2}\Big(1-F(t|x)\Big)^2\Big)L(t|x)\frac{1}{h}+O\left(\frac{g}{n^3h}\right). \qquad (5.25)$$

Replacing (5.24) and (5.25) in (5.23),

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

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 (5.20) 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

$$\sum_{i=1}^n \left(\widetilde{\varphi}_{n,i}(t,x) - E\Big[\widetilde{\varphi}_{n,i}(t,x)\Big]\right) \xrightarrow{d} N(0,s_0), \tag{5.26}$$

where

$$\begin{aligned}
s_0^2 \;=\; & a_1^2\frac{c_K}{m(x)}\Big(1-F(t+b|x)\Big)^2 L(t+b|x) + a_2^2\frac{c_K}{m(x)}\Big(1-F(t|x)\Big)^2 L(t|x) \\
& +4a_1a_2\frac{c_K}{m(x)}\Big(1-F(t|x)\Big)\Big(1-F(t+b|x)\Big)L(t|x).
\end{aligned} \tag{5.27}$$

We will now check Lindeberg's condition:

$$\lim_{n\to\infty}\frac{1}{s_0^2}E\left[\sum_{i=1}^n \left(\widetilde{\varphi}_{n,i}(t,x) - E\Big[\widetilde{\varphi}_{n,i}(t,x)\Big]\right)^2 \mathbb{1}_{n,i}\right] = 0 \tag{5.28}$$

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

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

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

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

and, consequently, η is also bounded:

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

Since η is bounded, K and $m(x)$ have compact support and $nh\to\infty$, the sequence $\Big\{\widetilde{\varphi}_{n,i}(t,x), i=1,...,n,\ n\in\mathbb{N}\Big\}$ is a random variables sequence bounded by a convergent to zero sequence. Hence, there exists $n_0\in\mathbb{N}$ such that for all $i=1,...,n$, $\mathbb{1}_{n,i}=0$ for all $n\geq n_0$ and 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 (5.28).

Furthermore, from the proof of Theorem 3.3 in Section 3.3,

$$E\Big(\varphi_{n,1}(t,x)\Big)=b_1(t,x)\frac{h^2}{n}+o\left(\frac{h^2}{n}\right),$$

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

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

where

$$\mu_1=\mu_0+C_h^{1/2}C_g^{4/2}\Big(a_1b_2(t+b,x)+a_2b_2(t,x)\Big).$$

Finally, using equation (5.18) 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

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

□

Chapter 6

Bootstrap bandwidth selection for the smoothed Beran's PD estimator

6.1 Introduction

In Chapters 2 and 5, Beran's estimator and the smoothed Beran's estimator of the probability of default were presented and the asymptotic properties and the performance of the estimators have been deeply studied in these previous pages. The results of the simulation studies carried out are promising, especially for the smoothed Beran's estimator. However, these analyses are based on the smoothing parameters that minimised the mean integrated squared error, MISE, obtained from the theoretical curve of PD and the asymptotic bias and variance expressions do not make it easy to get plug-in estimations of these theoretical bandwidths. The goal of this chapter is to propose resampling techniques to approximate them. Our approach follows the ideas of Li and Datta (2001), and it is based on the obvious bootstrap. Both Beran and smoothed Beran's estimators are bootstrapped in order to approximate their corresponding optimal bandwidths. The bootstrap is also useful to compute confidence regions.

A simulation study shows the behaviour of the PD estimators with bootstrap bandwidths. The issue of obtaining confidence regions based on Beran's and the smoothed Beran's estimator for the probability of default, $PD(t|x)$, for a fixed value of $x \in I \subseteq \mathbb{R}$ and t covering the interval $I_T \subseteq \mathbb{R}^+$ is addressed. Finally, these estimators with bootstrap bandwidths are used to analyse the probability of default function conditional on the credit scoring for the German credit dataset.

The content of this chapter is published in Peláez et al. (2022a).

6.2 Bandwidth selection for Beran's and the smoothed Beran's PD estimators

In this section, methods for the automatic selection of the bandwidths for Beran's estimator in (2.4) and the smoothed Beran's estimator in (5.2) of the probability of default are proposed. Consider the right censored random sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$ of (X, Z, δ)

6.2.1 Beran's estimator

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 is no ties in the sample values of the covariate. In this chapter, the following obvious bootstrap method combined with a smoothed bootstrap for the covariate is proposed for the automatic selection of the covariate bandwidth h of Beran's estimator, $\widehat{PD}_h^B(t|x)$, defined in (2.4). Here, this estimator is simply denoted by $\widehat{PD}_h(t|x)$.

Our goal is to estimate the probability of default function, $PD(t|x)$, for a fixed $x \in I$ and t covering the interval $I_T \subset \mathbb{R}$. Therefore, our goal is to get the bandwidth

$h_{MISE} \in I_1$ that minimises the mean integrated squared error given by

$$MISE_x(h) = E\left(\int_{I_T} \left(\widehat{PD}_h(t|x) - PD(t|x) \right)^2 dt \right) \tag{6.1}$$

whose bootstrap approximation is

$$MISE_x^*(h) = E^*\left(\int_{I_T} \left(\widehat{PD}_h^*(t|x) - \widehat{PD}_r(t|x) \right)^2 dt \right)$$

where $\widehat{PD}_r(t|x)$ is the estimator of the theoretical PD with pilot bandwidth r, using the sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$ and $\widehat{PD}_h^*(t|x)$ is the bootstrap estimator of PD with bandwidth h, using the bootstrap resample $\left\{(X_i^*, Z_i^*, \delta_i^*)\right\}_{i=1}^n$.

The resampling distribution of $\widehat{PD}_h^*(t|x)$ cannot be computed in a close form, so the Monte Carlo method is used. It is based on obtaining B bootstrap resamples and estimating $\widehat{PD}_h^*(t|x)$ for each of them. Thus, the distribution of $\widehat{PD}_h^*(t|x)$ is approximated by the empirical one of $\widehat{PD}_h^{*,1}(t|x), \ldots, \widehat{PD}_h^{*,B}(t|x)$, obtained from B bootstrap resamples and the bootstrap version of the estimation error made by Beran's estimator for any smoothing parameter h is given by

$$MISE_x^*(h) \simeq \frac{1}{B} \sum_{k=1}^{B} \left(\int_{I_T} \left(\widehat{PD}_h^{*,k}(t|x) - \widehat{PD}_r(t|x) \right)^2 dt \right). \tag{6.2}$$

Likewise, the integral is approximated by a Riemann sum.

Algorithm for bootstrap bandwidth selector for Beran's estimator

Let $x \in I$ be a fixed value of the covariate, $t \in I_T$ and $r \in I_1$:

1. Compute $\widehat{PD}_r(t|x)$ from the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$.

2. Obtain B bootstrap resamples of the form $\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\}_{i=1}^n$ with $k = 1, ..., B$ using the bootstrap technique based on Beran's estimator proposed in Subsection 4.2.1 with pilot bandwidth $r \in I_1$ and compute $\widehat{PD}_h^{*,k}(t|x)$ for each of them.

3. Approximate $MISE_x^*(h)$ according to (6.2).

4. Repeat Steps 1–3 for values of h in a grid of I_1.

5. Select the value of h that provides the smallest $MISE_x^*(h)$ as the bootstrap bandwidth h^*.

Concerning the auxiliary bandwidth $r \in I_1$, a preliminary analysis not shown here suggests that a good choice for it is

$$r = \frac{3}{4}\Big(Q_X(0.975) - Q_X(0.025)\Big)\left(\sum_{i=1}^{n}\delta_i\right)^{-1/3}, \tag{6.3}$$

where $Q_X(u)$ is the u quantile of the sample $\left\{X_i\right\}_{i=1}^n$. Equation (6.3) takes into account the variability of the covariate, $Q_X(0.975) - Q_X(0.025)$, and the number of uncensored data, $\sum_{i=1}^n \delta_i$. The exponent of this sample size, $-1/3$, is heuristically deduced from the asymptotic expression of the MISE of the PD estimators (see Chapter 2). It is typically the appropriate exponent in selection of the optimal bandwidth for estimating the distribution function (Azzalini (1981), Jones (1990)). This expression was derived after several attempts in the simulation studies.

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 $MISE_x^*(h)$ by $MSE_{t,x}^*(h)$, which is the bootstrap analogue of

$$MSE_{t,x}(h) = E\left(\left(\widehat{PD}_h(t|x) - PD(t|x)\right)^2\right).$$

6.2.2 The smoothed Beran's estimator

Given the good performance that the doubly smoothed PD estimator based on Beran's estimator showed in previous simulation studies, it is interesting to propose a method for automatic selection of the two-dimensional bandwidth on which it depends. Then, consider the smoothed Beran's estimator of the probability of default, $\widetilde{PD}_{h,g}^B(t|x)$, defined in (5.2). For simplicity of notation, the smoothed Beran's estimator of the PD is denoted by $\widetilde{PD}_{h,g}(t|x)$ in this chapter. A bootstrap method is proposed for the automatic selection of the bivariate bandwidth (h, g).

The optimal bivariate bandwidth, $(h_{MISE}, g_{MISE}) \in I_1 \times I_2$ is defined as the pair of bandwidths that minimises the mean integrated squared error given by

$$MISE_x(h,g) = E\left(\int_{I_T}\left(\widetilde{PD}_{h,g}(t|x) - PD(t|x)\right)^2 dt\right). \tag{6.4}$$

The bootstrap version of $MISE_x(h,g)$ is given by

$$MISE_x^*(h,g) = E^*\left(\int_{I_T}\left(\widetilde{PD}_{h,g}^*(t|x) - \widetilde{PD}_{r,s}(t|x)\right)^2 dt\right),$$

where $\widetilde{PD}_{r,s}(t|x)$ is the smoothed Beran's PD estimation 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 $\widetilde{PD}_{h,g}^*(t|x)$ is the bootstrap estimation of PD with bandwidths (h,g), using the bootstrap resample $\left\{(X_i^*, Z_i^*, \delta_i^*)\right\}_{i=1}^n$. Since the sampling distribution of $\widetilde{PD}_{h,g}^*(t|x)$ is unknown, the Monte Carlo method gives the following approximation

$$MISE_x^*(h,g) \simeq \frac{1}{B}\sum_{k=1}^{B}\left(\int_{I_T}\left(\widetilde{PD}_{h,g}^{*,k}(t|x) - \widetilde{PD}_{r,s}(t|x)\right)^2 dt\right), \tag{6.5}$$

based on the empirical distribution of $\widetilde{PD}_{h,g}^*(t|x)$ obtained from B bootstrap resamples. The integral is approximated by a Riemann sum.

Algorithm for bootstrap bandwidth selector for the smoothed Beran's estimator

Let x be a fixed value of the covariate, $t \in I_T$ and $(r,s) \in I_1 \times I_2$:

1. Compute $\widetilde{PD}_{r,s}(t|x)$ from the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$.
2. Obtain B bootstrap resamples of the form $\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\}_{i=1}^n$ with $k = 1, ..., B$ using the bootstrap technique based on the smoothed Beran's estimator proposed in Subsection 4.2.2 and compute $\widetilde{PD}_{h,g}^{*,k}(t|x)$ for each of them.
3. Approximate $MISE_x^*(h)$ according to (6.5).
4. Repeat Steps 1–3 for pairs of values (h,g) in a grid of $I_1 \times I_2$.

5. Obtain the pair (h, g) that provides the smallest $MISE_x^*(h, g)$ as the bootstrap bivariate bandwidth (h^*, g^*).

The auxiliary bandwidth $r \in I_1$ was defined in (6.3). The pilot bandwidth $s \in I_2$ for the time variable smoothing is chosen as

$$s = \frac{3}{4}\Big(Q_Z(0.975) - Q_Z(0.025)\Big)\left(\sum_{i=1}^{n} \delta_i\right)^{-1/7}, \tag{6.6}$$

where $Q_Z(u)$ is the u quantile of the sample $\left\{Z_i\right\}_{i=1}^{n}$. This expression considers the variability of the observed time variable, $Q_Z(0.975) - Q_Z(0.025)$, and the sample size of the uncensored population, $\sum_{i=1}^{n} \delta_i$.

6.3 Simulation study for bandwidth selection

A simulation study was conducted in order to show the behaviour of bootstrap bandwidth selectors for Beran's and smoothed Beran's estimators proposed in Section 6.2.

The simulation setup is similar to the one introduced in Section 2.4. Due to the computational cost of the resampling methods, only Models 2 and 3 in Section 2.4 and the low and medium censoring scenarios ($P(\delta = 0|x) = 0.2$ and $P(\delta = 0|x) = 0.5$) will be considered in this chapter. Model 2 considers a uniform distribution for the credit scoring and Weibull life and censoring times. The probability of default for this model is estimated at $x = 0.6$ in a time grid over the interval $I_T = (0, 0.8654)$. Model 3 also considers a uniform distribution for the credit scoring and exponential distributions for life time and censoring time. The probability of default for this model is estimated at $x = 0.8$ in a time grid over the interval $I_T = (0, 3.8211)$. The limited-memory algorithm for solving large nonlinear optimization problems presented in Section 3.4 is used to minimise the MISE error function.

6.3.1 Simulation study for Beran's estimator

In this subsection, the performance of the bootstrap bandwidth selector for Beran's estimator is analysed. For each model, the estimation error function $MISE_x(h)$ is approximated via Monte Carlo using 300 simulated samples. The bandwidth that minimises $MISE_x(h)$ is obtained and denoted by h_{MISE}. The values of h_{MISE} and $MISE_x(h_{MISE})$ are used as a benchmark.

In the simulation study, $N = 300$ simulated samples are used. For each sample, $B = 500$ bootstrap resamples are obtained to approximate the bootstrap MISE function, $MISE_x^*(h)$, and obtain the bootstrap bandwidth associated to each simulated sample h_j^*, $j = 1, 2, \ldots, N$. The mean value of the N bootstrap bandwidths and the standard deviation are defined as follows

$$\overline{h^*} = \frac{1}{N}\sum_{j=1}^{N} h_j^*, \quad sd\big(h^*\big) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\Big(h_j^* - \overline{h^*}\Big)^2}.$$

As a relative measure of the difference between the bootstrap bandwidth and the optimal one, we compute

$$H_j^* = \frac{h_j^* - h_{MISE}}{h_{MISE}},$$

with $j = 1, \ldots, N$. The mean of the absolute value of these relative deviations, $\overline{H^*} = \frac{1}{N}\sum_{j=1}^{N}|H_j^*|$, is a good measure of how close the bootstrap bandwidth is to the optimal one.

For each sample, the estimation error of Beran's estimator with the corresponding bootstrap bandwidth,

$$MISE_x(h_j^*) = E\bigg(\int_{I_T}\Big(\widehat{PD}_{h_j^*}(t|x) - PD(t|x)\Big)^2 dt\bigg),$$

and its square root, $RMISE_x(h_j^*)$, are approximated via Monte Carlo using 300 simulated samples. The mean of these estimation errors given by

$$\overline{RMISE_x(h^*)} = \frac{1}{N}\sum_{j=1}^{N} RMISE_x(h_j^*)$$

is used as a measure of the estimation error made by the bootstrap bandwidth, when compared with the estimation error made by the MISE bandwidth.

As a relative measure of the difference between the estimation errors using the bootstrap and the MISE bandwidths, the following ratios are defined:

$$R_j^* = \frac{RMISE_x(h_j^*) - RMISE_x(h_{MISE})}{RMISE_x(h_{MISE})}$$

satisfying $R_j^* \geq 0$ for all $j = 1, \ldots, N$. The mean of the R_j^* values with $j = 1, \ldots, N$ is denoted by $\overline{R^*} = \frac{1}{N}\sum_{j=1}^{N} R_j^*$. Small values (close to zero) of $\overline{H^*}$ and $\overline{R^*}$ indicate good behaviour of the bootstrap bandwidth. Values of the bootstrap bandwidths, estimation errors and relative measures for Models 2 and 3 are included in Table 6.1. The results show a good performance of the proposed bootstrap selector.

	Model 2		Model 3	
$P(\delta = 0 \mid X = x)$	0.2	0.5	0.2	0.5
h_{MISE}	0.37576	0.35909	0.09494	0.10959
$RMISE_x(h_{MISE})$	0.05520	0.11144	0.27942	0.49991
$\overline{h^*}$ (sd)	0.27856 (0.092)	0.30892 (0.110)	0.21763 (0.041)	0.23091 (0.068)
$\overline{H^*}$	0.31431	0.29306	1.29211	1.10692
$\overline{RMISE_x(h^*)}$	0.05700	0.11405	0.29671	0.50824
$\overline{R^*}$	0.03260	0.02336	0.06188	0.01666

Table 6.1: RMISE, average bootstrap bandwidths and estimation errors of Beran's PD estimator in each level of conditional censoring probability for Models 2 and 3. Numbers within brackets are standard deviations.

Figures 6.1 and 6.2 show the function $MISE_x(h)$ along with the Monte Carlo approximations of $MISE_x^*(h)$ for some simulated samples and the boxplots of H_j^* and R_j^* with $j = 1, \ldots, N$ for Models 2 and 3. The method tends to slightly underestimate the value of h^* with respect to h_{MISE} in Model 2 and overestimate its value in Model 3, which is reflected in the boxplots of H_j^*. Nevertheless, these figures show that the $MISE_x(h)$ curve is fairly flat and variations in the selection of h do not imply an important increase in the estimation error.

In order to illustrate the results, Figure 6.3 shows the theoretical probability of default function and Beran's estimation with the MISE and bootstrap bandwidths drawn for one sample from Model 2 and 3 when the conditional probability of censoring is 0.5. For large values of time, the performance of the estimator becomes worse, due to the fact that in that region there are few data, most of them censored, and therefore offering poor information.

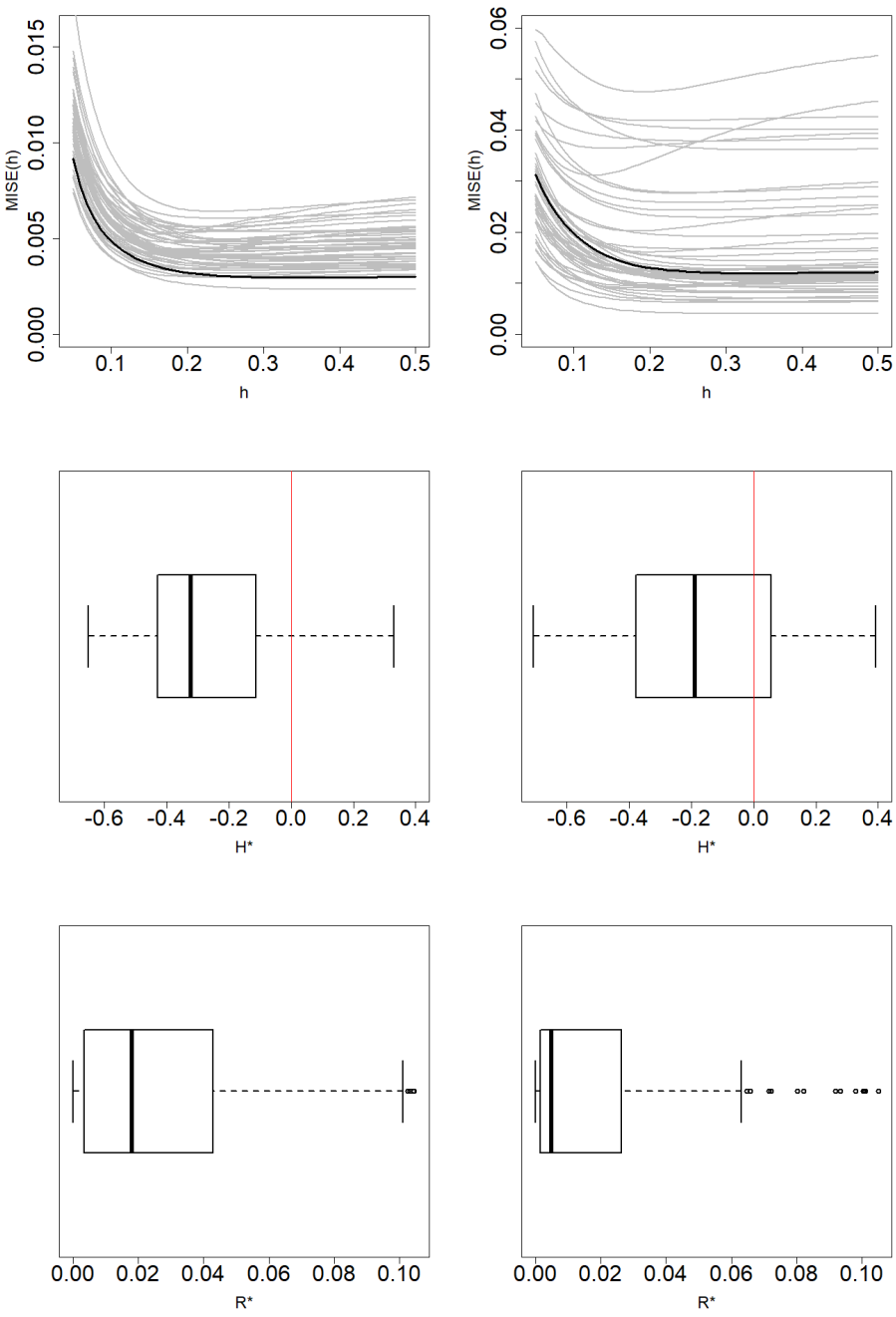

Figure 6.1: $MISE_x(h)$ function (black line) approximated via Monte Carlo and $MISE_x^*(h)$ functions (gray lines) for $N = 300$ samples (top), boxplot of $H_1^*, ..., H_N^*$ values (middle) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 2.

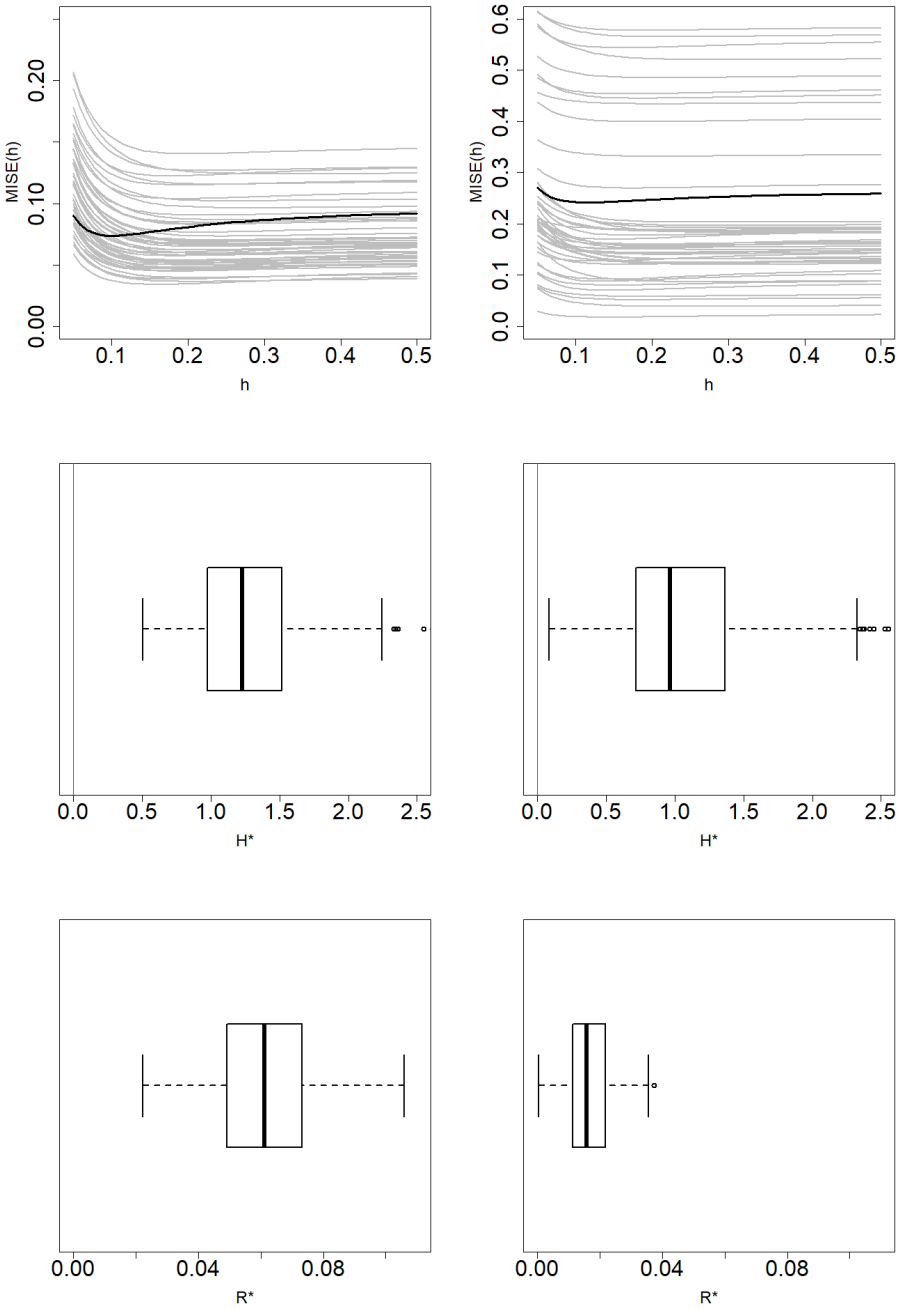

Figure 6.2: $MISE_x(h)$ function (black line) approximated via Monte Carlo and $MISE_x^*(h)$ functions (gray lines) for $N = 300$ samples (top), boxplot of $H_1^*, ..., H_N^*$ values (middle) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 3.

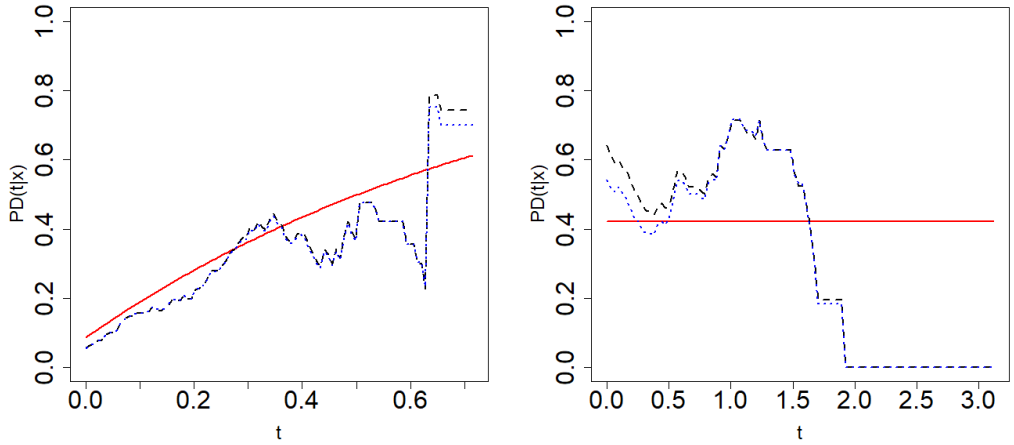

Figure 6.3: Theoretical probability of default function $PD(t|x)$ (solid line), Beran's estimation with MISE bandwidth (dotted line) and Beran's estimation with bootstrap bandwidth (dashed line) for one sample from Model 2 (left) and Model 3 (right) with $P(\delta = 0|x) = 0.5$.

6.3.2 Simulation study for the smoothed Beran's estimator

In this section, a simulation study on the bootstrap bandwidth selector of the smoothed Beran's estimator in (5.2) is carried out. The resampling technique and Monte Carlo approximation of the MISE presented in Subsection 6.2.2 are used.

For each model, the error function $MISE_x(h, g)$ is approximated via Monte Carlo from 300 simulated samples and the bivariate bandwidth that minimises $MISE_x(h, g)$ is obtained and denoted by (h_{MISE}, g_{MISE}). The values of (h_{MISE}, g_{MISE}) and $MISE_x(h_{MISE}, g_{MISE})$ are used as a benchmark.

In the study, $N = 300$ 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 value of the N bootstrap bandwidths and the standard deviation are the following:

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

$$sd\big(h^*\big) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \big(h_j^* - \overline{h^*}\big)^2}, \quad sd\big(g^*\big) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \big(g_j^* - \overline{g^*}\big)^2}.$$

In order to measure the distance of the bootstrap two-dimensional bandwidth of the j-th sample, (h_j^*, g_j^*), to the corresponding MISE bandwidth, (h_{MISE}, g_{MISE}), we consider the vector

$$D_j^* = \left(\frac{h_j^* - h_{MISE}}{h_{MISE}}, \frac{g_j^* - g_{MISE}}{g_{MISE}} \right) \in \mathbb{R}^2.$$

and its Euclidean norm denoted by $H_j^* = \|D_j^*\|_2$ with $j = 1, \ldots, N$. The mean value, $\overline{H^*} = \frac{1}{N} \sum_{j=1}^{N} H_j^*$ is a measure of how close the bootstrap bandwidths are to the MISE one.

For each sample, the estimation error of the smoothed Beran's estimator with the corresponding bootstrap bandwidth,

$$MISE_x(h_j^*, g_j^*) = E\left(\int_{I_T} \left(\widetilde{PD}_{h_j^*, g_j^*}(t|x) - PD(t|x) \right)^2 dt \right),$$

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

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

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

The ratio

$$R_j^* = \frac{RMISE_x(h_j^*, g_j^*) - RMISE_x(h_{MISE}, g_{MISE})}{RMISE_x(h_{MISE}, g_{MISE})}$$

is defined as a relative measure of the difference between the error made by the estimator with bootstrap bandwidth and MISE bandwidth. The mean of the positive values R_j^* with $j = 1, \ldots, N$ is denoted by $\overline{R^*} = \frac{1}{N} \sum_{j=1}^{N} R_j^*$. Values of the bootstrap bivariate bandwidths, estimation errors and relative measures for Models 2 and 3 are included in Table 6.2.

	Model 2		Model 3	
$P(\delta = 0\|X = x)$	0.2	0.5	0.2	0.5
h_{MISE}	0.21633	0.16735	0.11122	0.37551
g_{MISE}	0.09286	0.14612	1.27755	1.68878
$RMISE_x(h_{MISE}, g_{MISE})$	0.03710	0.05094	0.09829	0.12322
$\overline{h^*}$ (sd)	0.11736 (0.051)	0.11219 (0.057)	0.19813 (0.180)	0.16593 (0.218)
$\overline{g^*}$ (sd)	0.12647 (0.039)	0.19671 (0.054)	0.60005 (0.375)	1.45428 (0.711)
$\overline{H^*}$	0.68121	0.63604	1.22609	0.89164
$\overline{RMISE_x(h^*, g^*)}$	0.04620	0.06793	0.22135	0.28342
$\overline{R^*}$	0.24517	0.33357	1.25199	1.30003

Table 6.2: RMISE, average bootstrap bandwidths and estimation errors of the smoothed Beran's PD estimator in each level of conditional censoring probability for Models 2 and 3. Numbers within brackets are standard deviations.

Figures 6.4 and 6.5 show the $MISE_x(h, g)$ function for the smoothed Beran's estimator and its bootstrap approximation along with the corresponding contour plot for one sample of both Models 2 and 3 when the conditional probability of censoring is 0.5. It is approximated on a meshgrid of 50×50 values of (h, g). Note that both $MISE_x(h, g)$ and $MISE_x^*(h, g)$ curves, for each fixed h value, are quite similar in the region close to the minimum value of $MISE_x^*(h, g)$. Thus, the influence of covariate smoothing parameter h is weak when estimating the PD using values of bandwidth g close to the optimal one.

Figures 6.6 and 6.7 show the boxplots of H_j^* and R_j^* with $j = 1, \ldots, N$. In general, the selector tends to underestimate the value of the bandwidths. Due to the behaviour of the $MISE_x(h, g)$ curves mentioned above, this does not lead to a significant increase in the estimation error.

Figure 6.8 shows the theoretical probability of the default function and Beran's estimation with MISE and bootstrap bandwidths for one sample from Models 2 and 3 when the conditional probability of censoring is 0.5. Comparing this figure with the equivalent one for Beran's estimator shown in Figure 6.3, the improvement in estimation due to the double smoothing is remarkable.

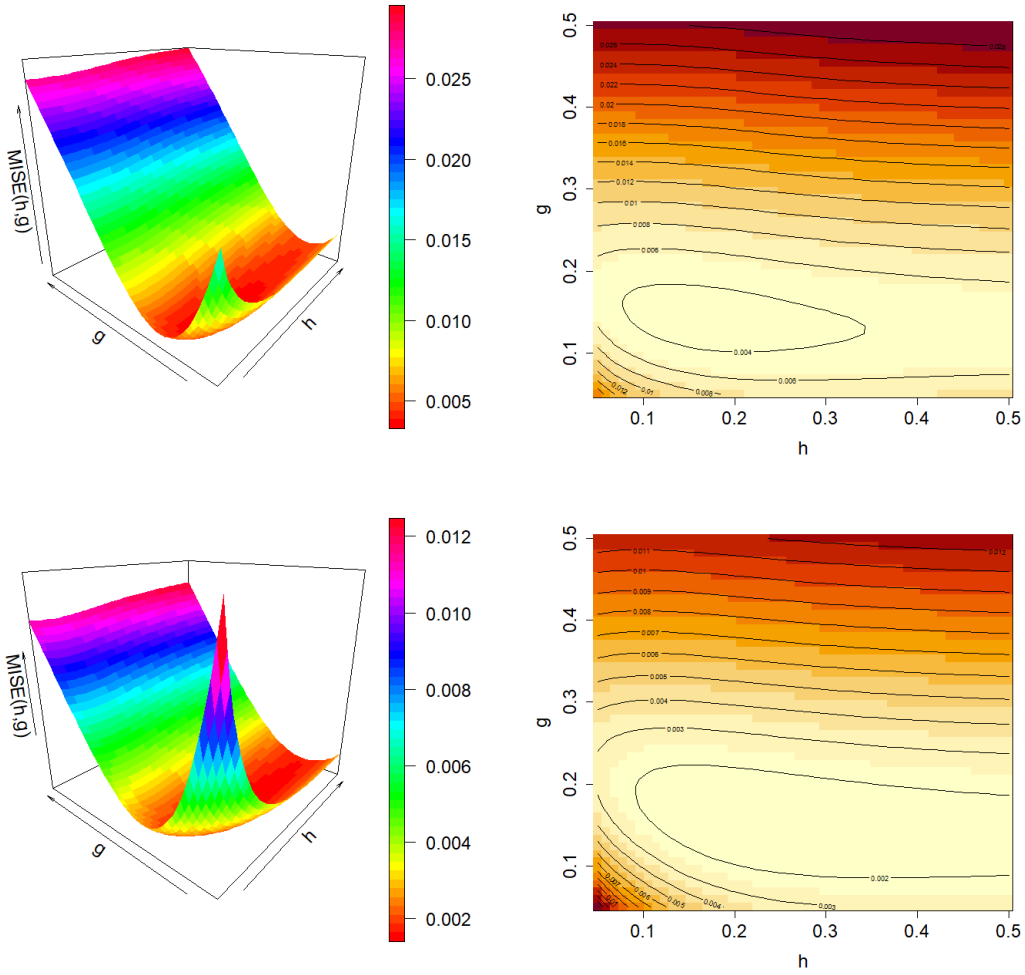

Figure 6.4: $MISE_x(h, g)$ function (top left) and contour plot of $MISE_x(h, g)$ (top right) and $MISE_x^*(h, g)$ function (bottom left) and contour plot of $MISE_x^*(h, g)$ (bottom right) for one sample from Model 2 when $P(\delta = 0|x) = 0.5$.

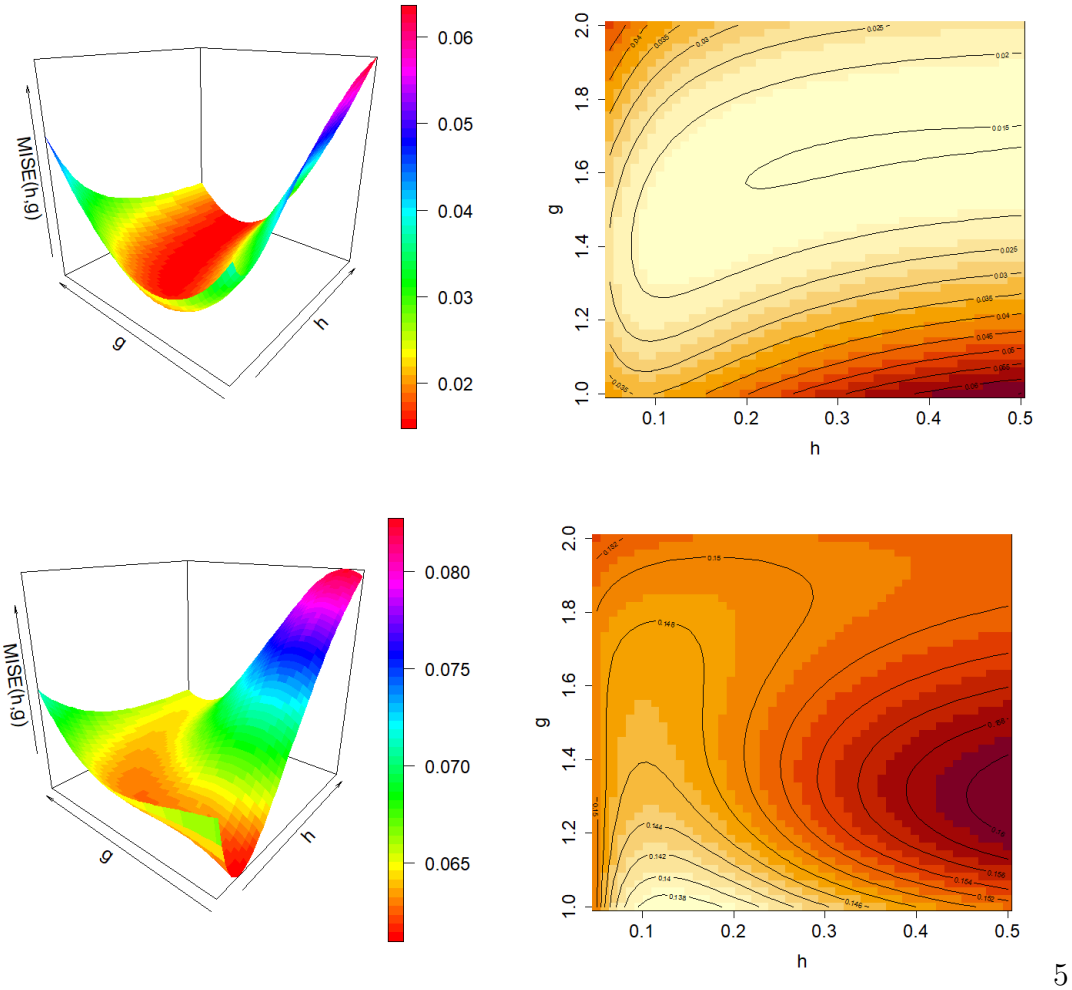

Figure 6.5: $MISE_x(h, g)$ function (top left) and contour plot of $MISE_x(h, g)$ (top right) and $MISE_x^*(h, g)$ function (bottom left) and contour plot of $MISE_x^*(h, g)$ (bottom right) for one sample from Model 3 when $P(\delta = 0|x) = 0.5$.

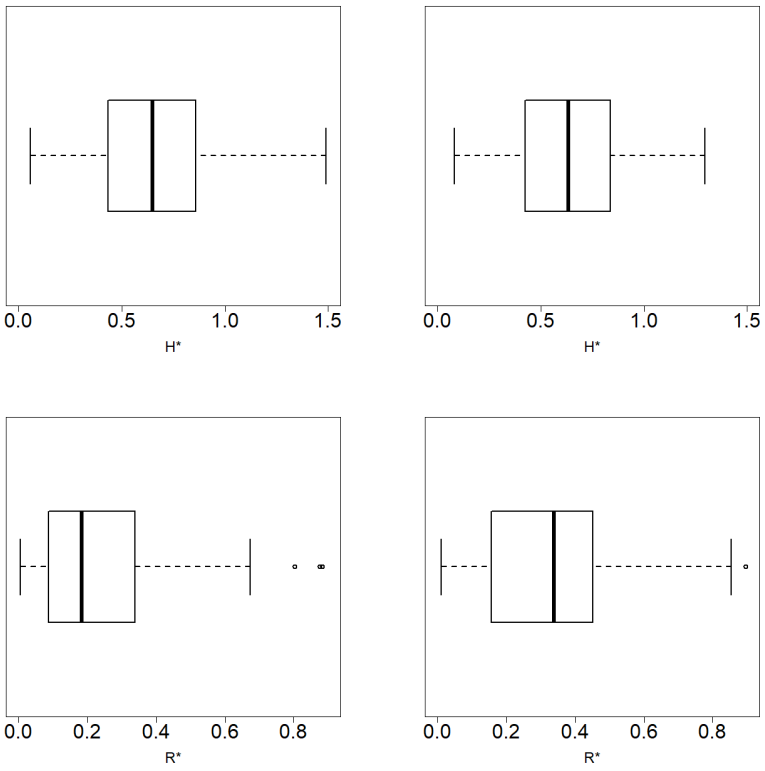

Figure 6.6: Boxplot of $H_1^*, ..., H_N^*$ values (top) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 2.

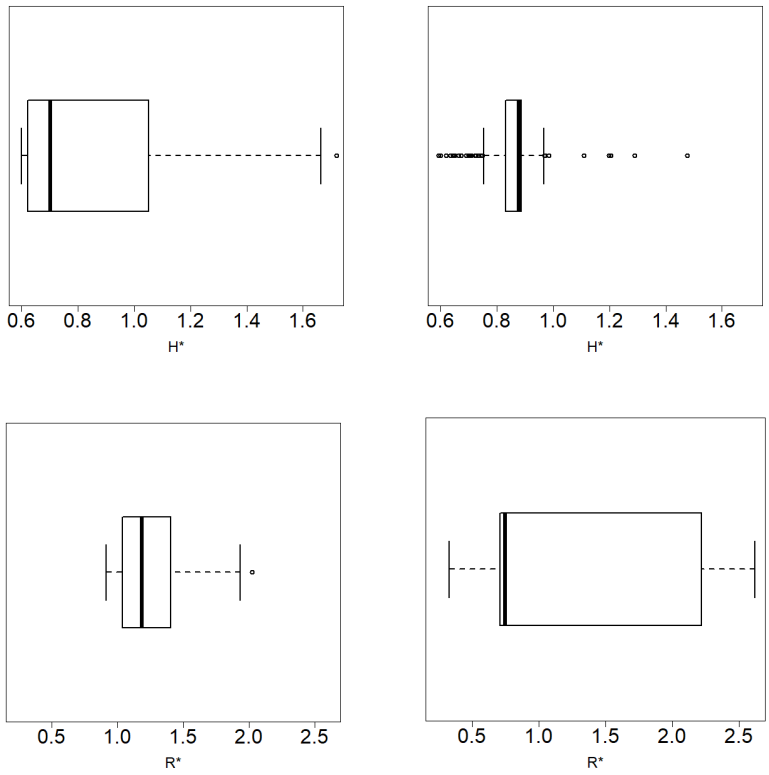

Figure 6.7: Boxplot of $H_1^*, ..., H_N^*$ values (top) and boxplot of $R_1^*, ..., R_N^*$ values (bottom) when the conditional probability of censoring is 0.2 (left) and 0.5 (right) in Model 3.

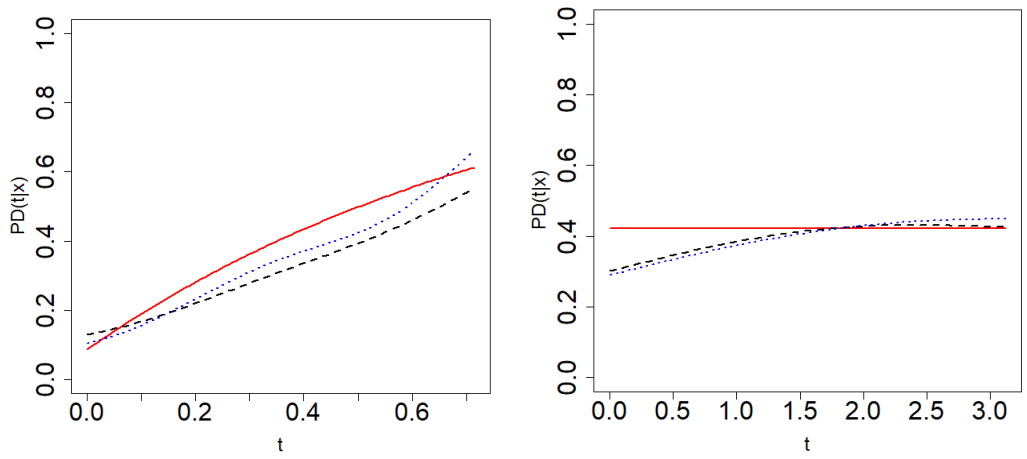

Figure 6.8: Theoretical probability of default function, $PD(t|x)$, (solid line), smoothed Beran's estimation with MISE bandwidth (dotted line) and smoothed Beran's estimation with bootstrap bandwidth (dashed line) for one sample from Model 2 (left), Model 3 (right) with $P(\delta = 0|x) = 0.5$.

The results showed in Tables 6.1 and 6.2 are summarized in Table 6.3 to compare the behaviour of Beran and the smoothed Beran's estimators for the PD and to evaluate whether the improvement that smoothing in the time variable provides for PD estimation is preserved when approximating the smoothing parameters by resampling techniques. Table 6.3 shows the estimation errors of Beran's and the smoothed Beran's estimators of the probability of default using bootstrap bandwidths. In order to measure the increase in estimation error resulting from using Beran's estimator, the following ratio is defined:

$$R_S = \frac{\overline{RMISE_x(h^*)} - \overline{RMISE(h^*, g^*)}}{\overline{RMISE(h^*, g^*)}}$$

and included in Table 6.3.

In Model 2, the estimation error of Beran's estimator is 20% larger than the error of the smoothed Beran's estimator when the conditional probability of censoring is 0.2 and 50% larger when the conditional probability of censoring is 0.5. In Model 3, these differences are even more significant: the estimation error increases up to 80% when using Beran's estimator with bootstrap bandwidth instead of the smoothed Beran's estimator.

		Model 2		Model 3	
$P(\delta = 0 \mid X = x)$		0.2	0.5	0.2	0.5
Beran	$\overline{RMISE_x(h^*)}$	0.05579	0.11206	0.28593	0.49916
SBeran	$\overline{RMISE(h^*, g^*)}$	0.04629	0.07216	0.20007	0.27611
R_S		0.20523	0.55294	0.42915	0.80783

Table 6.3: Comparative table of the estimation error of Beran's estimator and the smoothed Beran's estimator in Models 2 and 3.

6.4 Confidence regions using Beran's and the smoothed Beran's estimators

Let $x \in I$ be a fixed value of the covariate and consider $PD(t|x)$ the probability of default curve with $t \in I_T$. The curve $PD(t|x)$ belongs to the function space $\mathcal{F}(I_T)$ whose elements are real-valued functions with domain I_T. From the sample $\{(X_i, Z_i, \delta_i),\ i = 1, ..., n\}$, Beran's estimation of $PD(t|x)$, $\widehat{PD}_h(t|x)$, is obtained and a confidence region of $PD(t|x)$ at $1-\alpha$ confidence level associated to Beran's estimator can be constructed. A similar construction is done for the smoothed Beran's estimator. This confidence region of $PD(t|x)$ is a random subset of $I_T \times \mathcal{F}(I_T)$ denoted by R_α that satisfies

$$P\Big((t, PD(t|x)) \in R_\alpha, \forall t \in I_T\Big) = 1 - \alpha.$$

In this section, a method for constructing confidence regions, R_α, based on Beran and the smoothed Beran's estimator is developed.

First, Beran's estimator of the probability of default, $\widehat{PD}_h(t|x)$, given in (2.4) is used. This method follows the ideas of Cao et al. (2010) to obtain prediction regions. It is based on finding the value of $\lambda_\alpha \in \mathbb{R}^+$ such that

$$P\Big(|\widehat{PD}_h(t|x) - PD(t|x)| < \lambda_\alpha \sigma(t), \forall t \in I_T\Big) = 1 - \alpha$$

with $\sigma^2(t) = Var\Big(\widehat{PD}_h(t|x)\Big)$. Thus, the theoretical confidence region is defined by

$$R_\alpha = \Big\{(t, y) : t \in I_T, y \in \Big(\widehat{PD}_h(t|x) - \lambda_\alpha \sigma(t), \widehat{PD}_h(t|x) + \lambda_\alpha \sigma(t)\Big)\Big\}.$$

Since λ_α and $\sigma(t)$ are unknown, they are approximated by the bootstrap. These two values are calibrated in order that a certain confidence region has the desired $1-\alpha$ confidence level. The bootstrap confidence region is defined as follows:

$$R^*_\alpha = \Big\{(t, y) : t \in I_T, y \in \Big(\widehat{PD}^*_h(t|x) - \lambda^*_\alpha \sigma^*(t), \widehat{PD}^*_h(t|x) + \lambda^*_\alpha \sigma^*(t)\Big)\Big\}.$$

where $\widehat{PD}^*_h(t|x)$ is the bootstrap estimation of PD with bandwidth h and λ^*_α and $\sigma^*(t)$ are the bootstrap analogue of λ_α and $\sigma(t)$. The confidence region R^*_α satisfies

$$p(\lambda^*_\alpha) = P^*\Big((t, \widehat{PD}_r(t|x)) \in R^*_\alpha, \forall t \in I_T\Big) = 1 - \alpha. \tag{6.7}$$

From the original sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$, Beran's estimator of $PD(t|x)$ is obtained with an appropriate bandwidth h, $\widehat{PD}_h(t|x)$. The algorithm to obtain the bootstrap confidence region for $PD(t|x)$ at confidence level $1-\alpha$ associated to $\widehat{PD}_h(t|x)$ is explained below. The Monte Carlo method is used to approximate $\sigma^*(t)$, and an iterative method is used to approximate the value of λ_α^* so that the confidence region has a confidence level approximately equal to $1-\alpha$.

Confidence region based on Beran's estimator

1. Compute Beran's estimator $\widehat{PD}_r(t|x)$ from the original sample $\left\{(X_i, Z_i, \delta_i)\right\}_{i=1}^n$ using pilot bandwidth $r \in I_1$.

2. Generate B bootstrap resamples of the form $\left\{(X_i^{*,k}, Z_i^{*,k}, \delta_i^{*,k})\right\}_{i=1}^n$, $k = 1, \ldots, B$, by means of the resampling algorithm presented in Subsection 6.2.1 and pilot bandwidth r.

3. For $k = 1, \ldots, B$, compute $\widehat{PD}_h^{*,k}(t|x)$ with the k-th bootstrap resample and bandwidth h, obtaining $\left\{\widehat{PD}_h^{*,k}(t|x)\right\}_{k=1}^B$.

4. Approximate the standard deviation of $\widehat{PD}_h^*(t|x)$ by

$$\sigma^*(t) \simeq \left(\frac{1}{B}\sum_{k=1}^{B}\left(\widehat{PD}_h^{*,k}(t|x) - \frac{1}{B}\sum_{l=1}^{B}\widehat{PD}_h^{*,l}(t|x)\right)^2\right)^{1/2}, \quad t \in I_T.$$

5. Use an iterative method to obtain an approximation of the value λ_α^* defined in (6.7).

6. The confidence region is given by

$$\widehat{R}_\alpha = \left\{(t, y) : t \in I_T, y \in \left(\widehat{PD}_h(t|x) - \lambda_\alpha^*\sigma^*(t), \widehat{PD}_h(t|x) + \lambda_\alpha^*\sigma^*(t)\right)\right\}.$$

Iterative method to approximate λ_α^*

The iterative method to approximate the value of $\lambda_\alpha^* \in \mathbb{R}^+$ so that the confidence region R_α^* has a confidence level approximately equal to $1-\alpha$ is explained below.

It is based on the method proposed by Cao et al. (2010). This algorithm allows the parameter λ_α^* to be approximated quickly and efficiently.

Let $\left\{\widehat{PD}_h^{*,k}(t|x)\right\}_{k=1}^B$ be the Beran's estimations of the PD with bandwidth h over a set of B bootstrap resamples. Define the Monte Carlo approximation of $p(\lambda)$ in (6.7), for any $\lambda \in \mathbb{R}^+$, as follows:

$$p(\lambda) \simeq \frac{1}{B}\sum_{k=1}^{B} I\Big(\widehat{PD}_r(t|x) \in \Big(\widehat{PD}_h^{*,k}(t|x) - \lambda\sigma^*(t), \widehat{PD}_h^{*,k}(t|x) + \lambda\sigma^*(t)\Big), \forall t \in I_T\Big). \tag{6.8}$$

Let $\lambda_L, \lambda_H \in \mathbb{R}^+$ be such that $p(\lambda_L) \leq 1-\alpha \leq p(\lambda_H)$ and let $\zeta > 0$ be a tolerance, for example, $\zeta = 10^{-4}$.

1. Obtain $\lambda_M = \dfrac{\lambda_L + \lambda_H}{2}$ and compute Monte Carlo approximations of $p(\lambda_L)$, $p(\lambda_M)$ and $p(\lambda_H)$ according to (6.8).

2. If $p(\lambda_M) = 1-\alpha$ or $p(\lambda_H) - p(\lambda_L) < \zeta$, then $\lambda_\alpha^* = \lambda_M$. Otherwise,

   (a) If $1-\alpha < p(\lambda_M)$, then $\lambda_H = \lambda_M$ and return to Step 1.

   (b) If $p(\lambda_M) < 1-\alpha$, then $\lambda_L = \lambda_M$ and return to Step 1.

A preliminary analysis not shown here suggests the following choice for the pilot bandwidth:

$$r = \frac{3}{4}\Big(Q_X(0.975) - Q_X(0.025)\Big)\left(\sum_{i=1}^{n}\delta_i\right)^{-1/3}.$$

This method to obtain confidence regions for the curve $PD(t|x)$ for fixed $x \in I$ and t covering I_T based on Beran's estimator can be adapted to obtain confidence regions using the smoothed Beran's estimator. Simply replace Beran's estimator $\widehat{PD}_h(t|x)$ by the smoothed Beran's estimator $\widetilde{PD}_{h,g}(t|x)$ given in (5.2) where necessary, and obtain the analogous bootstrap approximations of λ_α and $\sigma(t)$. The confidence region is given by

$$\widetilde{R}_\alpha = \left\{(t,y) : t \in I_T, y \in \Big(\widetilde{PD}_{h,g}(t|x) - \lambda_\alpha^*\sigma^*(t), \widetilde{PD}_{h,g}(t|x) + \lambda_\alpha^*\sigma^*(t)\Big)\right\}.$$

Denote the lower and upper bounds of the confidence region by $l(t,x)$ and $u(t,x)$, respectively. It may happen that the lower bound of the confidence region is less than 0 or the upper bound is greater than one for some points (t_0, x_0). When this happens, we set $l(t_0, x_0) = 0$ or $u(t_0, x_0) = 1$, as appropriate.

The pilot bandwidths defined in (6.3) and (6.6) are used for the confidence region algorithm based on both Beran and smoothed Beran's estimators.

This is an analogous procedure to Method 1 presented in Section 4.4 to obtain confidence regions for the conditional survival function. Due to the variability that the PD estimations exhibit, especially those obtained without time smoothing, the confidence regions with constant width, such as those resulting from Method 2, do not have a promising behaviour in this context.

6.5 Simulation study for confidence regions

A simulation study is carried out to test the performance of bootstrap confidence regions proposed. Models 2 and 3 described in 2.4 are again considered in this study, with identical features. The methods shown in Section 6.4 are used for this purpose with both Beran's and smoothed Beran's estimators. When Beran's estimator is used, the bandwidth that minimises the mean integrated squared error, $h = h_{MISE}$, is used. Similarly, if the smoothed Beran's estimator is used, the two-dimensional bandwidth that minimises the mean integrated squared error, $(h, g) = (h_{MISE}, g_{MISE})$, is used. These bandwidths are unknown in practice, but they allow a fair comparison of methods in the simulation study.

The simulation setup is the one explained in Section 6.3. Two conditional probabilities of censoring are considered for each model: $P(\delta = 0|x) = 0.2$ and $P(\delta = 0|x) = 0.5$. The number of bootstrap resamples of each samples is $B = 500$, and $N = 300$ simulated samples of each model are obtained. The sample size is $n = 400$. The confidence level is $1 - \alpha$ with $\alpha = 0.05$.

Figure 6.9 shows Beran's estimations of the PD for $B = 500$ resamples from one sample of Model 3 when the conditional probability of censoring is 0.5. The theoretical probability of default is also plotted in the figure. The PD is estimated on a time grid $t_1 = 0 < t_2 < \cdots < t_{n_T}$ such that $t_{n_T} + b = F^{-1}(0.95|x)$. The information provided by the data in the right tail of such a time distribution is sparse due to heavy censoring. The method results in extremely wide confidence regions or degeneration to zero as in the case of Model 3 (see Figure 6.9). Therefore, the time grid in this section is restricted to the interval where sufficient information is available.

For this section, we consider the problem of obtaining the bootstrap confidence region for the probability of default in a time grid $t_1 = 0 < t_2 < \cdots < t_{n_T}$ such that $t_k \in I_T \subseteq \mathbb{R}^+$ for all $k = 1, \ldots, n_T$ and $t_{n_T} + b = F^{-1}(0.70|x)$, with b being approximately equal to 20% of the grid length.

For Model 2, having set the value of the covariate, $x = 0.6$, the time horizon is $b = 0.1$ (20% of the time range) and $t_{n_T} + b = F^{-1}(0.70|x = 0.6) = 0.55$. For Model 2, having set the value of the covariate, $x = 0.8$, the time horizon is $b = 0.3$ (20% of the time range) and $t_{n_T} + b = F^{-1}(0.70|x = 0.8) = 1.55$. Table 6.4 contains the bandwidths that minimise the MISE function for Beran's estimator and the smoothed Beran's estimator along this new time grid.

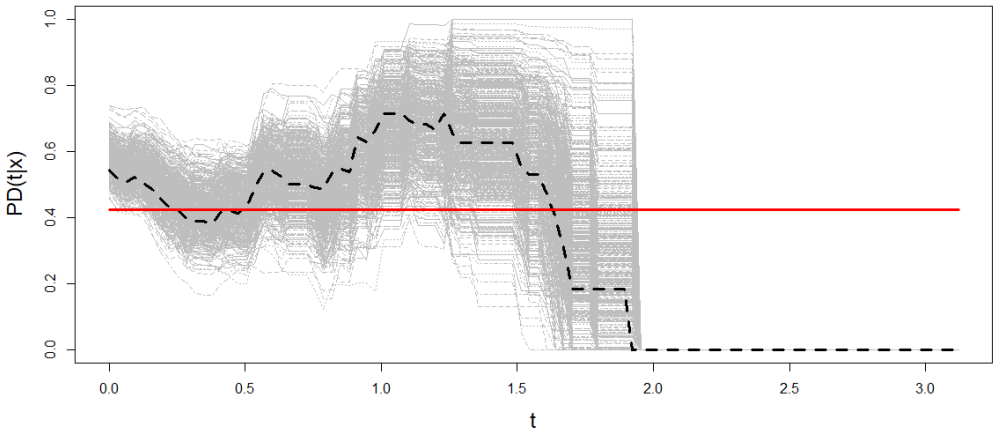

Figure 6.9: Theoretical $PD(t|x)$ (red solid line), Beran's estimation of $PD(t|x)$ with MISE bandwidths (black dashed line) and bootstrap versions of Beran's estimations of $PD(t|x)$ from $B = 500$ resamples (gray dashed lines) for one sample from Model 2 when $P(\delta = 0|x) = 0.5$.

		Model 2		Model 3	
$P(\delta = 0\|X = x)$		0.2	0.5	0.2	0.5
Beran	h_{MISE}	0.375510	0.320408	0.041837	0.057755
	$RMISE_x(h_{MISE})$	0.019403	0.025943	0.193334	0.220733
SBeran	h_{MISE}	0.230612	0.196939	0.094286	0.154490
	g_{MISE}	0.073673	0.083469	0.908163	1.071429
	$RMISE_x(h_{MISE}, g_{MISE})$	0.013658	0.018165	0.026161	0.029007

Table 6.4: MISE bandwidths and RMISE of Beran's and the smoothed Beran's estimator in each level of conditional censoring probability for Models 2 and 3 when $t_{n_T} + b = F^{-1}(0.70|x)$.

For each model, the confidence region is obtained according to the method explained in Section 6.4 using both Beran's estimator and the smoothed Beran's estimator. The criteria for comparing the methods are set out below.

A confidence region performs well if its coverage is close to the nominal one, in this case $1 - \alpha = 0.95$, and has a small area or average width. For each sample, $j = 1, \ldots, N$, denoting $l_j(t, x) = \widehat{PD}_h(t|x) - \lambda^*_\alpha \sigma^*(t)$ and $u_j(t, x) = \widehat{PD}_h(t|x) + \lambda^*_\alpha \sigma^*(t)$ when using Beran's estimator or $l_j(t, x) = \widetilde{PD}_{h,g}(t|x) - \lambda^*_\alpha \sigma^*(t)$ and $u_j(t, x) =$

$\widetilde{PD}_{h,g}(t|x) + \lambda^*_\alpha \sigma^*(t)$ when using the smoothed Beran's estimator, the following values measure the performance of the confidence region and allow for comparison of results.

Coverage is the proportion of bootstrap regions that contain the whole theoretical probability of default curve and it is defined as follows

$$\frac{1}{N}\sum_{j=1}^{N} I\Big\{PD(t_k|x) \in \Big(l_j(t_k,x), u_j(t_k,x)\Big), \forall k = 1, ..., n_T\Big\}.$$

The mean pointwise coverage is the mean of the proportion of time grid values for which the confidence region contains the theoretical probability of default curve. It is given by

$$\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_T}\sum_{k=1}^{n_T} I\Big\{PD(t_k|x) \in \Big(l_j(t_k,x), u_j(t_k,x)\Big)\Big\}\right).$$

Average width of the bootstrap confidence region is defined by

$$\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_T}\sum_{k=1}^{n_T}\Big(u_j(t_k,x) - l_j(t_k,x)\Big)\right).$$

Winkler score (see Winkler (1972)) is also used to compare the behaviour of the methods. For classical confidence or prediction intervals, it is defined as the length of the interval plus a penalty if the theoretical value is outside the interval. Thus, it combines width and coverage. For values that fall within the interval, the Winkler score is simply the length of the interval. So low scores are associated with narrow intervals. When the theoretical value falls outside the interval, the penalty is proportional to how far the observation is from the interval. The formula of the Winkler score (WS) as a function of the time and covariate variables is as follows:

$$\begin{aligned}\mathrm{WS}(t,x) &= u_j(t,x) - l_j(t,x) + \frac{2}{\alpha}(l_j(t,x) - PD(t|x))I\Big(PD(t|x) < l_j(t,x)\Big) \\ &\quad + \frac{2}{\alpha}(PD(t|x) - u_j(t,x))I\Big(PD(t|x) > u_j(t,x)\Big).\end{aligned}$$

Since we are working with confidence regions for fixed $x \in I$ and t varying over the interval I_T, the integrated Winkle score is proposed as a criteria for the

comparison of the confidence regions. It is defined by

$$IWS(x) = \int_{I_T} WS(t, x)dt.$$

and the lower the value of IWS, the better the performance of the confidence region.

The results obtained are shown in Table 6.5. The high values of pointwise coverage in all scenarios are remarkable. Furthermore, these coverage percentages are preserved when using double smoothing, while the average width of the confidence regions is halved. This is reflected in the IWS, which presents much larger values in the Beran's estimator-based confidence regions.

	Model 2				Model 3			
$P(\delta = 0\|X = x)$	0.2		0.5		0.2		0.5	
Estimator	Beran	SBeran	Beran	SBeran	Beran	SBeran	Beran	SBeran
Coverage (%)	96.33	90.67	90.00	85.33	97.33	83.00	91.46	98.00
Pointwise coverage(%)	99.94	98.05	99.63	96.85	99.88	98.53	99.65	99.85
Width	0.21997	0.09539	0.24827	0.10937	0.50514	0.17969	0.55581	0.33033
IWS	0.09869	0.04537	0.11218	0.05571	0.62590	0.22825	0.71009	0.40845

Table 6.5: Coverage, pointwise coverage, width and IWS of the 95% confidence regions by means of Beran's and the smoothed Beran's estimators using $N = 300$ simulated samples from Models 2 and 3.

This analysis is also illustrated in Figures 6.10 and 6.11, where the confidence region for the probability of default of one sample from Models 2 and 3 is shown. These graphs show the higher variability of the Beran's estimations in the resamples with respect to the smoothed Beran's estimations. This leads to much wider confidence regions, especially at the right tail of the time distribution.

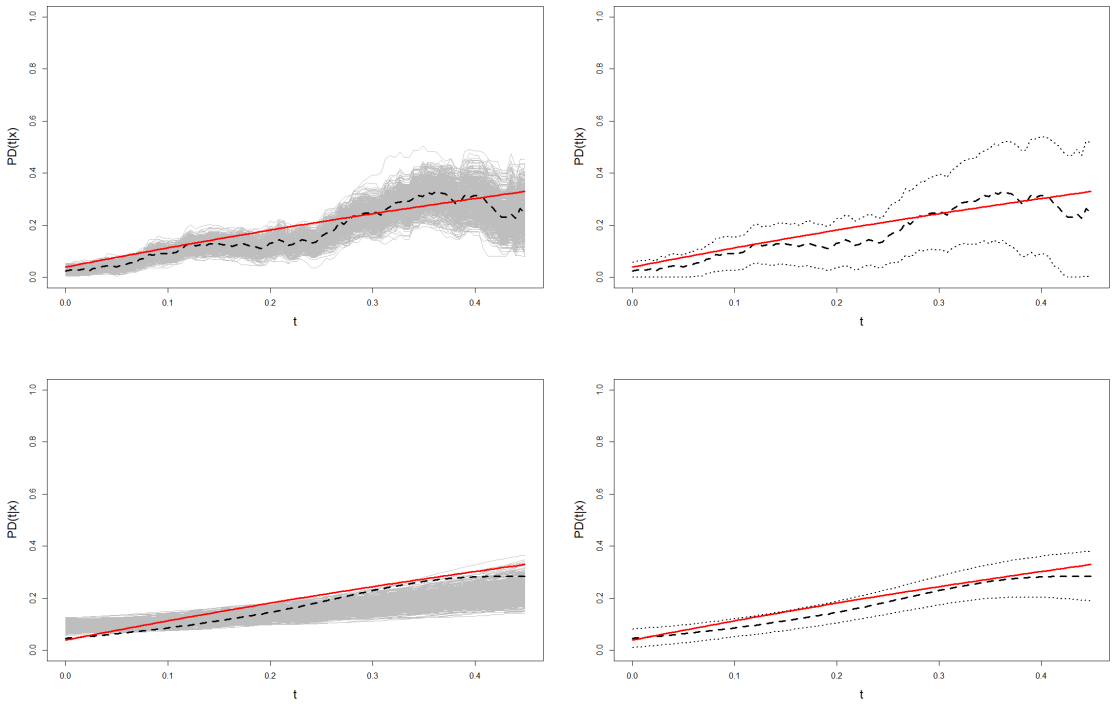

Figure 6.10: Theoretical $PD(t|x)$ (red solid line) and estimation with MISE bandwidths (black dashed line) along with the bootstrap estimations of $PD(t|x)$ from $B = 500$ resamples (gray dashed lines) in the left panel and 95% confidence region (black dotted lines) in the right panel by means of Beran's estimator (top) and the smoothed Beran's estimator (bottom) for one sample from Model 2 when $P(\delta = 0|x) = 0.5$.

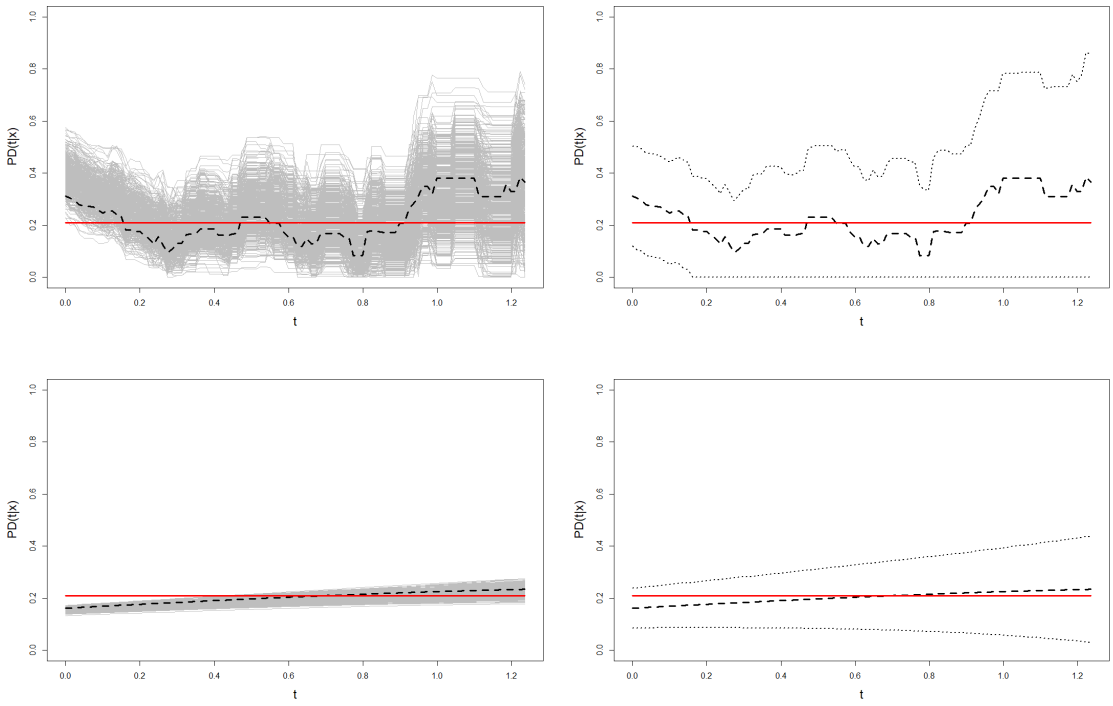

Figure 6.11: Theoretical $PD(t|x)$ (red solid line) and estimation with MISE bandwidths (black dashed line) along with the bootstrap estimations of $PD(t|x)$ from $B = 500$ resamples (gray dashed lines) in the left panel and 95% confidence region (black dotted lines) in the right panel by means of Beran's estimator (top) and the smoothed Beran's estimator (bottom) for one sample from Model 3 when $P(\delta = 0|x) = 0.5$.

An analysis of the computational times of these techniques could be of interest. Since the slowing part of these methods is the resampling and this is the same as that presented in Section 4.2, the computation times are similar to those shown in Section 4.6.

6.6 Application to real data

In this section, bandwidth selectors for Beran's and the smoothed Beran's estimators are applied to the German Credit dataset, and the confidence region of the probability of default is obtained. This dataset is publicly available on the webpage `http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)` (last ac-

cess on September 15th, 2021) and was previously analysed in Strzalkowska-Kominiak and Cao (2013). This data set includes information about 1000 credits, from which 293 were classified as bad credits and 707 as good credits. Then, the censoring percentage of this dataset is 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): X_1 is already a continuous variable denoting amount of credit in DM, $X_2 \in \{-0.05, 0.01, 0.25, 0\}$ denotes the amount of money in the checking account in thousands of DM, $X_3 \in \{0, 0.05, 0.25, 0.75, 1.25\}$ denotes the savings amount in thousands of DM and $X_4 \in \{0, 0.5, 2.5, 5.5, 8.5\}$ denotes the years of employment. The single-index method proposed in Strzalkowska-Kominiak and Cao (2013) is used to estimate $(1, \theta_2, \theta_3, \theta_4)$, obtaining the credit scoring $X = X_1 + 3.2091X_2 + 0.2312X_3 + 2.1891X_4$.

Figure 6.12 shows the scatter plot of credit scoring and follow-up time distinguishing between the censored and uncensored (and therefore defaulted) credits. A dependency relationship between the two variables can be identified in the plot.

The probability of default, $PD(t|x)$, is estimated when $x = 0.85$, which is a close value to the sample mean of the credit scoring, and $t \in [0, 60]$. The bandwidth selector presented in Section 6.2.1 is used to approximate the optimal bandwidth for Beran's estimator, obtaining $h^* = 0.500$. The bandwidth selector presented in Section 6.2.2 gives the bootstrap approximation of the optimal two-dimensional bandwidth for the smoothed Beran's estimator, $(h^*, g^*) = (0.102, 13.614)$. The estimations of the conditional survival function and the probability of default by means of Beran's and the smoothed Beran's estimator with the corresponding bootstrap bandwidths are shown in Figure 6.13. The poor behaviour of Beran's PD estimator for large values of time is evident. The results obtained by the smoothed Beran's

estimator seem to be more appropriate (Barnard (2017); dos Reis and Smith (2018)).

Figure 6.12: Scatter plot of credit scoring and duration of the credit in the censored group (red circles) and the uncensored group (blue triangles) for the German credit data.

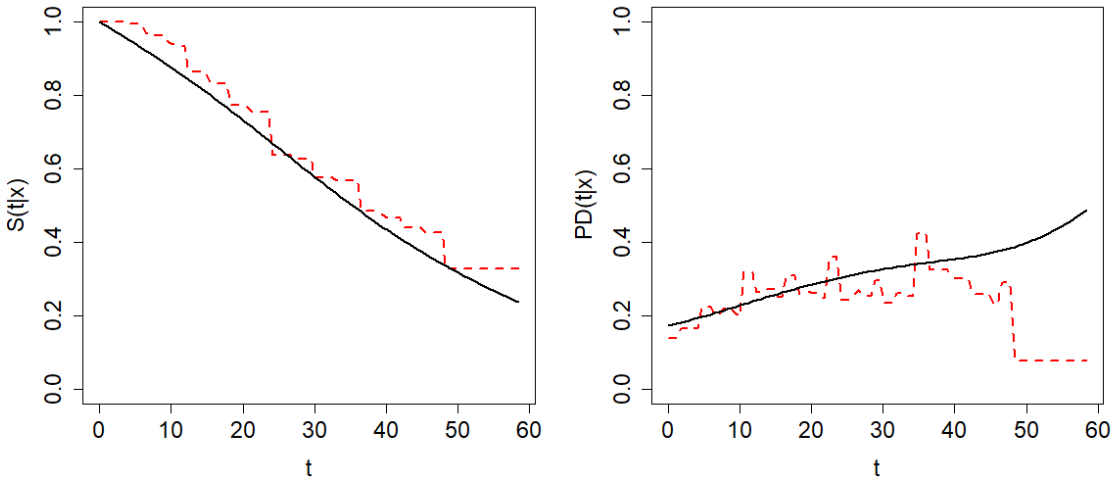

Figure 6.13: Conditional survival function estimation (left) and probability of default estimation (right) by means of Beran's estimator (dashed line) and the smoothed Beran's estimator (solid line) with bootstrap bandwidths when $x = 0.85$ for the German credit dataset.

Finally, the confidence region methods proposed in Section 6.4 are applied. Since the MISE bandwidths are unknown in this context, bootstrap bandwidths are used. The bootstrap resamples and the resulting confidence regions at confidence level 95% using each estimator are shown in Figure 6.14. The average width of the confidence region based on Beran's estimator is 0.5581, and the average width of the one based

on the smoothed Beran's estimator is 0.1438. Note that the confidence region of $PD(t|x)$ is computed over the time interval $[0, 40]$. Since the information provided in the right tail of the time distribution is sparse, Beran's estimator performs very poorly, leading to extremely wide confidence regions. However, this problem is not as severe for the smoothed Beran's estimator, so the confidence region is computable for higher values of time. Figure 6.15 shows the confidence region of $PD(t|x)$ based on the smoothed Beran's estimator with $t \in [0, 60]$. The average width of this confidence region is 0.2398.

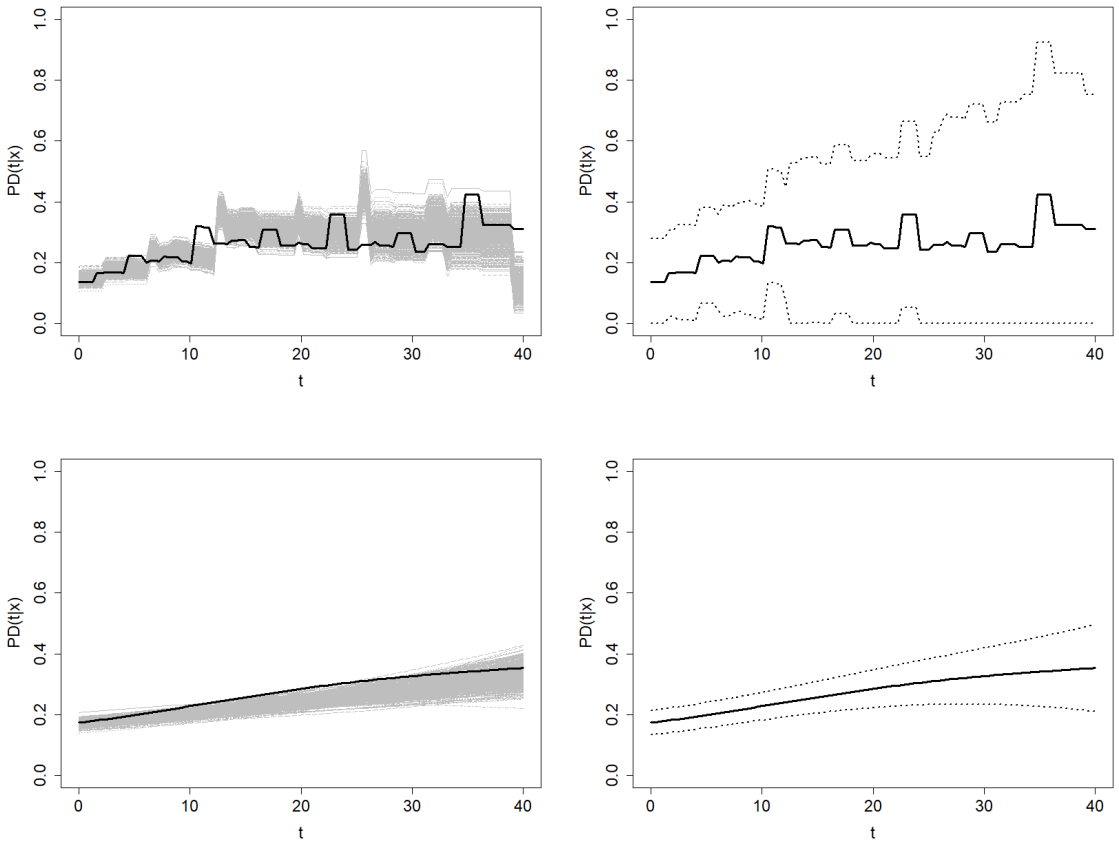

Figure 6.14: Estimation of $PD(t|x)$ with bootstrap bandwidths along with bootstrap estimations of PD from $B = 500$ resamples (left) and 95% confidence region (right) by Beran's estimator (top) and the smoothed Beran's estimator (bottom) when $x = 0.85$ and $t \in [0, 40]$ for the German credit data set.

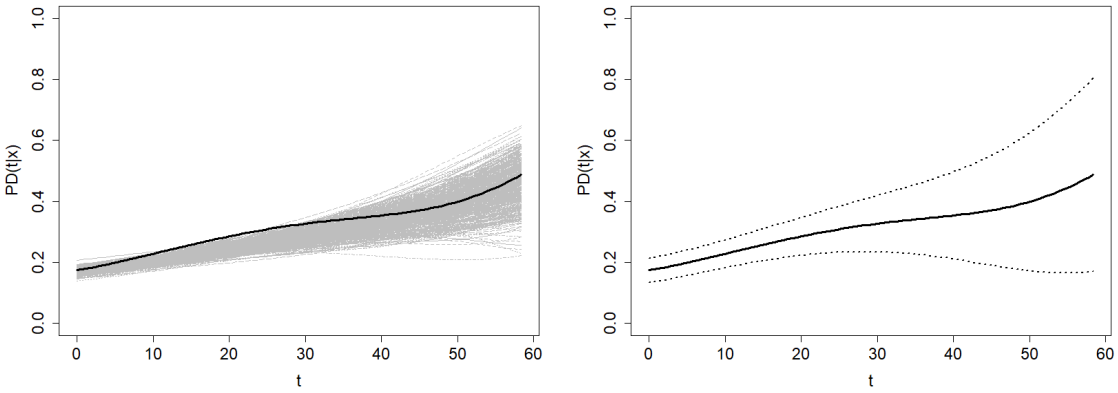

Figure 6.15: Estimation of $PD(t|x)$ with bootstrap bandwidths along with bootstrap estimations of PD from $B = 500$ resamples (left) and 95% confidence region (right) by the smoothed Beran's estimator when $x = 0.85$ and $t \in [0, 60]$ in the German credit data set.

Chapter 7

PD estimator based on cure models

7.1 Introduction

Time to default could face not only a problem of right censoring, but also the existence of cured individuals which 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.

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. We refer to Amico and Van Keilegom (2018), for an overview paper on this topic. However in the area of credit risks cure models have not been used much 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.

Instead of working with Beran's estimator (Beran (1981)), we will use another nonparametric estimator, that estimates separately the probability that a borrower will eventually default, 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.

In this chapter, a nonparametric estimator of the PD based on mixture cure models is proposed. Asymptotic properties of this PD estimator are presented and a simulation study shows the behaviour of the nonparametric cure model estimator and a comparison with Beran's estimator and other semiparametric estimators. The PD estimators are applied to a set of modified real data.

The content of this chapter has been submitted for possible publication and it is currently under revision in Peláez et al. (2022c).

7.2 Nonparametric cure model 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 follow-up time, 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:

Group 1: The group of individuals who are susceptible to default and censored. The default will eventually occur but it will not be observed. It corresponds to the situation $\nu = 0$, $\delta = 0$ and, consequentelly, $C < T < \infty$.

Group 2: The group of individuals who are susceptible to default and noncensored. The default happens and it is observed. It corresponds to the situation $\nu = 0$, $\delta = 1$. In this case, $T < C$.

Group 3: The group of individuals who are not susceptible to default. These individuals will never experience default no matter how long they are observed, so they are censored for sure. It corresponds to the situation $\nu = 1$, $\delta = 0$, and, therefore, $C < T = \infty$.

The event ν =1 and δ =1 are not compatible. It would correspond to an individual who is both non-susceptible (will never experience default since $T = \infty$) and uncensored (default is observed since $T < C$). 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

$$\begin{aligned} S(t|x) &= P(T > t|x) = P(T > t|\nu = 1, x)P(\nu = 1|x) \\ &\quad + P(T > t|\nu = 0, x)P(\nu = 0|x) \\ &= 1\Big(1 - P(\nu = 0|x)\Big) + P(T > t|\nu = 0, x)P(\nu = 0|x) \\ &= \Big(1 - P(\nu = 0|x)\Big) + P(T_0 > t|x)P(\nu = 0|x) \\ &= 1 - p(x) + S_0(t|x)p(x) \end{aligned}$$

where $1 - p(x)$ is the probability of being cured (nonsusceptible to default) and $S_0(t|x)$ the conditional survival function of the uncured population. The functions $p(x)$ and $S_0(t|x)$ are called the incidence and the latency, respectively.

Let x be a fixed value of the covariate X (typically, the scoring) and b a horizon time. 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, $\widehat{PD}_h(t|x)$, in this context. For

this purpose, a nonparametric survival estimator based on mixture cure models is considered.

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

$$\widehat{S}^{NPCM}_{h_1,h_2}(t|x) = 1 - \widehat{p}_{h_1}(x) + \widehat{p}_{h_1}(x)\widehat{S}_{0,h_2}(t|x). \tag{7.1}$$

The incidence estimator, $\widehat{p}_{h_1}(x)$, 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:

$$\widehat{p}_{h_1}(x) = 1 - \widehat{S}^{B}_{h_1}\Big(\max\{T_i : i = 1, ..., n, \delta_i = 1\}|x\Big).$$

The latency estimator depending on one single bandwidth, $\widehat{S}_{0,h_2}(t|x)$, proposed by López-Cheda et al. (2017a) is as follows:

$$\widehat{S}_{0,h_2}(t|x) = \frac{\widehat{S}^{B}_{h_2}(t|x) - \Big(1 - \widehat{p}_{h_2}(x)\Big)}{\widehat{p}_{h_2}(x)}.$$

Replacing (7.1) in (2.2), we obtain the nonparametric cure model estimator (NPCM) of the probability of default.

Note that the particular case $h_1 = h_2$ 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_1 and h_2 are substantially different in practice, although they have the same convergence rate. Choosing the best bandwidth h_1 for incidence and the best bandwidth h_2 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.

7.3 Asymptotic results for the NPCM estimator

Asymptotic properties of nonparametric incidence and latency estimators are already available in López-Cheda et al. (2017a) and López-Cheda et al. (2017b). In

this Section, asymptotic properties of the NPCM estimator of the probability of default are studied.

A number of notations used below were defined in Section 3.3.1. Additionally, let $R : \mathbb{R} \longrightarrow \mathbb{R}$ be any function and and given any constant $a \in \mathbb{R}$,

$$\widetilde{c}_R(a) = \int R(at)R(t)dt. \tag{7.2}$$

The following functions are required to state the results.

$$\begin{aligned}
\omega(Z,\delta,t,x) &= -\frac{S(t|x)}{p(x)}\xi(Z,\delta,t,x) - \frac{\Big(1-p(x)\Big)\Big(1-S(t|x)\Big)}{p^2(x)}\xi(Z,\delta,\infty,x),\\
\Phi_2(u,t,x) &= E\Big[\xi^2(Z,\delta,t,x)|X=u\Big],\\
B_1(t,x) &= \frac{d_K\Big(S_0(t|x)-1\Big)\Big(p(x)-1\Big)}{2m(x)}\frac{\partial^2}{\partial u^2}\Big(\Phi_\xi(u,t,x)m(u)\Big)|_{u=x},\\
B_2(t,x) &= -\frac{d_k S(t|x)}{2m(x)}\frac{\partial^2}{\partial u^2}\Big(\Phi_\xi(u,t,x)m(u)\Big)|_{u=x}\\
&\quad -\frac{d_K\Big(1-p(x)\Big)\Big(1-S(t|x)\Big)}{2p(x)m(x)}\frac{\partial^2}{\partial u^2}\Big(\Phi_\xi(u,\infty,x)m(u)\Big)|_{u=x},\\
\widetilde{B}_1(t,x) &= -\frac{1}{S(t|x)}B_1(t+b,x) + \frac{S(t+b|x)}{S^2(t|x)}B_1(t,x),\\
\widetilde{B}_2(t,x) &= -\frac{1}{S(t|x)}B_2(t+b,x) + \frac{S(t+b|x)}{S^2(t|x)}B_2(t,x),\\
D_\xi(u,t_1,t_2,x) &= Cov\Big[\xi(Z_1,\delta_1,t_1,x),\xi(Z_1,\delta_1,t_2,x)\Big|X_1=u\Big]m(u),\\
D_{\xi,\omega}(u,t_1,t_2,x) &= Cov\Big[\xi(Z_1,\delta_1,t_1,x),\omega(Z_1,\delta_1,t_2,x)\Big|X_1=u\Big]m(u),
\end{aligned}$$

$$
\begin{aligned}
C_1(t_1,t_2,x) &= \frac{c_K S(t_1|x)S(t_2|x)}{p^2(x)} D_\xi(x,t_1,t_2,x) \\
&+ \frac{c_K S(t_1|x)\Big(1-S(t_2|x)\Big)}{p^3(x)} D_\xi(x,t_1,\infty,x) \\
&+ \frac{c_K \Big(1-S(t_1|x)\Big)S(t_2|x)\Big(1-p(x)\Big)}{p^3(x)} D_\xi(x,\infty,t_2,x) \\
&+ \frac{c_K \Big(1-p(x)\Big)^2\Big(1-S(t_1|x)\Big)\Big(1-S(t_2|x)\Big)}{p^4(x)} \Phi_2(x,\infty,x)m(x), \\
V_1(t_1,t_2,x) &= \frac{\Big(S_0(t_1|x)-1\Big)\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)^2}{m(x)} c_K \Phi_2(x,\infty,x), \\
V_2(t_1,t_2,x) &= \frac{p^2(x)C_1(t_1,t_2,x)}{m^2(x)}, \\
V_3(t_1,t_2,x) &= \frac{\Big(S_0(t_1|x)-1\Big)\Big(p(x)-1\Big)p(x)}{m^2(x)} D_{\xi,\omega}(x,\infty,t_2,x) \\
&+ \frac{\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)p(x)}{m^2(x)} D_{\xi,\omega}(x,t_1,\infty,x).
\end{aligned}
$$

Assumptions A.1-A.10 listed in Section 3.3.1 are required to state the results. Additional assumptions are introduced in this section to study the asymptotic properties of the NPCM estimator. They were previously assumed in López-Cheda et al. (2017a) and López-Cheda et al. (2017b) to obtain almost sure representations of the incidence and latency estimators.

A.11. Let $G(t) = P(C \leq t)$ 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.12. The first derivatives with respect to t of the functions $S_0(t|x)$ and $G(t|x)$, i.e. $S_0'(t|x)$ and $G'(t|x)$ exist and are continuous on $[l,u] \times I_c$.

A.13. 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.14. The density function of T, $f(t)$ is bounded away from 0 on $[l,u]$.

A.15. $\int_0^\infty \frac{dH_1(t|x)}{\Big(1-H(t|x)\Big)^2} < \infty \quad \forall x \in I.$

A.16. The smoothing parameter $h_1 = h(n)$ satisfies $h_1 \to 0$, $\frac{nh_1^5}{\ln n} = O(1)$, $\frac{(\ln n)^3}{nh_1} \to 0$ and $(\ln n)^3 nh_1 \to \infty$.

A.17. The smoothing parameter $h_2 = h_2(n)$ satisfies $h_2 \to 0$, $\frac{nh_2^5}{\ln n} = O(1)$, $\frac{(\ln n)^3}{nh_2} \to 0$ and $(\ln n)^3 nh_2 \to \infty$.

A.18. The smoothing parameters $h_1 = h(n)$ and $h_2 = h_2(n)$ satisfy $\frac{nh_1^3}{h_2^2(\ln n)^3} \to \infty$ and $\frac{nh_2^3}{h_1^2(\ln n)^3} \to \infty$.

A.19. Let $(t,x) \in [l,u] \times I_c$. The second derivative of $m(u)$ exists at $u = x$. The second derivative of $\Phi_\xi(u,t,x)$ 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 $D_\xi(u,t_1,t_2,x)$ exists at $(x,t,t+b,x)$, (x,t,∞,x) and (x,∞,t,x). The second derivative of $D_{\xi,\omega}(u,t_1,t_2,x)$ exists at (x,t,∞,x) and (x,∞,t,x).

Assumptions A.11-A.15 are needed to bound some population functions. They require existence and continuity of population function derivatives. Bandwidths requirements are covered by Assumptions A.16, A.17 and A.18. Assumption A.19 refers to the differentiability of the functions previously defined in this section.

Lemma 7.1 (Almost sure representation of the NPCM estimator for the conditional survival function)**.** *Under Assumptions A.1-A.19, for fixed values* $(t,x) \in [l,u] \times I$*,*

$$\begin{aligned} &\widehat{S}^{NPCM}_{h_1,h_2}(t|x) - S(t|x) \\ &= \Big(S_0(t|x)-1\Big)\Big(p(x)-1\Big)\textstyle\sum_{i=1}^n w^A_{h_1,i}(x)\xi(Z_i,\delta_i,\infty,x) \\ &\quad + p(x)\textstyle\sum_{i=1}^n w^A_{h_2,i}(x)\omega(Z_i,\delta_i,t,x) + R^1_n(t|x) \quad a.s., \end{aligned} \tag{7.3}$$

where $w^A_{h,i}(x)$ *was defined in Lemma 3.1 and*

$$\sup_{(t,x)\in[l,u]\times I} |R^1_n(t|x)| = O_p\left(\frac{\ln n}{nh_1}\right)^{3/4} + O_p\left(\frac{\ln n}{nh_2}\right)^{3/4}.$$

Theorem 7.1 (Almost sure representation of the NPCM estimator for the PD)**.** *Under Assumptions A.1-A.19, for fixed values* $(t,x),(t+b,x)\in[l,u]\times I$*,*

$$\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)-PD(t|x) \;=\; \sum_{i=1}^{n}\Psi_{n,i}(t,x)+R^2_n(t|x) \quad a.s.,$$

where

$$\Psi_{n,i}(t,x)=-\frac{1}{S(t|x)}\zeta_{n,i}(t+b,x)+\frac{S(t+b|x)}{S^2(t|x)}\zeta_{n,i}(t,x),$$

$$\zeta_{n,i}(t,x)=\Big(S_0(t|x)-1\Big)\Big(p(x)-1\Big)w^A_{h_1,i}(x)\xi(Z_i,\delta_i,\infty,x)+p(x)w^A_{h_2,i}(x)\omega(Z_i,\delta_i,t,x)$$

and

$$\sup_{(t,x)\in[l,u]\times I}|R^2_n(t|x)|=O_p\left(\ln n\left(\frac{1}{nh_1}+\frac{1}{nh_2}\right)\right)^{3/4}.$$

Theorem 7.2 (Asymptotic bias and variance of the NPCM estimator for the PD)**.** *Under Assumptions A.1-A.19, 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* $\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)$ *are the following:*

$$ABias\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)\Big)=\widetilde{B}_1(t,x)h_1^2+\widetilde{B}_2(t,x)h_2^2+o(h_1^2)+o(h_2^2) \tag{7.4}$$

(i) If $C_{h_1,h_2}:=\lim_{n\to\infty}\dfrac{h_1}{h_2}\in(0,\infty)$*, then*

$$\begin{aligned}AVar\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)\Big) \;=\;& \Big(\widetilde{V}_1(t+b,t,x)+C_{h_1,h_2}\widetilde{V}_2(t+b,t,x)\\ &+C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})\widetilde{V}_3(t+b,t,x)\Big)\frac{1}{nh_1}\\ &+o\left(\frac{1}{nh_1}\right)+O\left(\frac{h_1}{n}\right)\end{aligned}$$

(ii) If $\lim_{n\to\infty}\dfrac{h_1}{h_2}=0$*, then*

$$AVar\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)\Big) \;=\; \widetilde{V}_1(t+b,t,x)\frac{1}{nh_1}+o\left(\frac{1}{nh_1}\right)+O\left(\frac{h_2}{n}\right)$$

(iii) If $\lim_{n\to\infty}\dfrac{h_2}{h_1}=0$*, then*

$$AVar\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)\Big) \;=\; \widetilde{V}_2(t+b,t,x)\frac{1}{ng}+o\left(\frac{1}{nh_2}\right)+O\left(\frac{h_1}{n}\right),$$

where

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

with $i=1,2,3$ *and* $\widetilde{c}_K$ *is defined in* (7.2).

Theorem 7.3 (Asymptotic normality of the NPCM estimator for the PD)**.** *Under Assumptions A.1-A.19, for fixed values* $(t,x),(t+b,x)\in[l,u]\times I$*, the limit distribution of* $\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)$ *is the following:*

(i) Assuming $C_{h_1}:=\lim_{n\to\infty}n^{1/5}h_1\in(0,\infty)$*,* $C_{h_2}:=\lim_{n\to\infty}n^{1/5}h_2\in(0,\infty)$*, then*

$$\sqrt{nh_1}\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)-PD(t|x)\Big)\xrightarrow{d}N(\mu,s),$$

where

$$\mu=C_{h_1}^{5/2}\widetilde{B}_1(t,x)+C_{h_2}^{5/2}\widetilde{B}_2(t,x)$$

and

$$s^2=\Big(\widetilde{V}_1(t+b,t,x)+C_{h_1,h_2}\widetilde{V}_2(t+b,t,x)+C_{h_1,h_2}\widetilde{c}_K(_{h_1,h_2})\widetilde{V}_3(t+b,t,x)\Big).$$

(ii) Assuming $C_{h_2}:=\lim_{n\to\infty}n^{1/5}h_2\in(0,\infty)$ *and* $\lim_{n\to\infty}n^{1/5}h_1=0$*, then*

$$\sqrt{nh_1}\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)-PD(t|x)\Big)\xrightarrow{d}N(\mu,s),$$

where

$$\mu=C_{h_2}^{5/2}\widetilde{B}_2(t,x)$$

and

$$s^2=\widetilde{V}_1(t+b,t,x).$$

(iii) Assuming $C_{h_1}:=\lim_{n\to\infty}n^{1/5}h_1\in(0,\infty)$*,* $\lim_{n\to\infty}n^{1/5}h_2=0$*, then*

$$\sqrt{nh_2}\Big(\widehat{PD}^{NPCM}_{h_1,h_2}(t|x)-PD(t|x)\Big)\xrightarrow{d}N(\mu,s),$$

where

$$\mu=C_{h_1}^{5/2}\widetilde{B}_1(t,x)$$

,

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

and $\widetilde{V}_i(t_1, t_2, x)$, $i = 1, 2, 3$ *are defined in Theorem 7.2.*

Proofs of the results presented here are included in Section 7.6.

7.4 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 nonzero probability of cure and the proportion of cured subjects and the survival distribution of uncured subjects are modeled 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 fulfill 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. For more details, see Section 2.4.

The conditional censoring probability of the uncured population is as follows:

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

In this scenario,

$$\begin{aligned} P(\delta = 0|X = x) &= P(\delta = 0|\nu = 0, X = x)P(\nu = 0|X = x) \\ &\quad + P(\delta = 0|\nu = 1, X = x)P(\nu = 1|X = x). \end{aligned}$$

Since $P(\nu = 0|x) = p(x)$, $P(\nu = 1|x) = 1 - p(x)$ and

$$P(\delta = 0|\nu = 1, X = x) = P(C < T|T = \infty, X = x) = 1,$$

the expression of the censoring conditional probability is as follows:

$$P(\delta = 0|X = x) = 1 - p(x) + p(x)P(\delta = 0|\nu = 0, X = x) \tag{7.5}$$

The unconditional probability of censoring is given by

$$P(\delta = 0) = \int_{-\infty}^{+\infty} P(\delta = 0|X = x)m(x)dx, \tag{7.6}$$

and the probability of censoring in Model 1 is 0.77151.

The conditional survival function and the probability of default are the following:

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

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

Figure 7.1 shows the incidence and latency functions in Model 1. The incidence function is decreasing: the higher the value of the scoring x, the lower the value of $p(x)$ and the higher the probability of being cured of the event falling into default. This is consistent in the context of credit risk.

Figure 7.2 shows the theoretical conditional survival function and the probability of default in Model 1. The non-zero tendency of the survival function under this cure model is clear. This plateau in the right tail informs about the proportion of cured individuals in this model.

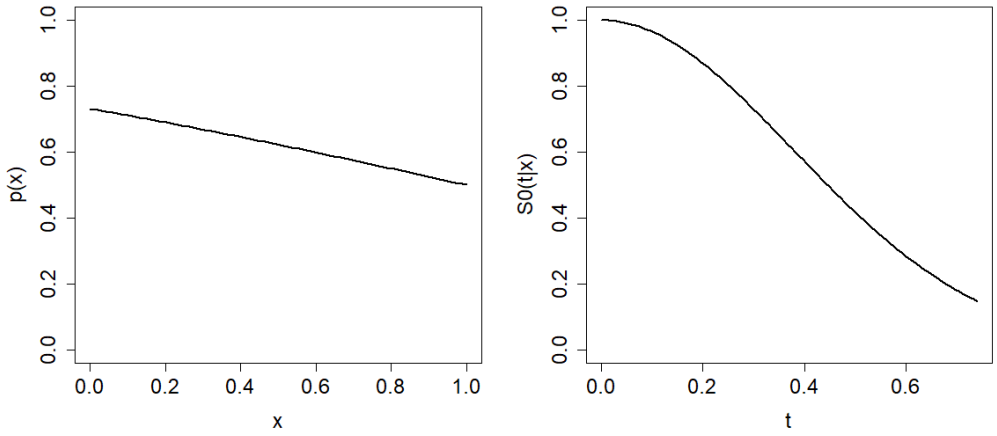

Figure 7.1: Theoretical incidence $p(x)$ (left) and latency $S_0(t|x = 0.5)$ (right) of Model 1.

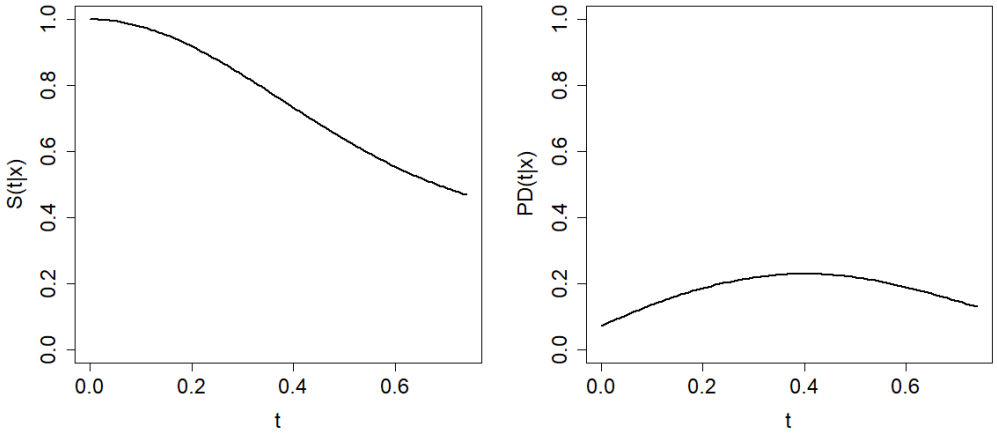

Figure 7.2: Theoretical conditional survival function $S(t|x = 0.5)$ (left) and probability of default $PD(t|x = 0.5)$ (right) of Model 1.

The conditional survival function and the probability of default of Model 1 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 about 20% of the time grid. For $x = 0.5$, $b = 0.18503$ and $t_{n_T} = 0.74013$.

In Model 2, the incidence is given by

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3)} \tag{7.7}$$

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 $Q(x) = 2 + 58x - 160x^2 + 107x^3$,

$$T_0|_{X=x} \sim Exp\Big(Q(x)\Big),$$

and the censoring time conditional to the credit scoring, $C_0|_{X=x}$, follows an exponential distribution with parameter $R(x) = 10 - \frac{55}{2}x + 20x^2$,

$$C_0|_{X=x} \sim Exp\Big(R(x)\Big).$$

The latency is given by

$$S_0(t|x) = e^{-Q(x)t}.$$

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. For more details, see the explanations in Section 2.4.

In this scenario, the conditional censoring probability of the uncured population results in:

$$P(\delta = 0|\nu = 0, X = x) \quad = \quad \frac{R(x)}{Q(x) + R(x)}.$$

The conditional survival function and the probability of default are the following:

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

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

Figure 7.3 shows the incidence and latency functions for Model 2. The decreasing incidence function is consistent with a credit risk context, since the higher the value

of the scoring x, the higher the probability of being cured of the event falling into default.

Figure 7.4 shows the theoretical conditional survival function and the probability of default in Model 2. The plateau in the right tail of the survival function under this cure model is clear.

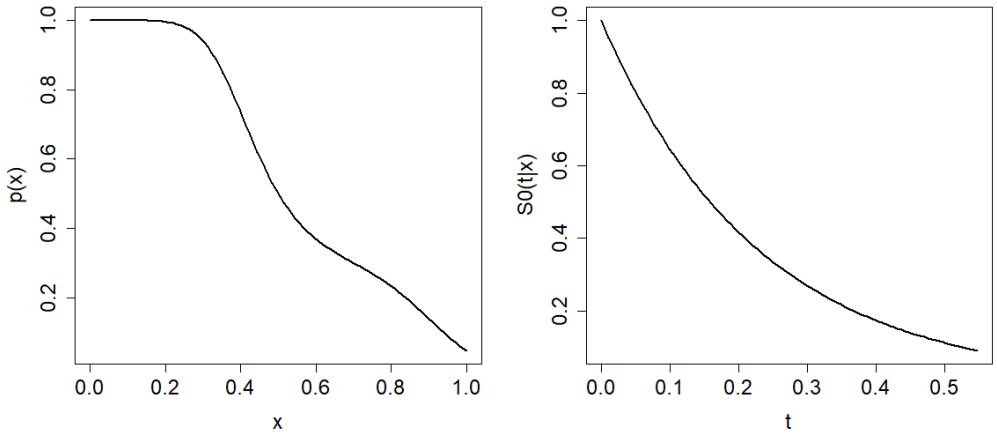

Figure 7.3: Theoretical incidence $p(x)$ (left) and latency $S_0(t|x = 0.5)$ (right) of Model 2.

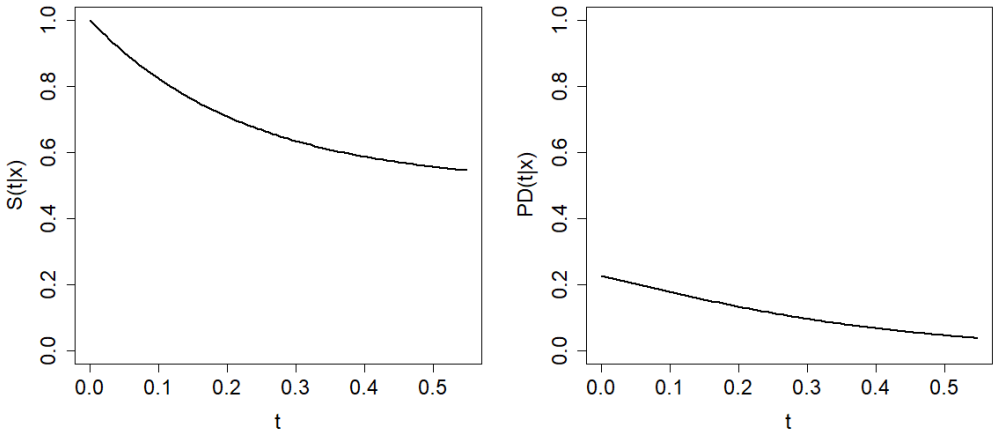

Figure 7.4: Theoretical conditional survival function $S(t|x = 0.5)$ (left) and probability of default $PD(t|x = 0.5)$ (right) of Model 2.

The conditional survival function and the probability of default of Model 2 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. For the previously set parameters and $x = 0.5$, one has $b = 0.13695$ (20% of the grid range) and $t_{n_T} = 0.54779$.

In Model 3, the incidence is given by (7.7) 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

$$S_0(t|x) = e^{-(B_1(x)t)^{k_1(x)}}.$$

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 paremeter of the Weibull distribution, $k_1(x)$, depends on x.

In this scenario, the conditional censoring probability of the uncured population is as follows:

$$P(\delta = 0|\nu = 0, X = x) \;=\; \frac{B_1(x)}{B_1(x) + B_2(x)}.$$

and the conditional survival function and the probability of default are the following:

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

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

The simulation analysis is conducted for different credit scoring values in each model. The probability of cured, the unconditional probability of censoring and the

probabilities of censoring conditional on each chosen value of x are shown in Table 7.1 for Models 1, 2 and 3.

		Model 1	Model 2	Model 3
	$P(\delta = 0)$	0.771510	0.656636	0.706833
$x = 0.2$	$1 - p(x)$	0.310026	0.004022	0.000014
	$P(\delta = 0\|x)$	0.835720	0.399251	0.483227
$x = 0.5$	$1 - p(x)$	0.377541	0.500000	0.500031
	$P(\delta = 0\|x)$	0.709519	0.611111	0.745433
$x = 0.8$	$1 - p(x)$	0.450166	0.767098	0.743473
	$P(\delta = 0\|x)$	0.730474	0.884726	0.870492

Table 7.1: Unconditional and conditional probabilities of censoring in Models 1, 2 and 3.

For comparison purposes, Beran's estimator and the smoothed Beran's estimator are included in this simulation study. They do not consider a priori the existence of a cured population, but may be able to detect a nonzero probability of cure from the sample data. The smoothed Beran's estimator showed a better performance than Beran's estimator in Chapter 5. In this section its performance is compared with the NPCM estimator.

Two other semiparametric 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 $1 - p(x)$ fiting a logistic model and the latency $S_0(t|x)$ fiting a proportional hazards model and an accelerated failure time model, respectively. The details of the methods can be consulted in Sy and Taylor (2000) and Sy and Taylor (2001).

Model 1 fits Cox and AFT cure models with logistic cure probability, meanwhile Model 2 and 3 move away from these semiparametric 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 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)). The semiparametric methods are implemented in the R-Package *smcure* (see Cai et al. (2012)).

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 in $[-50, 50]$ is used for the covariable smoothing in Beran's estimator.

The optimal value of the bandwidth involved in Beran's estimator, h_{MISE}, is chosen as the value that minimises a Monte Carlo approximation of the 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 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 $MISE_x(h_{MISE})$.

The optimal bivariate bandwidth (h_{MISE}, g_{MISE}) involved in the smoothed Beran's estimator is chosen, from a meshgrid of 50×50 values of (h, g), as the pair that minimises a Monte Carlo approximation of the MISE given by

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

based on $N = 100$ simulated samples. Then, $N = 300$ simulated samples are used to approximate $MISE_x(h_{MISE}, g_{MISE})$.

The optimal bivariate bandwidth (h_1^{MISE}, h_2^{MISE}) involved in the NPCM estimator is chosen, from a meshgrid of 50×50 values of (h_1, h_2), as the pair that minimises a Monte Carlo approximation of the MISE given by

$$MISE_x(h_1, h_2) = E\left(\int \left(\widehat{PD}_{h_1,h_2}^{NPCM}(t|x) - PD(t|x)\right)^2 dt\right)$$

based on $N = 100$ simulated samples. Then, $N = 300$ simulated samples are used to approximate $MISE_x(h_1^{MISE}, h_2^{MISE})$.

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. The value of $MISE$ and its square root, $RMISE$, are used as a measure of the estimation error of the PD estimators.

Figure 7.5 shows the contour plots of the $MISE_x(h, g)$ function for the NPCM estimator in Models 1, 2 and 3 when $x = 0.5$.

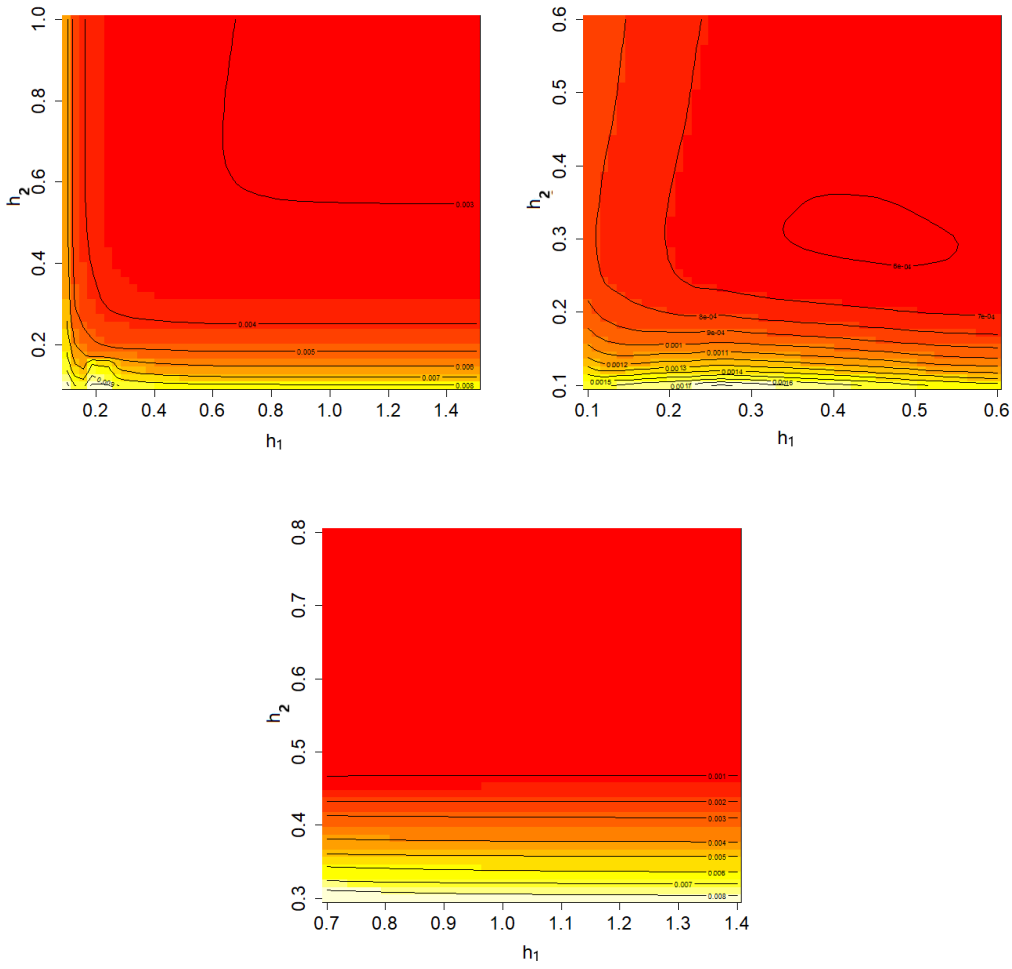

Figure 7.5: Contour plots of the approximation of $MISE_x(h)$ function of NPCM estimator using $N = 1000$ samples from Model 1 (top left), Model 2 (top right) and Model 3 (bottom) when $x = 0.5$.

Tables 7.2-7.4 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 $x = 0.8$. In some of the scenarios analysed, the MISE function turned out to be a decreasing function of the covariate bandwidth. For this reason, the MISE bandwidth selected was a high but reasonable value, considering that the variable X moves in the interval $[0, 1]$. This is the case of the optimal of Beran's estimator and the NPCM estimator in Model 1.

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 the 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. The smoothed Beran's estimator provides the smaller error in all scenarios. Its good performance is remarkable.

Beran's and the smoothed Beran's estimators make no assumptions about the survival function, but use 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. This is also confirmed by Figures 7.6, 7.7 and 7.8.

		SBeran	Beran	NPCM	PHCM	AFTCM
$x = 0.2$	Bandwidths	(1.000000, 0.853061)	0.522449	(0.926531, 0.871429)	—	—
	$RMISE_x$	0.047494	0.135059	0.134939	0.139143	0.096897
$x = 0.5$	Bandwidths	(0.155102, 0.632653)	0.559184	(1.000000, 0.724490)	—	—
	$RMISE_x$	0.042042	0.058921	0.058925	0.054809	0.050675
$x = 0.8$	Bandwidths	(0.320408, 0.111020)	0.430612	(1.000000, 0.687755)	—	—
	$RMISE_x$	0.028910	0.037749	0.037591	0.045671	0.045183

Table 7.2: Optimal bandwidths and $RMISE$ of the probability of default estimators when $x = 0.2$, $x = 0.5$ and $x = 0.8$ in Model 1.

		SBeran	Beran	NPCM	PHCM	AFTCM
$x = 0.2$	Bandwidths	(0.069388, 0.088775)	0.108163	(0.127551, 0.375510)	—	—
	$RMISE_x$	0.050586	0.089049	0.076577	0.093894	0.102628
$x = 0.5$	Bandwidths	(0.263265, 0.146939)	0.185714	(0.457143, 0.302041)	—	—
	$RMISE_x$	0.011431	0.025038	0.025178	0.029877	0.030471
$x = 0.8$	Bandwidths	(0.108163, 0.961225)	0.146939	(0.263265, 0.632653)	—	—
	$RMISE_x$	0.046706	0.066779	0.055069	0.051907	0.052058

Table 7.3: Optimal bandwidths and $RMISE$ of the probability of default estimators when $x = 0.2$, $x = 0.5$ and $x = 0.8$ in Model 2.

		SBeran	Beran	NPCM	PHCM	AFTCM
$x = 0.2$	Bandwidths	(0.059184, 0.090816)	0.077551	(0.160204, 0.178571)	—	—
	$RMISE_x$	0.060799	0.068428	0.068027	0.101598	0.159328
$x = 0.5$	Bandwidths	(0.320408, 0.030204)	0.353061	(0.928571, 0.565306)	—	—
	$RMISE_x$	0.022604	0.023605	0.023751	0.029312	0.058107
$x = 0.8$	Bandwidths	(0.191867, 0.050408)	0.178571	(0.614286, 0.814286)	—	—
	$RMISE_x$	0.015308	0.016911	0.025107	0.028312	0.046132

Table 7.4: Optimal bandwidths and $RMISE$ of the probability of default estimators when $x = 0.2$, $x = 0.5$ and $x = 0.8$ in Model 3.

Figures 7.6, 7.7 and 7.8 show the survival and PD estimations obtained by Beran's estimator, the smoothed Beran's estimator and the NPCM estimator in one sample from Models 1, 2 and 3, respectively, using the optimal MISE bandwidths from Tables 7.2, 7.3 and 7.4 with $x = 0.5$. All estimators, including Beran's one, capture the nonzero tendency of survival, which is a characteristic of cure models.

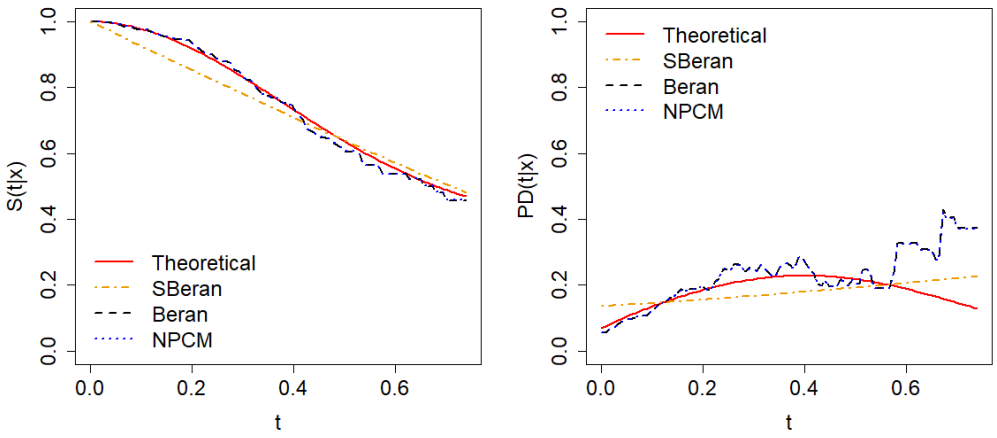

Figure 7.6: Theoretical survival function and estimations (left) and probability of default curve and estimations (right). Theoretical curve (solid line), smoothed Beran's estimation (dash-dotted line), Beran's estimation (dashed line) and NPCM estimation (dotted line) using MISE bandwidths in one sample from Model 1 when $x = 0.5$.

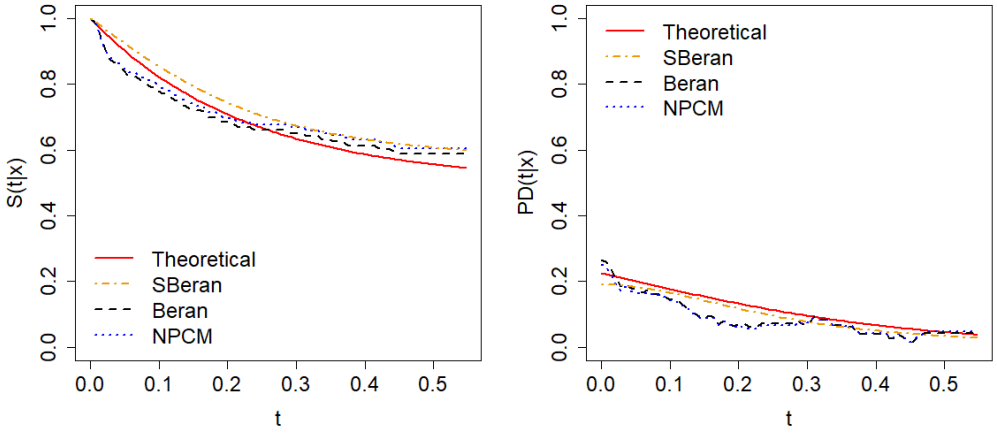

Figure 7.7: Theoretical survival function and estimations (left) and probability of default curve and estimations (right). Theoretical curve (solid line), smoothed Beran's estimation (dash-dotted line), Beran's estimation (dashed line) and NPCM estimation (dotted line) using MISE bandwidths in one sample from Model 2 when $x = 0.5$.

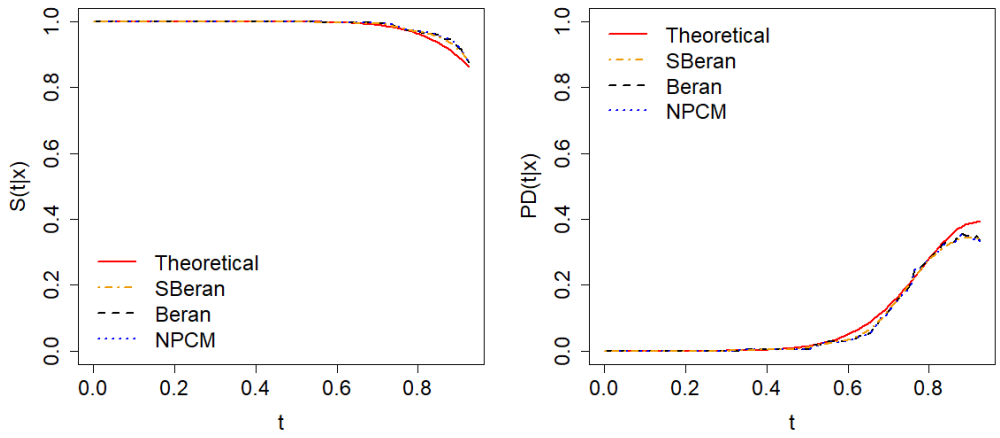

Figure 7.8: Theoretical survival function and estimations (left) and probability of default curve and estimations (right). Theoretical curve (solid line), smoothed Beran's estimation (dash-dotted line), Beran's estimation (dashed line) and NPCM estimation (dotted line) using MISE bandwidths in one sample from Model 3 when $x = 0.5$.

Since computation time is an important aspect to be considered in the comparison of the estimators, a small study of CPU time is addressed in this section. Table 7.5 shows the CPU times in seconds needed to estimate the PD for a single sample of different sizes with the four studied estimators. Table 7.6 shows the CPU times in seconds needed to approximate the optimal bandwidths to estimate the PD from $N = 100$ simulated samples of different sizes with Beran's estimator and the NPCM estimator. The estimators based on PH cure model and AFT cure model do not depend on any smoothing parameter.

According to Table 7.5, Beran's 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. The smoothed Beran's estimator is somewhat more time consuming than Beran's and NPCM estimators. The semiparametric methods are slower; in particular, the AFTCM estimator. However, the optimal bandwidth approximation is what slows down nonparametric methods as opposed to semiparametric methods, which do not depend on bandwidth parameters. Table 7.6 shows the computation times required to obtain the MISE bandwidths for the nonparametric estimators.

In practice, these bandwidths will be obtained by some resampling technique and the time required will be different from the one shown here. However, it gives an idea of the disadvantage of the smoothed Beran's estimator that depends on one smoothing bandwidth in the covariate and one in the time variable. The same applies to the NPCM estimator which depends on two bandwidths for the covariate. The error functions in both cases must be minimised in two dimensions to obtain these bandwidths.

n	100	400	800	1600	2400
Beran	0.01	0.01	0.01	0.02	0.02
SBeran	0.03	0.05	0.07	0.18	0.32
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 7.5: CPU time (in seconds) for the estimation of $PD(t|x)$ in a time grid of size 100 and $x = 0.5$ for one sample of size n with Beran's estimator, the smoothed Beran's estimator, the NPCM estimator, the PHCM estimator and the AFTCM estimator.

n	100	400	800	1600	2400
Beran	5.01	4.14	13.34	33.65	44.92
SBeran	53.55	197.85	415.50	954.37	3020.22
NPCM	20.44	65.38	35.03	94.97	37.76

Table 7.6: CPU time (in seconds) for the approximation of the optimal bandwidth from $N = 100$ samples of size n to estimate $PD(t|x)$ in a time grid of size 100 and $x = 0.5$ with Beran's estimator, the smoothed Beran's estimator and the NPCM estimator.

7.5 Application to real data

In this section we apply the above PD estimators to the German Credit data set which was previously analysed in Section 6.6. 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 credit scoring (X) and the default indicator (δ).

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 survival function and the probability of default conditional on the credit scoring are estimated using the estimators presented in the simulation study and the result is shown in Figure 7.9. The estimations of these curves are obtained at $x = 0.85$ through empirically chosen bandwidths based on visual inspection of the PD curves and considering the ranges in which the variables lie: $h = 0.5$ for Beran's estimator, $(h, g) = (0.5, 5)$ for the smoothed Beran's estimator and $(h_1, h_2) = (0.3, 0.5)$ for the NPCM estimator.

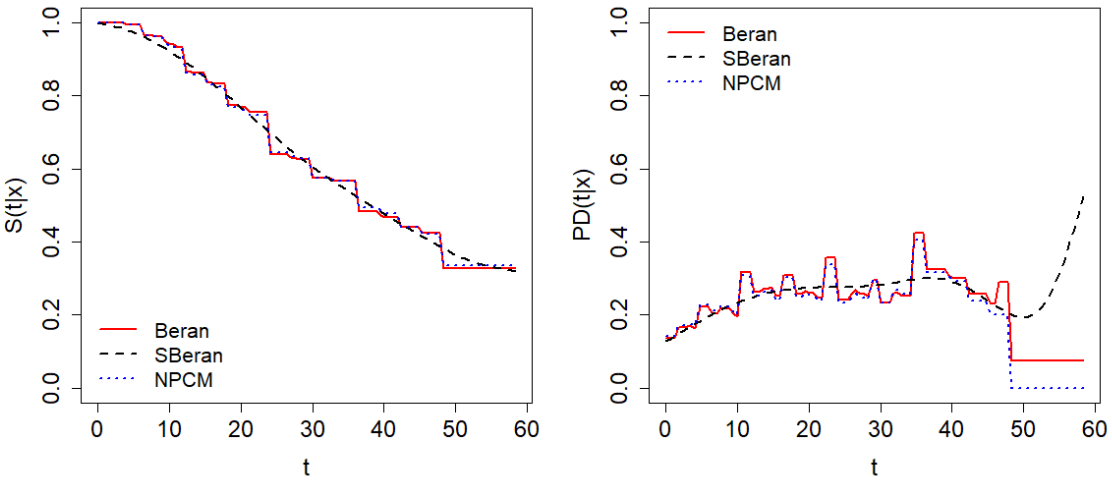

Figure 7.9: Estimation of the conditional survival function (left) and the probability of default (right) for $x = 0.85$ by means of Beran's estimator (solid line), the smoothed Beran's estimator (dashed line) and the NPCM estimator (dotted line) in the German credit data set.

7.6 Proofs

Lemma 7.2. *Denote* $\Phi_\xi(u,t,x) = E\Big[\xi(Z,\delta,t,x)|X=u\Big]$ *with* $\xi(Z,\delta,t,x)$ *defined in Section 7.3. Under Assumptions A.8 and A.19, then*

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

Proof.

Using a Taylor expansion for $\Phi_\xi(u,t,x)m(u)$ when $u = x - hv$ around $u = x$ and Assumption A.8:

$$\begin{aligned}
&E\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t,x)\right]\\
&= E\left[K\left(\frac{x-X_1}{h}\right)E\Big[\xi(Z_1,\delta_1,t,x)\Big|X_1\Big]\right] = \int_{-\infty}^{+\infty} K\left(\frac{x-u}{h}\right)\Phi(u,t,x)m(u)du\\
&= (-h)\int_{+\infty}^{-\infty} K(v)\Phi(x-hv,t,x)m(x-hv)dv\\
&= \int_{-\infty}^{+\infty} hK(v)\Bigg(\Phi(x,t,x)m(x) - hv\frac{\partial}{\partial u}\Big(\Phi(u,t,x)m(u)\Big)\Big|_{u=x}\\
&\quad + \frac{h^2v^2}{2}\frac{\partial^2}{\partial u^2}\Big(\Phi(u,t,x)m(u)\Big)\Big|_{u=x} + o(h^2)\Bigg)dv\\
&= \Phi(x,t,x)m(x)h + \frac{d_K}{2}\frac{\partial^2}{\partial u^2}\Big(\Phi(u,t,x)m(u)\Big)\Big|_{u=x}h^3 + o(h^3).
\end{aligned}$$

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

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

Lemma 7.3. *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 7.3. Under Assumptions A.8 and A.19, then*

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

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.8:

$$E\left[K^2\left(\frac{x-X_1}{h}\right)\xi^2(Z_1,\delta_1,t,x)\right] = \int_{-\infty}^{+\infty} K^2\left(\frac{x-u}{h}\right)\Phi_2(u,t,x)m(u)du$$

$$= c_K\Phi_2(x,t,x)m(x)h + \frac{d_{K^2}}{2}\frac{\partial^2}{\partial u^2}\Big(\Phi_2(u,t,x)m(u)\Big)\Big|_{u=x}h^3 + o(h^3).$$

From Lemma 7.2,

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

Then,

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

$$= c_K\Phi_2(x,t,x)m(x)h + \frac{d_{K^2}}{2}\frac{\partial^2}{\partial u^2}\Big(\Phi_2(u,t,x)m(u)\Big)\Big|_{u=x}h^3 + o(h^3).$$

□

Lemma 7.4. *Denote* $D_\xi(u,t_1,t_2,x) = Cov\Big[\xi(Z_1,\delta_1,t_1,x),\xi(Z_1,\delta_1,t_2,x)|X_1 = u\Big]$ *and* $B(u,t_1,t_2,x) = \Phi_\xi(u,t_1,x)\Phi_\xi(u,t_2,x)m(u)$. *Under Assumptions A.8 and A.19, then*

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

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

Proof.

Using the Law of total covariance,

$$
\begin{aligned}
& Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_2,x)\right] \\
= \; & E\left[Cov\left[K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h}\right)\xi(Z_1,\delta_1,t_2,x)\Big|X_1\right]\right] \\
& +E\left[K^2\left(\frac{x-X_1}{h}\right)\Phi_\xi(X_1,t_1,x)\Phi_\xi(X_1,t_2,x)\right] \\
& -E\left[K\left(\frac{x-X_1}{h}\right)\Phi_\xi(X_1,t_1,x)\right]E\left[K\left(\frac{x-X_1}{h}\right)\Phi_\xi(X_1,t_2,x)\right] = S_1+S_2-S_3.
\end{aligned} \tag{7.8}
$$

Using a Taylor expansion for $D_\xi(u,t_1,t_2,x)m(u)$ when $u=x-hv$ around $u=x$ and Assumption A.8:

$$
\begin{aligned}
S_1 &= \int_{-\infty}^{+\infty} K^2\left(\frac{x-u}{h}\right)Cov\Big[\xi(Z_1,\delta_1,t_1,x),\xi(Z_1,\delta_1,t_2,x)|X_1=u\Big]m(u)du \\
&= (-h)\int_{+\infty}^{-\infty} K^2(v)D_\xi(x-hv,t_1,t_2,x)dv \\
&= h\int_{-\infty}^{+\infty} K(v)\Big(D_\xi(x,t_1,t_2,x)-hvD'_\xi(x,t_1,t_2,x) \\
&\quad +\frac{h^2v^2}{2}D''_\xi(x,t_1,t_2,x)+o(h^2)\Big)dv \\
&= c_K D_\xi(x,t_1,t_2,x)h+\frac{d_{K^2}}{2}D''_\xi(x,t_1,t_2,x)h^3+o(h^3).
\end{aligned}
$$

Using a Taylor expansion for $B(u,t_1,t_2,x)$ when $u=x-hv$ around $u=x$ and Assumption A.8:

$$
\begin{aligned}
S_2 &= \int_{-\infty}^{+\infty} K^2\left(\frac{x-u}{h}\right)\Phi(u,t_1,x)\Phi(u,t_2,x)m(u)du \\
&= \int_{-\infty}^{+\infty} K^2\left(\frac{x-u}{h}\right)B(u,t_1,t_2,x)du \\
&= \int_{+\infty}^{-\infty}(-h)K^2(v)B(x-hv,t_1,t_2,x)dv=\int_{-\infty}^{+\infty} hK^2(v)B(x-hv,t_1,t_2,x)dv \\
&= \int_{-\infty}^{+\infty} hK^2(v)\Big(B(x,t_1,t_2,x)-hvB'(x,t_1,t_2,x) \\
&\quad +\frac{h^2v^2}{2}B''(x,t_1,t_2,x)+o(h^2)\Big)dv \\
&= c_K B(x,t_1,t_2,x)h+\frac{d_{K^2}}{2}B''(x,t_1,t_2,x)h^3+o(h^3).
\end{aligned}
$$

Since $\Phi(x,t,x)=0 \quad \forall (t,x)\in[0,\infty)\times I$, $B(x,t_1,t_2,x)=0$ for all $t_1,t_2\in[0,\infty)$. Then,

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

Finally, from Lemma 7.2,

$$E\left[K\left(\frac{x-X_1}{h}\right)\Phi_\xi(X_1,t,x)\right]=O(h^3).$$

Then, $S_3=O(h^6)$, and replacing S_1, S_2 and S_3 in (7.8), the lemma is proved.

□

Proof of Lemma 7.1

Let us denote $\widehat{S}_{h_1,h_2}(t|x) := \widehat{S}^{NPCM}_{h_1,h_2}(t|x)$. According to the definition of the NPCM estimator in (7.1),

$$\begin{aligned}
&\widehat{S}_{h,g}(t|x)-S(t|x)\\
&= 1-\widehat{p}_h(x)+\widehat{p}_{h_1}(x)\widehat{S}_{0,h_2}(t|x)-\Big(1-p(x)+p(x)S_0(t|x)\Big)\\
&= p(x)-\widehat{p}_{h_1}(x)+\widehat{p}_{h_1}(x)\widehat{S}_{0,h_2}(t|x)-p(x)S_0(t|x)+\widehat{p}_{h_1}(x)S_0(t|x)-\widehat{p}_{h_1}(x)S_0(t|x)\\
&= p(x)-\widehat{p}_{h_1}(x)+S_0(t|x)\Big(\widehat{p}_{h_1}(x)-p(x)\Big)+\widehat{p}_{h_1}(x)\Big(\widehat{S}_{0,h_2}(t|x)-S_0(t|x)\Big)\\
&= \Big(S_0(t|x)-1\Big)\Big(\widehat{p}_{h_1}(x)-p(x)\Big)+\Big(\widehat{p}_{h_1}(x)+p(x)-p(x)\Big)\Big(\widehat{S}_{0,h_2}(t|x)-S_0(t|x)\Big)\\
&= \Big(S_0(t|x)-1\Big)\Big(\widehat{p}_{h_1}(x)-p(x)\Big)+p(x)\Big(\widehat{S}_{0,h_2}(t|x)-S_0(t|x)\Big)\\
&\quad +\Big(\widehat{p}_{h_1}(x)-p(x)\Big)\Big(\widehat{S}_{0,h_2}(t|x)-S_0(t|x)\Big)
\end{aligned} \tag{7.9}$$

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:

$$\widehat{p}_{h_1}(x)-p(x)=\Big(p(x)-1\Big)\sum_{i=1}^{n}w^A_{h_1,i}(x)\xi(Z_i,\delta_i,\infty,x)+R_n(x), \tag{7.10}$$

$$\widehat{S}_{0,h_2}(t|x)-S_0(t|x)=\sum_{i=1}^{n}w^A_{h_2,i}(x)\omega(Z_i,\delta_i,t,x)+R_n(t|x), \tag{7.11}$$

with

$$\sup_{x\in I}|R_n(x)| = O\left(\frac{\ln n}{nh_1}\right)^{3/4} \quad \text{a.s.} \quad \text{and} \quad \sup_{(t,x)\in[l,u]\times I}|R_n(t|x)| = O\left(\frac{\ln n}{nh_2}\right)^{3/4} \quad \text{a.s.}$$

Replacing (7.10) and (7.11) in (7.9), the almost sure representation of the NPCM survival estimator will follow from expression:

$$\begin{aligned}
&\widehat{S}_{h_1,h_2}(t|x) - S(t|x) \\
&= \Big(S_0(t|x)-1\Big)\Big(p(x)-1\Big)\sum_{i=1}^{n} w^A_{h_1,i}(x)\xi(Z_i,\delta_i,\infty,x) \\
&\quad + p(x)\sum_{i=1}^{n} w^A_{h_2,i}(x)\omega(Z_i,\delta_i,t,x) + \Big(S_0(t|x)-1\Big)R_n(x) + p(x)R_n(t|x) \\
&\quad + \Big(\widehat{p}_{h_1}(x)-p(x)\Big)\Big(\widehat{S}_{0,h_2}(t|x)-S_0(t|x)\Big).
\end{aligned}$$

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

$$\widehat{p}_{h_1}(x) - p(x) = O_p\left(\frac{1}{\sqrt{nh_1}}\right),$$

and

$$\widehat{S}_{0,h_2}(t|x) - S_0(t|x) = O_p\left(\frac{1}{\sqrt{nh_2}}\right).$$

Then,

$$\Big(\widehat{p}_{h_1}(x)-p(x)\Big)\Big(\widehat{S}_{0,h_2}(t|x)-S_0(t|x)\Big) = O_p\left(\frac{1}{n\sqrt{h_1h_2}}\right)$$

and

$$\begin{aligned}
\widehat{S}_{h_1,h_2}(t|x) - S(t|x) &= \Big(S_0(t|x)-1\Big)\Big(p(x)-1\Big)\sum_{i=1}^{n} w^A_{h_1,i}(x)\xi(Z_i,\delta_i,\infty,x) \\
&\quad + p(x)\sum_{i=1}^{n} w^A_{h_2,i}(x)\omega(Z_i,\delta_i,t,x) + R^1_n(t|x),
\end{aligned}$$

where

$$\begin{aligned}
R^1_n(t|x) &= \Big(S_0(t|x)-1\Big)R_n(x) + p(x)R_n(t|x) + O_p\left(\frac{1}{n\sqrt{h_1h_2}}\right) \\
&= O_p\left(\frac{\ln n}{nh_1}\right)^{3/4} + O_p\left(\frac{\ln n}{nh_2}\right)^{3/4} + O_p\left(\frac{1}{n\sqrt{h_1h_2}}\right)
\end{aligned}$$

Under Assumptions A.16 and A.17,

$$R_n^1(t|x) \;=\; O_p\left(\ln n\left(\frac{1}{nh_1}+\frac{1}{nh_2}\right)\right)^{3/4}$$

and Lemma 7.1 is proved.

□

Proof of Theorem 7.1

Let us denote $\widehat{PD}_{h_1,h_2}(t|x) := \widehat{PD}^{NPCM}_{h_1,h_2}(t|x)$ and $\widehat{S}_{h_1,h_2}(t|x) := \widehat{S}^{NPCM}_{h_1,h_2}(t|x)$. Consider the function

$$W_{h_1,h_2}(t,t+b,x)=\frac{S(t|x)\Big(\widehat{S}_{h_1,h_2}(t+b|x)-S(t+b|x)\Big)-S(t+b|x)\Big(\widehat{S}_{h_1,h_2}(t|x)-S(t|x)\Big)}{\widehat{S}_{h_1,h_2}(t|x)S(t|x)}.$$

Since

$$\frac{\widehat{S}_{h_1,h_2}(t+b|x)}{\widehat{S}_{h_1,h_2}(t|x)}-\frac{S(t+b|x)}{S(t|x)} \;=\; -\Big(\widehat{PD}_{h_1,h_2}(t|x)-PD(t|x)\Big)$$

and

$$\begin{aligned}
&\frac{\widehat{S}_{h_1,h_2}(t+b|x)}{\widehat{S}_{h_1,h_2}(t|x)}-\frac{S(t+b|x)}{S(t|x)}= \\
&= \frac{\widehat{S}_{h_1,h_2}(t+b|x)S(t|x)-S(t+b|x)\widehat{S}_{h_1,h_2}(t|x)-S(t+b|x)S(t|x)+S(t+b|x)S(t|x)}{\widehat{S}_{h_1,h_2}(t|x)S(t|x)} \\
&= \frac{S(t|x)\Big(\widehat{S}_{h_1,h_2}(t+b|x)-S(t+b|x)\Big)-S(t+b|x)\Big(\widehat{S}_{h_1,h_2}(t|x)-S(t|x)\Big)}{\widehat{S}_{h_1,h_2}(t|x)S(t|x)} \\
&= W_{h_1,h_2}(t,t+b,x)\left(\frac{\widehat{S}_{h_1,h_2}(t|x)}{S(t|x)}+1-\frac{\widehat{S}_{h_1,h_2}(t|x)}{S(t|x)}\right) \\
&= \frac{1}{S(t|x)}\Big(\widehat{S}_{h_1,h_2}(t+b|x)-S(t+b|x)\Big)-\frac{S(t+b|x)}{S^2(t|x)}\Big(\widehat{S}_{h_1,h_2}(t|x)-S(t|x)\Big) \\
&\quad +W_{h_1,h_2}(t,t+b,x)\left(1-\frac{\widehat{S}_{h_1,h_2}(t|x)}{S(t|x)}\right),
\end{aligned}$$

we have

$$\begin{aligned}
&\widehat{PD}_{h_1,h_2}(t|x)-PD(t|x) \\
&= a_1\Big(\widehat{S}_{h_1,h_2}(t+b|x)-S(t+b|x)\Big)+a_2\Big(\widehat{S}_{h_1,h_2}(t|x)-S(t|x)\Big) \\
&\quad +W_{h_1,h_2}(t,t+b,x)\left(\frac{\widehat{S}_{h_1,h_2}(t|x)-S(t|x)}{S(t|x)}\right)
\end{aligned} \tag{7.12}$$

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 $\widehat{S}_{h_1,h_2}(t+b|x)$ from Lemma 7.1 in (7.12) and considering the functions $\zeta_{n,i}(t|x)$ defined in the statement of Theorem 7.1, the almost sure representation of $\widehat{PD}_{h_1,h_2}(t|x)$ follows:

$$\begin{aligned}\widehat{PD}_{h_1,h_2}(t|x) - PD(t|x) &= a_1 \textstyle\sum_{i=1}^n \zeta_{n,i}(t+b|x) + a_2 \sum_{i=1}^n \zeta_{n,i}(t|x) + R_n^2(t|x) \\ &= \textstyle\sum_{i=1}^n \Psi_{n,i}(t,x) + R_n^2(t|x),\end{aligned} \tag{7.13}$$

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

$$\begin{aligned}R_n^2(t|x) &= -\frac{1}{S(t|x)}R_n^1(t+b|x) + \frac{S(t+b|x)}{S^2(t|x)}R_n^1(t|x) \\ &\quad + W_{h,g}(t,t+b,x)\left(\frac{\widehat{S}_{h_1,h_2}(t|x) - S(t|x)}{S(t|x)}\right).\end{aligned}$$

From Equation (7.3) in Lemma 7.1, we have $\widehat{S}_{h_1,h_2}(t|x) - S(t|x) = \tau_1 + \tau_2 + \tau_3$ where

$$\tau_1 = \Big(S_0(t|x) - 1\Big)\Big(p(x) - 1\Big)\sum_{i=1}^n w_{h_1,i}^A(x)\xi(Z_i, \delta_i, \infty, x),$$

$$\tau_2 = p(x)\sum_{i=1}^n w_{h_2,i}^A(x)\omega(Z_i, \delta_i, t, x),$$

$$\tau_3 = O_p\left(\ln n\left(\frac{1}{nh_1} + \frac{1}{nh_2}\right)\right)^{3/4}.$$

Lemmas 7.2 and 7.3 and straightforward but tedious calculations give

$$\tau_1 = O_p\left({h_1}^2 + \frac{1}{\sqrt{nh_1}}\right)$$

and

$$\tau_2 = O_p\left(h_2^2 + \frac{1}{\sqrt{nh_2}}\right).$$

Under Assumption A.18, τ_3 is proved to be negligible with respect to τ_1 and τ_2. Then,

$$W_{h_1,h_2}(t,t+b,x)\left(\frac{\widehat{S}_{h_1,h_2}(t|x) - S(t|x)}{S(t|x)}\right) = O_p\left(h_1^4 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2}\right).$$

Therefore,

$$R_n^2(t|x) = O_p\left(\ln n \left(\frac{1}{nh_1} + \frac{1}{nh_2} \right) \right)^{3/4} + O_p\left(h_1^4 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2} \right).$$

Using Assumptions A.16 and A.17, the second term in $R_n^2(t|x)$ is negligible with respect to the first one and Theorem 7.1 is proved.

□

Proof of Theorem 7.2

According to the almost sure representation of $\widehat{PD}_{h_1,h_2}(t|x) := \widehat{PD}_{h_1,h_2}^{NPCM}(t|x)$, the asymptotic expression for the bias is obtained from its dominant term. Then,

$$\begin{aligned} E\left[\sum_{i=1}^{n} \Psi_{n,i}(t,x) \right] &= \sum_{i=1}^{n} E\Big[\Psi_{n,i}(t,x) \Big] = nE\Big[\Psi_{n,1}(t,x) \Big] \\ &= na_1 E\Big[\zeta_{n,1}(t+b,x) \Big] + na_2 E\Big[\zeta_{n,1}(t,x) \Big]. \end{aligned} \tag{7.14}$$

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

The expression of $E\Big[\zeta_{n,1}(t,x) \Big]$ in (7.14) is then calculated:

$$\begin{aligned} E\Big[\zeta_{n,1}(t,x) \Big] &= \Big(S_0(t|x) - 1 \Big) \Big(p(x) - 1 \Big) E\Big[w_{h_1,1}^A(x) \xi(Z_1, \delta_1, \infty, x) \Big] \\ &\quad + p(x) E\Big[w_{h_2,1}^A(x) \omega(Z_1, \delta_1, t, x) \Big] \\ &= \Big(S_0(t|x) - 1 \Big) \Big(p(x) - 1 \Big) \frac{1}{nh_1 m(x)} E\left[K\left(\frac{x - X_1}{h_1} \right) \xi(Z_1, \delta_1, \infty, x) \right] \\ &\quad - S(t|x) \frac{1}{nh_2 m(x)} E\left[K\left(\frac{x - X_1}{h_2} \right) \xi(Z_1, \delta_1, t, x) \right] \\ &\quad - \frac{\Big(1 - p(x) \Big) \Big(1 - S(t|x) \Big)}{p(x)} \frac{1}{nh_2 m(x)} E\left[K\left(\frac{x - X_1}{h_2} \right) \xi(Z_1, \delta_1, \infty, x) \right]. \end{aligned}$$

Using Lemmas 7.2 and 7.3:

$$E\Big[\zeta_{n,1}(t,x) \Big] = B_1(t,x) \frac{{h_1}^2}{n} + B_2(t,x) \frac{h_2^2}{n} + o\left(\frac{{h_1}^2}{n} \right) + o\left(\frac{h_2^2}{n} \right). \tag{7.15}$$

Replacing expression (7.15) in (7.14), the bias part of the theorem is proved:

$$E\left[\sum_{i=1}^{n} \Psi_{n,i}(t,x) \right] = \widetilde{B}_1(t,x) {h_1}^2 + \widetilde{B}_2(t,x) g^2 + o({h_1}^2) + o(h_2^2),$$

where $\widetilde{B}_1(t,x)$ and $\widetilde{B}_1(t,x)$ were defined in Section 7.3.

The asymptotic expression for the variance of $\widehat{PD}_{h_1,h_2}(t|x)$ is obtained from the variance of the dominant term of its almost sure representation:

$$\begin{aligned} Var\Big[\sum_{i=1}^{n}\Psi_{n,i}(t,x)\Big] &= \sum_{i=1}^{n} Var\Big[\Psi_{n,1}(t,x)\Big] = nVar\Big[\Psi_{n,1}(t,x)\Big] \\ &= na_1^2 Var\Big[\zeta_{n,1}(t+b,x)\Big] + na_2^2 Var\Big[\zeta_{n,1}(t,x)\Big] \\ &\quad +2na_1a_2 Cov\Big[\zeta_{n,1}(t+b,x),\zeta_{n,1}(t,x)\Big]. \end{aligned} \tag{7.16}$$

Consider

$$A_1 = Var\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,\infty,x)\right],$$

$$A_2 = Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,\infty,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right], \tag{7.17}$$

$$A_3 = Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,\infty,x)\right] \tag{7.18}$$

and

$$A_4 = Cov\left[K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right].$$

To find the asymptotic expression for $Cov\Big[\zeta_{n,1}(t+b,x),\zeta_{n,1}(t,x)\Big]$, some calculations lead to

$$\begin{aligned} &Cov\Big[\zeta_{n,1}(t_1,x),\zeta_{n,1}(t_2,x)\Big] \\ &= \Big(S_0(t_1|x)-1\Big)\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)^2\frac{1}{n^2{h_1}^2m^2(x)}A_1 \\ &\quad +\Big(S_0(t_1|x)-1\Big)\Big(p(x)-1\Big)p(x)\frac{1}{n^2h_1h_2m^2(x)}A_2 \\ &\quad +\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)p(x)\frac{1}{n^2h_1h_2m^2(x)}A_3 + p^2(x)\frac{1}{n^2h_2^2m^2(x)}A_4. \end{aligned} \tag{7.19}$$

First, from Lemma 7.3,

$$A_1 = h_1\Phi_2(x,\infty,x)m(x)c_K + O({h_1}^3). \tag{7.20}$$

Second, using Lemmas 7.3 and 7.4,

$$A_4 \;=\; C_1(t_1,t_2,x)h_2 + O(h_2^3). \tag{7.21}$$

where $C_1(t_1,t_2,x)$ is defined in Section 7.3.

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

$$Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right]$$

is obtained by distinguishing three different cases:

(i) If $C_{h_1,h_2} := \lim_{n\to\infty} \frac{h_1}{h_2} \in (0,\infty)$:

$$\begin{aligned} &Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right] \\ &\quad\simeq\; Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_1/C_{h_1,h_2}}\right)\omega(Z_1,\delta_1,t_2,x)\right] \\ &\quad=\; S_1 + S_2 - S_3. \end{aligned}$$

where

$$S_1 = E\left[Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_1/C_{h_1,h_2}}\right)\omega(Z_1,\delta_1,t_2,x)\Big|X_1\right]\right],$$

$$S_2 = E\left[K\left(\frac{x-u}{h_1}\right)K\left(C_{h_1,h_2}\frac{x-u}{h_1}\right)\Phi_\xi(X_1,t_1,x)\Phi_\omega(X_1,t_2,x)\right]$$

and

$$S_3 = E\left[K\left(\frac{x-X_1}{h_1}\right)\Phi_\xi(X_1,t_1,x)\right]E\left[K\left(C_{h_1,h_2}\frac{x-u}{h_1}\right)\Phi_\omega(X_1,t_2,x)\right].$$

Considering the function $D_{\xi,\omega}(u,t_1,t_2,x)$ defined in Section 7.3 and its Taylor expansion when $u = x - hv$ around $u = x$:

$$\begin{aligned} S_1 \;&=\; \int_{-\infty}^{+\infty} K\left(\frac{x-u}{h_1}\right)K\left(C_{h_1,h_2}\frac{x-u}{h_1}\right)D_{\xi,\omega}(u,t_1,t_2,x)du \\ &=\; h_1\int_{-\infty}^{+\infty} K(v)K(C_{h_1,h_2}v)\Big(D_{\xi,\omega}(x,t_1,t_2,x) \\ &\quad -h_1 v D'_{\xi,\omega}(x,t_1,t_2,x) + O({h_1}^2)\Big)dv. \end{aligned}$$

Since K is symmetric, $K(C_{h_1,h_2}v) = K(-C_{h_1,h_2}v)$ and the function $K(v)K(C_{h_1,h_2}v)$ is also even. Consequently,

$$\int_{-\infty}^{+\infty} K(v)K(C_{h_1,h_2}v)v dv = 0.$$

Then,

$$S_1 \;=\; \widetilde{c}_K(C_{h_1,h_2})D_{\xi,\omega}(x,t_1,t_2,x)h_1 + O({h_1}^3). \tag{7.22}$$

Defining $B_\omega(u,t_1,t_2,x) = \Phi_\xi(u,t_1,x)\Phi_\omega(u,t_2,x)m(u)$ and using a Taylor expansion for $B_\omega(u,t_1,t_2,x)$ when $u = x - hv$ around $u = x$:

$$\begin{aligned}
S_2 &= \int_{-\infty}^{+\infty} K\left(\frac{x-u}{h_1}\right)K\left(C_{h_1,h_2}\frac{x-u}{h_1}\right)\Phi_\xi(u,t_1,x)\Phi_\omega(u,t_2,x)m(u)du \\
&= \int_{-\infty}^{+\infty} K\left(\frac{x-u}{h}\right)K\left(C_{h,g}\frac{x-u}{h}\right)B_\omega(u,t_1,t_2,x)du \\
&= \int_{-\infty}^{+\infty} hK\left(\frac{x-u}{h}\right)K\left(C_{h,g}\frac{x-u}{h}\right)B_\omega(x-hv,t_1,t_2,x)dv \\
&= \widetilde{c}_K(C_{h_1,h_2})B_\omega(x,t_1,t_2,x)h_1 + O({h_1}^3).
\end{aligned}$$

Since $\Phi_\xi(x,t,x) = 0$ for all $(t,x) \in [0,\infty) \times I$, then $B_\omega(x,t_1,t_2,x) = 0$ for all $t_1,t_2 \in [0,\infty)$, $x \in I$, and, consequently:

$$S_2 \;=\; O({h_1}^3). \tag{7.23}$$

From Lemma 7.2,

$$E\left[K\left(\frac{x-X_1}{h_1}\right)\Phi_\xi(X_1,t,x)\right] = O({h_1}^3).$$

Now, using a Taylor expansion for $\Phi_\omega(u,t,x)m(u)$ when $u = x - hv$ around $u = x$, gives

$$\begin{aligned}
&E\left[K\left(C_{h_1,h_2}\frac{x-X_1}{h_1}\right)\Phi_\omega(X_1,t,x)\right] \\
&\qquad = \left(\int_{-\infty}^{+\infty} K\Big(C_{h_1,h_2}v\Big)dv\right)\Phi_\omega(x,t,x)m(x)h_1 + O({h_1}^3).
\end{aligned}$$

Considering the definition of the function $\omega(Z,\delta,t,x)$, given in Section 7.3, and Lemma 7.2, $\Phi_\omega(x,t,x) = 0$ for all $(t,x) \in [0,\infty) \times I$ and

$$E\left[K\left(C_{h_1,h_2}\frac{x-X_1}{h_1}\right)\Phi_\omega(X_1,t,x)\right] = O({h_1}^3).$$

Therefore,

$$S_3 = O(h^6). \tag{7.24}$$

Using the expresions for S_1 in (7.22), S_2 in (7.23) and S_3 in (7.24),

$$Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right]$$
$$= \widetilde{c}_K(C_{h_1,h_2})D_{\xi,\omega}(x,t_1,t_2,x)h_1 + O({h_1}^3).$$

Therefore, recalling (7.17) and (7.18),

$$\begin{aligned} A_2 &= Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,\infty,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right] \\ &= \widetilde{c}_K(C_{h_1,h_2})D_{\xi,\omega}(x,\infty,t_2,x)h_1 + O({h_1}^3) \end{aligned} \tag{7.25}$$

and

$$\begin{aligned} A_3 &= Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x), K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,\infty,x)\right] \\ &= \widetilde{c}_K(C_{h_1,h_2})D_{\xi,\omega}(x,t_1,\infty,x)h_1 + O({h_1}^3). \end{aligned} \tag{7.26}$$

Replacing (7.20), (7.21), (7.25) and (7.26) in (7.19) and assuming $\lim_{n\to\infty}\frac{h_1}{h_2} = C_{h_1,h_2} \in (0,+\infty)$, we have

$$Cov\Big[\zeta_{n,1}(t_1,x), \zeta_{n,1}(t_2,x)\Big]$$

$$\begin{aligned} = &\frac{\Big(S_0(t_1|x)-1\Big)\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)^2}{m(x)}c_K\Phi_2(x,\infty,x)\frac{1}{n^2h_1} \\ &+C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})\frac{\Big(S_0(t_1|x)-1\Big)\Big(p(x)-1\Big)p(x)}{m^2(x)}D_{\xi,\omega}(x,\infty,t_2,x)\frac{1}{n^2h_1} \\ &+C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})\frac{\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)p(x)}{m^2(x)}D_{\xi,\omega}(x,t_1,\infty,x)\frac{1}{n^2h_1} \\ &+C_{h_1,h_2}\frac{p^2(x)C_1(t_1,t_2,x)}{m^2(x)}\frac{1}{n^2h_1} + o\left(\frac{1}{n^2h_1}\right) + O\left(\frac{h_1}{n^2}\right). \end{aligned}$$

Considering the functions V_1, V_2 and V_3, defined in Section 7.3:

$$
\begin{aligned}
&Cov\Big[\zeta_{n,1}(t_1,x),\zeta_{n,1}(t_2,x)\Big] \\
&= \Big(V_1(t_1,t_2,x)+C_{h_1,h_2}V_2(t_1,t_2,x)+C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})V_3(t_1,t_2,x)\Big)\frac{1}{n^2h_1} \\
&\quad +o\left(\frac{1}{n^2h_1}\right)+O\left(\frac{h_1}{n^2}\right).
\end{aligned} \tag{7.27}
$$

Using Equation (7.27) with $t_1 = t_2 = t+b$ and $t_1 = t_2 = t$, the expressions of $Var\Big[\zeta_{n,1}(t+b,x)\Big]$ and $Var\Big[\zeta_{n,1}(t,x)\Big]$ are also available. Therefore, Case (i) of the Theorem is proved by pluggin (7.27) in (7.16):

$$
\begin{aligned}
&Var\Big[\textstyle\sum_{i=1}^n \Psi_{n,i}(t,x)\Big] \\
&= \Big[a_1^2V_1(t+b,t+b,x)+a_2^2V_1(t,t,x)+2a_1a_2V_1(t+b,t,x) \\
&\quad +C_{h_1,h_2}\Big(a_1^2V_2(t+b,t+b,x)+a_2^2V_2(t,t,x)+2a_1a_2V_2(t+b,t,x)\Big) \\
&\quad +C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})\Big(a_1^2V_3(t+b,t+b,x)+a_2^2V_3(t,t,x) \\
&\quad +2a_1a_2V_3(t+b,t,x)\Big)\Big]\frac{1}{nh_1}+o\left(\frac{1}{nh_1}\right)+O\left(\frac{h_1}{n}\right) \\
&= \Big(\widetilde{V}_1(t+b,t,x)+C_{h_1,h_2}\widetilde{V}_2(t+b,t,x)+C_{h_1,h_2}\widetilde{c}_K(C_{h,g})\widetilde{V}_3(t+b,t,x)\Big)\frac{1}{nh_1} \\
&\quad +o\left(\frac{1}{nh_1}\right)+O\left(\frac{h_1}{n}\right).
\end{aligned}
$$

(ii) If $\lim_{n\to\infty}\dfrac{h_1}{h_2}=0$:

From Lemma 7.3 and Equation (7.21) when $t_1 = t_2$, we have

$$
Var\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x)\right] = h_1c_K\Phi_2(x,t_1,x)m(x)+O({h_1}^3),
$$

$$
Var\left[K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right] = C_1(t_2,t_2,x)h_2+O(h_2^3).
$$

Then, using the Cauchy-Schwarz inequality:

$$
\begin{aligned}
&Cov\left[K\left(\frac{x-X_1}{h_1}\right)\xi(Z_1,\delta_1,t_1,x),K\left(\frac{x-X_1}{h_2}\right)\omega(Z_1,\delta_1,t_2,x)\right] \\
&\leq \sqrt{h_1h_2c_K\Phi_2(x,t_1,x)m(x)C_1(t_2,t_2,x)+O(h_1h_2^3)+O(h_2{h_1}^3)}.
\end{aligned} \tag{7.28}
$$

Therefore,

$$A_2 = O\Big((h_1h_2)^{1/2}\Big), \quad A_3 = O\Big((h_1h_2)^{1/2}\Big). \tag{7.29}$$

Plugging (7.20), (7.21) and (7.29) in (7.19), we have

$$\begin{aligned} & Cov\Big[\zeta_{n,1}(t_1,x),\zeta_{n,1}(t_2,x)\Big] \\ = \; & \frac{\Big(S_0(t_1|x)-1\Big)\Big(S_0(t_2|x)-1\Big)\Big(p(x)-1\Big)^2}{m(x)} c_K \Phi_2(x,\infty,x)\frac{1}{n^2h_1} \\ & + \frac{p^2(x)C_1(t_1,t_2,x)}{m^2(x)}\frac{1}{n^2h_2} + O\left(\frac{h_1}{n^2}\right) + O\left(\frac{h_2}{n^2}\right) + O\left(\frac{\sqrt{h_1h_2}}{n^2h_1h_2}\right). \end{aligned} \tag{7.30}$$

Assuming $\lim_{n\to\infty}\frac{h_1}{h_2} = 0$ and considering the function $V_1(t_1,t_2,x)$, we have

$$Cov\Big[\zeta_{n,1}(t_1,x),\zeta_{n,1}(t_2,x)\Big] \;=\; V_1(t_1,t_2,x) + o\left(\frac{1}{n^2h_1}\right) + O\left(\frac{g}{n^2}\right). \tag{7.31}$$

Using Equation (7.31) with $t_1 = t_2 = t+b$ and $t_1 = t_2 = t$, the expressions of $Var\Big[\zeta_{n,1}(t+b,x)\Big]$ and $Var\Big[\zeta_{n,1}(t,x)\Big]$ are also available. Therefore, Case (ii) of the Theorem is proved by replacing (7.31) in (7.16):

$$\begin{aligned} & Var\Big[\textstyle\sum_{i=1}^n \Psi_{n,i}(t,x)\Big] \\ = \; & \Big(a_1^2V_1(t+b,t+b,x) + a_2^2V_1(t,t,x) + 2a_1a_2V_1(t+b,t,x)\Big)\frac{1}{nh_1} \\ & + o\left(\frac{1}{nh_1}\right) + O\left(\frac{h_2}{n}\right) \\ = \; & \widetilde{V}_1(t+b,t,x)\frac{1}{nh_1} + o\left(\frac{1}{nh_1}\right) + O\left(\frac{h_2}{n}\right). \end{aligned}$$

(iii) If $\lim_{n\to\infty}\frac{h_2}{h_1} = 0$:

From Equation (7.30) and assuming that $\lim_{n\to\infty} h_2/h_1 = 0$, we have

$$Cov\Big[\zeta_{n,1}(t_1,x),\zeta_{n,1}(t_2,x)\Big] \;=\; V_2(t_1,t_2,x)\frac{1}{n^2h_2} + o\left(\frac{1}{n^2h_2}\right) + O\left(\frac{h_1}{n^2}\right). \tag{7.32}$$

Considering the expression for $Cov\Big[\zeta_{n,1}(t_1,x),\zeta_{n,1}(t_2,x)\Big]$ in (7.32) with $t_1 = t_2 = t+b$ and $t_1 = t_2 = t$, the expressions for $Var\Big[\zeta_{n,1}(t+b,x)\Big]$ and

$Var\Big[\zeta_{n,1}(t,x)\Big]$ are also available. Therefore, Case (iii) of the Theorem is proved by replacing (7.32) in (7.16):

$$\begin{aligned} Var\Big[\sum_{i=1}^{n}\Psi_{n,i}(t,x)\Big] &= \Big(a_1^2V_2(t+b,t+b,x)+a_2^2V_2(t,t,x) \\ &\quad +2a_1a_2V_2(t+b,t,x)\Big)\frac{1}{ng}+o\left(\frac{1}{ng}\right)+O\left(\frac{h}{n}\right) \\ &= \widetilde{V}_2(t+b,t,x)\frac{1}{ng}+o\left(\frac{1}{ng}\right)+O\left(\frac{h}{n}\right). \end{aligned}$$

□

Proof of Theorem 7.3

(i) Assuming $C_{h_1} := \lim_{n\to\infty} n^{1/5}h_1 \in (0,\infty)$, $C_{h_2} := \lim_{n\to\infty} n^{1/5}h_2 \in (0,\infty)$:

From Equation (7.13) in the proof of Lemma 7.1 we have

$$\sqrt{nh_1}\Big(\widehat{PD}_{h_1,h_2}(t|x)-PD(t|x)\Big)=\sqrt{nh_1}\sum_{i=1}^{n}\Psi_{n,i}(t,x)+\widetilde{R}_n^2(t|x), \qquad (7.33)$$

where $\Psi_{n,i}(t,x) = a_1\zeta_{n,i}(t+b|x)+a_2\zeta_{n,i}(t|x)$ with $a_1 = -\dfrac{1}{S(t|x)}$, $a_2 = \dfrac{S(t+b|x)}{S^2(t|x)}$ and $\widetilde{R}_n^2(t|x)=\sqrt{nh_1}R_n^2(t|x)$ with $R_n^2(t|x)$ defined in the statement of Theorem 7.1. The variables $\Psi_{n,i}(t,x)$ are independent and identically distributed for all $i=1,...,n$. The remainder term is as follows:

$$\widetilde{R}_n^2(t|x) = \sqrt{nh_1}R_n^2(t|x)=\sqrt{nh_1}O_P\left(\ln n\left(\frac{1}{nh_1}+\frac{1}{nh_2}\right)\right)^{3/4}.$$

Using Assumptions A.16, A.17 and $\lim_{n\to\infty} h_1/h_2 \in (0,+\infty)$ from Theorem 7.3, the remainder term $\widetilde{R}_n^2(t|x)$ is negligible with respect to the dominant term of (7.33).

On the other hand, from Case (i) of Theorem 7.2 and Equation (7.33), the

variance of the dominant term is finite, since it is given by:

$$\begin{aligned} Var\Big[\sqrt{nh_1}\sum_{i=1}^{n}\Psi_{n,i}(t,x)\Big] &= nh_1\Big(\widetilde{V}_1(t+b,t,x)+C_{h_1,h_2}\widetilde{V}_2(t+b,t,x) \\ &\quad +C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})\widetilde{V}_3(t+b,t,x)\Big)\frac{1}{nh_1} \\ &\quad +nh_1\,o\left(\frac{1}{nh_1}\right)+nh_1\,O\left(\frac{h_1}{n}\right) \\ &= O(1). \end{aligned}$$

Therefore, the asymptotic distribution of $\sqrt{nh_1}\Big(\widehat{PD}_{h_1,h_2}(t|x)-PD(t|x)\Big)$ is the same as the asymptotic distribution of $\sqrt{nh_1}\sum_{i=1}^{n}\Psi_{n,i}(t,x)$. If Lindeberg's condition for triangular arrays (see Theorem 7.2 in Billingsley (1968)) is satisfied, then

$$\sum_{i=1}^{n}\left(\sqrt{nh_1}\Psi_{n,i}(t,x)-E\Big[\sqrt{nh_1}\Psi_{n,i}(t,x)\Big]\right)\xrightarrow{d}N(0,s), \tag{7.34}$$

where

$$s^2 = \widetilde{V}_1(t+b,t,x)+C_{h_1,h_2}\widetilde{V}_2(t+b,t,x)+C_{h_1,h_2}\widetilde{c}_K(C_{h_1,h_2})\widetilde{V}_3(t+b,t,x).$$

Lindeberg's condition is now checked. The condition is:

$$\lim_{n\to\infty}\frac{1}{s^2}E\left[\sum_{i=1}^{n}\left(\sqrt{nh_1}\Psi_{n,i}(t,x)-E\Big[\sqrt{nh_1}\Psi_{n,i}(t,x)\Big]\right)^2\mathbb{1}_{n,i}\right]=0 \tag{7.35}$$

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

$$\mathbb{1}_{n,i} = \mathbb{1}\Big(\Big|\sqrt{nh_1}\Psi_{n,i}(t,x)-E[\sqrt{nh_1}\Psi_{n,i}(t,x)]\Big|>\varepsilon s\Big).$$

Using Assumption A.3d, $\xi(Z,\delta,t,x)$ can be proved 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, ω is also bounded:

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

Since ω is bounded, K and $m(x)$ have compact support and $nh_1\to\infty$, $\Big\{\Psi_{n,i}(t,x)-E\Big[\Psi_{n,i}(t,x)\Big], i=1,...,n,\ n\in\mathbb{N}\Big\}$ is a sequence of random variables which is bounded by a convergent to zero nonrandom sequence, $\frac{\varepsilon s}{\sqrt{nh_1}}$.

Hence, there exists $n_0 \in \mathbb{N}$ such that for all $i = 1, ..., n$, $\mathbb{1}_{n,i} = 0$ for all $n \geq n_0$ and accordingly,

$$\lim_{n\to\infty} \frac{1}{s^2} E\left[\sum_{i=1}^{n} \left(\sqrt{nh_1}\Psi_{n,i}(t,x) - E\left[\sqrt{nh_1}\Psi_{n,i}(t,x)\right] \right)^2 \mathbb{1}_{n,i} \right] = 0,$$

which proves Lindeberg's condition in (7.35).

Finally, assuming $h_1 = C_{h_1} n^{-1/5}$ and $h_2 = C_{h_2} n^{-1/5}$ and considering Equation (7.4), we have

$$\sqrt{nh_1} \sum_{i=1}^{n} \Psi_{n,i}(t,x) \xrightarrow{d} N(\mu, s),$$

where $\mu = C_{h_1}^{5/2} \widetilde{B}_1(t,x) + C_{h_2}^{5/2} \widetilde{B}_2(t,x)$.

(ii) Assuming $C_{h_2} := \lim_{n\to\infty} n^{1/5} h_2 \in (0, \infty)$ and $\lim_{n\to\infty} n^{1/5} h_1 = 0$:

Consider (7.33). Under Assumptions A.16, A.17 and $\lim_{n\to\infty} h_1/h_2 = 0$ from Case (ii) in Theorem 7.3, the remainder term $\widetilde{R}_n^2(t|x)$ is found to be negligible with respect to the dominant term in (7.33). Furthermore, the variance of this dominant term is finite, since, from the proof of Theorem 7.2,

$$\begin{aligned} Var\left[\sqrt{nh_1} \textstyle\sum_{i=1}^{n} \Psi_{n,i}(t,x)\right] &= nh_1 \left(\widetilde{V}_1(t+b,t,x) \frac{1}{nh_1} + o\left(\frac{1}{nh_1}\right) + O\left(\frac{h_1}{n}\right) \right) \\ &= O(1). \end{aligned}$$

Therefore, the asymptotic distribution of $\sqrt{nh_1}\left(\widehat{PD}_{h_1,h_2}(t|x) - PD(t|x)\right)$ is the same as the asymptotic distribution of $\sqrt{nh_1} \sum_{i=1}^{n} \Psi_{n,i}(t,x)$. If Lindeberg's condition given in (7.35) is satisfied, then

$$\sum_{i=1}^{n} \left(\sqrt{nh_1}\Psi_{n,i}(t,x) - E\left[\sqrt{nh_1}\Psi_{n,i}(t,x)\right] \right) \xrightarrow{d} N(0, s), \tag{7.36}$$

where $s^2 = \widetilde{V}_1(t+b,t,x)$.

Lindeberg's condition is proved here following the same argument used for Case i. Finally, assuming $h_2 = C_{h_2} n^{-1/5}$ and $n^{1/5} h_1 \to 0$ and considering Equation (7.4),

$$\sqrt{nh_1} \sum_{i=1}^{n} \Psi_{n,i}(t,x) \xrightarrow{d} N(\mu, s),$$

where $\mu = C_{h_2}^{5/2} \widetilde{B}_2(t,x)$.

(iii) Assuming $C_{h_1} := \lim_{n\to\infty} n^{1/5}h_1 \in (0,\infty)$, $\lim_{n\to\infty} n^{1/5}h_2 = 0$:

From Equation (7.13) in the proof of Theorem 7.2 we have

$$\sqrt{nh_2}\Big(\widehat{PD}_{h_1,h_2}(t|x) - PD(t|x)\Big) \;=\; \sqrt{nh_2}\sum_{i=1}^{n}\Psi_{n,i}(t,x) + \widetilde{R}_n^2(t|x) \quad (7.37)$$

where $\Psi_{n,i}(t,x)$ are independent and identically distributed for all $i = 1,...,n$ and $\widetilde{R}_n^2(t|x) = \sqrt{nh_2}R_n^2(t|x)$ with $R_n^2(t|x)$ defined in the statement of Theorem 7.1. The remainder term is as follows:

$$\begin{aligned}\widetilde{R}_n^2(t|x) &= \sqrt{nh_2}R_n^2(t|x)\\ &= \sqrt{nh_2}O_P\left(\ln n\left(\frac{1}{nh_1}+\frac{1}{nh_2}\right)\right)^{3/4}.\end{aligned}$$

Using Assumptions A.16, A.17 and $\lim_{n\to\infty} h_2/h_1 = 0$ from Theorem 7.3, the remainder term $\widetilde{R}_n^2(t|x)$ is negligible with respect to the dominant term of (7.37).

The remainder term $\widetilde{R}_n^2(t|x)$ is then found to be negligible with respect to the dominant term in (7.33). Furthermore, the variance of this dominant term is finite, since, from the proof of Theorem 7.2,

$$\begin{aligned}&Var\Big[\sqrt{nh_2}\sum_{i=1}^{n}\Psi_{n,i}(t,x)\Big]\\ &\quad= nh_2\left(\widetilde{V}_2(t+b,t,x)\frac{1}{nh_2} + o\left(\frac{1}{nh_2}\right) + O\left(\frac{h_1}{n}\right)\right) = O(1).\end{aligned}$$

Therefore, the asymptotic distribution of $\sqrt{nh_2}\Big(\widehat{PD}_{h_1,h_2}(t|x) - PD(t|x)\Big)$ is the same as the asymptotic distribution of $\sqrt{nh_2}\sum_{i=1}^{n}\Psi_{n,i}(t,x)$. If Lindeberg's condition given by

$$\lim_{n\to\infty}\frac{1}{s^2}E\left[\sum_{i=1}^{n}\left(\sqrt{nh_2}\Psi_{n,i}(t,x) - E\Big[\sqrt{nh_2}\Psi_{n,i}(t,x)\Big]\right)^2 \mathbb{1}_{n,i}\right] = 0$$

is satisfied, then

$$\sum_{i=1}^{n}\left(\sqrt{nh_2}\Psi_{n,i}(t,x) - E\Big[\sqrt{nh_2}\Psi_{n,i}(t,x)\Big]\right) \xrightarrow{d} N(0,s), \quad (7.38)$$

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

Similar arguments to those used for Case i prove Lindeberg's condition. Finally, assuming $h_1 = C_{h_1} n^{-1/5}$ and $n^{1/5} h_2 \to 0$ and considering Equation (7.4), we have

$$\sqrt{nh_2} \sum_{i=1}^{n} \Psi_{n,i}(t,x) \xrightarrow{d} N(\mu, s),$$

where $\mu = C_{h_1}^{5/2} \widetilde{B}_1(t,x)$.

□

Chapter 8

Conclusions and future research lines

In the context of credit risk, one is often interested in modelling and estimating the probability of default (PD) measuring the probability of a client to run into arrears on his or her credit obligation. Since the work of Naraim (1992), new techniques based on survival analysis have been developed to solve credit risk issues. Examples of this have been widely cited throughout this dissertation. Here, we have proposed a novel method to estimate the probability of default in a time horizon $t + b$ from a maturity time t using nonparametric estimators.

After an in-depth study of several nonparametric estimators of the probability of default obtained from conditional survival estimators, it became clear the convenience of proposing a doubly smoothed estimator of the conditional survival function both in the covariate and in the time variable. This doubly smoothed estimator showed desirable asymptotic properties and promising simulation performance. This allowed us to propose a doubly smoothed estimator of the probability of default based on the generalised product-generalised limit estimator of Beran (1981). This time variable smoothing of the PD estimator resulted in a remarkable decrease of the estimation error. In addition, asymptotic expressions for the bias and variance of the smoothed PD estimator based on Beran's estimator for the sur-

vival function and its limit distribution were proved. Automatic bandwidth selectors and confidence regions algorithms based on bootstrap for Beran's estimator and the smoothed Beran's estimator of the conditinal survival function and the probability of default proposed in this report exhibited a reasonable behaviour. Since a group of cured individuals who will never experience the default could potentially exists, a nonparametric estimator of the probability of default based on mixture cure models was also proposed. The asymptotic bias and variance and the asymptotic normality of the cure model probability of default estimator were proved. The performance of the smoothed Beran's estimator of the PD remained however competitive with this proposal.

Interesting challenges remain to be dealt in the future:

There is a clear practical need for an automatic selector of the bandwidths involved in the cure model-based nonparametric estimator. In addition, a time variable smoothing of this estimator could be considered, as it was done with the smoothed Beran's estimator.

Beyond what has been seen in this work, a usefulness of cure models is that they allow estimating not only the PD, but also the probability of cure. In fact, a hypothesis test to confirm whether or not the probability of cure is zero could be very interesting in this context.

Using cure models when the cure status is partially known is also an appealing idea to be considered for future research. A nonparametric view along the lines similar to Safari et al. (2020) can be used to propose a nonparametric estimator which incorporates information from individuals for whom the cure is observed. A multi-state model that allows individuals to move among a finite number of states could also be interesting.

Throughout this work, the purpose has been to estimate the probability of default curve over time for fixed values of the covariate, so the mean integrated squared error was used as estimation error and the optimal bandwidths were obtained by minimising this function or its bootstrap approximation. Local smoothing parame-

ters could be obtained by a k-nearest neighbours criterion. An error criterion could also be established to choose the best parameters to estimate the $PD(t|x)$ curve for a fixed time and x variable. Similarly, a global error measure would allow both t and x to vary.

A critical point of this work lies in the low global (though not pointwise) coverage that the confidence regions have in some scenarios. The auxiliary bandwidths could be more influential than expected and their choice may need to be further refined. Another proposal for improvement would be the use of multiple or Bonferroni-type tests. The main limitation of proposed resampling methods is their high computational cost. Future work could include the analysis of subsampling techniques for optimising these methods.

In a financial context, one-dimensional credit scoring typically summarises several interesting features of clients in order to measure their creditworthiness. However, this work could be extended to the case of having a multidimensional covariate $(X_1, \ldots, X_q)$ where each X_i is a feature of the individual. Methods such as single-index may be useful for this purpose to avoid the curse of dimensionality. An approach along the lines similar to Strzalkowska-Kominiak and Cao (2013) could be used.

In this dissertation, and previous papers on this topic, it is assumed that life and censoring times are conditionally independent given the covariate. In some biomedical scenarios, this assumption is known not to be entirely realistic and research on that is already underway (see Deresa and Van Keilegom (2021)). An engaging open problem would be to extend the existing results to the estimation of the probability of default when the survival data are subject to dependent censoring.

For the time being, comparison tests of probability of default curves or classification tests based on the default probability are also left out of the picture. This is however a field we would like to explore in the future.

A practical use of the techniques developed here could be the calculation of risk appetite. Risk appetite is the amount of risk that a financial institution is willing

to take in order to achieve its strategic objectives. It would be interesting to know the probability of default at horizon b conditional on numerous scoring values, so that the scoring value associated with this risk appetite could be estimated.

Finally, we would like to make the software developed in this thesis publicly accessible to any interested user. For this reason, we plan to implement an R package that will be available on CRAN.

Appendix A

Resumen en castellano

Esta tesis pretende recoger los estudios realizados y los resultados obtenidos a lo largo del proceso de doctorado. Este trabajo se centra en estimar la curva de probabilidad de mora a lo largo del tiempo, condicionalmente a la puntuación crediticia. Con este fin, se proponen estimadores no paramétricos basados en el análisis de supervivencia y los modelos de curación. A continuación se expone un resumen del contenido más relevante de esta disertación.

En la Introducción se motiva la necesidad de conocer y estimar la curva de probabilidad de mora. El riesgo de crédito se define como la posible pérdida que asume un agente económico en caso de que la contraparte incumpla sus obligaciones contractuales y es un área de investigación importante dentro de las finanzas cuantitativas. Las deudas procedentes de clientes con créditos impagados tienen un importante impacto en la solvencia de los bancos y otras entidades financieras. Por esta razón, el Comité de Basilea para la Supervisión Bancaria del Banco de Pagos Internacionales estableció en 2004 un conjunto de mecanismos estándar para la medición del riesgo de crédito en las instituciones financieras. Según este Acuerdo de Basilea, uno de los elementos de mayor influencia en el riesgo de crédito es la probabilidad de mora, denotada por PD, por sus siglas en inglés *probability of default*. Es por ello que para la entidad bancaria resulta importante determinar la probabilidad de que un crédito caiga en mora, convirtiéndose en un crédito moroso. Para un tiempo fijo, t, y un

horizonte de mora, b, la PD puede definirse como la probabilidad de que un crédito que ha sido pagado hasta el instante t, caiga en mora no más tarde del instante $t+b$. Para estimar la PD, los bancos e instituciones financieras utilizan covariables que contienen información sobre los créditos y los correspondientes clientes. La información contenida en estas covariables suele resumirse mediante alguna combinación lineal de las mismas que mide la capacidad de clientes o futuros clientes para hacer frente a una posible deuda que contraigan con el banco a través de un préstamo. Esta combinación lineal recibe el nombre de puntuación crediticia o *credit scoring* en inglés. En este escenario, la variable de interés es el tiempo hasta la caída en mora. Esta variable no es completamente observable: sólo es posible conocer el tiempo de vida de un crédito hasta que el cliente deja de pagarlo cuando la caída en mora tiene lugar durante el tiempo de observación de los créditos; en otro caso, el dato es censurado y el tiempo observado es el tiempo hasta la censura. La puntuación crediticia juega el papel de variable predictora, proporcionando información acerca del tiempo hasta la caída en mora. Se asume, por tanto, que la probabilidad de mora es una curva $PD(t|x)$ a lo largo del tiempo, t, que depende de la puntuación crediticia, x.

El Capítulo 1 está dedicado a introducir al lector en el contexto en el que se desarrolla este trabajo. Se explica en qué consiste un escenario de censura por la derecha y se exponen los conceptos básicos del análisis de supervivencia y, en particular, de los modelos de curación útiles en dicho escenario. El análisis de supervivencia es un conjunto de procedimientos estadísticos que permiten describir y estudiar datos cuando la variable de interés es el tiempo que transcurre hasta que se produce cierto evento. La censura tiene lugar cuando una proporción de tiempos hasta el evento de interés son desconocidos y se denomina censura por la derecha cuando el motivo de la misma es que el estudio finalice antes de que todos los individuos hayan experimentado el evento de interés. Los modelos de curación son modelos de supervivencia que incorporan explícitamente la posibilidad de que un sujeto nunca experimente dicho evento. Se introduce también la estimación no paramétrica de curvas. Se trata de estimadores flexibles que requieren, a lo sumo, condiciones de continuidad

y diferenciabilidad para la curva subyacente. Esta sección se centra principalmente en la estimación de la función de supervivencia condicional, presentando el estimador empírico, el estimador límite-producto para el caso censurado y el estimador límite-producto generalizado para el caso con covariables. Se incluye una revisión de los principales métodos de remuestreo basados en bootstrap en los que se considera la presencia de censura y/o covariables y se detalla brevemente su utilidad en la selección de parámetros de suavizado mediante la minimización de alguna medida de error global como el error cuadrático medio integrado.

En el Capítulo 2 se proponen modelos de supervivencia que permiten estimar la probabilidad de mora como función de la puntuación crediticia en créditos personales. Sea T la variable que denota el tiempo hasta la caída en mora y X la puntuación crediticia. Dados x un valor fijo de la puntuación crediticia y b un horizonte de mora, la probabilidad de mora se escribe formalmente del siguiente modo:

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

donde $S(t|x)$ es la función de supervivencia condicional del tiempo hasta la mora, T. Así, empleando técnicas del análisis de supervivencia para proponer estimadores de $S(t|x)$ se presentan en este capítulo varios estimadores no paramétricos de la probabilidad de mora.

Se consideran cuatro estimadores para la función de supervivencia condicional: el estimador de Beran (1981), el estimador de Van Keilegom-Akritas (Van Keilegom and Akritas (1999)), el estimador lineal local ponderado propuesto por Cai (2003) y el estimador de Nadaraya-Watson ponderado presentado en Peláez et al. (2021b). El estimador de Beran es la generalización del estimador límite-producto al caso con covariables. El estimador de Van Keilegom-Akritas se basa en ajustar un modelo de regresión heterocedástico donde la variable respuesta es el tiempo hasta la mora y la variable predictora la puntuación crediticia. El estimador lineal local ponderado también se construye a partir de un modelo de regresión realizando, en este caso, un ajuste lineal local del mismo. Un ajuste constante de dicho modelo de regresión permite obtener el estimador de Nadaraya-Watson ponderado. Estos estimadores de

la función de supervivencia se transforman de acuerdo a la expresión dada en (A.1) para obtener los correspondientes estimadores de la PD. En este capitulo se demuestra un resultado general acerca de las propiedades asintóticas de los estimadores de la PD así construidos. Conocidas las expresiones asintóticas del sesgo, varianza y covarianza de los estimadores de la supervivencia, mediante este resultado se obtienen las expresiones análogas para el estimador de la PD resultante. Se trata de expresiones complejas que dependen de varios parámetros poblacionales desconocidos, por lo que dificultan la obtención de una aproximación del MISE. Para analizar el comportamiento de los estimadores de la PD basados en los estimadores de Beran, Van Keilegom-Akritas, lineal local ponderado y Nadaraya-Watson ponderado de la supervivencia, se lleva a cabo un estudio de simulación basado en varios modelos con diferentes escenarios de censura. Se incluye también el método paramétrico basado en el modelo de Cox como referencia. Los resultados obtenidos muestran que el estimador de Beran es el que ofrece las mejores aproximaciones de la PD, en cuanto a que proporciona un menor error cuadrático medio integrado en la mayor parte de los escenarios analizados. Además, requiere un menor tiempo de computación. Los distintos métodos se ilustran mediante su aplicación a un conjunto de créditos bancarios concedidos por una entidad financiera española entre 2004 y 2006. El contenido de este capítulo se encuentra publicado en Peláez et al. (2021b).

En el Capítulo 3 se propone un estimador doblemente suavizado de la función de supervivencia condicional. Las estimaciones de la probabilidad de mora obtenidas mediante los estimadores presentados en el Capítulo 2 son muy razonables, pero tienen una variabilidad excesiva y son curvas muy rugosas. El origen de esta variabilidad es el cociente de supervivencias que debe hacerse para estimar la PD (veáse (A.1)). Este cociente magnifica los saltos que caracterizan los estimadores de la función de supervivencia, estimadores suaves en la covariable, pero funciones a saltos en la variable tiempo. Por ello, en este capítulo se propone un estimador no paramétrico de la función de supervivencia condicional doblemente suavizado, tanto en la covariable como en la variable temporal. La técnica aquí presentada es general y permite suavizar en la variable tiempo numerosos estimadores ya conocidos de la función de

supervivencia condicional. Sin embargo, este capítulo se centra principalmente en el estimador suavizado de la supervivencia basado en el estimador clásico de Beran. Se demuestran propiedades asintóticas del estimador no paramétrico con doble suavizado asociado al estimador de Beran. Se obtiene una representación casi segura del estimador de Beran suavizado, expresiones asintóticas del sesgo y varianza y la distribución límite del mismo. Mediante estudios de simulación, se comprueba que este suavizado supone una considerable reducción del error de estimación. También se estudia la influencia de los dos parámetros de ventana involucrados en el estimador suavizado. Un pequeño análisis de las funciones de supervivencia de los tiempos de hospitalización de pacientes de COVID-19 en Galicia proporcionados por el Servicio Gallego de Salud (SERGAS) muestra las diferencias entre el estimador clásico de Beran y la propuesta suavizada en la variable tiempo. El contenido de este capítulo se encuentra publicado en Peláez et al. (2022b).

En el Capítulo 4 se proponen métodos de selección automática del parámetro de suavizado en la covariable para el estimador de Beran y de los parámetros de suavizado en la covariable y en la variable tiempo para el estimador de Beran suavizado de la función de supervivencia condicional. Las técnicas de remuestreo propuestas se basan en combinar un bootstrap suavizado con un bootstrap con covariables. Las ventanas bootstrap se obtienen mediante la minimización de la aproximación bootstrap del error cuadrático medio integrado. Un estudio de simulación basado en varios modelos y diferentes niveles de censura muestra el comportamiento de los estimadores de la función de supervivencia con ventanas bootstrap. También se aborda la obtención de regiones de confianza para la función de supervivencia condicional, $S(t|x)$, para un valor fijo de $x \in I \subseteq \mathbb{R}$ y t variando en el intervalo $I_T \subseteq \mathbb{R}^+$, utilizando los estimadores de Beran y de Beran suavizado. Se proponen dos métodos diferentes basados en bootstrap para la construcción de estas regiones de confianza. El primero de ellos da lugar a regiones de confianza de amplitud constante y el segundo de ellos a regiones de confianza de amplitud variable. La necesidad práctica de estos métodos de selección de la ventana se hace patente en el análisis de tiempos de hospitalización de pacientes de COVID-19 en Galicia, España.

Las técnicas desarrolladas se emplean para estudiar en profundidad la relación entre la edad, el sexo y ciertas patologías previas, como obesidad o EPOC, con el tiempo de recuperación de pacientes infectados con SARS-CoV-2. Los datos fueron proporcionados por el Servicio Gallego de Salud y corresponden a ingresos en hospitales gallegos durante las primeras semanas de la pandemia.

En el Capítulo 5 se presenta un estimador no paramétrico de la probabilidad de mora con doble suavizado que pretende solucionar el problema de variabilidad observado en las estimaciones de la PD obtenidas hasta el momento. Este estimador deriva del estimador suavizado de la supervivencia que se introduce en el Capítulo 3. Por tanto, se trata de un estimador de la PD general, que puede obtenerse a partir de cualquier estimador doblemente suavizado de la función de supervivencia. Sin embargo, este capítulo se centra, principalmente, en el estimador de la PD basado en el estimador suavizado de Beran. A partir de las propiedades asintóticas del estimador suavizado de Beran para la función de supervivencia se obtienen las expresiones asintóticas del sesgo y la varianza del estimador de la probabilidad de mora. También se demuestra la normalidad asintótica de este estimador. El estudio de simulación realizado muestra que el suavizado de la variable temporal reduce significativamente el error cometido en la estimación de la PD. Esta técnica implica un aumento considerable del tiempo de cálculo. Sin embargo, la variabilidad y la rugosidad de las estimaciones se ven claramente reducidas. Además, las simulaciones llevadas a cabo permiten concluir que otros estimadores de la PD también presentan estas mejoras cuando se suavizan en la variable temporal. También se analiza la relación entre los dos parámetros de suavizado involucrados en el estimador y su influencia en el error cuadrático medio integrado. La utilidad del doble suavizado se ilustra mediante el análisis del conjunto de créditos bancarios concedidos por cierta entidad financiera española correspondientes al periodo 2004-2006, donde se observan las diferencias entre los estimadores de Beran y Beran suavizado de la PD en un problema real. El contenido de este capítulo se encuentra publicado en Peláez et al. (2021a).

En el Capítulo 6 se proponen métodos de selección automática del parámetro

de suavizado en la covariable para el estimador de Beran y de los parámetros de suavizado en la covariable y en la variable tiempo para el estimador de Beran suavizado de la PD. Las ventanas bootstrap se obtienen mediante la minimización de la aproximación bootstrap del error cuadrático medio integrado. Las técnicas de remuestreo empleadas para ello son las presentadas en el Capítulo 4. El rendimiento de las técnicas propuestas se analiza mediante simulaciones en diferentes escenarios, obteniendo resultados prometedores. El método para obtener regiones de confianza de amplitud variable basadas tanto en el estimador de Beran como en el estimador de Beran suavizado para la función de supervivencia también se extiende en este capítulo para obtener regiones de confianza de la curva de probabilidad de mora, $PD(t|x)$, para un valor fijo de $x \in I \subseteq \mathbb{R}$ y t cubriendo el intervalo $I_T \subseteq \mathbb{R}^+$. Un estudio de simulación nos permite analizar el comportamiento de estas regiones de confianza basadas en bootstrap y comparar los resultados obtenidos mediante los estimadores de Beran y Beran suavizado. Por último, los selectores automáticos de la ventana y los estimadores con ventanas bootstrap se utilizan para analizar la función de probabilidad de mora condicional a la puntuación crediticia para un conjunto de datos de créditos alemanes públicamente accesible. El contenido de este capítulo se encuentra publicado en Peláez et al. (2022a).

En el Capítulo 7 se discuten técnicas para estimar la PD basadas en modelos de curación. El tiempo hasta la caída en mora podría enfrentarse no sólo a un problema de censura por la derecha, sino también a la presencia de curación. Podrían existir clientes que nunca caen en mora, es decir, no importa cuánto tiempo se observe a tales individuos, nunca experimentarán el evento de interés. Los modelos de curación consideran la existencia de un grupo de individuos curados que no son susceptibles de caer en mora. Se considera el estimador no paramétrico de la función de supervivencia propuesto por López-Cheda et al. (2017a) y López-Cheda et al. (2017b) basado en modelos de curación de tipo mixtura y se transforma para obtener el estimador no paramétrico de modelos de curación (NPCM) de la PD. Se analizan las propiedades asintóticas del estimador NPCM de la PD: se obtiene una representación casi segura del estimador y expresiones asintóticas del sesgo y la va-

rianza, así como su normalidad asintótica. Su comportamiento se compara mediante simulación con el estimador de Beran de la PD, el estimador de Beran doblemente suavizado y con métodos paramétricos basados en modelos de curación como el de riesgos proporcionales y el de tiempo de fallo acelerado. Los resultados obtenidos muestran que el estimador NPCM proporciona buenas estimaciones de la PD, reduciendo el error cometido por las alternativas semiparamétricas. El estimador de Beran de la PD es competitivo con el estimador NPCM en la mayoría de los escenarios. El estimador de Beran doblemente suavizado es, de nuevo, la alternativa que proporciona un menor error de estimación en todos los escenarios analizados. Para ilustrar el uso de los estimadores de Beran, Beran suavizado y el estimador NPCM, se realiza un análisis estadístico del conjunto de préstamos alemanes citado previamente. El contenido de este capítulo se encuentra en revisión para su publicación en Peláez et al. (2022c).

En el Capítulo 8 se resumen las principales conclusiones del trabajo elaborado a lo largo de esta tesis doctoral y se detallan las líneas de trabajo que se pretenden abordar en el futuro. Consideramos que la principal contribución de esta tesis es la propuesta del estimador suavizado en la covariable y en la variable tiempo basado en el estimador de Beran para la función de supervivencia que permite estimar la curva de la PD reduciendo el error de estimación de forma notable. El análisis en profundidad de las propiedades asintóticas de este estimador, así como la propuesta de selectores automáticos para los parámetros de suavizado constituyen elementos importantes de esta disertación. En el futuro trataremos de resolver el problema de la selección automática de las ventanas involucradas en el estimador no paramétrico basado en modelos de curación. También se analizará su comportamiento cuando este estimador se suaviza en la variable tiempo. El uso de modelos de curación con cura parcialmente conocida y la extensión de los resultados de esta memoria al caso multidimensional mediante técnicas, por ejemplo, de single-index, son ideas atractivas para futuras investigaciones. Una línea interesante que podría tener utilidad práctica es la construcción de contrastes de clasificación basados en la PD o contrastes de hipótesis para la comparación de curvas de la PD. Finalmente, nos

gustaría que el software desarrollado a lo largo de esta tesis estuviese disponible públicamente, por lo que en un futuro inmediato tenemos la intención de elaborar un paquete de R.

Bibliography

Akritas, M. (1986). Bootstrapping the Kaplan-Meier estimator. *Journal of American Statistical Association*, 81(396):1032–1039.

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

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and its Application*, 5(1):311–342.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328.

Bagkavos, D. and Ioannides, D. (2021). Fixed design local polynomial smoothing and bandwidth selection for right censored data. *Computational Statistics and Data Analysis*, 153(107064).

Banasik, J., Crook, J., and Thomas, L. (1999). Not if but when will borrowers default. *Journal of Operational Research Society*, 50(12):1185–1190.

Bank for International Settlements (2009). 79th annual report: April 2008-march 2009. *Bank for International Settlements*.

Barbeito, I. and Cao, R. (2019). Nonparametric curve estimation and bootstrap bandwidth selection. *WIREs Computational Statistics*, 12(3):e1488.

Barnard, B. (2017). Rating migration and bond valuation: A historical interest rate and default probability term structures. *University of the Witwatersrand, Wits Business School*.

Basel Committee on Banking Supervision (1999). Credit risk modelling: current practices and applications. *Bank for International Settlements, Basel*.

Basel Committee on Banking Supervision (2001a). The internal ratings-based approach. *Bank for International Settlements, Basel*.

Basel Committee on Banking Supervision (2001b). The new Basel Capital Accord. *Bank for International Settlements, Basel*.

Basel Committee on Banking Supervision (2004). International convergence of capital measurement and capital standards. *Bank for International Settlements, Basel*.

Basel Committee on Banking Supervision (2005a). An explanatory note on the Basel ii IRB risk weight functions. *Bank for International Settlements, Basel*.

Basel Committee on Banking Supervision (2005b). Studies on the validation of internal rating systems. *Bank for International Settlements, Basel*.

Beran, J. and Djaïdja, A. (2007). Credit risk modeling based on survival analysis with immunes. *Statistical Methodology*, 4(3):251–276.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report, University of California*.

Bessis, J. (2002). *Risk management in banking*. John Wiley and Sons, New York.

Billingsley, P. (1968). Convergence of probability measure. Wiley Series in Probability and Mathematical Statistics: Tracts on probability and statistics. *John Wiley and Sons, New York*, 9.

Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, 11(1):15–44.

Broyden, C. G. (1970). The convergence of a cass of double-rank minimization algorithms 1. General gonsiderations. *Journal of Applied Mathematics*, 6(1):76–90.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). *An R-package for estimating semi-parametric mixture cure models*. `https://cran.r-project.org/web/packages/smcure/smcure.pdf`.

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

Cao, R. (1993). Bootstrapping the mean integrated squared error. *Journal of Multivariate Analysis*, 45(1):137–160.

Cao, R., Francisco-Fernández, M., and Quinto, E. (2010). A random effect multiplicative heteroscedastic model for bacterial growth. *BMC Bioinformatics*, 11(77).

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.

Cecchetti, S. G., Kohler, M., and Upper, C. (2009). *Financial crises and economic activity*. Bank for International Settlements, Basel.

Corbière, F., Commenges, D., Taylor, J., , and Joly, P. (2009). A penalized likelihood approach for mixture cure models. *Statistics in Medicine*, 28(3):510–524.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–202.

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

Deresa, N. W. and Van Keilegom, I. (2021). On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. *Biometrika*, 108(4):965–979.

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

Dirick, L., Bellotti, T., Claeskens, G., and Baesens, B. (2019). Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. *Journal of Business and Economic Statistics*, 37(1):40–53.

Dirick, L., Claeskens, G., and Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, 241(2):449–457.

dos Reis, G. and Smith, G. (2018). Robust and consistent estimation of generators in credit risk. *Quantitative Finance*, 18(6):983–1001.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (1981). Censored data and the bootstrap. *Journal of American Statistical Association*, 76(374):312–319.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall, London.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, 13(3):317–322.

Földes, A., Rejtø, L., and Winter, B. B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data, ii: estimation of density and failure rate. *Periodica Mathematica Hungarica*, 12(1):15–29.

Gannoun, A., Saracco, J., and Yu, K. (2007). Comparison of kernel estimator of conditional distribution function and quantile regression under censoring. *Statistical Modelling*, 7(4):329–344.

Gannoun, A., Saracco, J., Yuan, A., and Bonney, G. (2005). Nonparametric quantile regression with censored data. *Scandinavian Journal of Statistics*, 32(4):527–550.

Geerdens, C., Acar, E. F., and Janssen, P. (2017). Conditional copula models for right-censored clustered event time data. *Biostatistics*, 19(2):247–262.

Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, 141(3):333–387.

Girón, A. (1998). Crisis financieras. *Universidad de Marne-La-Vallèe, Paris*.

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(5):923–947.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26.

González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized kaplan-meier estimator with applications. *Nonparametric Statistics*, 4(1):65–78.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer, New York.

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

Haybittle, J. L. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *The British Journal of Radiology*, 32(383):725–733.

Haybittle, J. L. (1965). A two-parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association*, 60(309):16–26.

Hollander, M., Wolfe, D. A., and Chicken, E. (1999). *Nonparametric statistical methods*. John Wiley and Sons, New York.

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.

Jones, M. C. (1990). The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, 9(2):129–132.

Kalbfleisch, J. and Prentice, R. (1980). *The statistical analysis of failure time data*. John Wiley and Sons, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*, 53(282):457–481.

Laska, E. and Meisner, M. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, 48(4):1223–1234.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. John Wiley and Sons, New York.

Leconte, E., Poiraud-Casanova, S., and Thomas-Agnan, C. (2002). Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Analysis*, 8(3):229–246.

Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *The Institute of Statistical Mathematics*, 53(4):708–729.

Liu, H. and Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association*, 104(487):1168–1178.

Lo, S. H. and Singh, K. (1986). The product-limit estimator and the bootstrap: Some asymptotic representations. *Probability Theory and Related Fields*, 71(3):455–465.

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

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

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

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

Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors*. Wiley, Chichester - UK.

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

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

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

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

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

Peláez, R., Cao, R., and Vilar, J. M. (2022b). Nonparametric estimation of the conditional survival function with double smoothing. *To appear in Journal of Nonparametric Statistics*.

Peláez, R., Van Keilegom, I., Cao, R., and Vilar, J. M. (2022c). Probability of default estimation in credit risk using mixture cure models. *Technical Report, Universidade da Coruña*.

Portier, F. and Segers, J. (2018). On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, 166(11):160–181.

Pyle, D. (1997). Bank risk management: theory. *Research Programme in Finance Working Papers, University of California, Berkeley*, (RDF-272).

R. A. Maller, S. Z. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, 79(4):731–739.

Rao, B. P. (1983). *Nonparametric Functional Estimation*. Academic Press, New York.

Reid, N. (1981). Estimating the median survival time. *Biometrika*, 68(3):1601–608.

Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

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

Samreen, A., Zaidi, F., and Sarwar, A. (2013). 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*, 2(5):1–26.

Saunders, A. and Cornett, M. (2008). *Financial institutions management: A risk management approach*. McGraw-Hill Irwin, Singapore.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. In *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Srinivasan, V. and Kim, Y. H. (1987). Credit granting: a comparative analysis of clasification procedures. *Journal of Finance*, 42(3):665–681.

Steenackers, A. and Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics*, 8(1):31–34.

Strzalkowska-Kominiak, E. and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *Journal of Multivariate Analysis*, 114(7):74–98.

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

Sy, J. P. and Taylor, J. M. G. (2001). Standard errors for the cox proportional hazards cure model. *Mathematical and Computer Modelling*, 33(12):1237–1251.

Therneau, T. (2015). *A package for survival analysis in R.* `https://CRAN.R-project.org/package=survival`.

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

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 and Data Analysis*, 35(4):487–500.

Van Keilegom, I. and Veraverbeke, N. (1996). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles. *Communications in Statistics, Theory Methods*, 25(2):2251–2265.

Van Keilegom, I. and Veraverbeke, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Annals of the Institute of Statistical Mathematics*, 49(3):467–491.

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(3):757–770.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191.

Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *The Canadian Journal of Statistics*, 42(1):1–17.

Yakovlev, A. and Tsodikov, A. (1996). *Stochastic models of tumor latency and their biostatistical application*. World Scientific, Singapore.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560.

Zhun, T. and Politis, D. (2017). Kernel estimates of nonparametric functional autoregression models and their bootstrap approximation. *Electronic Journal of Statistics*, 11(2):2876–2906.