Nonparametric inference in mixture cure models

Ana López Cheda

PhD Thesis

University of A Coruña

2018

Nonparametric inference in mixture cure models

Ana López Cheda

PhD Thesis

2018

PhD advisors:
Ricardo Cao Abad
Mª Amalia Jácome Pumar

Doctoral Programme in Statistics and Operations Research
Department of Mathematics
University of A Coruña

The undersigned certify that they are the advisors of the Doctoral Thesis entitled "Nonparametric inference in mixture cure models", developed by Ana López Cheda at the University of A Coruña (Department of Mathematics), as part of the interuniversity PhD program (UDC, USC and UVigo) of Statistics and Operational Research, and hereby give their consent to the author to proceed with the thesis presentation and the subsequent defense.

Los abajo firmantes hacen constar que son los directores de la Tesis Doctoral titulada "Nonparametric inference in mixture cure models", realizada por Ana López Cheda en la Universidade da Coruña (Departamento de Matemáticas) en el marco del programa interuniversitario (UDC, USC y UVigo) de doctorado en Estadística e Investigación Operativa, dando su consentimiento para que la autora proceda a su presentación y posterior defensa.

Os abaixo asinantes fan constar que son os directores da Tese de Doutoramento titulada "Nonparametric inference in mixture cure models", desenvolta por Ana López Cheda na Universidade da Coruña (Departamento de Matemáticas) no marco do programa interuniversitario (UDC, USC e UVigo) de doutoramento en Estatística e Investigación de Operacións, dando o seu consentimento para que a autora proceda á súa presentación e posterior defensa.

A Coruña, March 6th, 2018.

Advisors:

Dr. Ricardo Cao Abad

Dr. Mª Amalia Jácome Pumar

PhD student:

Ana López Cheda

Acknowledgments

This thesis concludes four years of intensive research, and it is the result of the effort and both direct and indirect support of many people.

To Ricardo Cao, for giving me the opportunity to start with you this PhD project four years ago, for your dedication, for your enthusiasm in research and for guiding me through all the process.
To Amalia Jácome, for your inexhaustible patience and your constant motivation. I will never forget your friendly guidance, careful supervision and commitment.
I have been extremely lucky to have both of you as supervisors. I have infinite gratitude and admiration for you.

To the Predefense Committee, Mario Francisco-Fernández, Wenceslao González-Manteiga and Mª Carmen Iglesias-Pérez; to those who participated in the predefense, Mª Luz Calle, Luís Meira-Machado and Yingwei Peng; and to the external reviewers, Paul Janssen and Ingrid Van Keilegom, for all of your guidance through this process. Your discussion, ideas and feedback have substantially improved this thesis.

To Ingrid Van Keilegom, for your warm welcome and your helpful suggestions and comments during my two research stays at the Université catholique de Louvain, in Belgium. Your remarks have been absolutely invaluable. I would also like to thank all the people I met in Louvain-la-Neuve, who made me feel at home during those five months.

To the professors of the PhD program, in particular to Germán Aneiros, with whom I started collaborating in the lab while I was studying the master degree, and to the professors with whom I shared some teaching, especially to the coordinators of the Statistics courses: Graciela Estévez, at the Faculty of Sciences, and Manuel Antonio Presedo, at the Faculty of Computer Sciences.

To my PhD colleagues from the University of A Coruña, University of Santiago and University of Vigo, and to all the people I met in conferences and other courses during these years, with whom I shared memorable moments. All of you have contributed to make the process of development of my thesis more pleasant. Special thanks to Paula Raña, with whom I walked together from the very beginning of the PhD, for the talks about academic and non-academic aspects; to Miguel Reyes, for being close even in spite of the distance; to Laura Borrajo, for the warm company every day; and to Laura del Río, for the good moments inside and outside the lab. In addition, I would also like to thank Luis Rodríguez for the willingness to give me always technical support.

To my parents, for being my foundation, for their constant support and for being always there in all those things of life beyond a PhD. I cannot adequately express how thankful I am. To my family, for believing in me. To my couple, for accompanying me through this entire process of the thesis, and to my friends, for making me forget about it.

Thank you for everything that helped me get to this day.

The PhD student's research was sponsored by the Spanish FPU (Formación de Profesorado Universitario) Grant from MECD (Ministerio de Educación, Cultura y Deporte) with reference FPU13/01371. The work has been partially carried out during two visits at the Université catholique de Louvain. The first stay was financed by INDITEX and the second one was supported by the research group MODES (Modelización, Optimización e Inferencia Estadística).

The doctoral student acknowledges partial support by the MICINN (Ministerio de Ciencia e Innovación) Grant MTM2011-22392, the MINECO (Ministerio de Economía y Competitividad) Grant MTM2014-52876-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva CN2012/130, ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF (European Regional Development Fund).

The author is grateful to Dr. Sonia Pértega and Dr. Salvador Pita, at the University Hospital of A Coruña, for providing the colorectal cancer dataset, and to Dr. Ángel Díaz Lagares, at the Translational Medical Oncology (OMT) group, the Health Research Institute of Santiago (IDIS) and the University Hospital of Santiago (CHUS), for providing the sarcomas dataset.

To my parents

The journey is the reward
Chinese proverb

Abstract

A completely nonparametric method for the estimation of mixture cure models is proposed. An incidence estimator is extensively studied and a latency estimator is presented. These estimators, which are based on the Beran estimator of the conditional survival function, are proven to be the local maximum likelihood estimators. Two i.i.d. representations for the incidence and the latency estimators are obtained. Moreover, an asymptotic expression for the mean squared error of the latency estimator is derived, and its asymptotic normality is proven. In addition, bootstrap bandwidth selection methods for each nonparametric estimator are introduced. The proposed nonparametric estimators are compared with existing semiparametric approaches in simulation studies, in which the performance of the bootstrap bandwidth selectors are also assessed. The nonparametric incidence and latency estimators are applied to a dataset of colorectal cancer patients from the University Hospital of A Coruña (CHUAC).

Furthermore, a nonparametric covariate significance test for the incidence is proposed. The method is extended to non continuous covariates: binary, discrete and qualitative, and also to contexts with a large number of covariates. The efficiency of the procedure is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap. The test is applied to a sarcomas dataset.

Resumen

Se propone un método completamente no paramétrico para la estimación de modelos de curación de tipo mixtura. Se estudia ampliamente un estimador para la incidencia y se presenta un estimador para la latencia. Se demuestra que estos estimadores, basados en el estimador de Beran de la función de supervivencia condicional, son los estimadores máximo verosímiles locales. Se obtienen representaciones i.i.d. de los estimadores de la incidencia y de la latencia. Además, se halla una expresión asintótica para el error cuadrático medio del estimador de la latencia y se demuestra su normalidad asintótica. También se presentan métodos de selección de la ventana, de tipo bootstrap, para cada estimador no paramétrico. Los estimadores no paramétricos propuestos se comparan con otros enfoques semiparamétricos existentes en la literatura en estudios de simulación, en donde también se evalúa el comportamiento de los selectores de la ventana. Los estimadores no paramétricos de la incidencia y la latencia se aplican a una base de datos de pacientes de cáncer colorrectal del Complejo Hospitalario Universitario de A Coruña (CHUAC).

Además, se propone un test no paramétrico de significación de covariables. El método se extiende a covariables no continuas: binarias, discretas y cualitativas, y también a contextos con un gran número de covariables. Se evalúa su eficiencia en un estudio de simulación de Monte Carlo, en el cual la distribución del test es aproximada por bootstrap. Se aplica el método a una base de datos de pacientes con sarcomas.

Resumo

Proponse un método completamente non paramétrico para a estimación de modelos de curación de tipo mixtura. Estúdase ampliamente un estimador para a incidencia e preséntase un estimador para a latencia. Demóstrase que estes estimadores, baseados no estimador de Beran da función de supervivencia condicional, son os estimadores máximo verosímiles locais. Obtéñense representacións i.i.d. dos estimadores da incidencia e da latencia. Ademais, áchase unha expresión asintótica para o erro cadrático medio do estimador da latencia e demóstrase a súa normalidade asintótica. Tamén se presentan métodos de selección da ventá, de tipo bootstrap, para cada estimador non paramétrico. Compáranse os estimadores non paramétricos propostos con outros enfoques semiparamétricos existentes na literatura en estudos de simulación, onde tamén se avalía o comportamento dos selectores da ventá. Aplícanse os estimadores non paramétricos da incidencia e da latencia a unha base de datos de doentes de cancro colorrectal do Complexo Hospitalario Universitario de A Coruña (CHUAC). Ademais, proponse un test non paramétrico de significación de covariables. O método é extendido a covariables non continuas: binarias, discretas e cualitativas, e tamén a contextos cun gran número de covariables. Avalíase a súa eficiencia nun estudo de simulación de Monte Carlo, no que a distribución do test é aproximada por bootstrap. Aplícase o método a unha base de datos de doentes con sarcomas.

Contents

Preface

This work intends to summarize all the study developed along the PhD trajectory. Mainly, it is focused on estimation and covariate significance tests for nonparametric mixture cure models. Specifically, the methodology is applied to two medical datasets: one related to colorrectal cancer patients from the University Hospital of A Coruña (CHUAC), and other related to patients with sarcomas from the University Hospital of Santiago (CHUS).

Chapter 1 is devoted to introduce the reader to the context in which the thesis is developed: cure models. It begins with a presentation of survival analysis and the different types of censoring in Section 1.1. Since the work is carried out in a nonparametric context, a small introduction to nonparametric curve estimation, including some definitions and a small review of the basic concepts is given in Section 1.2. It consists of an overview of the distribution function estimation (including the study of the Kaplan-Meier and Beran estimators and some asymptotic properties of both of them), and to the density function and regression estimation. In Section 1.3 a detailed description of the bootstrap method is provided, considering also the case with censored data and an explanation of the bootstrap bandwidth selection method. Cure models are presented in Section 1.4, which includes a small bibliographical review about parametric and semiparametric methods, and presents the need of nonparametric techniques in this context.

In Chapter 2, the main results for the nonparametric estimator of the probability of cure are introduced. Section 2.2 presents the notation which will be used along the thesis, and Section 2.3 introduces the asymptotic results. Moreover, the bootstrap bandwidth selection method is detailed in Section 2.4. In order to evaluate the performance of the proposed nonparametric estimator and to assess the practical behavior of the bootstrap bandwidth selector, a simulation study is carried out in Section 2.5. Coming up next, in Section 2.6 the methodology is applied to a dataset of colorectal cancer patients from CHUAC.

The nonparametric latency estimator is deeply studied in Chapter 3. Asymptotic results for the estimator considering two different bandwidths and considering only one bandwidth are included in Sections 3.2 and 3.3, respectively. A bootstrap bandwidth selector for the latency estimator is proposed in Section 3.4. In Section 3.5 the results of a simulation study are presented. It consists of three different parts: in the first one, it is shown that little efficiency is lost when considering only one bandwidth in the nonparametric latency estimator; in the second one, the good practical behavior of the proposed estimator is evaluated; and in the third one, the performance of the bootstrap bandwidth selector is assessed. The proposed methods are applied to the colorrectal cancer dataset from CHUAC in Section 3.6.

In Chapter 4, a covariate significance test for the probability of cure is presented in Sections 4.1 and 4.2. Furthermore, the method is extended to non continuous covariates: binary, discrete and qualitative. In Section 4.3, the case with only one covariate (that is, when it is tested if the cure rate, as a function of that covariate, can be considered as a constant value) is introduced. Section 4.4 presents the case where, under the null hypothesis, the probability of cure depends on a one-dimensional covariate. Under the alternative hypothesis, the same probability depends on an m-dimensional covariate, with $m > 1$. In Section 4.5, this approach is also extended to contexts with a large number of covariates. An extensive simulation study is included in Section 4.6. To conclude this chapter, in Section 4.7, the methodology is applied to two medical datasets: one related to colorrectal cancer patients from CHUAC and, for the high dimensional case, a database related to patients with sarcomas from CHUS.

Some comments about future work are given in Chapter 5. The possibility of applying the proposed methodology to high dimensional datasets which include analysis of images, related to cancer for medical diagnosis, is considered. Another plan is to develop an R package with all the techniques studied. Moreover, some other problems for future work are introduced: studying the presmoothed estimator of the probability of cure, extending the methods to cases with truncated and/or interval censored data, using single-index models in survival analysis for censored data, and proving the consistency of the bootstrap methods, studying the limit convergence of the bootstrap.

In Appendices A.1 and A.2, the proofs for the theoretical results in Chapters 2 and 3 are collected.

Chapter 1

State of the art

1.1 Survival analysis (censored data)

Survival analysis is a collection of statistical techniques used to describe and quantify time to event data. The term *failure* is used to specify the occurrence of the event of interest, and the *survival time* refers to the length of time from the beginning of the study until the occurrence of the event. In biomedical applications, this time may represent the survival time of a living organism or the time until a disease is cured. Survival analysis can also be applied to data from different areas: social sciences (time for doing some task); economics (time looking for employment) or engineering (time to a failure of some electronic component).

The main goals of survival analysis are:

- Estimating and comparing the survival functions of different groups. These cumulative survival functions are defined as $S(t) = P(\mathcal{T} > t) = 1 - F(t)$.
- Assessing the relationship of covariates to time-to-event.

The distinguishing feature of survival analysis is that it may incorporate censoring: the exact survival time is only known for those individuals who show the event of interest during the follow-up period. The other individuals (those who are disease free at the end of the observation period and those that were lost) are called censored observations. Among common cases in censoring we highlight:

- *Right censoring*: it occurs when a subject leaves the study before the event happens, or the study ends before the event has occurred. It arises often in medical and biological applications. For example, an objective of the study of the colorectal cancer patients dataset presented in Section 2.6 is to study the

lifetimes. The study ends after 19 years. In this case, censoring may occur in the form of loss to follow-up, drop-out or termination of the study (that is, those patients who are alive by the end of the 19th year are censored). Therefore, the event is only observed if it occurs prior to some specified time.

- *Left censoring*: the lifetime is considered to be left censored if it is known that the failure occurs some time before the recorded follow-up period. That is, the event of interest has already occurred for the individual before the observed time. For example, onset of a pre-symptomatic illness such as cancer, or infection with a sexually-transmitted disease such as HIV/AIDS.

- *Interval censoring*: a subject is interval censored if it is known that the event occurs between two times, but the exact time of failure is not known. Such interval censoring occurs, for example, when patients in a clinical trial or longitudinal study have periodic follow-up.

Furthermore, the reasons which cause an observation to be censored can be random or controlled. It leads to distinguish between two types of censoring:

- *Type I Censoring*: the event is observed if it occurs before a fixed time C. In this case, C is a constant predetermined for all the sample. This type of censoring is common when, for different reasons, the study ends before all the subjects have experienced the event of interest.

- *Type II Censoring*: the study ends when a fixed number of events amongst the subjects has occurred. The observed lifetimes are the r smallest values of the data.

Note that some subjects may experience other occurrences (independent of the event of interest), which cause their dropout of the study. It is called random censoring, where the censoring variable is supposed to be independent of the variable of interest.

In this thesis we assume that each individual is subject to random right censoring. Let us denote by Y the time to occurrence of the event, and denote by C the censoring time. In the presence of random right censoring, for each subject it is only observed the pair (T, δ), where $T = \min(Y, C)$ is the observed time and $\delta = I(Y \leq C)$ is the uncensoring indicator. Note that $I(A)$ is the indicator function of the event A. Associated to this censoring model, some functions can be defined:

- The distribution function of Y is denoted by $F(t) = P(Y \leq t)$, and its survival function is denoted by $S(t) = 1 - F(t) = P(Y > t)$.

- The distribution function of C is denoted by $G(t) = P(C \leq t)$, and its survival function is $1 - G(t) = P(C > t)$.
- The distribution function of the observed variable T is denoted by $H(t) = P(T \leq t)$ and its corresponding survival function is $1 - H(t) = P(T > t) = P(\min(Y, C) > t) = P(Y > t, C > t)$. Under the assumption that Y and C are independent, it is straightforward that $1 - H(t) = (1 - F(t))(1 - G(t))$.

1.2 Nonparametric curve estimation

Nonparametric curve estimation has been one of the most active fields in statistics during the last decades. Methods for nonparametric curve estimation allow to analyze data without much prior information about them, that is, without any parametric assumption on the distribution of the underlying variables.

1.2.1 Distribution function estimation

If the data are not censored, $T_i \equiv Y_i$, the survival function estimation would be simply the empirical survival function (see, for example, Andersen et al., 1993), that is, the proportion of subjects alive at time t:

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

Unlike the context with complete data, in the presence of censored subjects we do not observe a sample of size n of the variable Y, but we do observe the sample $\{(T_i, \delta_i)\}_{i=1}^n$, where T_i is equal to Y_i only when $\delta_i = 1$. If we estimate $S(t)$ using the reduced sample $\{(T_i, \delta_i) : \delta_i = 1\}$, besides being working with a biased estimator, we would be working with a complete dataset with a smaller sample size and the information provided by the censored data would not be considered.

Kaplan & Meier (1958) extended the estimate in (1.1) to censored data:

$$\hat{S}(t) = \prod_{i:T_{(i)} \leq t}^{n} \left(1 - \frac{\delta_{[i]}}{n - i + 1}\right), \tag{1.2}$$

where $\delta_{[i]}$ is the corresponding uncensoring indicator concomitant of $T_{(i)}$, and $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$ are the ordered T_i's. Note that if $t < T_{(1)}$ then $\hat{S}(t) = 1$. The Kaplan-Meier estimator, also known as the product-limit (PL) estimator, is the survival function estimator mostly used for random right censored data. This estimator has some relevant properties:

- It is very simple to calculate.
- If there are no censored data, it is equivalent to the classical empirical estimator. From (1.2) and considering that when there are no censored observations we have that $\delta = 1$ and $T = Y$, then:

$$\begin{aligned}\hat{S}(t) &= \prod_{i:Y_{(i)}\leq t}\left(\frac{n-i}{n-i+1}\right)\\ &= \frac{n-1}{n}\frac{n-2}{n-1}\frac{n-3}{n-2}\cdots\frac{n-k}{n-k+1} = \frac{n-k}{n} = 1-\sum_{i=1}^{n}\frac{I(Y_i\leq t)}{n},\end{aligned}$$

  where $k = \max\{i : Y_{(i)} \leq t\}$ (and clearly $k = \sum_{i=1}^{n} I(Y_i \leq t)$).
- It uses the censored data, which the reduced sample estimator does not consider.
- It presents a stair-step shape with jumps only at the uncensored observations, and weights which depend on the number of censored observations among them. The size of the jump at T_i can be expressed in the following and equivalent ways:

$$\hat{S}(T_{(i-1)}) - \hat{S}(T_{(i)}) = \frac{\delta_{[i]}}{n-i+1}\prod_{k=1}^{i-1}\left(1-\frac{1}{n-k+1}\right)^{\delta_{[k]}} = \frac{\delta_{[i]}}{n-i+1}\hat{S}(T_{(i-1)}).$$

- It is the nonparametric maximum likelihood estimator of $S(t)$ for censored data. The proof is shown in Section 5 of Kaplan & Meier (1958), see also Johansen (1978) and Wang (1987).
- Since a censored observation of Y_i corresponds to an uncensored observation of C_i, then the Kaplan-Meier estimator for the distribution function G is the following:

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

- It is always monotone and has smaller variance than the reduced sample estimator.

It is important to note that the Kaplan-Meier estimator relies on the independence between the censoring and the survival times. Specifically, it may overestimate the survival function of Y if the survival time and the censoring time are positively

correlated, and underestimate the survival function if the times are negatively correlated. Therefore, if independence does not hold, the estimator may be biased (see Kaplan & Meier, 1958).

We will introduce some asymptotic properties of this estimator.

Property 1. *(Theorem 1 in Phadia & Shao, 1999)*
Let $\hat{S}$ be the Kaplan-Meier estimator of the survival function S based on (T_i, δ_i), $i = 1, 2, \ldots, n$. Then the $k-$th moment of $\hat{S}$ is given by:

$$E[\hat{S}^k(t)] = \sum_{i=0}^{n-1} \frac{n!}{(n-i)!} \bar{H}^{n-i}(t) \int_0^t \int_0^{t_i} \cdots \int_0^{t_2} \prod_{j \leq i} d\phi_j(t_j), \tag{1.3}$$

where $\bar{H}(t) = 1 - H(t) = S(t)(1 - G(t))$, $0 < t_1 < t_2 < \cdots < t_i \leq t$ and

$$\phi_j(t) = \left[H(t) - H^1(t) + H^1(t) \left(\frac{n-j}{n-j+1} \right)^k \right],$$

with $H^1(t) = P(T \leq t | \delta = 1)$. The product is taken over $j = 1, 2, \ldots, i$. For $i = 0$, the product is defined to be 1.

Gill (1980) suggested to estimate $1-F(t)$ for $t > T_{(n)}$ with $\hat{S}(T_{(n)}) = 1-\hat{F}(T_{(n)})$, that is, considering that the individual will fail at time $t = \infty$. It is easy to see that for Gill's version the summation in the above expression (1.3) extends to n. However, the above exact formula is not readily amenable to practical applications. Therefore, Phadia & Shao (1999) propose the following approximation.

Property 2. *(Theorem 2 in Phadia & Shao, 1999)*
If we approximate each $\phi_j(x)$ defined above by its linear component $[(\phi_j(t)-\phi_j(0))/t]x$ on the interval $(0, t)$ and substitute in the above expression (1.3), we get:

$$E\left[\hat{S}^k(t)\right] \approx \sum_{i=0}^{n-1} \binom{n}{i} \bar{H}^{n-i}(t) \prod_{j \leq i} \left[H(t) - H^1(t) + H^1(t) \left(\frac{n-j}{n-j+1} \right)^k \right].$$

Breslow & Crowley (1974) obtained the asymptotic normality.

Property 3. *(Asymptotic normality in Breslow & Crowley, 1974)*
Let $F(\cdot)$ and $G(\cdot)$ be two continuous distribution functions, and the distribution function $H(\cdot)$ is defined by:

$$1 - H(t) = (1 - F(t))(1 - G(t)).$$

Moreover, let $b_H = \sup\{t > 0 : H(t) < 1\}$ be the right endpoint of the distribution function $H(\cdot)$. Then,

1. *For all* $0 < t < b_H$,

$$\sqrt{n}\left(\hat{F}(t) - F(t)\right) \xrightarrow{d} N(0, \sigma(t)),$$

where

$$\sigma^2(t) = (1 - F(t))^2 \int_0^t (1 - H(v))^{-2} dH^1(v).$$

2. *The stochastic process* $\sqrt{n}(\hat{F} - F)$ *converges globally in* $\mathcal{D}[0, T]$, *for each* $T < b_H$, *to a Gaussian process* $Z(\cdot)$*:*

$$\sqrt{n}(\hat{F} - F) \xrightarrow{d} Z(\cdot),$$

with mean equal to 0 *and the following covariance function:*

$$Cov(Z(x), Z(t)) = (1 - F(x))(1 - F(t)) \int_0^{\min(x,t)} (1 - H(v))^{-2} dH^1(v),$$

where $\mathcal{D}[0, T] = \{H$ *a function from* $[0, T]$ *in* $\mathbb{R}$*: H is right continuous, with discontinuities, at most, of jump*$\}$.

Földes & Rejtö (1981), among others, studied the consistency of the Kaplan-Meier estimator.

Property 4. *(Strong uniform consistency in Földes & Rejtö, 1981)*

1. *Let* $0 < T < b_H$. *Therefore,*

$$|\hat{F} - F| \to 0 \text{ a.s. uniformly in } [0, T].$$

2. *Besides that, if* $G(b_F^-) > 0$, *where* $b_F = \sup\{t : F(t) < 1\}$, *then:*

$$|\hat{F} - F| \to 0 \text{ a.s. uniformly in } \mathbb{R}.$$

Lo & Singh (1986) obtained a strong uniform approximation of the difference between the Kaplan-Meier estimator, $\hat{F}(\cdot)$, and the theoretical distribution function, $F(\cdot)$, as a mean of independent and identically distributed (i.i.d.) bounded random variables, plus a neglibible term of known order. These approaches allow us to work with a sum of i.i.d. variables, easier to manipulate than the product which defines the estimator $\hat{F}(\cdot)$. Departing from it, properties such as the asymptotic normality and the convergence of the process can be easily studied.

Property 5. *(Almost sure representation in Lo & Singh, 1986)*
Under the hypothesis that $F(\cdot)$ and $G(\cdot)$ are continuous, then for all $t \leq T < b_H$:

$$\hat{F}(t) - F(t) = n^{-1} \sum_{i=1}^{n} \xi(T_i, \delta_i, t) + r_n(t),$$

where

$$\xi(T, \delta, t) = (1 - F(t)) \left[g(\min(T, t)) + \frac{1}{1 - H(T)} I(T \leq t, \delta = 1) \right],$$

with

$$g(t) = \int_0^t (1 - H(v))^{-2} d(1 - H^1(v))$$

and

$$\sup_{0 \leq t \leq T} |r_n(t)| = O\left(\left(\frac{\ln n}{n} \right)^{3/4} \right) \quad a.s.$$

Furthermore, for all $\alpha \geq 1$,

$$\sup_{0 \leq t \leq T} E|r_n(t)|^{\alpha} = O\left(\left(\frac{\ln n}{n} \right)^{3\alpha/4} \right).$$

The order of the term $r_n(\cdot)$ is sufficient to prove most of the properties of the Kaplan-Meier estimator.

The nonparametric estimation of the conditional survival function with censored data has been studied by different authors. In order to estimate the conditional survival function for a continuous covariate X, Beran (1981) introduces the conditional PL estimator, also known as the generalized Kaplan-Meier estimator:

$$\hat{S}_h(t|x) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{\delta_{[i]} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)} \right), \tag{1.4}$$

where

$$B_{h(i)}(x) = \frac{K_h(x - X_{[i]})}{\sum_{j=1}^{n} K_h(x - X_{[j]})}$$

are the Nadaraya-Watson (NW) weights with $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$ the rescaled kernel with bandwidth $h \to 0$. Here $X_{[i]}$ is the covariate concomitant of $T_{(i)}$. We will also denote $\hat{F}_h(t|x) = 1 - \hat{S}_h(t|x)$ the Beran estimator of $F(t|x)$.

The Beran estimator of the conditional survival function has been deeply studied in the literature. Depending on the design (fixed or random) and on the weights,

different properties of the estimator are obtained. Dabrowska (1989), in Theorem 2.1, shows its asymptotic unbiasedness, considering Nadaraya-Watson weights. Furthermore, using Gasser-Müller weights, González-Manteiga & Cadarso-Suárez (1994) give an almost sure representation for the estimator as a sum of independent variables plus a remainder term, and Van Keilegom & Veraverbeke (1997b) prove an asymptotic representation for the bootstrapped estimator and obtain a strong consistency of the bootstrap approximation for the conditional distribution function.

We now summarize the most interesting properties of the conditional product-limit estimator under censoring.

Property 6. *Following Beran (1981), the PL estimator in (1.4) is equal to:*

- *The kernel type estimator of the conditional survival function (for complete data) if there are no censored observations.*
- *The Kaplan-Meier estimator if there are no covariates.*

Property 7. *The product of the PL estimators for the survival functions of the variables Y and C is equal to the nonparametric survival estimator of the variable T:*

$$(1-\hat{F}_h(t|x))(1-\hat{G}_h(t|x)) = 1-\hat{H}_h(t|x). \tag{1.5}$$

Detailing the equations of the left hand side in (1.5),

$$\begin{aligned} &\prod_{i:T_{(i)}\leq t}\left(1-\frac{\delta_{[i]}B_{h(i)}(x)}{\sum_{r=i}^{n}B_{h(r)}(x)}\right)\prod_{i:T_{(i)}\leq t}\left(1-\frac{(1-\delta_{[i]})B_{h(i)}(x)}{\sum_{r=i}^{n}B_{h(r)}(x)}\right) \\ = &\prod_{i:T_{(i)}\leq t}\left(1-\frac{B_{h(i)}(x)}{\sum_{r=i}^{n}B_{h(r)}(x)}\right) = 1-\sum_{i=1}^{n}B_{h(i)}(x)I(T_{(i)}\leq t). \end{aligned}$$

Property 8. *The PL estimator in (1.4) is a step function, with jumps at the uncensored observations. Departing from Efron (1967), who works in an unconditional context, the magnitude of the jump at T_i in our case is equal to:*

$$d\hat{F}_h(T_i|x) = \frac{\delta_i B_{h,i}(x)}{1-\hat{H}_h(T_i|x)}\left(1-\hat{F}_h(T_i^-|x)\right), \tag{1.6}$$

where T_i^- is the left limit of T_i. Note that Equation (1.6) is equivalent to

$$d\hat{F}_h(T_{(i)}|x) = \frac{\delta_{[i]}B_{h(i)}(x)}{1-\hat{H}_h(T_{(i)}|x)}\left(1-\hat{F}_h(T_{(i-1)}|x)\right).$$

Property 9. *The PL estimator in (1.4) for the survival function of the censoring variable C contitional on X, $1 - G(t|x)$, is the following:*

$$1 - \hat{G}_h(t|x) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{(1 - \delta_{[i]})B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right),$$

since a censored observation of Y_i corresponds with an uncensored observation of C_i and vice versa.

Property 10. *(Asymptotic normality obtained by González-Manteiga & Cadarso-Suárez, 1994, and Van Keilegom & Veraverbeke, 1997b).*
Under certain conditions,

$$(nh)^{1/2}[\hat{F}_h(\cdot|x) - F(\cdot|x)] \xrightarrow{d} N(\cdot|x) \text{ in } \mathcal{D}[0, T],$$

where $N(\cdot|x)$ is a Gaussian process with mean equal to 0 and covariance function:

$$\Gamma(y, t|x) = (1 - F(y|x))(1 - F(t|x)) \left(\int K^2(z)dz\right) \left(\int_0^{\min(y,t)} \frac{dH_1(u|x)}{(1 - H(u|x))^2}\right),$$

and $\mathcal{D}[0, T] = \{f$ a function from $[0, T]$ in $\mathbb{R}$: f is right continuous, with discontinuities, at most, of jump$\}$, considering the topology of Skorohod (see Billingsley, 1968, pg. 111).

The weak convergence of the process $(nh)^{1/2}[\hat{F}_h(\cdot|x) - F(\cdot|x)]$ is studied in Van Keilegom & Veraverbeke (1997a).

1.2.2 Density function estimation

Let $(X_1, X_2, \ldots, X_n)$ be an independent and identically distributed sample drawn from some unknown distribution function F with density function f. Suppose we are interested in estimating the density function. The most common nonparametric estimator for the density function of a random variable X is the kernel type estimator proposed by Parzen (1962) and Rosenblatt (1956) which, in the complete data setup (uncensored case), is:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$, K is a kernel function (bounded and symmetric real function with $\int_{-\infty}^{\infty} K(x)dx = 1$), and $h > 0$ is the smoothing parameter.

In the presence of right censoring, Diehl & Stute (1988) obtained an i.i.d. representation of the kernel estimator of f under censoring:

$$f_h(t) = \frac{1}{h}\int K\left(\frac{t-x}{h}\right)\hat{F}_h(dx),$$

where $(h) = (h_n)$ is a sequence of bandwidths tending to zero at appropriate rates, K is a smooth kernel and $1-\hat{F}_h(x)$ is the Kaplan-Meier estimator of the survival function $1-F(x)$. The main results for the kernel estimator of the density under censoring are obtained in Földes et al. (1981), Mielniczuk (1986), Diehl & Stute (1988), Lo et al. (1989) and Ghorai & Pattanaik (1990), among others. Furthermore, in Cai (1998), dependent data are also considered.

1.2.3 Regression estimation

Nonparametric regression smoothing includes many techniques to estimate the regression function without making any assumption about its shape.

Let (X,Y) be a two-dimensional random variable such that $E(|Y|) < \infty$. We will estimate the regression function $r(x) = E(Y|X=x)$, that can be also expressed as:

$$\begin{aligned} r(x) &= \int y f_{2|1}(y|x)dy = \int y\frac{f(x,y)}{f_1(x)}dy = \frac{\int y f(x,y)dy}{f_1(x)} \\ &= \frac{\int y f_{1|2}(x|y)f_2(y)dy}{f_1(x)} = \frac{\Phi(x)}{f_1(x)}, \end{aligned}$$

where $f_1(x)$ is the marginal density function of X and

$$\Phi(x) = \int y f_{1|2}(x|y)f_2(y)dy = E(Y f_{1|2}(x|Y)).$$

Note that $E(Y f_{1|2}(x|Y))$ can be estimated by $\frac{1}{n}\sum_{i=1}^n Y_i f_{1|2}(x|Y_i)$, and the conditional density $f_{1|2}(x|Y_i)$ can be estimated with a kernel type estimator and considering only one observation X_i $(n=1)$. Therefore, both functions $f_1(x)$ and $\Phi(x)$ can be estimated by the kernel method:

$$\hat{f}_{1,h}(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) \text{ and } \hat{\Phi}_h(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)Y_i,$$

which lead to the Nadaraya-Watson kernel estimator (Nadaraya (1964) and Watson (1964)):

$$\hat{r}_h(x) = \frac{\hat{\Phi}_h(x)}{\hat{f}_{1,h}(x)} = \frac{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)Y_i}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} = \sum_{i=1}^n B_{hi}(x)Y_i, \tag{1.7}$$

where

$$B_{hi}(x) = \frac{\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)}{\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)} = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)}. \tag{1.8}$$

Intuitively, $\hat{r}_h(x)$ is a weighted local average, that is, $r(x)$ is estimated by a weighted mean of Y_i, where the weights, $B_{hi}(x)$, take into account the distance between the values X_i and x, and the smoothing parameter, h, adjusts the size of the weights near x.

Following Nadaraya (1964) and Watson (1964), the asymptotic distribution of (1.7) can be obtained under moment and regularity conditions and also under the necessary conditions for consistency of the estimator ($h \to 0$, $nh \to \infty$ if $n \to \infty$). If these conditions are fulfilled and the bandwidth h is chosen as $c_0 n^{-1/5}$, then:

$$\sqrt{nh}\left(\hat{r}_h(x) - r(x)\right) \xrightarrow{d} N(B(x), V(x)),$$

with

$$B(x) = \frac{1}{2}c_0^{5/2}d_K\frac{r''(x)f(x) + 2r'(x)f'(x)}{f(x)} \text{ and } V(x) = c_K\frac{\sigma^2(x)}{f(x)},$$

where

$$d_K = \int v^2 K(v)dv, \tag{1.9}$$

$$c_K = \int K^2(v)dv, \tag{1.10}$$

$f(x)$ is the marginal density function of X and $\sigma^2(x) = Var(Y|X = x)$ is the conditional variance of Y given $X = x$. Furthermore, it can be proven that the asymptotically optimal value of h regarding the mean squared error (MSE) has the form: $h = c_0 n^{-1/5}$. The smoothing parameter can be selected using automatic data-dependent methods. The bandwidth selection techniques that will be mentioned in this thesis are plug-in, cross validation (CV) and bootstrap methods.

1.3 Bootstrap

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. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a simple random sample of the unknown distribution F. The bootstrap method approximates $R(\mathbf{Y}, F)$ by $R^* = R(\mathbf{Y}^*, \hat{F})$ using the following procedure:

1. From the empirical distribution function:

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y),$$

obtain bootstrap resamples $(Y_1^*, Y_2^*, \ldots, Y_n^*)$, that is, $Y_i^* = Y_j$ with probability $1/n$, for $j = 1, \ldots, n$.

2. Approximate the sampling distribution $R(\mathbf{Y}, F)$ by the distribution in the resampling of $R^* = R(\mathbf{Y}^*, \hat{F})$. Specifically, the theoretical distribution is replaced by the empirical distribution, and the observed sample is replaced by the bootstrap resample.

Note that in order to approximate the distribution of R^*, the Monte Carlo method is used. We generate B bootstrap resamples of size n from the distribution $\hat{F}(\cdot)$, that is, we obtain $\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_B^*$ random bootstrap resamples drawn from $\hat{F}$. Then, with the values $R(\mathbf{Y}_1^*, \hat{F})$, $R(\mathbf{Y}_2^*, \hat{F})$, $\ldots$, $R(\mathbf{Y}_B^*, \hat{F})$, we obtain the approximation of R.

A noticeable disadvantage of this method is that it is computationally expensive, since it is based on Monte Carlo. Apart from that, the most important benefit is that the bootstrap method does not need any hypothesis about the mechanism which generates the data (see Efron, 1979; Hall, 1992; Efron & Tibshirani, 1993, among others).

From now on, we will use the notation E^* and P^* for bootstrap expectation and probability, i.e., conditionally on the original observations.

1.3.1 Bootstrap with censored data

The resampling needs to be adaptable to each context. With censored data, Efron (1981) introduced two equivalent resampling methods known as *simple bootstrap* and *obvious bootstrap* (see also Reid, 1981; Akritas, 1986). Both procedures approximate the distribution of an statistic $R(\mathbf{T}, \boldsymbol{\delta})$, where $\mathbf{T} = (T_1, T_2, \ldots, T_n)$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$.

The simple bootstrap consists of the following steps:

1. From the sample $\{(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n)\}$, obtain the two-dimensional empirical distribution, $\hat{F}^{T,\delta}$.

2. From the empirical distribution in Step 1, obtain $\{(T_1^*, \delta_1^*), (T_2^*, \delta_2^*), \ldots, (T_n^*, \delta_n^*)\}$ considering:

$$P((T^*, \delta^*) = (T_i, \delta_i)) = \frac{1}{n}, \text{ for } i = 1, 2, \ldots, n.$$

3. Evaluate the statistic $R^* = R(\mathbf{T}^*, \boldsymbol{\delta}^*)$, with $\mathbf{T}^* = (T_1^*, T_2^*, \ldots, T_n^*)$ and $\boldsymbol{\delta}^* = (\delta_1^*, \delta_2^*, \ldots, \delta_n^*)$.

4. Approximate the sampling distribution of the statistic $R(\mathbf{T}, \boldsymbol{\delta})$ by the resampling distribution of the corresponding bootstrap statistic $R(\mathbf{T}^*, \boldsymbol{\delta}^*)$.

Furthermore, the obvious bootstrap can be applied following the steps:

1. Obtain the Kaplan-Meier estimators of the variable of interest, $\hat{F}(t)$, and the censoring variable, $\hat{G}(t)$.

2. Obtain independent observations, Y_i^* with distribution $\hat{F}$ and C_i^* with distribution $\hat{G}$, for $i = 1, 2, \ldots, n$.

3. For $i = 1, 2, \ldots, n$, define $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \leq C_i^*)$, and consider the bootstrap resample $(\mathbf{T}^*, \boldsymbol{\delta}^*)$, with $\mathbf{T}^* = (T_1^*, T_2^*, \ldots, T_n^*)$ and $\boldsymbol{\delta}^* = (\delta_1^*, \delta_2^*, \ldots, \delta_n^*)$.

4. Approximate the sampling distribution of the statistic $R(\mathbf{T}, \boldsymbol{\delta})$ by the resampling distribution of the corresponding bootstrap statistic $R(\mathbf{T}^*, \boldsymbol{\delta}^*)$.

Note that this method is computationally more expensive than the simple bootstrap.

It is straightforward to prove that the simple and the obvious bootstrap are equivalent when there are no ties between censored and uncensored observations (see Efron, 1981). Then, the bootstrap distribution of (T^*, δ^*) is the same for both methods.

1.3.2 Bootstrap with covariates

Li & Datta (2001) propose a two-stage bootstrap method for nonparametric regression with right censored data. Its asymptotic validity is established using counting process techniques and martingale central limit theory. The method is applied to construct confidence intervals and bands for the conditional survival function estimate.

The proposed resampling is carried out in two steps. In the first stage, they resample with replacement from the set $\{X_1, ..., X_n\}$ to obtain the bootstrap sample for the covariate $\{X_1^*, ..., X_n^*\}$. Then, in the second stage, they generate a pair (T_i^*, δ_i^*) for each X_i^*; using ideas similar to that of Efron (1981). The additional phase of covariate resampling also introduces a technical challenge for establishing the asymptotic validity of the bootstrap.

Li & Datta (2001) give two equivalent resampling algorithms for bootstrapping the Beran estimate of the conditional survival function: the simple weighted bootstrap and the obvious bootstrap.

The simple weighted bootstrap with covariates is detailed in the following steps:

1. Generate $X_1^*, \ldots, X_n^*$ i.i.d. from the empirical distribution of $\{X_1, \ldots, X_n\}$.
2. For each i, generate a pair (T_i^*, δ_i^*) from the weighted empirical distribution $\hat{F}_h(\cdot, \cdot | X_i^*)$ of $\{(T_1, \delta_1), \ldots, (T_n, \delta_n)\}$, where
$$\hat{F}_h(u, v|x) = \sum_{i=1}^{n} B_{hi}(x) I(T_i \leq u, \delta_i \leq v),$$
and $B_{hi}(x)$ is defined in (1.8).
3. The bootstrap resample is formed as $\{(T_1^*, \delta_1^*, X_1^*), \ldots, (T_n^*, \delta_n^*, X_n^*)\}$.

Regarding the obvious bootstrap with covariates, let us recall $\hat{S}_h(t|x)$ and $\hat{G}_h(t|x)$ as the Beran estimates of $S(t|x)$ and $G(t|x)$, respectively, using the same weight function $B_{hi}(x)$ in (1.8). Li & Datta (2001) force $\hat{S}_h(t|X_i^*)$ and $\hat{G}_h(t|X_i^*)$ to 0 beyond the larger of the last jump points of the two step functions to make both proper survival functions in order to sample the failure and censoring times described below:

1. Generate $X_1^*, \ldots, X_n^*$ i.i.d. from the empirical distribution of $\{X_1, \ldots, X_n\}$.
2. For each i, generate Y_i^* from $\hat{S}_h(t|X_i^*)$ and C_i^* from $\hat{G}_h(t|X_i^*)$ independently, and define:
$$T_i^* = \min(Y_i^*, C_i^*) \qquad \text{and} \qquad \delta_i^* = I(Y_i^* \leq C_i^*).$$
3. The bootstrap resample is formed as $\{(T_1^*, \delta_1^*, X_1^*), \ldots, (T_n^*, \delta_n^*, X_n^*)\}$.

The equivalence of the two methods is parallel to that of Efron (1981) resampling procedures for the unconditional setting.

1.3.3 Bootstrap bandwidth selection

The choice of the bandwidth is a crucial issue in kernel estimation, since it controls the trade-off between bias and variance. Suppose that we want to study a generic function, ϕ. Note that when choosing the bandwidth, a large parameter value will oversmooth the data, which will lead to a biased estimate of ϕ. On the contrary, a small parameter value will undersmooth the data and it will give an estimator for ϕ with high variability. Most of the methods for smoothing parameter selection in nonparametric curve estimation look for a small error when approximating the underlying curve by the smooth estimate.

The general idea of a bootstrap bandwidth selector consists of estimating, by resampling, the mean squared error ($MSE_x(h_x)$) or the mean integrated squared error ($MISE_x(h_x)$), and obtain the bandwidth which minimizes its bootstrap version.

For a variable X with distribution function F, and considering the classical notation for the bootstrap method, it is important to know the distribution of the random variable:

$$R_h(\mathbf{X}, F) = \int \left(\hat{\phi}_h(v) - \phi(v)\right)^2 w(v)dv,$$

where h is the smoothing parameter, ϕ is a generic function depending on F, $\hat{\phi}_h$ is a smooth estimator of ϕ, and w is a weight function. The bootstrap method approximates the distribution of $R_h(\mathbf{X}, F)$ by the bootstrap distribution of $R^* = R_h(\mathbf{X}^*, \hat{F})$, that is, the theoretical distribution F is replaced by an estimation $\hat{F}$, and the observed sample $\mathbf{X}$ is replaced by the bootstrap resample $\mathbf{X}^*$, drawn from $\hat{F}(\cdot)$. The most commonly used technique for approximating the bootstrap distribution of R^* is by means of Monte Carlo, by generating B resamples of size n: $\mathbf{X_1}^*, \ldots, \mathbf{X_B}^*$, from $\hat{F}(\cdot)$ and approximating the resampling distribution of R^* by the empirical distribution of $R_1^* = R_h(\mathbf{X_1}^*, \hat{F}), \ldots, R_B^* = R_h(\mathbf{X_B}^*, \hat{F})$.

A common error criterion for bandwidth selection is $MISE(h) = E[R_h(\mathbf{X}, F)]$, which can be estimated by its bootstrap version $MISE^*(h) = E^*[R_h(\mathbf{X}^*, \hat{F})]$, that can be approximated by Monte Carlo using $\frac{1}{B}\sum_{j=1}^{B} R_h(\mathbf{X}_j^*, \hat{F})$.

In this thesis, for the bootstrap bandwidth selector for the cure probability estimator, we consider the simple weighted bootstrap, without resampling the covariate X, and we depart from the simple weighted bootstrap proposed by Li & Datta

(2001). Regarding the bootstrap bandwidth selector for the estimator of the survival function of the uncured patients, we consider an unconditional censoring bootstrap resampling, assuming that $G(t|x) = G(t)$. The resampling method is based on the obvious bootstrap proposed by Li & Datta (2001).

1.4 Cure models

In the last two decades there has been a remarkable progress in cancer treatments, which led to longer patient survival and improved their quality of life. Consequently, data coming from cancer studies typically have heavy censoring (due to long term survival) at the end of the follow-up period, and a standard survival model is inappropriate. To accommodate for the cured or insusceptible proportion of subjects, a cure fraction can be explicitly incorporated into survival models and, as a consequence, cure models arise.

These models are a useful tool to analyze and describe cancer survival data, since they express and predict the prognosis of a patient considering, as a novelty, the real possibility that the subject may never experience the event of interest. They generally require long-term follow-up and large sample sizes, as well as empirical and biological evidence of a nonsusceptible subpopulation (Farewell, 1986). In Figure 1.1 we can see a standard survival function and a survival function with a fraction of cured patients.

Although cure models were originally proposed to model long-term survival of cancer patients, as in this thesis, they can be applied to any survival context in which it is assumed that a group of individuals will not experiment the event of interest, no matter how long they are followed. For example, credit scoring where a proportion of borrowers will not default during the loan term.

Let us introduce a real example in which we do not consider the cure possibility. We work with a dataset related to colorectal cancer patients from CHUAC (Complexo Hospitalario Universitario de A Coruña), Spain. It contains 414 observations on 8 variables. The event of interest is the follow-up time, in months, since the diagnostic until death. Censoring is caused by death by a different cause, dropout, or end of the study. The dataset is described in detail in Section 2.6. Figure 1.2 shows the Kaplan-Meier estimator for the survival function for the complete dataset, $\hat{S}$, to-

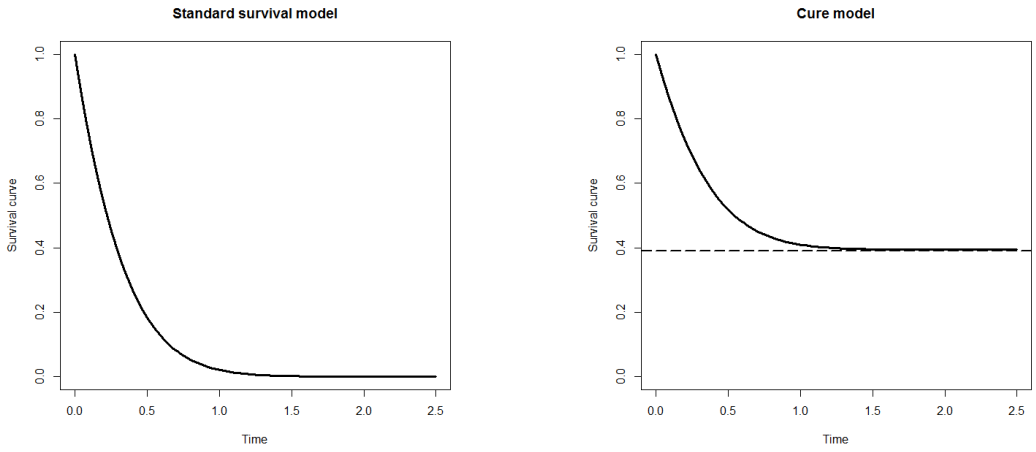

Figure 1.1: Standard survival function (left) and survival function with a fraction of cured patients (right).

gether with the survival function of the uncured individuals, $\hat{S}_0$. We can appreciate that the survival curve of $\hat{S}$ has a plateau at the end of the study. This non-zero asymptote could be taken as an estimator of the cure rate, that is, the proportion of patients who will not die from colorectal cancer (so they can be considered as "cured"). Specifically, the estimated cure probability is $1-\hat{p} = 0.16$ for this dataset. Since a standard survival model does not take into account the proportion of cured patients, it is not an appropriate way to analyze the data. On the contrary, a cure model may be a suitable alternative.

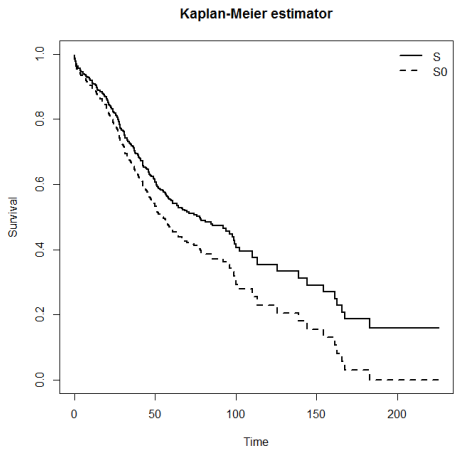

Figure 1.2: Standard survival function (solid line) and survival function with a fraction of cured patients (dashed line).

There are two main classes of cure models: mixture and non-mixture models. The first papers in non-mixture models were due to Haybittle (1959, 1965). One category, belonging to this group, is the proportional hazards (PH) cure model, also known as the promotion time cure model, proposed by Yakovlev & Tsodikov (1996):

$$\bar{S}(t,x) = G(-a(\theta,x)F_0(t)),$$

where a is a known link function depending on the unknown parameter θ, and G is a known transformation function. The unknown terms in this model can be estimated parametrically (Yakovlev et al., 1994; Chen et al., 1999, 2002; Yin, 2005; Chi & Ibrahim, 2007) or semiparametrically (Tsodikov, 1998, 2003; Zeng et al., 2006). Betensky & Schoenfeld (2001) applied a competing risks model and showed it to be equivalent to a mixture model. Li et al. (2001) investigated the identifiability of a standard cure model based on a mixture distribution and based on a non-mixture proportional hazards model with long-term survivors. Furthermore, they derived an estimator for the variance of the cumulative incidence function. Moreover, Tsodikov (2001) proposed a nonparametric estimator of the incidence, but it cannot handle continuous covariates. Liu & Shen (2009) presented a semiparametric nonmixture cure model for the regression analysis of interval-censored time-to-event data. They developed semiparametric maximum likelihood estimation for the model using the expectation-maximization (EM) method for interval-censored data.

The model belonging to the other category of cure models, called two-component mixture cure model, is the one studied in this work. Mixture cure models were proposed by Boag (1949) and they consider the survival function as a mixture of two groups of subjects: the susceptible group and the cured group. More specifically, they allow to estimate the probability of cure, also known as *incidence*, and the survival function of the uncured population, denoted by *latency*. The model can be formulated as follows:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t|\mathbf{x}),$$

where $\mathbf{x}$ is a set of covariates, $S(t|\mathbf{x})$ is the survival function of all the (cured and uncured) patients, $1 - p(\mathbf{x})$ is the incidence and $S_0(t|\mathbf{x})$ is the latency. An important benefit of this model is that it allows covariates to have different influence on cured and uncured patients. Maller & Zhou (1996) provided a detailed review of this model. More recently, Corbière et al. (2009) presented the main approaches so far for mixture cure models. In these models, the incidence is usually assumed to have a logistic form and the latency is usually estimated parametrically (Farewell,

1982, 1986; Cantor & Shuster, 1992; Ghitany et al., 1994; Denham et al., 1996) or semiparametrically (Kuk & Chen, 1992; Peng et al., 1998; Peng & Dear, 2000; Sy & Taylor, 2000; Zhang & Peng, 2007). Farewell (1982, 1986) studied the logistic Weibull mixture cure model when the survival function of the uncured population is the Weibull distribution and covariates are related to the cure probability by a logistic expression. Yamaguchi (1992) used an accelerated failure time model with generalized Gamma distribution as the latency. Kuk & Chen (1992) proposed to model the effect of covariates on the failure time of uncured patients by the proportional hazards assumption. They applied a marginal likelihood approach and used an estimation method involving Monte Carlo simulation. Taylor (1995) implemented an EM algorithm for the parameter inference in the mixture cure model by a Kaplan-Meier type estimator of the survival function of the susceptible population. Peng et al. (1998) proposed to use a generalized F distribution for the survival function of uncured patients. Peng & Dear (2000), Sy & Taylor (2000) and Sy & Taylor (2001) employed an EM algorithm for the Cox's proportional hazard cure model. Li & Taylor (2002) introduced a semiparametric accelerated failure-time cure model, where the latency distribution is determined by an AFT model with unspecifiied baseline distribution. The approach by Peng (2003) is based on recognizing that the M-step of the proposed EM algorithm consists of fitting a proportional hazards model and a logistic model, respectively, with some fixed coefficients. Lu & Ying (2004) proposed a class of semiparametric transformation models incorporating cure fractions, which included the aforementioned mixture cure rate models as special cases. To take account for unobserved heterogeneity which cannot be explained via observed covariates, Peng & Zhang (2008) included a frailty into Cox's proportional hazards mixture cure model. Othus et al. (2009) extended their model to allow for time dependent covariates and dependent censoring. Some recent topics covered in the mixture cure models literature include multivariate survival data (Yu & Peng, 2008), clustered survival data (Lai & Yau, 2010) and accelerated models (Zhang et al., 2013). More recently, Lu (2010) proposed an accelerated failure time model with cure fraction, where the unknown error density was estimated by the kernel method.

Due to the fact that the effects of the covariate on the cure rate and on the latency cannot always be well approximated using parametric or semiparametric methods, a nonparametric approach is needed. In the literature, some nonparametric methods for the estimation of the cure rate have been studied: Maller & Zhou (1992) proposed a consistent nonparametric estimator of the incidence, but it cannot handle covariates. In order to overcome this drawback, Laska & Meis-

ner (1992) proposed another nonparametric estimator of the cure rate, but it only works for discrete covariates. They study the nonparametric generalized maximum likelihood product limit point estimators and confidence intervals for a cure model with random censorship. They also developed one-, two- and K-sample likelihood ratio tests for inference on the cure rates. Furthermore, Wang et al. (2012) proposed a cure model with a nonparametric form in the cure probability. To ensure model identifiability, they assumed a nonparametric proportional hazards model for the hazard function, whose relative risk part also takes a flexible nonparametric form, different from the traditional semiparametric proportional hazards model. The estimation was carried out by an EM algorithm for a penalized likelihood. They defined the smoothing spline function estimates as the minimizers of the penalized likelihood, which consists of the negative log likelihood representing the goodness-of-fit, a roughness penalty enforcing smooth conditions, and a smoothing parameter balancing the tradeoff. In Van Keilegom (2013), the problem of goodness-of-fit tests for regression models with cured data was briefly considered. More recently, a completely nonparametric approach to the mixture cure model was firstly addressed by Xu & Peng (2014), proposing a nonparametric incidence estimator which allows for a continuous covariate, and proving its consistency and asymptotic normality.

Although the aforementioned papers have a nonparametric flavor, they fail to consider a completely nonparametric mixture cure model which works for discrete and continuous covariates in both the incidence and the latency. To overcome this problem, in Chapters 2 and 3 a two-component mixture cure model is proposed with nonparametric forms for both the cure probability and the survival function of the uncured individuals. Even though it is considered only one covariate, the method can be directly extended to a case with multiple covariates (using, for example, the Nadaraya-Watson kernel estimator, obtaining $\hat{S}_h(t|\mathbf{x})$, and facing the problem of high dimensional estimation in nonparametric contexts). This enables the mixture cure model with covariates to be addressed in a completely nonparametric way.

To the best of our knowledge, no nonparametric significance testing has been proposed yet in cure models. To fill this important gap, a covariate significance test for the incidence is presented in Chapter 4. The method is based on the significance test by Delgado & González-Manteiga (2001). Its efficiency is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap.

1.5 Content of the thesis

The rest of the thesis is organized as follows. In Chapter 2 some notation and a completely nonparametric method for the estimation of mixture cure models are introduced. The nonparametric incidence estimator by Xu & Peng (2014) is extensively studied and an i.i.d. representation for it is obtained. As a consequence, an asymptotically optimal bandwidth is found. Moreover, a bootstrap bandwidth selection method is proposed. The nonparametric estimator is compared with existing semiparametric approaches in a simulation study, in which the performance of the bootstrap bandwidth selector is also assessed. Finally, the method is applied to a dataset of colorectal cancer from the University Hospital of A Coruña (CHUAC). The nonparametric latency estimator is deeply studied in Chapter 3. An i.i.d. representation is obtained, the asymptotic mean squared error of the latency estimator is found, and its asymptotic normality is proven. A bootstrap bandwidth selector for the latency estimator is introduced and its efficiency is evaluated in a simulation study. The proposed nonparametric latency estimator, together with the bandwidth selector, are applied to the colorectal cancer dataset. In Chapter 4, a covariate significance test for the incidence is presented. The method is extended to non continuous covariates: binary, discrete and qualitative, and also to contexts with a large number of covariates. The behavior of the test is assessed in a simulation study, and finally it is applied to two datasets: the colorrectal cancer patients dataset from CHUAC and a sarcomas dataset from CHUS (University Hospital of Santiago). The proofs for the theoretical results in Chapters 2 and 3 are collected in Appendices A.1 and A.2.

Chapter 2

Nonparametric incidence estimator

2.1 Introduction

In this chapter, we introduce some notation and we study the nonparametric incidence estimator by Xu & Peng (2014). We address the identifiability problem and we obtain an i.i.d. representation for the estimator. We also find an asymptotic expression of the mean squared error of the incidence estimator. Furthermore, we propose a bootstrap bandwidth selection method. We assess the performance of both the nonparametric estimator and the bootstrap bandwidth selector in a simulation study. Finally, we apply the methodology to a dataset of colorectal cancer patients from CHUAC. The results of this chapter have been published in López-Cheda et al. (2017a).

2.2 Notation

Let ν be a binary variable where $\nu = 0$ indicates if the individual belongs to the susceptible group (the individual will eventually experience the event of interest if followed for long enough) and $\nu = 1$ indicates if the subject is cured (the individual will never experience the event). The proportion of cured patients and the survival function in the group of uncured patients can depend on certain characteristics of the subject, represented by a set of covariates $\mathbf{X}$. Let $p(\mathbf{x}) = P(\nu = 0|\mathbf{X} = \mathbf{x})$ be the conditional probability of not being cured, and let Y be the time to occurrence of the event. When $\nu = 1$ it is assumed that $Y = \infty$.

As mentioned in Chapter 1, we will denote the conditional distribution function of Y as $F(t|\mathbf{x}) = P(Y \leq t|\mathbf{X} = \mathbf{x})$. Note that the corresponding survival function, $S(t|\mathbf{x})$, is improper when cured patients exist, since $\lim_{t\to\infty} S(t|\mathbf{x}) = 1 - p(\mathbf{x}) > 0$. The conditional survival function of Y given that the subject is not cured is denoted as

$$S_0(t|\mathbf{x}) = P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 0).$$

Total probability theorem gives:

$$\begin{aligned} P(Y > t|\mathbf{X} = \mathbf{x}) &= P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 1)P(\nu = 1|\mathbf{X} = \mathbf{x}) \\ &+ P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 0)P(\nu = 0|\mathbf{X} = \mathbf{x}). \end{aligned}$$

Then, the mixture cure model can be written as:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t|\mathbf{x}), \tag{2.1}$$

where $1-p(\mathbf{x})$ is the incidence and $S_0(t|\mathbf{x})$ is the latency. We assume that each individual is subject to random right censoring and that the censoring time, C, with distribution function G, is independent of Y given the covariates $\mathbf{X}$. Let $T = \min(Y, C)$ be the observed time with distribution function H and $\delta = I(Y \leq C)$ the uncensoring indicator. Observe that $\delta = 0$ for all the cured patients, and it also happens for uncured patients with censored lifetime. Without loss of generality, let X be a univariate continuous covariate with density function $m(x)$. Therefore, the observations will be $\{(X_i, T_i, \delta_i), i = 1, \ldots, n\}$, i.i.d. copies of the random vector (X, T, δ).

In order to introduce the nonparametric approach in mixture cure models, we consider the generalized Kaplan-Meier estimator by Beran (1981) to estimate the conditional survival function with covariates:

$$\hat{S}_h(t|x) = \prod_{i:T_{(i)}\leq t} \left(1 - \frac{\delta_{[i]} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right), \tag{2.2}$$

where

$$B_{h(i)}(x) = \frac{K_h(x - X_{[i]})}{\sum_{j=1}^{n} K_h(x - X_{[j]})} \tag{2.3}$$

are the Nadaraya-Watson (NW) weights with $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ the rescaled kernel with bandwidth $h > 0$. In the case of fixed design, the Gasser-Müller (GM) weights (Gasser & Müller, 1984) are more common. Here $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$ are the ordered T_i's, and $\delta_{[i]}$ and $X_{[i]}$ are the corresponding uncensoring indicator and covariate concomitants. We will also denote $\hat{F}_h(t|x) = 1 - \hat{S}_h(t|x)$ for the Beran

estimator of $F(t|x)$. The estimator (2.2) can be extended to the case of multiple covariates $\mathbf{X} = (X_1, \ldots, X_q)$ using, for example, the product kernel (Simonoff, 1996). Discrete covariates can also be included by splitting the sample into subsamples corresponding to the different category combination of the discrete covariates, for each subsample conducting a nonparametric Beran estimator on the continuous covariates. Note that the previous approach requires enough data. Another possibility is smoothing the discrete covariates with certain kernel functions (Li & Racine, 2004).

It is worth mentioning that the Beran estimator can be written in terms of the original (unordered) sample:

$$\hat{S}_h(t|x) = \prod_{i:T_i \leq t} \left(1 - \frac{\delta_i B_{hi}(x)}{\sum_{r=1}^{n} B_{hr}(x) I(T_r \geq T_i)}\right),$$

where

$$B_{hi}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)}.$$

Starting from the Beran estimator, Xu & Peng (2014) introduced the following kernel type incidence estimator:

$$1 - \hat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right) = \hat{S}_h(T^1_{\max}|x), \tag{2.4}$$

where $T^1_{\max} = \max_{i:\delta_i=1}(T_i)$ is the largest uncensored failure time. These authors also proved the consistency and asymptotic normality of the estimator in (2.4).

The identifiability of a cure model is needed to obtain unique estimates of the model functions. In a cure model, all observed uncensored lifetimes ($\delta_i = 1$) correspond necessarily to uncured subjects ($\nu_i = 0$); but it is impossible to distinguish if a subject with a censored time ($\delta_i = 0$) belongs to the susceptible group ($\nu_i = 0$) or to the non-susceptible group ($\nu_i = 1$), because some censored subjects may experience failures beyond the study period. This leads to difficulties in making a distinction between models with high incidence and long tails of the latency distribution, and low incidence and short tails of the latency distribution. Lemma 2.2.1 addresses this problem.

Lemma 2.2.1. *Let D be the support of X. Model (2.1), with $p(x)$ and $S_0(t|x)$ unspecified, is identifiable if $S_0(t|x)$ is a proper survival function for $x \in D$.*

The proof of Lemma 2.2.1 is included in Appendix A.1. Note that Lemma 2.2.1 also holds in a context with no covariates.

2.3 Asymptotic results

The nonparametric estimator of the incidence function in (2.4) is strongly based on the Beran estimator in (2.2). The Beran estimator of the conditional survival function has been deeply studied in the literature. Dabrowska (1989), in Theorem 2.1, shows its asymptotic unbiasedness, considering NW weights. Furthermore, using GM weights, González-Manteiga & Cadarso-Suárez (1994) give an almost sure i.i.d. representation for the estimator, and Van Keilegom & Veraverbeke (1997b) prove an asymptotic representation for the bootstrapped estimator and obtain the strong consistency of the bootstrap approximation for the conditional distribution function.

Let

$$\hat{H}_h(t|x) = \sum_{i=1}^{n} B_{hi}(x) I(T_i \leq t) \tag{2.5}$$

and

$$\hat{H}_h^1(t|x) = \sum_{i=1}^{n} B_{hi}(x) I(T_i \leq t, \delta_i = 1) \tag{2.6}$$

be the empirical estimators of

$$H(t|x) = P\left(T \leq t | X = x\right) \quad \text{and} \quad H^1(t|x) = P\left(T \leq t, \delta = 1 | X = x\right), \tag{2.7}$$

respectively. Moreover, $\tau_H(x) = \sup\{t : H(t|x) < 1\}$, $\tau_{S_0}(x) = \sup\{t : S_0(t|x) > 0\}$ and $\tau_G(x) = \sup\{t : G(t|x) < 1\}$. Since $S(t|x)$ is an improper survival function, then $S(t|x) > 0$ for any $t \in [0, \infty)$, and $1 - H(t|x) = S(t|x) \times \bar{G}(t|x)$ with $\bar{G}(t|x) = 1 - G(t|x)$ the proper conditional survival function of the censoring time C, we have $\tau_H(x) = \tau_G(x)$.

Let $\tau_0 = \sup_{x \in D} \tau_{S_0}(x)$. As in Xu & Peng (2014), we assume

$$\tau_0 < \tau_G\left(x\right), \forall x \in D. \tag{2.8}$$

This condition states that the support of the censoring variable is not contained in the support of Y, which guarantees that censored subjects beyond the largest observable failure time are cured. Hence, our estimator does not overestimate the true cure rate. A similar assumption was used by Maller & Zhou (1992, 1994) in unconditional cases. As pointed out in Laska & Meisner (1992), if the censoring variable takes values always below a time $\tau_G < \tau_0$, for example in a clinical trial with a fixed maximum follow-up period, the largest uncensored observation $T^1_{\max}$ may occur at a time not larger than τ_G and therefore always before τ_0. In such a case, for a large sample size, the estimator in (2.4) is an estimator of $1 - p$, which is

strictly larger than $1-p$. This comment shows the need of considering the length of follow-up in the design of a clinical trial carefully, so that $S_0(\tau_G)$ is sufficiently small to take the estimator (2.4) as a good estimator of $1-p(x)$ for practical purposes. The simulations in Xu & Peng (2014) show that if the censoring distribution $G(t|x)$ has a heavier tail than $S_0(t|x)$, the estimates using the proposed method will tend to have small biases regardless the value of $\tau_{S_0}(x)$.

Maller & Zhou (1992) dealt with the problem of testing a similar condition to (2.8) in an unconditional setting. They proposed to test $H_0 : \tau_0 > \tau_G$ versus the alternative $H_1 : \tau_0 \leq \tau_G$. One of the weak points of this approach is to include condition (2.8) in the alternative hypothesis. Since this is a neutral assumption, it seems more reasonable to keep (2.8) if there are no strong evidences against it. In that sense, it is more natural to include (2.8) in the null hypothesis. This test is depply studied by Maller & Zhou (1994), with the difference that the neutral assumption in set in the null hypothesis. Apart from that, the ideas by Maller & Zhou (1992) can be extended to a conditional setting as follows. Let us consider $\Pi(t) = E(\delta|T=t)$ and define $\underline{\tau}_G = \inf_{x\in D}\tau_G(x)$. Condition (2.8) implies that $\exists a < \underline{\tau}_G$ such that $\Pi(t) = 0\ \forall t \geq a$. Consequently, this condition can be checked in practice by the following hypothesis test:

$$\begin{cases} H_0 : \exists a < \underline{\tau}_G \ : \ \Pi(t) = 0,\ \forall t \geq a \\ H_1 : \forall a < \underline{\tau}_G,\ \exists t \geq a \ : \ \Pi(t) > 0 \end{cases}.$$

This can be tested by means of nonparametric regression estimators of $\Pi(t)$ based on the sample $\{(T_1,\delta_1),\ldots,(T_n,\delta_n)\}$. To do that, it is necessary to estimate $\underline{\tau}_G$ in a nonparametric way. This can be done by just estimating the support of $G(t|x)$, via the Beran estimator.

We need to consider the following assumptions, to be used in the asymptotic results for the incidence estimator:

(A1) X, Y and C are absolutely continuous random variables.

(A2) Condition (2.8) holds.

(A3) (a) Let $I = [x_1, x_2]$ be an interval contained in the support of m, and $I_\delta = [x_1-\delta, x_2+\delta]$ for some $\delta > 0$ such that

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

and $0 < \delta\Gamma < 1$. For all $x \in I_\delta$ the random variables Y and C are conditionally independent given $X = x$.

(b) There exist $a, b \in \mathbb{R}$, with $a < b$ satisfying $1 - H(t|x) \geq \theta > 0$ for $(t, x) \in [a, b] \times I_\delta$.

(A4) The first derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the first derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A5) The second derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the second derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A6) The first derivatives with respect to t of the functions $G(t|x)$, $H(t|x)$, $H^1(t|x)$ and $S_0(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A7) The second derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A8) The second partial derivatives with respect to t and x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded for $(t, x) \in [0, \infty) \times D$.

(A9) Let us define $H_{c,1}(t) = P(T < t|\delta = 1)$. The first and second derivatives of the distribution and subdistribution functions $H(t)$ and $H_{c,1}(t)$ are bounded and bounded away from zero in $[a, b]$. Moreover, $H'_{c,1}(\tau_0) > 0$.

(A10) The functions $H(t|x)$, $S_0(t|x)$ and $G(t|x)$ have bounded second-order derivatives with respect to x for any given value of t.

(A11) The kernel function, K, is a symmetric density vanishing outside $(-1, 1)$ and the total variation of K is less than some $\lambda < \infty$.

(A12) The density function of T, f_T, is bounded away from 0 in $[a, b]$.

(A13) $\displaystyle\int_0^\infty \frac{dH^1(t|x)}{(1 - H(t|x))^2} < \infty \ \forall x \in I.$

Assumptions $(A1), (A3)$-$(A9)$ and $(A11)$-$(A12)$ are necessary in Theorem 2.3.2 and in Theorem 3.3.2 because their proofs are strongly based on Theorem 2 of Iglesias-Pérez & González-Manteiga (1999). Assumptions $(A2)$ and $(A10)$ are needed to prove Lemma A.1.1. Similar results in the literature are stated for a fixed t such that $1 - H(t|x) \geq \theta > 0$ in $(t, x) \in [a, b] \times I_\delta$. Moreover, assumptions $(A2)$ and $(A10)$ are needed to prove these results for a random value $t = T^1_{\max}$. Assumptions $(A4)$-$(A8)$ and $(A10)$ are regularity conditions for the functions involved in the proofs and in the asymptotic results. Assumption $(A13)$ is necessary to bound the result of an

integral in Lemma A.1.4.

Next theorem states that the proposed nonparametric incidence estimator is the local maximum likelihood estimator of $1-p(x)$. Its proof is in Appendix A.1.

Theorem 2.3.1. *The estimator $1-\hat{p}_h(x)$ given in (2.4) is the local maximum likelihood estimator of $1-p(x)$ for the mixture cure model (2.1), for any $x \in D$.*

An i.i.d. representation for the incidence estimator is obtained in the next theorem, whose proof is also included in Appendix A.1.

Theorem 2.3.2. *Under assumptions (A1)-(A13), for any sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$ and $\ln n/(nh) \to 0$, then*

$$(1-\hat{p}_h(x)) - (1-p(x)) = (1-p(x)) \sum_{i=1}^{n} \tilde{B}_{hi}(x) \xi(T_i, \delta_i, \infty, x) + R_n(x),$$

where

$$\tilde{B}_{hi}(x) = \frac{\frac{1}{nh} K\left(\frac{x-X_i}{h}\right)}{m(x)}, \tag{2.9}$$

$$\xi(T_i, \delta_i, t, x) = \frac{I(T_i \le t, \delta_i = 1)}{1-H(T_i|x)} - \int_0^t \frac{I(u \le T_i) dH^1(u|x)}{(1-H(u|x))^2} \tag{2.10}$$

and

$$\sup_{x \in I} |R_n(x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

Observe that

$$\xi(T_i, \delta_i, \infty, x) = \frac{\delta_i}{1-H(T_i|x)} - \int_0^{T_i} \frac{dH^1(u|x)}{(1-H(u|x))^2}. \tag{2.11}$$

Finally, from the representation in Theorem 2.3.2 and following Iglesias-Pérez & González-Manteiga (1999), an asymptotic expression of the mean squared error of the incidence estimator,

$$MSE_x(h_x) = E[(\hat{p}_{h_x}(x) - p(x))^2], \tag{2.12}$$

is obtained in Corollary 2.3.1.

Corollary 2.3.1. *An asymptotic expression of the mean squared error of the incidence estimator is given by:*

$$AMSE_x(h) = \frac{1}{nh}(1-p(x))^2 c_K \sigma^2(x) + \left[h^2 \frac{1}{2} d_K (1-p(x))\mu(x)\right]^2, \tag{2.13}$$

where the first term corresponds to the asymptotic variance and the second one to the asymptotic squared bias, with d_K in (1.9) and c_K in (1.10) and, following a notation similar to that in Dabrowska (1992):

$$\sigma^2(x) = \frac{1}{m(x)} \int_0^\infty \frac{dH^1(t|x)}{(1 - H(t|x))^2}$$

and

$$\mu(x) = \frac{2\Phi'(x, \infty, x)m'(x) + \Phi''(x, \infty, x)m(x)}{m(x)}, \tag{2.14}$$

where

$$\Phi(y, t, x) = \int_0^t \frac{dH^1(v|y)}{1 - H(v|x)} - \int_0^t (1 - H(v|y)) \frac{dH^1(v|x)}{(1 - H(v|x))^2}, \tag{2.15}$$

with $\Phi'(y, t, x) = \partial\Phi(y, t, x)/\partial y$ and $\Phi''(y, t, x) = \partial^2\Phi(y, t, x)/\partial y^2$. Note that the AMSE denotes the MSE of the dominant part of the almost sure representation of the incidence estimator. If the censoring distribution does not depend on the covariate, then $\mu(x)$ can also be written as follows:

$$\mu(x) = \frac{1}{m(x)} \left([p(x)\, m(x)]'' - p(x)\, m''(x) \right) \left(1 - \int_0^\infty \frac{S(dv|x)}{S(v|x)^2} \right). \tag{2.16}$$

The straightforward calculations to obtain the AMSE for the incidence estimator are detailed in Appendix A.1.

2.4 Bandwidth selection

The choice of the bandwidth is a crucial issue in kernel estimation, since it controls the trade-off between bias and variance. Most of the methods for smoothing parameter selection in nonparametric curve estimation look for a small error when approximating the underlying curve by the smooth estimate. The asymptotically optimal local bandwidth to estimate the cure rate, $1 - p(x)$, in the sense of minimizing the asymptotic expression of the MSE_x in (2.13), is given by:

$$h_{AMSE}(x) = \left(\frac{c_K \sigma^2(x)}{d_K^2 \mu^2(x)} \right)^{1/5} n^{-1/5}, \tag{2.17}$$

which is an asymptotic approximation of the bandwidth $h_{MSE}(x)$ that minimizes the MSE_x. The optimal bandwidth $h_{AMSE}(x)$ depends on unknown functions through $\mu(x)$ and $\sigma^2(x)$. It is worth mentioning that the bandwidth $h_{AMSE}(x)$ fulfills the conditions on the bandwidth required in Theorem 2.3.2.

Considering Dabrowska (1989), a plug-in bandwidth selector can be obtained by replacing those unknown functions by consistent nonparametric estimates that need additional smoothing parameters. This gives rise to a never-ending process, which seems even harder than the original problem of incidence estimation. We drew some simulation studies following the proposal by Dabrowska (1989), together with the approach by Härdle & Marron (1985) for the choice of the pilot bandwidths. We did not obtain good results, since the second derivative of a function in the term $\mu^2(x)$ is very hard to estimate. On the other hand, the finite-sample behavior of the cross validation (CV) bandwidth selector in this context turned out to be disappointing. We followed Iglesias-Pérez (2007) and we also tested a few modifications of this criterion. Unfortunately, the CV bandwidth in this context is highly variable and tends to undersmooth.

2.4.1 Bootstrap bandwidth selector

Another way to select the bandwidth is to use the bootstrap method. It consists of minimizing a bootstrap estimate of the mean squared error, $MSE_x(h_x)$.

The bootstrap bandwidth is the minimizer of the bootstrap version of $MSE_x(h_x)$ in (2.12),

$$MSE^*_{x,g_x}(h_x) = E^*[(\hat{p}^*_{h_x,g_x}(x) - \hat{p}_{g_x}(x))^2], \tag{2.18}$$

which consists of replacing the original sample by the bootstrap resample, the kernel incidence estimator based on the sample by its bootstrap version and the theoretical incidence function by the estimated incidence based on a pilot bandwidth, g_x. Equation (2.18) can be approximated, using Monte Carlo, by:

$$MSE^*_{x,g_x}(h_x) \simeq \frac{1}{B}\sum_{b=1}^{B}(\hat{p}^{*b}_{h_x,g_x}(x) - \hat{p}_{g_x}(x))^2, \tag{2.19}$$

where $\hat{p}^{*b}_{h_x,g_x}(x)$ is the kernel estimator of $p(x)$ using bandwidth h_x and based on the b-th bootstrap resample generated from $\hat{F}_{g_x}$, and $\hat{p}_{g_x}(x)$ is the kernel estimator of $p(x)$ computed with the original sample and pilot bandwidth g_x.

Considering a bandwidth search grid $\{h_1, \ldots, h_L\}$, the procedure for obtaining the bootstrap bandwidth selector for a fixed covariate value, x, is as follows:

1. Generate B bootstrap resamples of the form:
   $\left\{ \left(X_1^{(b)}, T_1^{*(b)}, \delta_1^{*(b)}\right), \ldots, \left(X_n^{(b)}, T_n^{*(b)}, \delta_n^{*(b)}\right) \right\}$, $b = 1, \ldots, B$.

2. For the b-th bootstrap resample ($b = 1, ..., B$), compute the nonparametric estimator $\hat{p}^{*b}_{h_l,g_x}(x)$ with bandwidth h_l, $l = 1, 2, \ldots, L$.

3. With the original sample and the pilot bandwidth g_x, compute $\hat{p}_{g_x}(x)$.

4. For each bandwidth h_l in the grid, compute the Monte Carlo approximation of $MSE^*_{x,g_x}(h_l)$, given by (2.19).

5. The bootstrap bandwidth, h^*_x, is the minimizer of the Monte Carlo approximation of $MSE^*_{x,g_x}(h_l)$ over the grid of bandwidths $\{h_1, \ldots, h_L\}$.

In Step 1, we consider the simple weighted bootstrap, without resampling the covariate X, which is equivalent to the simple weighted bootstrap proposed by Li & Datta (2001). For fixed x and $i = 1, \ldots, n$, we set $X^*_i = X_i$ and generate a pair (T^*_i, δ^*_i) from the weighted empirical distribution $\hat{F}_{g_x}(\cdot, \cdot|X^*_i)$, where

$$\hat{F}_{g_x}(u, v|x) = \sum_{j=1}^{n} B_{g_x j}(x) I(T_j \leq u, \delta_j \leq v)$$

and $B_{g_x j}(x)$ is the NW weight in (2.3) with pilot bandwidth g_x. The resulting bootstrap resample is $\{(X_1, T^*_1, \delta^*_1), \ldots, (X_n, T^*_n, \delta^*_n)\}$. It is easy to generate (T^*_i, δ^*_i) using the marginal distribution of T^*_i, $P^*(T^* \leq u) = \sum_{j=1}^{n} B_{g_x,j}(X_i) I(T_j \leq u)$ and the conditional probability $P^*(\delta^* = 1|T^* = T_\zeta) = \delta_\zeta$, if there are no ties in the T_j's.

Based on the results in Van Keilegom & Veraverbeke (1997b, 1997a) for fixed design with GM weights, the optimal pilot bandwidth, g_x, could be chosen so that it minimizes (2.18) for a given sample. To the best of our knowledge, there are no similar results for random design. However, preliminary studies (see results below) showed that the choice of the pilot bandwidth has a small effect on the final bootstrap bandwidth. Consequently, a simple rule is proposed to select g_x as a global pilot bandwidth of order $n^{-1/9}$ (see Equation (2.25) in Section 2.5.3).

Remark: The bandwidth sequence $g_x = g_n$ has to be typically asymptotically larger than $h_x = h_n$. This oversmoothing pilot bandwidth is required for the bootstrap bias and variance to be asymptotically efficient estimators for the bias and variance terms. The order $n^{-1/9}$ for this asymptotically optimal pilot bandwidth satisfies the conditions in Theorem 1 of Li & Datta (2001), and it coincides with the order obtained by Cao & González-Manteiga (1993) for the uncensored case in nonparametric regression.

2.5 Simulation study

The software used in all the simulation studies in this thesis is R, a free environment for statistical computing and graphics. The procedures, coded in R language, were drawn in the computers of the Department of Mathematics, at the Faculty of Computer Sciences in the University of A Coruña (UDC).

In this section we compare the proposed nonparametric incidence estimator with the semiparametric incidence estimator by Peng & Dear (2000), implemented in the *smcure* package in R (Cai et al., 2012), which fits a semiparametric PH mixture cure model. The cure probability part is estimated by a generalized linear model which allows the logit link function, and the latency part follows a PH model. The semiparametric estimation procedures are based on the EM algorithm for both models. Specifically, the logit link function that Peng & Dear (2000) assume for the probability of uncure is given by

$$p(\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\boldsymbol{\beta}^T\mathbf{x})},$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters, used to model the effect of the covariates $\mathbf{x}$.

We carry out a simulation study with two purposes. First, we evaluate the finite sample performance of the nonparametric estimator, $1-\hat{p}_{h_x}$, computed in a grid of bandwidths, and we compare the results with those of the semiparametric estimator. Second, the practical behavior of the bootstrap bandwidth selector is assessed. We consider two different models. For both, the censoring times are generated according to an exponential distribution with mean $1/0.3$ and the covariate X is $U(-20,20)$.

The Epanechnikov kernel, defined in (2.20), optimal in a mean square error sense, is used for all the simulation studies in this thesis:

$$K(u) = \frac{3}{4}(1-u^2)I(|u|\leq 1). \tag{2.20}$$

Note that the choice of the kernel, K, is a problem of little importance, since the shape of the estimator is not very sensitive to the choice of K, and different kernels produce good estimates. Due to this reason, results using other kernels are omitted.

Model 1 For comparison reasons, this simulated setup is the same as the so-called mixture cure (MC) model considered in Xu & Peng (2014). The data are generated

from a logistic-exponential MC model, where the probability of not being cured is

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

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$, and the survival function of the uncured subjects is:

$$S_0(t|x) = \begin{cases} \dfrac{\exp(-\lambda(x)t) - \exp(-\lambda(x)\tau_0)}{1 - \exp(-\lambda(x)\tau_0)} & \text{if } t \leq \tau_0 \\ 0 & \text{if } t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda(x) = \exp((x + 20)/40)$. The percentage of censored data is 54% and of cured data is 47%. In Figure 2.1 (top) we show the shape of the theoretical incidence (cure rate) and latency functions. Note that in this model the incidence is a logistic function and the latency is very close to fulfill the proportional hazards model, that has been truncated to guarantee condition (2.8). Therefore, the semiparametric estimators are expected to give very good results in this model.

Model 2 The data are generated from a cubic logistic-exponential mixture model, where the incidence (cure rate) is:

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

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$ and $\beta_3 = 0.0020$, and the latency is:

$$S_0(t|x) = \frac{1}{2}\left(\exp(-\alpha(x)t^5) + \exp(-100t^5)\right),$$

with

$$\alpha(x) = \frac{1}{5}\exp((x + 20)/40).$$

The percentages of censored and cured data are 62% and 53%, respectively. Figure 2.1 (bottom) shows the theoretical incidence and latency in this model. The incidence is not a logistic function and the effect of the covariate on the failure time of the uncured patients does not fit a PH model anymore. So, the results will show the gain of using the proposed nonparametric estimators, that do not require any parametric or semiparametric assumptions, with respect to the semiparametric ones.

2.5.1 Preliminary studies for the pilot bandwidth selection

We considered different pilot bandwidths. For the sake of brevity, only the results of two approaches are shown. In the first one, we work with a constant pilot bandwidth:

$$g = \frac{X_{(n)} - X_{(1)}}{10}, \tag{2.21}$$

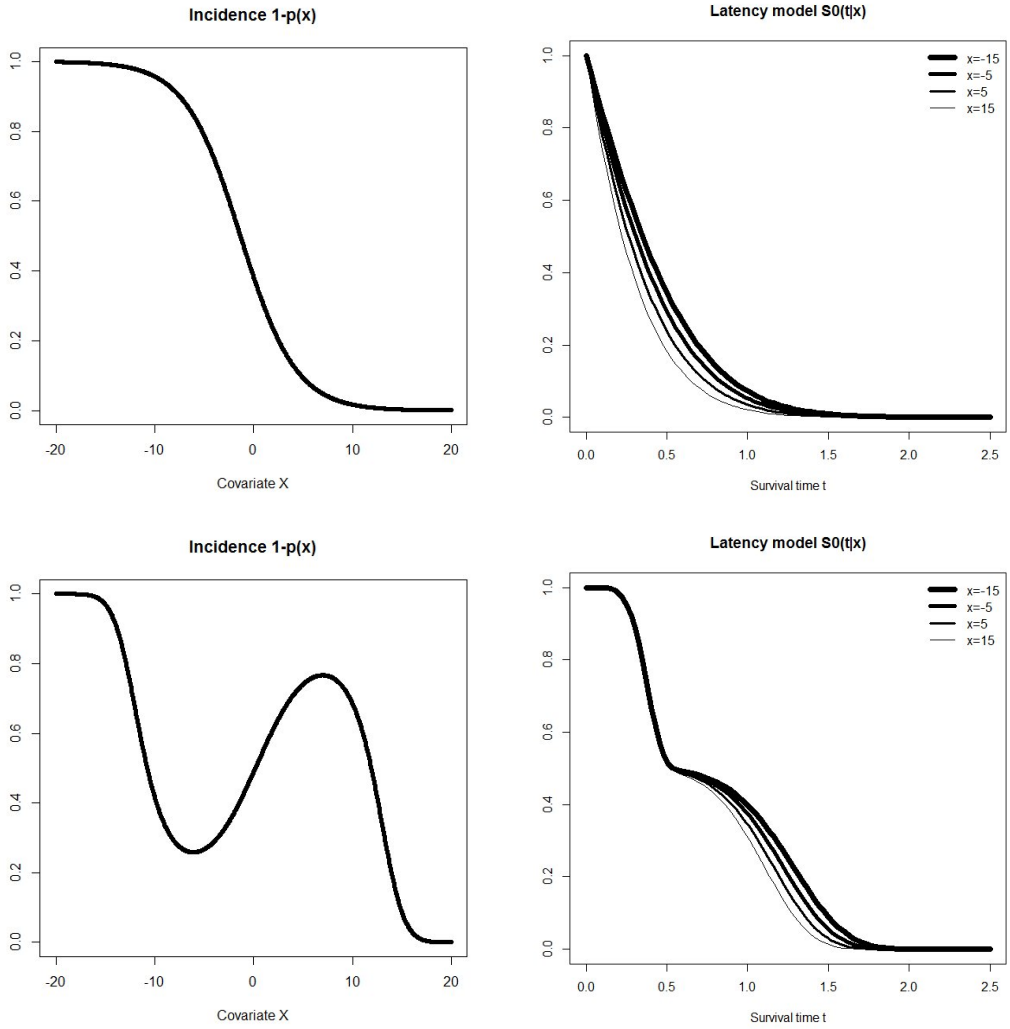

Figure 2.1: Theoretical incidence (left) and latency (right) functions for Model 1 (top) and Model 2 (bottom).

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

The second approach consists of working with the pilot bandwidth which is optimal in the sense of providing efficient estimates of $\ddot{H}(t|x) = \dfrac{\partial^2}{\partial x^2} H(t|x)$, the second derivative of $H(t|x)$ in (2.7), using the second derivative of the smoothed estimator given in (2.5),

$$\widehat{\ddot{H}}_g(t|x) = \sum_{i=1}^{n} B^{(2)}_{g(i)}(x) I(T_i \leq t), \tag{2.22}$$

with

$$B^{(2)}_{g(i)}(x) = \frac{\partial^2}{\partial x^2} B_{g(i)}(x) = \frac{\partial^2}{\partial x^2} \left(\frac{K\left(\frac{x-X_i}{g}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{g}\right)} \right).$$

Cao (1991) obtains the MSE and the MISE for the estimator $\widehat{\ddot{H}}_g(t|x)$ in (2.22). Departing from these results, the optimal global bandwidth for the estimation of $\widehat{\ddot{H}}_g(t|x)$ is

$$g\left(x\right) = \left(\frac{5c_{K''} \left(\int_{-\infty}^{\infty} \sigma^2\left(t|x\right) dt \right) m\left(x\right)}{\int_{-\infty}^{\infty} \left(\frac{\partial^4}{\partial x^4} \left(H\left(t|x\right) m\left(x\right) \right) - \frac{\partial^2}{\partial x^2} \left(H\left(t|x\right) m''\left(x\right) \right) \right)^2 dt} \right)^{1/9} n^{-1/9}. \quad (2.23)$$

In this preliminary study, we obtain two bootstrap bandwidths: one computed with the global pilot bandwidth in (2.21) and the other computed with the local pilot bandwidth in (2.23). Both bootstrap bandwidths are compared with the optimal bandwidth, h_{MSE}, in Figure 2.2. Furthermore, in Figure 2.3, the incidence computed with the optimal bandwidth is compared with the incidence computed with the two bootstrap bandwidths. Note that for simplicity, we consider sample size $n = 100$, $\kappa = 100$ trials, $B = 100$ bootstrap resamples and Model 2, detailed in Section 2.5. Figures 2.2 and 2.3 show that there are no important differences in the bootstrap bandwidth, h^*_x, and in the estimation of $1 - p(x)$, computed with the different pilot bandwidths. Therefore, we decided to work with a constant pilot bandwidth, keeping the optimal order, $n^{-1/9}$ (see Equation (2.25)).

2.5.2 Efficiency of the nonparametric incidence estimator

A total of $\kappa = 1000$ samples of size $n = 100$ are drawn to approximate, by Monte Carlo, the mean squared error (MSE) of the incidence estimator evaluated at 41 values $\{-20, -19, \ldots, 19, 20\}$ of the covariate X, and for a grid of 100 bandwidths in a logarithmic scale, from $h_1 = 1.2$ to $h_{100} = 20$. The results for both models are shown in Figure 2.4.

Regarding the MSE of the incidence estimators, Figure 2.4 shows that in Model 1 there is a range of bandwidths, from $h_{50} = 4.83$ to $h_{70} = 8.53$ (light blue lines) for which the nonparametric estimator is quite competitive with respect to the semi-parametric estimator in values x of the covariate near the endpoints of the support of X, and it works much better when the value of the covariate is around 0. In

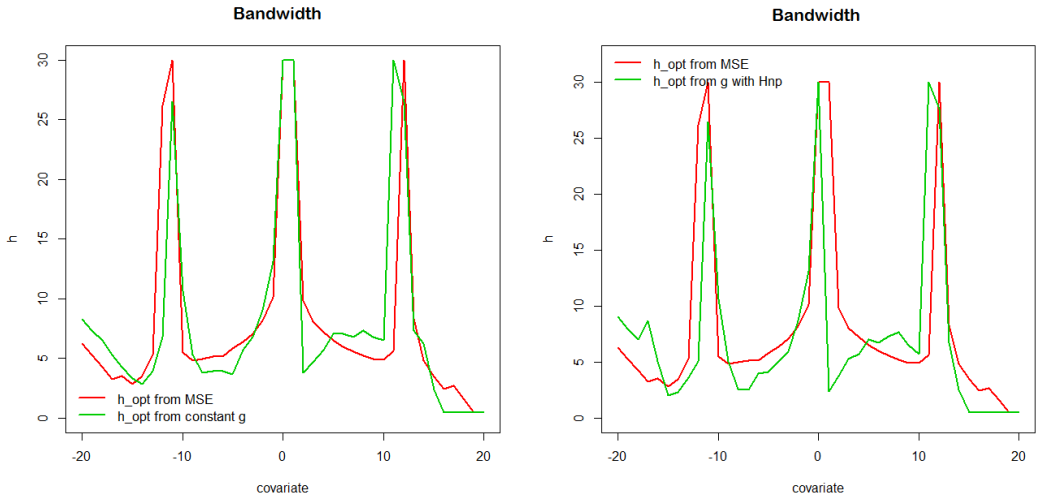

Figure 2.2: Theoretical optimal bandwidth, h_{MSE}, (red) for Model 2. On the left, h_x^* obtained using a constant pilot bandwidth (green). On the right, h_x^* obtained using a pilot bandwidth estimated with the nonparametric approach (green).

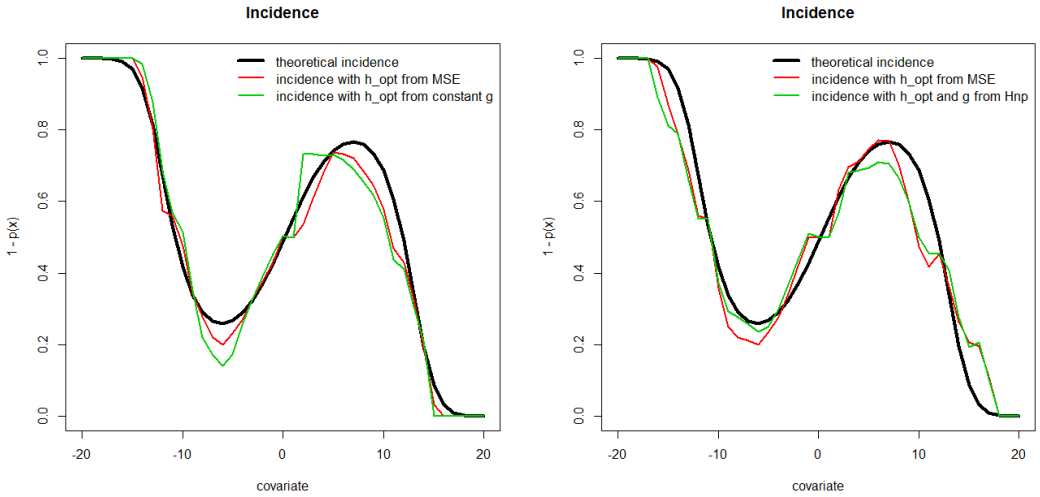

Figure 2.3: Theoretical incidence, $1 - p(x)$, (black), and incidence estimators computed with h_{MSE} (red), with h_x^* (green) obtained using a constant pilot bandwidth (left), and using a pilot bandwidth estimated with the nonparametric approach (right).

Model 2, as expected, the nonparametric estimator outperforms the semiparametric one for a wide range of bandwidths except for 3 singular values of the covariate X.

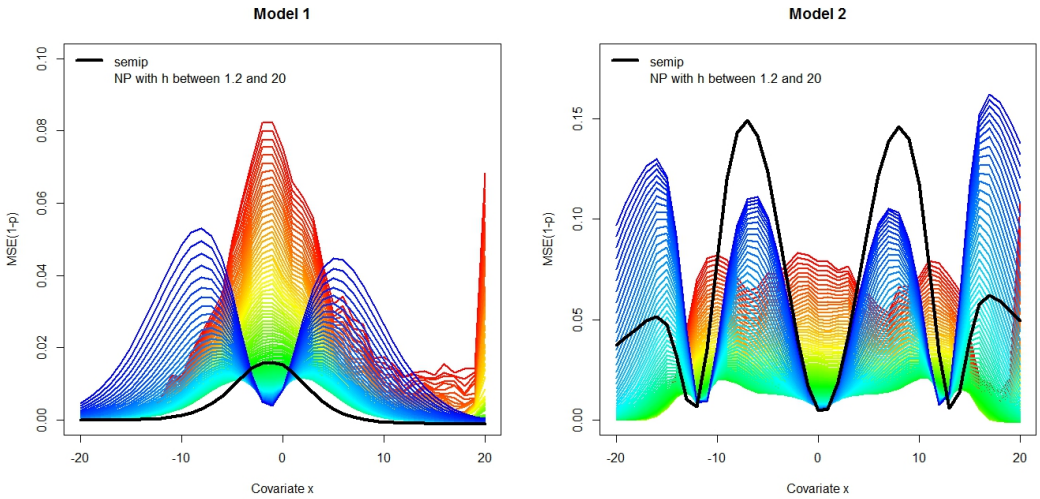

Figure 2.4: MSE for the semiparametric (black line) and the nonparametric estimator of $1-p(x)$ computed with different bandwidths: from $h_1 = 1.2$ (red line) to $h_{100} = 20$ (dark blue line). The data were generated from Model 1 (left) and Model 2 (right).

Note that, from Corollary 2.3.1, the dominant term of the bias of $p(x)$ is

$$\frac{1}{2}h^2 d_K \left(1 - p(x)\right) \mu(x), \tag{2.24}$$

where $\mu(x)$, defined in (2.14), can be expressed as

$$\mu(x) = \frac{1}{m(x)} \left([p(x)\, m(x)]'' - p(x)\, m''(x)\right) \left(1 - \int_0^\infty \frac{S(dv|x)}{S(v|x)^2}\right)$$

if the censoring distribution does not depend on the covariate.

In this simulation study, since the distribution of the covariate X is uniform and the distribution of the censoring variable does not depend on the covariate, then the expression of $\mu(x)$ in (2.16) reduces to

$$\mu(x) = p''(x) \left(1 - \int_0^\infty \frac{S(dv|x)}{S(v|x)^2}\right),$$

and the dominant term of the bias of $p(x)$ in Equation (2.24) is zero in points where the second derivative of $p(x)$ is also zero. Therefore, the dominant part of the AMSE is the asymptotic variance, which tends to zero when h tends to infinity.

Moreover, the denominator of the asymptotic expression of the bandwidth which minimizes the AMSE in (2.17) is also equal to 0, which makes the h_{AMSE} bandwidth

to tend to infinity. Consequently, the h_{MSE} has peaks in these points (see Figure 2.8). Specifically, in Model 1, if the covariate is uniform in $[-20, 20]$, then $m'(x) = 0$. Therefore, $2p'(x)m'(x) + p''(x)m(x)$ is given by $\frac{1}{40}p''(x)$, and the asymptotic bias of $\hat{p}_h(x)$ is zero at $x_0 = -1.3296$. If the density of the covariate is not constant, but $m(x) = \frac{3}{80000}x^2 + 0.02$, then the asymptotic bias of $\hat{p}_h(x)$ is zero at $x_0 = -1.505$. Figure 2.5 shows the MSE for Model 1 with $m(x) = \frac{3}{80000}x^2 + 0.02$.

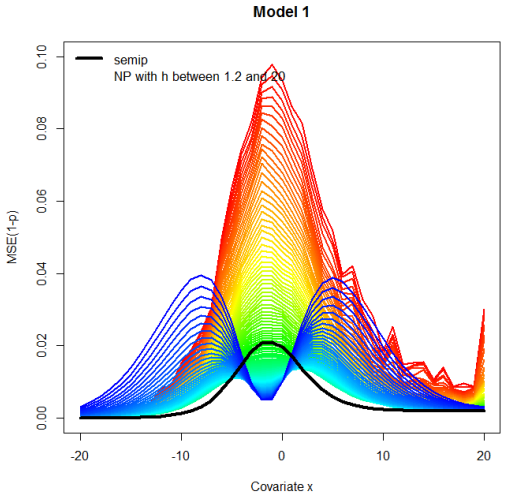

Figure 2.5: MSE for the semiparametric (black line) and the nonparametric estimator of $1 - p(x)$ computed with different bandwidths: from $h_1 = 1.2$ (red line) to $h_{100} = 20$ (dark blue line). The data were generated from Model 1 with $m(x) = 3x^2/80000 + 0.02$.

2.5.3 Efficiency of the bootstrap bandwidth selector

In this simulation study, we consider sample sizes of $n = 50, 100$ and 200. For $\kappa = 1000$ trials, we approximate the MSE_x and the optimal bandwidth, $h_{x,MSE}$, of the proposed nonparametric estimator of the incidence. The $MSE_x(h^*_{x,g_x})$ and the bootstrap bandwidth h^*_{x,g_x} are also approximated.

Note that minimizing $MSE^*_x(h_x)$ in h_x for each value, x, of the covariate, is a computationally expensive algorithm. For that reason, we carry out a two-step method with a double search in each stage. In the first step, we draw $B = 80$ bootstrap resamples and consider a number of 21 bandwidths equispaced on a logarithmic scale, from $h_1 = 0.2$ to $h_{21} = 50$ in the first search, whereas in the second search the grid is centered around the optimal bandwidth obtained in the first search. Then, we carry out the second step with also a double search in a similar way we did for the

first step, but now with two differences: we draw $B = 1000$ bootstrap resamples and we consider a finer smaller grid of 5 bandwidths in both the first and second search. It is important to highlight that for all the simulation studies which involve the use of a smoothing parameter, a logarithmic scale on the bandwidth is considered, since it acts as a multiplicative form.

In view of the fact that the choice of g_x has a low effect on the final bootstrap bandwidth, we propose to use a naive selector, keeping the $n^{-1/9}$ optimal order. Since the distribution of the covariate is uniform, we consider the following global pilot bandwidth, that does not depend on the value x for which the estimation is to be carried out:

$$g = \frac{X_{(n)} - X_{(1)}}{10^{7/9}} n^{-1/9}, \tag{2.25}$$

where $(X_{(1)}, \ldots, X_{(n)})$ is the ordered sample of covariates. Note that, for $X \in U(-20, 20)$, when $n = 100$ the value of the global pilot bandwidth g is $(X_{(n)} - X_{(1)})/10 \simeq 4$. Similarly, $g \simeq 4.32$ ($g \simeq 3.70$) when $n = 50$ ($n = 200$). For a naive pilot bandwidth selector if the distribution of X can not be assumed uniform, see Section 2.6. The MSE for the semiparametric incidence estimator, together with that of the nonparametric estimator computed with the optimal bootstrap bandwidth, are shown in Figure 2.6. It is important to highlight the similarity of the shape between both MSE curves in Model 1. In Model 2, the nonparametric estimator with the bootstrap bandwidth outperforms the semiparametric estimator for a wide range of covariate values.

Figure 2.7 shows the MSE_x evaluated at the median, 25th and 75th percentiles of the proposed bootstrap bandwidth, along the $\kappa = 1000$ simulated samples. The value of the MSE_x for the nonparametric estimator, approximated by Monte Carlo and evaluated at the MSE bandwidth, $h_{x,MSE}$, is also given as reference. We observe that the median, 25th and 75th percentiles of the bootstrap bandwidths have an MSE close to the optimal value. As expected, the similarity increases with the sample size. Moreover, we can also check how $MSE_x(h_{x,MSE})$ and $MSE_x(h^*)$ decrease as n becomes larger.

The performance of the bootstrap bandwidth for Models 1 and 2 is shown in Figure 2.8. The optimal $h_{x,MSE}$, approximated by Monte Carlo, is displayed together with the median and the 25th and 75th percentiles of the 1000 bootstrap bandwidths, h_x^*. We can appreciate how the bootstrap bandwidth, h_x^*, approaches $h_{x,MSE}$, adapting properly to the shape of $h_{x,MSE}$ for the three sample sizes. The

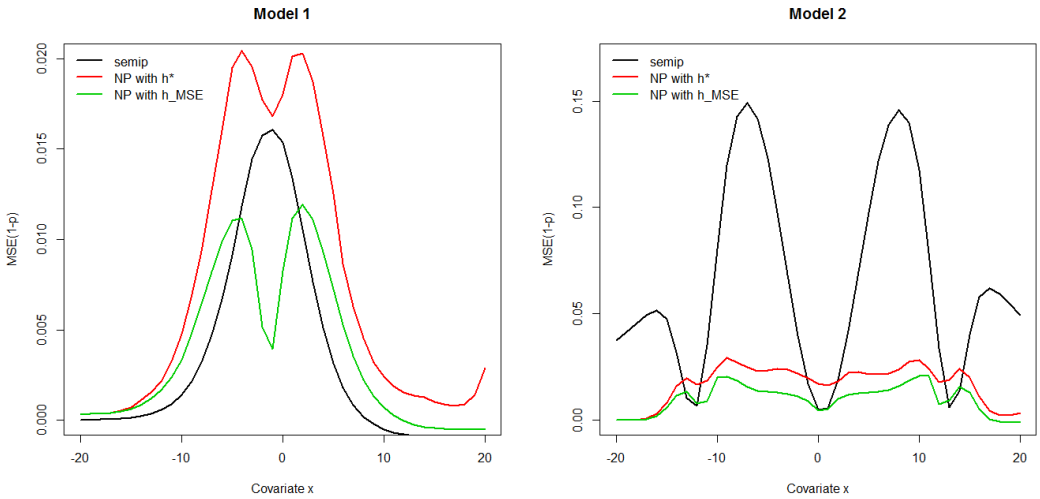

Figure 2.6: MSE for the semiparametric (black line) and the nonparametric estimator of $1-p(x)$ computed with the bootstrap bandwidth (red line) and with the optimal bandwidth (green line). The data were generated from Model 1 (left) and Model 2 (right), with sample size $n = 100$.

optimal bandwidth, $h_{x,MSE}$, has got peaks at the values x of the covariate for which $p''(x) = 0$. Those peaks only occur at points x for which $\mu(x) = 0$, and the dominant term of the AMSE is of order $1/(nh)$. As a consequence, the asymptotically optimal bandwidth is infinitely large and the best choice is to smooth as much as possible, that is, the best local fit is a global fit. Note that if such large bandwidths are used, those values of x correspond to the values where the MSE_x shows deep valleys, that is, there is a noticeable improvement in the incidence estimation.

Regarding the computational times, the method for the simulation study related to Figure 2.4 is considerably fast (taking less than 20 seconds for each model, with sample size $n = 100$). The simulations for the bootstrap bandwidth selector (Figures 2.6, 2.7 and 2.8) are more computationally expensive, even though they are drawn using a two-step method. For obtaining the Monte Carlo approximation of the theoretical functions in each model, the algorithm lasts around 20, 35 and 70 minutes, with $n = 50$, $n = 100$ and $n = 200$, respectively. Additionally, the method used to obtain the bootstrap bandwidth, together with the bootstrap MSE, takes 136.81 hours with $n = 50$, 242.20 hours with $n = 100$ and 462.91 hours with $n = 200$.

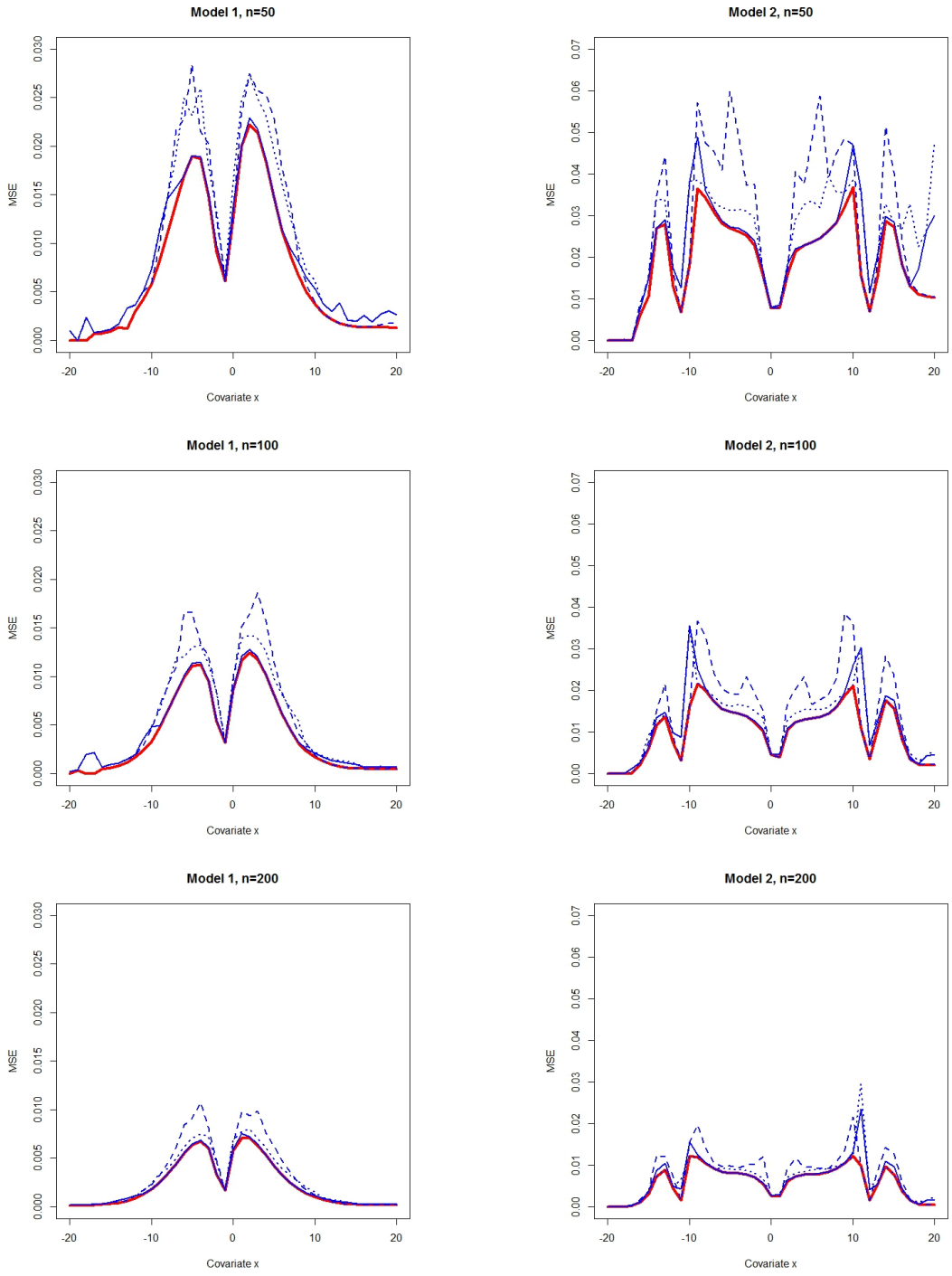

Figure 2.7: MSE_x of the nonparametric estimator of the incidence evaluated at $h_{x,MSE}$ (red line), and MSE_x evaluated at the median (solid blue line), 25th (dotted blue line) and 75th (dashed blue line) percentiles of the bootstrap bandwidth, h_x^*, along $\kappa = 1000$ samples of sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom), for Model 1 (left) and Model 2 (right).

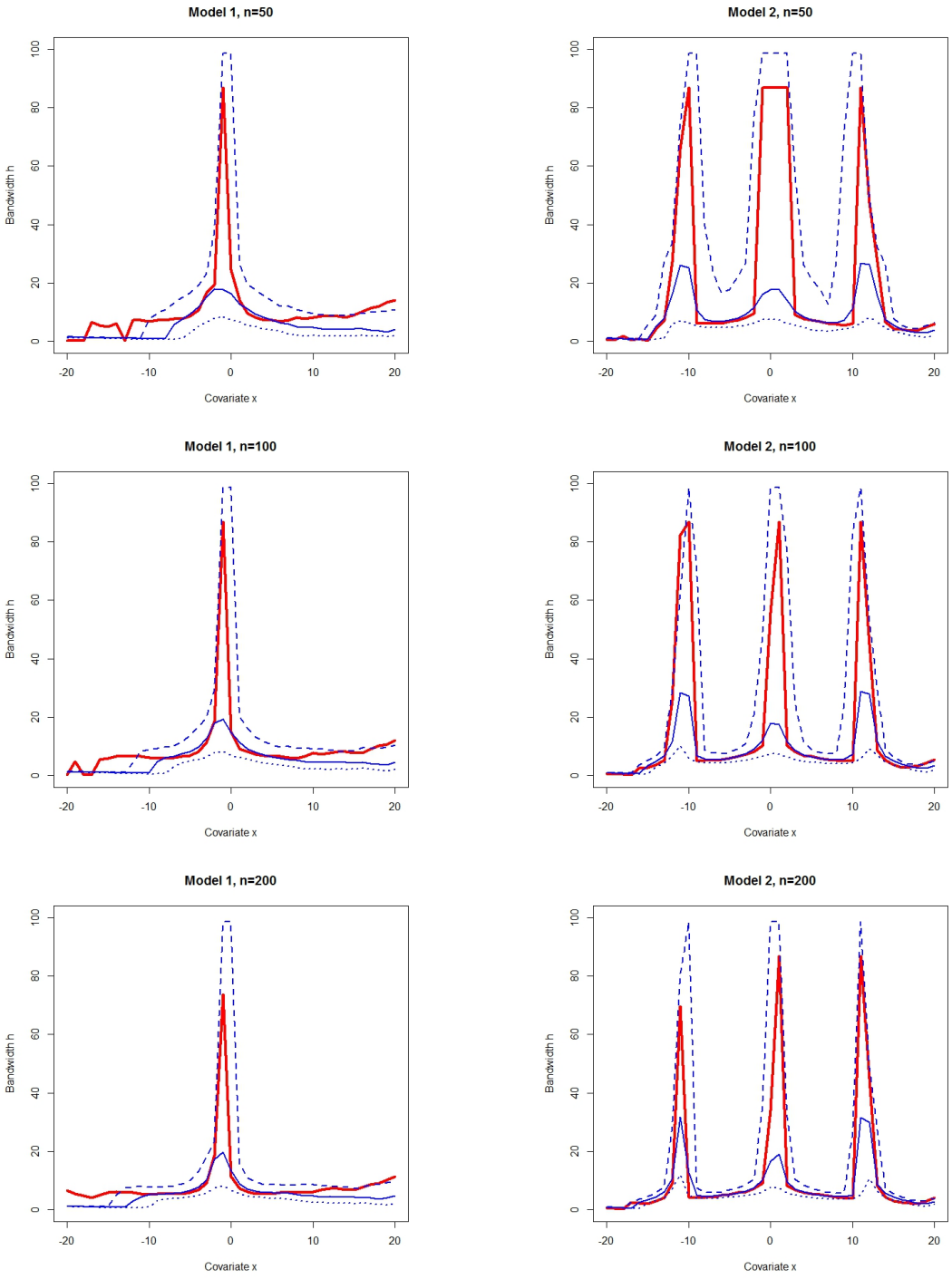

Figure 2.8: Optimal $h_{x,MSE}$ (red line), median (solid blue line), 25th (dotted blue line) and 75th (dashed blue line) percentiles of the bootstrap bandwidth, h_x^*, along $\kappa = 1000$ samples of sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom), for Model 1 (left) and Model 2 (right).

2.6 Application to real data

We applied both the semiparametric and the nonparametric estimators to a real dataset of 414 colorectal cancer patients from CHUAC, Spain. It consists of 8 variables:

- The censoring indicator.
- The observed survival time. The variable of interest, Y, is the time, in months, since the diagnostic until death. In this dataset, the follow-up time is almost 19 years.
- The location: colon or rectum.
- The age: from 23 to 102 years.
- The stage TNM, which is the main determinant in prognosis of these patients. The stage has 3 components: T, which describes the size of the tumor and whether it has invaded nearby tissue; N, which measures the lymph nodes that are involved; and M, which evaluates the presence (or not) of metastasis. The information of these 3 aspects can be combined and it lets us classificate each patient in a unique (and numeric) stage. That is, the variable stage, which takes values from 1 to 4, is defined from the variables T, N and M.

An individual is considered cured if he or she will not die because of colorrectal cancer. Censoring is caused by "cure", death due to any other cause different to colorrectal cancer, dropout, or end of the study. We work with two covariates listed above: the stage and the age. Note that since no important differences are expected in the survival depending on the location, we work with the complete dataset, regardless the location value in each patient. About 50% of the observations are censored, with the percentage of censoring varying from 30% to almost 71%, depending on the stage. In Table 2.1 we show a summary of the dataset.

The incidence is estimated with both the semiparametric logistic and the nonparametric estimators. The age of the patients has been considered as a continuous covariate, and the data have been split into four groups according to the categorical covariate stage.

Note that in order to obtain the bootstrap bandwidth selector, $B = 1000$ bootstrap resamples are used. In a similar way as we did in Section 2.5.3, we carry out a one-step procedure with a double search. We consider a number of 21 bandwidths

Stage	Number of patients	Number of censored data	% Censoring	Age	
				Min.	Max.
1	62	44	70.97	23	84
2	167	92	55.09	36	102
3	133	53	39.85	30	88
4	52	16	30.77	43	88
	414	205	49.52		

Table 2.1: Colorectal cancer patients from CHUAC.

equispaced on a logarithmic scale in both searches. The first search is performed between 0.2 and the empirical range of X. The second one is carried out using a narrower grid centered around the optimal bandwidth obtained in the first search.

For the bandwidth selector of the nonparametric estimator of $1 - p(x)$, a naive pilot bandwidth has been proposed in (2.25) if the distribution of X is uniform. The idea is to provide a data-driven pilot bandwidth which only depends on both the sample size and on the distribution of the covariate, keeping the $n^{-1/9}$ optimal order. Taking into account that, in this case, the distribution of the covariate is not uniform (see Figure 2.9), we propose to use the following local pilot bandwidth:

$$g_x = \frac{d_k^+(x) + d_k^-(x)}{2} 100^{1/9} n^{-1/9},$$

where $d_k^+(x)$ is the distance from x to the k-th nearest neighbor on the right, $d_k^-(x)$ the distance from x to the k-th nearest neighbor on the left, and k a suitable integer depending on the sample size. If there are not at least k neighbors on the right (or left), we use $d_k^+(x) = d_k^-(x)$ (or $d_k^-(x) = d_k^+(x)$, respectively). Our numerical experience shows that a good choice is to consider $k = n/4$. Note that when $n = 100$ the value of the local pilot bandwidth g_x is the mean distance to the 25th nearest neighbor on both the left and right sides.

Alongside the bootstrap bandwidth, we have also used a smoothed bootstrap bandwidth for the incidence estimator. We followed Cao et al. (2001), who applied a method for smoothing local bandwidths for a kernel-type estimator of the relative density (or grade density). Let $(X_{(1)}, \ldots, X_{(n)})$ be the ordered observations of X. The bootstrap bandwidths have been computed in the equispaced grid $x_0 < x_1 < \ldots < x_m$ of the interval $[X_{(1)}, X_{(n)}]$ given by $x_i = X_{(1)} + \Delta i, i = 0, 1, 2, \ldots, m$ where $\Delta = (X_{(n)} - X_{(1)})/m$ and $m + 1$ is the number of points considered in the grid of values of X in each stage. The smoothed bootstrap bandwidth for the covariate

value x_l is computed as follows:

$$h_{x_l}^{*\,smooth} = \begin{cases} \frac{\sum_{j=0}^{l+5} h_{x_j}^*}{l+6}, & l = 0,1,2,3,4 \\ \frac{\sum_{j=l-5}^{l+5} h_{x_j}^*}{11}, & l = 5,6,7,\ldots,m-5 \\ \frac{\sum_{j=l-5}^{m} h_{x_j}^*}{m-l+6}, & l = m-4, m-3, m-2, m-1, m \end{cases} .$$

Figure 2.9 shows the estimations of the probability of being cured for the different stages with respect to the age of the patients. We can see that the effect of the age on the incidence changes with the stage. The cure probabilities in Stages 1 and 2 are higher than in Stages 3 and 4. The reason is that, in initial stages, most of the surgeries have healing purposes, whereas in advanced stages, surgeries are usually palliative treatments, and therefore the incidence for these patients is lower. For example, using the nonparametric incidence estimation, in Stage 1, patients have a probability of survival between 25% and 65%, depending on the age; whereas in Stage 3, for patients above 60, in a 10 years gap that probability decreases considerably from 40% to almost 0%. It is important to highlight the difference between the nonparametric and the semiparametric curves, that seems to indicate that the logistic model is not valid for the data. The results in Stage 4 deserve some comments. A total of 11 in the 12 greatest lifetimes in Stage 4, including the largest lifetime, are uncensored and, consequently, uncured. This causes that the nonparametric estimation of the probability of being cured is equal to 0. Although it should not be stated that it is impossible for a patient with Stage 4 colorectal cancer to survive, this estimation reinforces the assertion that long-term survival in patients with Stage 4 colorectal cancer is uncommon (Miyamoto et al., 2015). This fact, far from being a weakness of the nonparametric method, is an important advantage, since it allows to detect situations in which introducing the possibility of cure does not contribute to improve the model.

We show the resulting bootstrap bandwidths, with the corresponding local pilot bandwidths, for the different values of the covariate age in Figure 2.10. The reason why the bootstrap bandwidth is larger than the pilot bandwidth for almost all the covariate values in the four stages seems to be that the number of data is limited. It is assumed that for larger sample sizes, the pilot bandwidth will increase to become as expected, asymptotically larger than the bootstrap bandwidth.

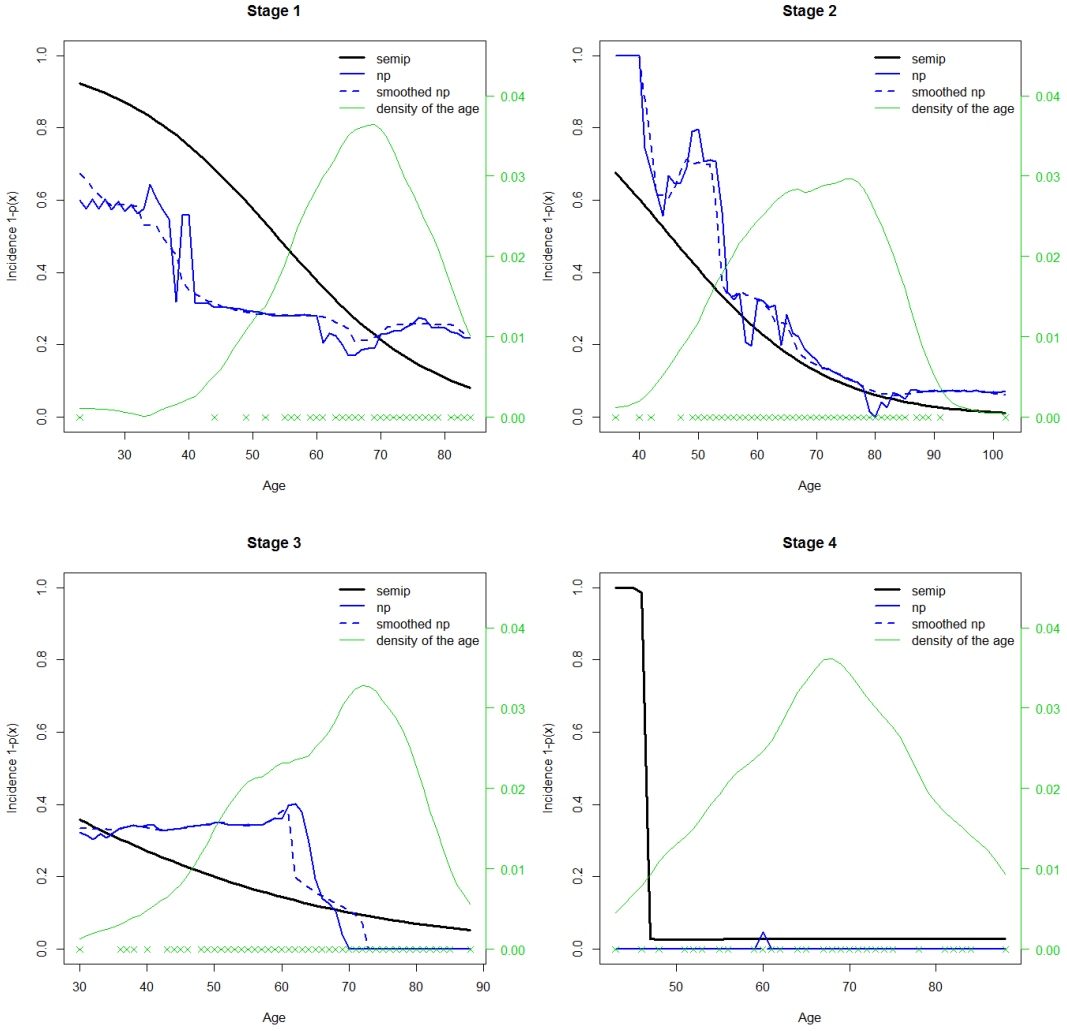

Figure 2.9: Semiparametric (black line) and nonparametric estimations of the incidence in Stages 1-4 depending on the age, computed with the bootstrap bandwidth h_x^* (solid blue line) and with the smoothed bootstrap bandwidth $h_x^{*\,smoothed}$ (dashed blue line). The green line represents the Parzen-Rosenblatt kernel density estimations of the covariate age, using Sheather and Jones' plug-in bandwidth.

2.6.1 Test by Maller & Zhou (1992)

As we studied in Section 2.3, for the incidence estimator to be consistent, assumption (2.8) should hold. In order to check this condition, Maller & Zhou (1992) propose a test for an unconditional context. Let $T^1_{\max}$ be the largest uncensored failure time, and $T_{(n)}$ the largest (censored or uncensored) time. If $T_{(n)}$ is not censored, then the estimator of the cure probability is zero. Therefore, we can assume that $T^1_{\max} < T_{(n)}$ and, on the interval $(T^1_{\max}, T_{(n)}]$, the survival estimator evaluated at $T^1_{\max}$ takes a

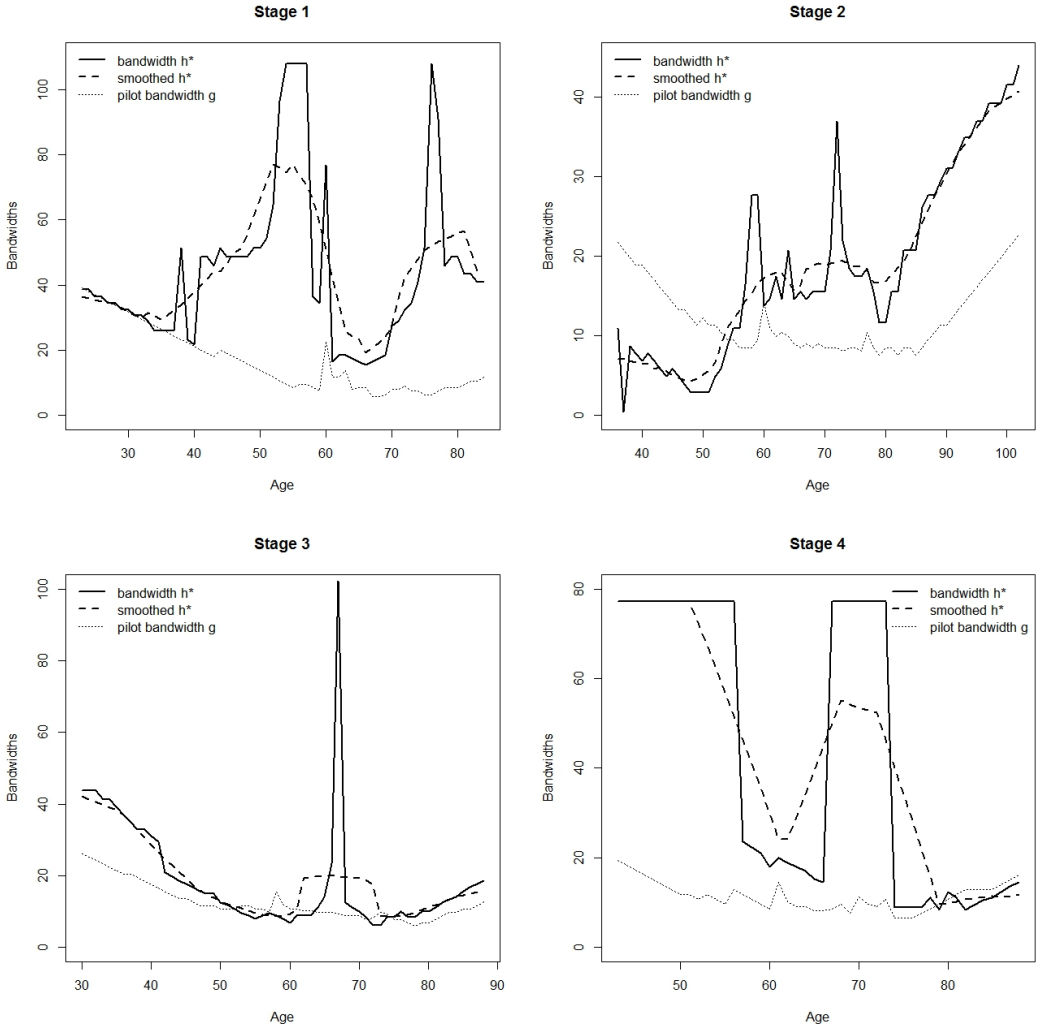

Figure 2.10: Bootstrap bandwidth h_x^* (solid line), smoothed bootstrap bandwidth $h_x^{*\,smoothed}$ (dashed line) and local pilot bandwidth g_x (dotted line) used for the nonparametric incidence estimator for patients in Stages 1-4.

constant value, equal to $1 - \hat{p}_h(T^1_{\max})$. The length of this final plateau is indicative of whether $\tau_0 < \tau_G$ or not.

The method by Maller & Zhou (1992) consists of testing the hypothesis $H_0 : \tau_0 > \tau_G$. For that purpose, the length of the interval of constancy at the right extreme of the survival estimator, $T_{(n)} - T^1_{\max}$, is considered. If $T_{(n)} - T^1_{\max}$ is too large, the hypothesis will be rejected. The p-value for the test is

$$P(T_{(n)} - T^1_{\max} > \varsigma_n) = P(T^1_{\max} < T_{(n)} - \varsigma_n), \tag{2.26}$$

where ς_n is the observed value of $T_{(n)} - T^1_{\max}$. Note that the largest censored observation can not exceed τ_G, and therefore (2.26) is bounded above by

$$P(T^1_{\max} < \tau_G - \varsigma_n). \tag{2.27}$$

Furthermore, the distribution of $T^1_{\max}$ can be obtained by

$$\begin{aligned} q(a) &= P(T^1_{\max} \leq a) = \{P(T \leq a) + P(T > a, Y > C)\}^n \\ &= \{1 - P(a < Y \leq C)\}^n \text{ for } a > 0, \end{aligned}$$

where Y is the failure time, C is the censoring variable and T is the observed time. The observed value of $T^1_{\max}$ in the dataset, w_n, will be replaced by τ_G in (2.27), which should be a close value given that for large sample sizes, w_n is similar to τ_G under the null hypothesis. Then, $q(a)$ can be estimated by

$$\hat{q}_n(a) = \left(1 - \frac{\text{number of uncensored observations} > a}{n}\right)^n,$$

and an approximation to the upper bound (2.27) for the p-value, Maller & Zhou (1992) propose

$$\alpha_n = \hat{q}_n\{\min((w_n - \varsigma_n), 0)\}.$$

Consequently, the null hypothesis should be rejected if $\alpha_n < \alpha$, where α is the significance level, since the observed value of the difference $T_{(n)} - T^1_{\max}$ is then improbable under H_0. Note that the value $\hat{q}_n(w_n - \varsigma_n)$ is the number of uncensored observations which belong to the interval $(2T^1_{\max} - T_{(n)}, T^1_{\max}]$, due to the fact that there are no uncensored observations greater than $T^1_{\max}$.

We apply the test by Maller & Zhou (1992) to the colorrectal cancer dataset. The largest uncensored failure time, $T^1_{\max}$ is 183.05; and the largest time (censored or uncensored), $T_{(n)}$, is 225.64. Then, $\varsigma = T_{(n)} - T^1_{\max} = 42.59$, and the interval we have to consider is $(183.05 - 42.59, 183.05] = (140.46, 183.05]$. Examining the dataset, there are 7 uncensored observations within this interval. Therefore,

$$\alpha_n = \left(1 - \frac{7}{414}\right)^{414} = 0.0008589.$$

Since $\alpha_n < \alpha = 0.05$, we reject the null hypothesis and then condition (2.8) holds.

Chapter 3

Nonparametric latency estimator

3.1 Introduction

In this chapter, we thoroughly study the nonparametric latency estimator. We start introducing a general estimator using two different bandwidths, b_1 and b_2. Then, we consider the nonparametric latency estimator for only one bandwidth, $b_1 = b_2$. We obtain an i.i.d. representation, we find the asymptotic mean squared error and we prove the asymptotic normality. Then, we focus on the nonparametric latency estimator which considers only one bandwidth. Similarly as for the incidence, we propose a bootstrap bandwidth selection method. We assess the performance of the nonparametric latency estimator and we evaluate the efficiency of the bandwidth selector in a simulation study. Finally, we apply these methods to the colorectal cancer dataset. The main results of this chapter have been published in López-Cheda et al. (2017b).

Departing from (2.1), a general form for the nonparametric latency estimator is:

$$\hat{S}_{0,b_1,b_2}(t|x) = \frac{\hat{S}_{b_2}(t|x) - (1 - \hat{p}_{b_1}(x))}{\hat{p}_{b_1}(x)}, \tag{3.1}$$

where $\hat{S}_{b_2}(t|x)$ is the Beran estimator of $S(t|x)$ in (2.2) and $1-\hat{p}_{b_1}(x)$ is the estimator by Xu & Peng (2014) in (2.4). Two different bandwidths are considered since the optimal bandwidth for $\hat{S}_b(t|x)$ needs not to be the optimal bandwidth for $\hat{p}_b(x)$. The latency estimator in (3.1) does not yield necessarily a proper survival function, since its limit as t tends to infinity is not required to be zero. In fact, it is not even guaranteed to be non negative. On the other hand, as it will be shown in Section

3.5.1, the optimal values for b_1 and b_2 in (3.1) are nearly equal. As a consequence, in this thesis we focus on the nonparametric latency estimator, which depends on one unique bandwidth, $b = b_1 = b_2$:

$$\hat{S}_{0,b}(t|x) = \frac{\hat{S}_b(t|x) - (1 - \hat{p}_b(x))}{\hat{p}_b(x)}, \tag{3.2}$$

where b is a smoothing parameter not necessarily equal to h in (2.4). For the estimator in (3.2) it is clear that $\hat{S}_{0,b}(t|x)$ is decreasing in t, $0 \leq \hat{S}_{0,b}(t|x) \leq 1$ for all x and t, $\lim_{t\to-\infty} \hat{S}_{0,b}(t|x) = \hat{S}_{0,b}(0|x) = 1$ and $\lim_{t\to+\infty} \hat{S}_{0,b}(t|x) = 0$. So $\hat{S}_{0,b}(t|x)$ is a proper survival function in t.

3.2 Asymptotic results considering two different bandwidths

Theorem 3.2.1 gives an i.i.d. representation of $\hat{S}_{0,b_1,b_2}(t|x)$, the nonparametric latency estimator considering two different bandwidths, b_1 and b_2, defined in Equation (3.1).

Additionally to Assumptions $(A1)$-$(A13)$ introduced in Section 2.3, we need to consider the following condition, to be used in the asymptotic results for the latency estimator:

(A14) $b_i \to 0$, $\frac{\ln n}{nb_i} \to 0$, $\frac{nb_i^5}{\ln n} = O(1)$, $\frac{(\ln\ln n)^4}{(\ln n)^3}\frac{b_i}{nb_j^2} = O(1)$ and $\frac{(\ln\ln n)^2}{(\ln n)^3}\frac{nb_i^{11}}{b_j^2} = O(1)$, for $i, j = 1, 2,\ i \neq j$.

This condition needs to hold in order to obtain the i.i.d. representation (Theorems 3.2.1 and 3.3.2), to find the asymptotic mean squared error (Theorems 3.2.2 and 3.3.3) and to prove the asymptotic normality of the nonparametric latency estimator (Theorems 3.2.4 and 3.3.4). If the bandwidths are $b_1 = c_1 n^{-a} + o(n^{-a})$ and $b_2 = c_2 n^{-b} + o(n^{-b})$, then the second condition in $(A14)$, $\frac{\ln n}{nb_i} \to 0$, implies that $a > 0$ and $b > 0$; the third condition, $\frac{nb_i^5}{\ln n} = O(1)$, implies that $a < 1$ and $b < 1$; the fourth condition, $\frac{(\ln\ln n)^4}{(\ln n)^3}\frac{b_i}{nb_j^2} = O(1)$, implies that $2b - 1 \leq a \leq \frac{1}{2}(b+1)$; and finally, the fifth condition, $\frac{(\ln\ln n)^2}{(\ln n)^3}\frac{nb_i^{11}}{b_j^2} = O(1)$, implies that, $\frac{1}{11}(2b+1) \leq a \leq \frac{1}{2}(11b-1)$. In the particular case that both bandwidths have the same order $a = b = 1/5$, all the conditions are fulfilled.

Theorem 3.2.1. *Under assumptions* $(A1)$-$(A13)$*, and for two sequences of bandwidths satisfying* $(A14)$*, then the i.i.d. representation of the nonparametric latency estimator in (3.1) is:*

$$\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) = \sum_{i=1}^{n} \eta_{b_1,b_2}(T_i, \delta_i, X_i, t, x) + O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4} + \left(\frac{\ln n}{nb_2}\right)^{3/4}\right) a.s.,$$

where

$$\begin{aligned} \eta_{b_1,b_2}(T_i, \delta_i, X_i, t, x) &= -\frac{S(t|x)}{p(x)} \tilde{B}_{b_2 i}(x) \xi(T_i, \delta_i, t, x) \\ &- \frac{(1-p(x))(1-S(t|x))}{p(x)^2} \tilde{B}_{b_1 i}(x) \xi(T_i, \delta_i, \infty, x), \end{aligned} \tag{3.3}$$

with $\xi(T_i, \delta_i, t, x)$ *defined in (2.10) and* $\tilde{B}_{b_j i}(x)$*,* $j = 1, 2$ *in (2.9).*

From Theorem 3.2.1, the asymptotic expression for the MSE of the nonparametric latency estimator can be obtained.

Theorem 3.2.2. *Under assumptions* $(A1)$-$(A13)$*, and for two sequences of bandwidths satisfying* $(A14)$*, then the mean squared error of the latency estimator satisfies*

$$\begin{aligned} MSE(\hat{S}_{0,b_1,b_2}(t|x)) &= AMSE(\hat{S}_{0,b_1,b_2}(t|x)) \\ &+ o(b_2^4) + o(b_1^4) + o\left(b_1^2 b_2^2\right) + O\left(\frac{b_2}{n}\right) + O\left(\frac{b_1}{nb_2}\right), \end{aligned}$$

where

$$\begin{aligned} AMSE(\hat{S}_{0,b_1,b_2}(t|x)) &= \left(\frac{b_2^2}{2} d_K B_1(t,x) + \frac{b_1^2}{2} d_K B_2(t,x)\right)^2 + \frac{1}{nb_2} V_1(t,x)\, c_K \\ &+ \frac{1}{nb_1} V_2(t,x)\, c_K + 2\frac{1}{nb_1} V_3(t,x) \int K(u)\, K\left(\frac{b_2}{b_1} u\right) du, \end{aligned}$$

and

$$B_1(t,x) = \frac{S(t|x)}{p(x)m(x)} \left(\Phi''(x,t,x)\, m(x) + 2\Phi'(x,t,x)\, m'(x)\right), \tag{3.4}$$

$$\begin{aligned} B_2(t,x) &= \frac{(1-p(x))(1-S(t|x))}{p^2(x)m(x)} \\ &\times \left(\Phi''(x,\infty,x)\, m(x) + 2\Phi'(x,\infty,x)\, m'(x)\right), \end{aligned} \tag{3.5}$$

$$\Phi(y,t,x) = \int_0^t \frac{dH^1(v|y)}{1-H(v|x)} - \int_0^t (1-H(v|y)) \frac{dH^1(v|x)}{(1-H(v|x))^2},$$

where Φ' *and* Φ'' *are the partial derivatives of* $\Phi(y,t,x)$ *with respect to* y*. Furthermore,*

$$V_1(t,x) = \left(\frac{S(t|x)}{p(x)}\right)^2 \frac{1}{m(x)} \int_0^t \frac{dH^1(v|x)}{(1-H(v|x))^2}, \tag{3.6}$$

$$V_2(t,x) = \left(\frac{(1-p(x))(1-S(t|x))}{p^2(x)}\right)^2 \frac{1}{m(x)} \int_0^\infty \frac{dH^1(v|x)}{(1-H(v|x))^2} \text{ and} \tag{3.7}$$

$$V_3(t,x) = \frac{(1-p(x))S(t|x)(1-S(t|x))}{p^3(x)m(x)} \int_0^t \frac{dH^1(v|x)}{(1-H(v|x))^2}, \tag{3.8}$$

with d_K *in (1.9) and* c_K *in (1.10).*

Note that, except for some constants, $B_1(t,x)$ in (3.4) and $B_2(t,x)$ in (3.5) are the dominant terms of the asymptotic bias of the estimators $\hat{S}_{b_2}$ and $1-\hat{p}_{b_1}$ in (2.2) and (2.4), respectively. Similarly, the terms $V_1(t,x)$ in (3.6) and $V_2(t,x)$ in (3.7) are the dominant terms of the corresponding asymptotic variances of $\hat{S}_{b_2}$ and $1-\hat{p}_{b_1}$. Finally, $V_3(t,x)$ in (3.8) accounts for the covariance of both estimators.

Remark: The expression $AMSE(\hat{S}_{0,b_1,b_2}(t|x))$ in Theorem 3.2.2 denotes the MSE of the almost sure dominant term of the estimator $\hat{S}_{0,b_1,b_2}(t|x)$ as shown in Theorem 3.2.1.

Departing from Theorem 3.2.1 and Theorem 3.2.2, the optimal bandwidths which minimize the AMSE of the latency estimator are obtained in Theorem 3.2.3.

Theorem 3.2.3. *The bandwidths which minimize the asymptotic expression of* $MSE(\hat{S}_{0,b_1,b_2}(t|x))$ *are*

$$\widehat{b}_1(t,x) = \left(\frac{\frac{c_K}{L_n(t,x)}V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\int K(u)K(L_n(t,x)u)du}{d_K^2\left(L_n^2(t,x)B_1(t,x) + B_2(t,x)\right)^2}\right)^{1/5} n^{-1/5}$$

and

$$\widehat{b}_2(t,x) = L_n(t,x)\widehat{b}_1(t,x),$$

where

$$L_n(t,x) = \arg\min_{L>0} \psi(t,x,L)$$

and

$$\begin{aligned}\psi(t,x,L) &= \left(L^2 B_1(t,x) + B_2(t,x)\right) \\ &\quad \times \left(\frac{c_K}{L}V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\int K(u)K(Lu)du\right)^2,\end{aligned} \tag{3.9}$$

with $B_1(t,x)$ *in (3.4),* $B_2(t,x)$ *in (3.5),* $V_1(t,x)$ *in (3.6),* $V_2(t,x)$ *in (3.7) and* $V_3(t,x)$ *in (3.8).*

In Theorem 3.2.4, the asymptotic normality of the latency estimator is proven.

Theorem 3.2.4. *Under assumptions* (A1)-(A13)*, if* $b_i \to 0$ *for* $i = 1, 2$ *and* $\left((\ln n)^3 / nb_i\right) \times \left(b_j/(b_1+b_2)\right)^2 \to 0$ *for* $i, j = 1, 2$ *with* $i \neq j$*, it follows that*

a) If $nb_i^5 \frac{b_j}{b_1+b_2} \to 0$ *for* $i, j = 1, 2$ *and* $i \neq j$*, then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(0, \sigma_a^2(t,x)\right),$$

where

$$\sigma_a^2(t,x) = \begin{cases} V_2(t,x)\, c_K, & \text{if } b_1/b_2 \to 0 \\ V_1(t,x)\, c_K, & \text{if } b_2/b_1 \to 0 \\ \frac{C_1}{C_1+C_2}\left(V_1(t,x)\, c_K + 2V_3(t,x) \int K(u) K\left(\frac{C_1}{C_2}u\right) du\right) + \frac{C_2}{C_1+C_2} V_2(t,x)\, c_K, & \\ \qquad \text{if } b_1 = C_1 n^{-\alpha} + o(n^{-\alpha}),\ b_2 = C_2 n^{-\alpha} + o(n^{-\alpha}),\ \text{with } \alpha > \frac{1}{5} \end{cases}$$

with $V_1(t,x)$ *in (3.6),* $V_2(t,x)$ *in (3.7) and* $V_3(t,x)$ *in (3.8).*

b) If $nb_1^5 \to 0$ *and* $nb_2^5 \to C_2^5 > 0$*, then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(0, \sigma_b^2(t,x)\right),$$

with $\sigma_b^2(t,x) = V_2(t,x)\, c_K$.

c) If $nb_1^5 \to C_1^5 > 0$ *and* $nb_2^5 \to 0$*, then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(0, \sigma_c^2(t,x)\right),$$

with $\sigma_c^2(t,x) = V_1(t,x)\, c_K$.

d) If $nb_1^5 \to C_1^5 > 0$ *and* $nb_2^5 \to C_2^5 > 0$*, then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(b(t,x), \sigma_d^2(t,x)\right),$$

where

$$b(t,x) = \frac{1}{2} d_K \left(\frac{C_1 C_2}{C_1+C_2}\right)^{1/2} \left(C_2^2 B_1(t,x) + C_1^2 B_2(t,x)\right)$$

and

$$\begin{aligned} \sigma_d^2(t,x) &= \frac{C_1}{C_1+C_2}\left(V_1(t,x)\, c_K + 2V_3(t,x) \int K(u) K\left(\frac{C_1}{C_2}u\right) du\right) \\ &+ \frac{C_2}{C_1+C_2} V_2(t,x)\, c_K. \end{aligned}$$

In Section 3.3, similar asymptotic results are easily obtained for the nonparametric latency estimator in (3.2), $\hat{S}_b(t|x)$, which considers only one bandwidth $b = b_1 = b_2$.

3.3 Asymptotic results considering one bandwidth

As we mentioned in Section 3.1, the latency estimator in (3.1) does not yield necessarily a proper survival function (indeed, it is not even guaranteed to be non negative). Since the optimal values for b_1 and b_2 in (3.1) are nearly equal, as it will be shown in Section 3.5.1, we focus on the nonparametric latency estimator using one bandwidth.

Assumptions $(A1)$-$(A12)$ presented in Section 2.3 are needed to prove the asymptotic results for the latency estimator. The following theorems are introduced for the latency estimator considering one bandwidth, b.

In the next theorem we show that the proposed nonparametric latency estimator is the local maximum likelihood estimator of $S_0(t|x)$. Its proof is in Appendix A.1.

Theorem 3.3.1. *The estimator $\hat{S}_{0,b}(t|x)$, given in (3.2) is the local maximum likelihood estimator of $S_0(t|x)$ for the mixture cure model (2.1), for any $x \in D$ and $t \geq 0$.*

In Theorem 3.3.2 we obtain an i.i.d. representation for $\hat{S}_{0,b}(t|x)$ in (3.2).

Theorem 3.3.2. *Suppose that conditions $(A1)$-$(A13)$ hold. If $\frac{\ln n}{nb} \to 0$ and $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$, then we have an i.i.d. representation for the nonparametric latency estimator for any $t \in [a, b]$:*

$$\hat{S}_{0,b}(t|x) - S_0(t|x) = \sum_{i=1}^{n} \eta_b(T_i, \delta_i, X_i, t, x) + O\left(\left(\frac{\ln n}{nb}\right)^{3/4}\right) a.s.,$$

with

$$\begin{aligned} \eta_b(T_i, \delta_i, X_i, t, x) &= -\frac{S(t|x)}{p(x)} \tilde{B}_{bi}(x) \xi(T_i, \delta_i, t, x) \\ &- \frac{(1 - p(x))(1 - S(t|x))}{p^2(x)} \tilde{B}_{bi}(x) \xi(T_i, \delta_i, \infty, x), \end{aligned}$$

where $\xi(T_i, \delta_i, t, x)$ has been defined (2.10) and $\tilde{B}_{bi}(x)$ in (2.9).

From Theorem 3.3.2, important properties of the nonparametric latency estimator can be obtained: the first one is an asymptotic expression of the mean squared error (MSE) given in Theorem 3.3.3, and the second one is the asymptotic normality, shown in Theorem 3.3.4.

Theorem 3.3.3. *Suppose that conditions* $(A1)$-$(A13)$ *hold. If* $\frac{\ln n}{nb} \to 0$ *and* $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$, *then the mean squared error of the latency estimator is*

$$MSE(\hat{S}_{0,b}(t|x)) = \frac{b^4}{4} d_K^2 B^2(t,x) + \frac{c_K}{nb} V(t,x) + o(b^4) + O\left(\frac{1}{n}\right),$$

where d_K *and* c_K *have been defined in (1.9) and (1.10), respectively, and*

$$B(t,x) = B_1(t,x) + B_2(t,x), \tag{3.10}$$

$$V(t,x) = V_1(t,x) + V_2(t,x) + 2V_3(t,x), \tag{3.11}$$

with $t \in [a,b]$, $B_1(t,x)$, $B_2(t,x)$, $V_1(t,x)$, $V_2(t,x)$ *and* $V_3(t,x)$ *in (3.4)-(3.8).*

Theorem 3.3.4. *Suppose that conditions* $(A1)$-$(A13)$ *hold. If* $b \to 0$ *and* $\frac{(\ln n)^3}{nb} \to 0$, *it follows that, for any* $t \in [a,b]$*:*

a) If $nb^5 \to 0$, *then*

$$\sqrt{nb}\left(\hat{S}_{0,b}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(0, V(t,x)\, c_K\right).$$

b) If $nb^5 \to C^5 > 0$, *then*

$$\sqrt{nb}\left(\hat{S}_{0,b}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(B(t,x)\, C^{5/2} d_K, V(t,x)\, c_K\right).$$

3.4 Bandwidth selection

From Theorem 3.3.3, the asymptotic mean integrated squared error of the latency estimator is:

$$AMISE(\hat{S}_{0,b}(\cdot|x)) = \frac{1}{4} d_K^2 b^4 \int B^2(t,x)\, dt + \frac{c_K}{nb} \int V(t,x)\, dt,$$

where $B(t,x)$ and $V(t,x)$ were defined in (3.10) and (3.11). The bandwidth which minimizes the asymptotic mean integrated squared error is

$$b_{AMISE}(x) = \left(\frac{c_K \int V(t,x)dt}{d_K^2 \int B^2(t,x)dt}\right)^{1/5} n^{-1/5},$$

which depends on plenty of unknown functions that are very hard to estimate. Consequently we propose to select the bandwidth using the bootstrap method.

3.4.1 Bootstrap bandwidth selector

The bootstrap bandwidth selector is the minimizer of the bootstrap version of the mean integrated squared error (MISE), that can be approximated, using Monte Carlo, by:

$$MISE^*_{x,g}(b) \simeq \frac{1}{B}\sum_{j=1}^{B}\int \left(\hat{S}^{*(j)}_{0,b}(t|x) - \hat{S}_{0,g}(t|x)\right)^2 w(t)dt, \tag{3.12}$$

where w is an appropriate weight function, $\hat{S}^{*(j)}_{0,b}(t|x)$ is the kernel estimator of $S_0(t|x)$ in (3.2) using bandwidth b and based on the j-th bootstrap resample, and $\hat{S}_{0,g}(t|x)$ is the same estimator computed with the original sample and pilot bandwidth g.

We consider an unconditional censoring bootstrap resampling, assuming that $G(t|x) = G(t)$, $\forall x, t$. Note that this resampling method is equivalent to the one for the incidence presented in Section 2.4.1. The procedure for obtaining the bootstrap bandwidth selector for a fixed covariate value, x, is as follows:

1. For $i = 1, 2, \ldots, n$, generate C^*_i from the product-limit estimator $\hat{G}$.
2. For $i = 1, 2, \ldots, n$, fix the bootstrap covariates $X^*_i = X_i$ and generate Y^*_i from $\hat{S}_{0,g}(\cdot|X^*_i)$ with probability $\hat{p}_g(X^*_i)$, and $Y^*_i = \infty$ otherwise.
3. Finally, define $T^*_i = \min(Y^*_i, C^*_i)$ and $\delta^*_i = I\left(Y^*_i \leq C^*_i\right)$ for $i = 1, 2, \ldots, n$.
4. Repeat Steps 1-3 above B times to generate bootstrap resamples of the form $\left\{(X^{(j)}_1, T^{*(j)}_1, \delta^{*(j)}_1), \ldots, (X^{(j)}_n, T^{*(j)}_n, \delta^{*(j)}_n)\right\}$, $j = 1, \ldots, B$.
5. For the j-th bootstrap resample ($j = 1, 2 \ldots, B$), compute $\hat{S}^{*(j)}_{0,b}(t|x)$ with bandwidth $b_l \in \{b_1, \ldots, b_L\}$.
6. With the original sample and pilot bandwidth g, compute $\hat{S}_{0,g}(t|x)$.
7. For each bandwidth $b_l \in \{b_1, \ldots, b_L\}$, compute the Monte Carlo approximation of $MISE^*_{x,g}(b_l)$ as in (3.12).
8. Find $b^*_x = \underset{b_l \in \{b_1, \ldots, b_L\}}{\arg\min}\ MISE^*_{x,g}(b_l)$.

Similarly as for the nonparametric incidence estimator, the effect of the pilot bandwidth, g, on the bootstrap bandwidth, b^*_x, is very weak. Preliminary studies (see related results below) showed that a good choice would be to consider the same

naive pilot bandwidth selector as in Chapter 2, $g = C(X_{(n)} - X_{(1)})n^{-1/9}$, with $C = 0.75$, and where $X_{(n)}$ ($X_{(1)}$) is the maximum (minimum) observed value of the covariate X.

3.5 Simulation study

There are three purposes of this simulation study: firstly, to show that little efficiency is lost if we consider only one bandwidth, $b = b_1 = b_2$ in the nonparametric latency estimator. Secondly, to evaluate the good practical behavior of the proposed estimator in (3.2), computed in a grid of bandwidths with the Epanechnikov kernel, the MISE of $\hat{S}_{0,b_x}$ is compared with the MISE of the semiparametric latency estimator by Peng & Dear (2000). Note that this estimator, implemented in the "smcure" package, considers a proportional hazards assumption for modeling the effect of covariates on the failure time of patients who are not cured,

$$S_0(t|\mathbf{x}) = U_0(t)^{\exp(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x})},$$

where $U_0(t)$ is a basal function. The third objective is to assess the performance of the bootstrap bandwidth selector for the nonparametric estimator and the weak effect of the pilot bandwidth. We worked with the same two models considered in Section 2.5, where X is $U(-20, 20)$. The results are obtained in the grid of 41 equispaced values of X given by $\{-20, -19, \ldots, 19, 20\}$.

3.5.1 Results for the latency considering two different bandwidths

We present some results for the latency estimator in (3.1), that is, when two different bandwidths are considered: b_1 for the incidence and b_2 for the improper survival function, S. Note that, for the sake of brevity, we only work with Model 1 and sample size $n = 100$, considering $\kappa = 1000$ samples.

Figure 3.1 provides the theoretical MISE bandwidths (approximated by Monte Carlo), (b_1, b_2), as a function of x. Note that for most of the covariate values both optimal bandwidths are very similar, being even equal for the values of x larger than 5.

The MISE of the nonparametric latency estimator, $\hat{S}_{0,b_1,b_2}(t|x)$, in (3.1), as a function of (b_1, b_2), is shown in Figure 3.2 for some fixed covariate values: $x = 5$, $x = 10$, $x = 15$ and $x = 18$. The MISE for other values of x is similar, but not shown here. We can see that for all cases, the minimum MISE (purple color) is reached

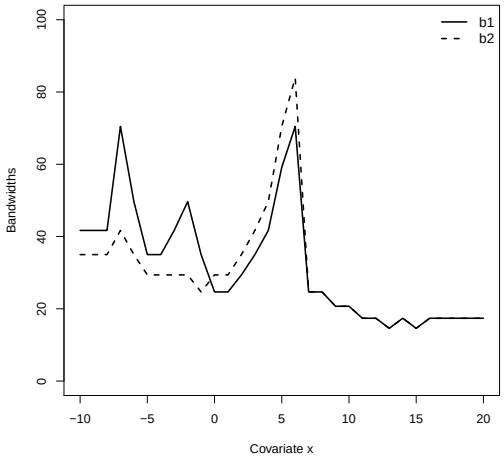

Figure 3.1: Optimal b_1 bandwidth (solid line) and b_2 bandwidth (dashed line), in terms of MISE.

around the diagonal, that is, when $b_1 = b_2$. Therefore, little efficiency is lost when considering one only bandwidth $b_1 = b_2$ to estimate S_0, while this guarantees that the resulting estimator is a proper survival function, as pointed out in Section 3.1.

3.5.2 Efficiency of the nonparametric latency estimator

Figure 3.3 shows the MISE of the latency estimator using one bandwidth, b, approximated by Monte Carlo, for a grid of 100 values equispaced in a logarithmic scale, from $b_1 = 10$ (red line) to $b_{100} = 40$ (dark blue line). A total of $\kappa = 1000$ samples of size $n = 100$ were drawn. Note that the MISE, as a function of the bandwidth b, has a U-shape, since it starts being very large for small bandwidths (red, orange and yellow lines) and then, for medium bandwidths (green and light blue lines), the MISE function decreases considerably, reaching its minimum. Finally, the MISE function becames larger when using the largest bandwidths (dark blue colors). For further clarification, Figure 3.4 shows the MISE for the nonparametric latency estimator depending on the bandwidth b, for four different values of the covariate.

It is noteworthy that in Model 1, for values of the covariate from $x = -20$ to $x = 10$, there is also a very wide range of bandwidths, specifically, between $b_{30} = 15.01$ (light green lines) and $b_{100} = 40$ (dark blue lines), for which the MISE of the nonparametric estimator is smaller than the MISE of the semiparametric estimator. In Model 2, the nonparametric estimator of the latency, computed with

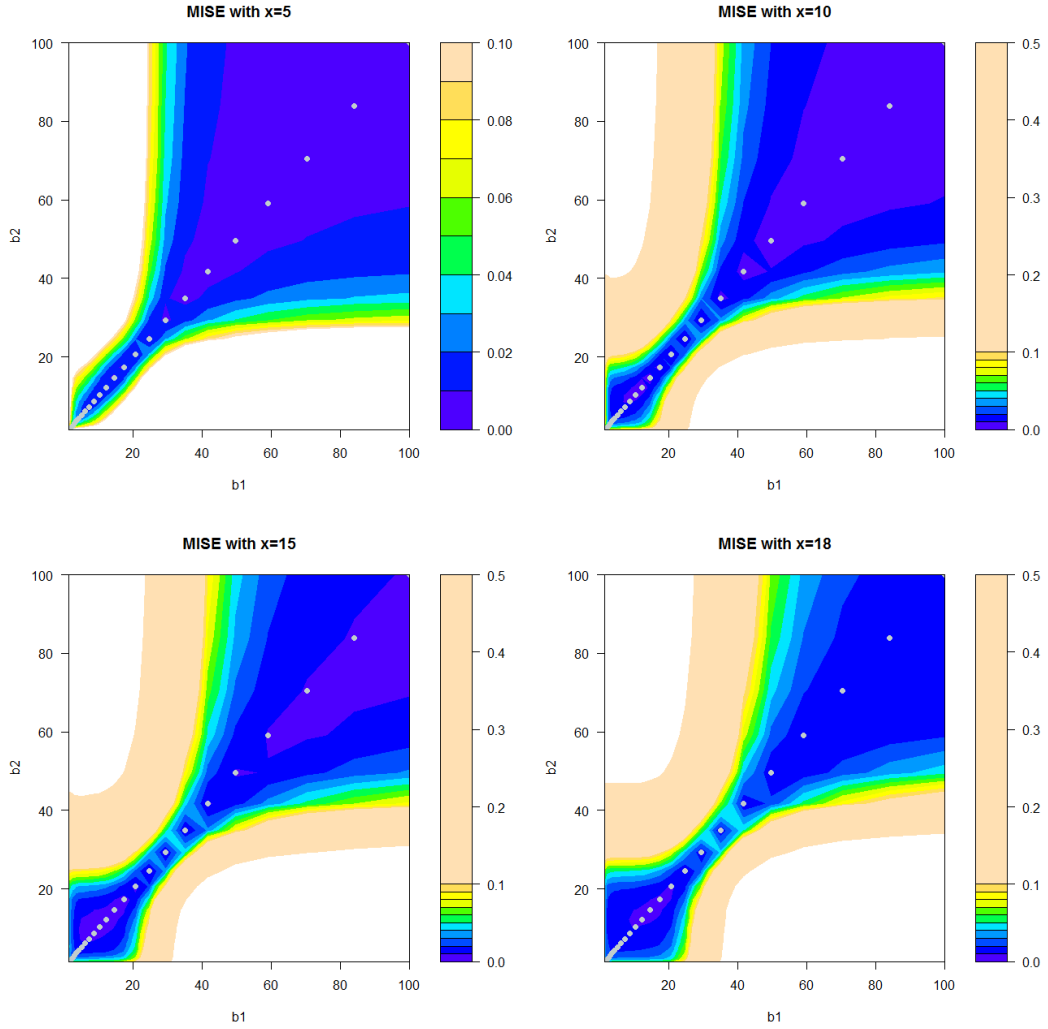

Figure 3.2: $MISE(b_1, b_2)$ of $\hat{S}_{0,b_1,b_2}$ for $x = 5$ (top left), $x = 10$ (top right), $x = 15$ (bottom left) and $x = 18$ (bottom right). The grid of bandwidths (equispaced on a logarithmic scale), where $b_1 = b_2$, is represented with light blue dots.

bandwidths between $b_{20} = 13.05$ (yellow lines) to $b_{100} = 40$ (dark blue lines), outperforms the semiparametric estimator for all the covariate values, except for $x \in [4, 9]$, where the semiparametric estimator is very competitive.

In short, both the nonparametric incidence (in Chapter 2) and latency estimators are quite comparable to the semiparametric ones in situations where the latter are expected to give better results, as in Model 1, and they outperform the semiparametric estimators when the incidence is not a logistic function and the latency does not fit a PH model (Model 2). The efficiency of the nonparametric estimators depends

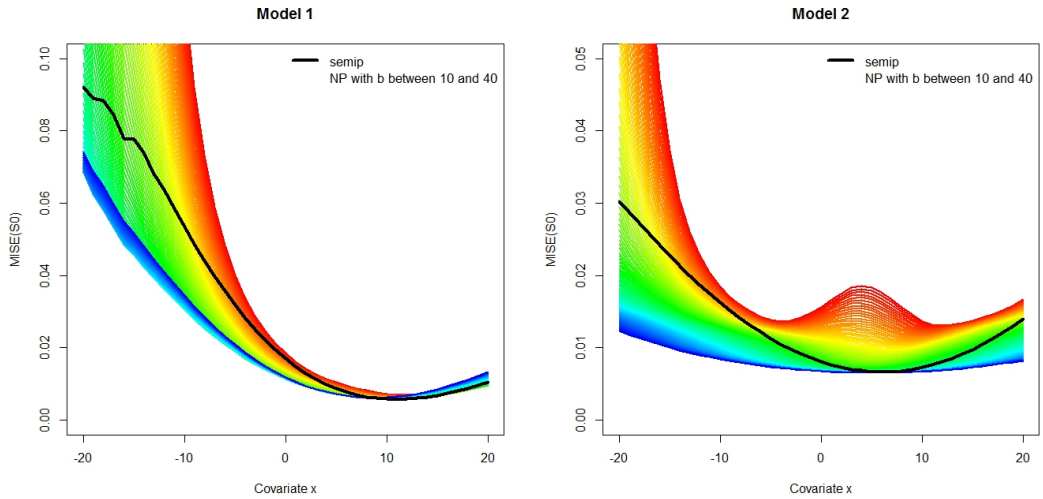

Figure 3.3: MISE for the semiparametric (black line) and the nonparametric estimators of $S_0(t|x)$ computed with different bandwidths: from $b_1 = 10$ (red line) to $b_{100} = 40$ (dark blue line). The data, with sample size $n = 100$, were generated from Model 1 (left) and Model 2 (right).

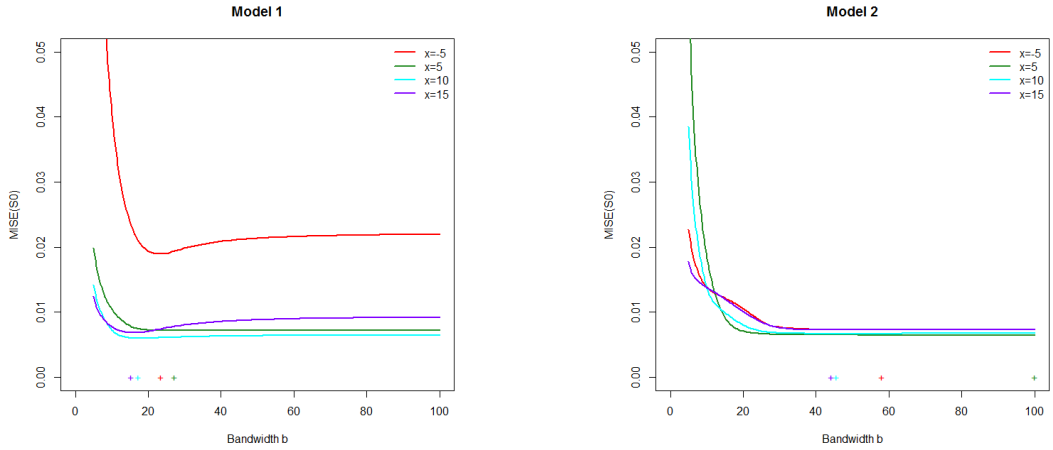

Figure 3.4: MISE for the nonparametric latency estimator depending on the bandwidth, for four different values of the covariate, $x \in \{-5,\ 5,\ 10,\ 15\}$, with sample size $n = 100$, for Model 1 (left) and Model 2 (right). The value of the bandwidth where the minimum MISE is reached for each covariate is marked with crosses of the corresponding color.

on the choice of the bandwidth, but although the optimal value of the bandwidth remains unknown, the simulations show that, for quite wide ranges of bandwidths, the proposed nonparametric methods outperform the existing semiparametric estimator by Peng & Dear (2000).

3.5.3 Efficiency of the bootstrap bandwidth selector

A total of $\kappa = 1000$ trials and $B = 200$ bootstrap resamples of sizes $n = 50$, $n = 100$ and $n = 200$ were drawn to approximate the bootstrap version of the $MISE_x$ for the nonparametric latency estimator, $\hat{S}_{0,b}$. A grid of 35 bandwidths (from 5 to 100), equispaced on a logarithmic scale, is considered. Note that, although the covariate X has distribution $U[-20, 20]$, we only work with $x \in [-10, 20]$. The reason is that $p(x) \simeq 0$ for $-20 \leq x \leq -10$ (see Figure 2.1, left). This implies that almost all the subjects are cured for $x \in [-20, -10)$, and therefore the estimation of the survival function of the uncured population can not be obtained.

Firstly, we compare the values of the resulting bootstrap bandwidth, b_x^*, and the optimal $b_{MISE,x}$ bandwidth by means of the ratio $b_x^*/b_{MISE,x}$ (see Figures 3.5 and 3.6). It is important to highlight that the bootstrap bandwidth b_x^* might be larger (smaller) than $b_{MISE,x}$ in Model 1 (Model 2). Note that in Model 1, the median of the ratio $b_x^*/b_{MISE,x}$ is closer to 1 for covariate values $x \in [-10, 8]$, whereas in Model 2, the covariate values in which the bootstrap bandwidth is more similar to the optimal bandwidth are $x \in [9, 20]$.

This difference between the bootstrap and optimal bandwidths can also be seen in Figure 3.7, where the density of the bootstrap bandwidths, b_x^*, is compared with the optimal $b_{MISE,x}$ bandwidth. The $MISE$ values obtained considering these bandwidths are also shown. We can appreciate how the bootstrap bandwidth might be larger (smaller) than $b_{MISE,x}$ in Model 1 (Model 2), for most of the covariate values. However, it is important to highlight that this slight difference between b_x^* and $b_{MISE,x}$ implies very little difference in terms of MISE between the estimates with the optimal and the bootstrap bandwidths. The reason is that $MISE(\hat{S}_{0,b}(\cdot|x))$, and consequently $MISE_{x,g}^*(b)$, is almost constant in a very wide interval around its minimizer (see Figure 3.4). This feature implies that very different bandwidths could yield very similar good estimates in terms of MISE. In order to check this, the performance of the bootstrap bandwidth, b_x^*, with respect to $b_{MISE,x}$ has been assessed in terms of MISE using the ratio

$$\frac{MISE(b_x^*) - MISE(b_{MISE,x})}{MISE(b_{MISE,x})}.$$

For both models, Figures 3.8 and 3.9 show that there is very little difference in terms of MISE between the estimates with the optimal and the bootstrap bandwidths. The relative error when using b_x^* instead of $b_{MISE,x}$ is generally less than 10%, except with Model 1, when the covariate value is close to 20, then the relative

error is slightly higher.

Similarly as in Section 2.5, we detail the computational time needed in each simulation study. For the latency estimator computed with two bandwidths (Figures 3.1 and 3.2), the simulations take 9.5 hours to be completed. Regarding the efficiency of the latency estimator (Figure 3.3), the method lasts around 40 minutes for each model, considering $n = 100$. Finally, for evaluating the efficiency of the bootstrap bandwidth selector (Figures 3.5, 3.6, 3.7, 3.8 and 3.9), the method is computationally more expensive: 6.5 hours, 7.3 hours and 11.8 hours for each model, with sample sizes $n = 50$, $n = 100$ and $n = 200$, respectively.

Preliminary studies for the pilot bandwidth selection

Only some results of the preliminary studies are shown here. Similarly as for the incidence, we decided to use a global pilot bandwidth for the nonparametric latency estimator. Specifically, we consider the following constant values for the pilot bandwidth: $g = 5,\ 10,\ 25$ and 40. Moreover, we consider a pilot bandwidth which depends on the support of the covariates and which keeps the optimal order:

$$g = (X_{(n)} - X_{(1)})n^{-1/9}.$$

Figure 3.10 shows the theoretical MISE for the nonparametric and semiparametric estimators, and MISE for the nonparametric estimator computed with the bootstrap bandwidth and using the different pilot bandwidths introduced above. Note that for simplicity, we consider sample size $n = 100$, $\kappa = 50$ trials, $B = 50$ bootstrap resamples and Models 1 and 2, detailed in Section 2.5.

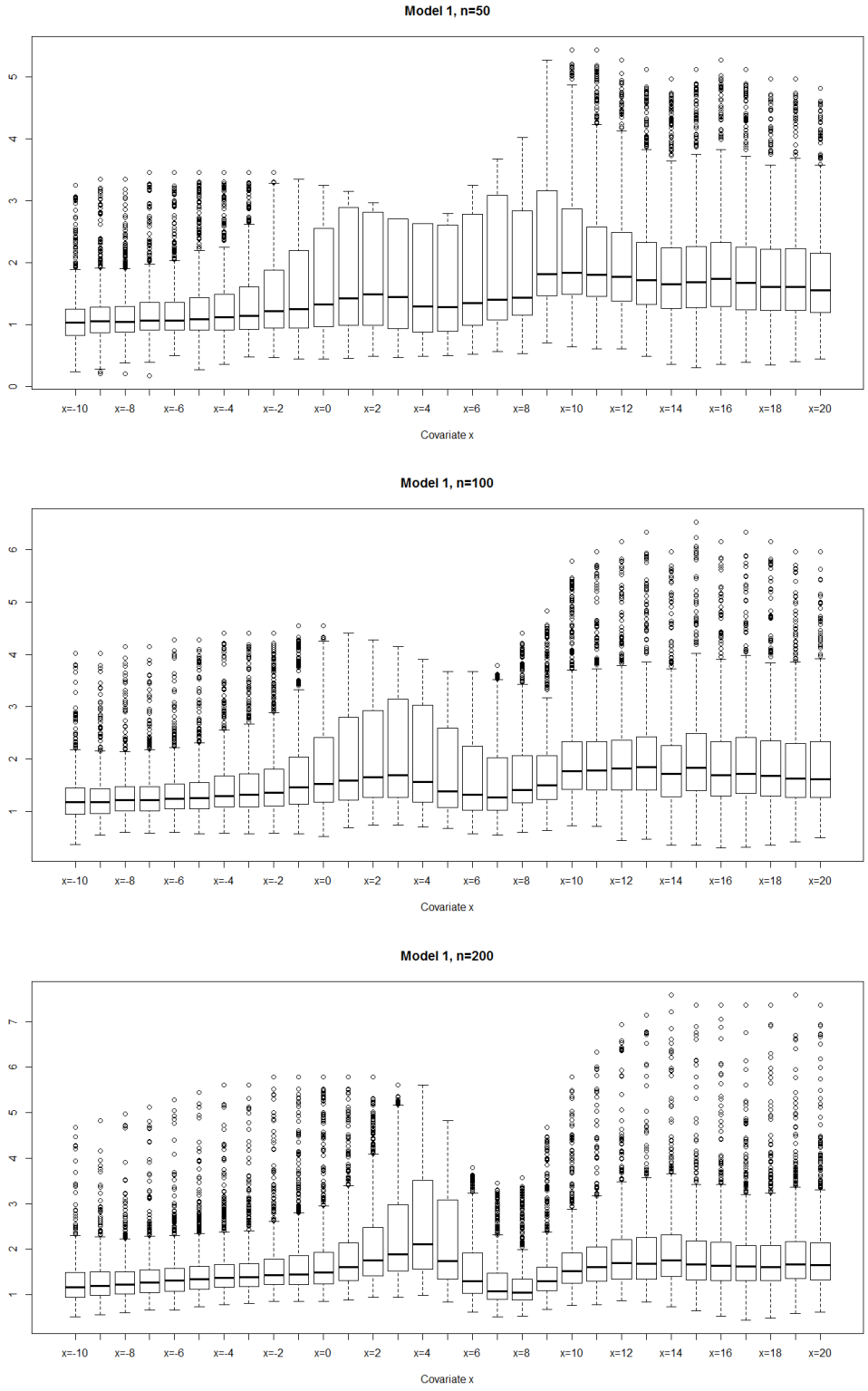

Figure 3.5: Boxplot of the ratio $b_x^*/b_{MISE,x}$ depending on the covariate, for Model 1, with sample sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom).

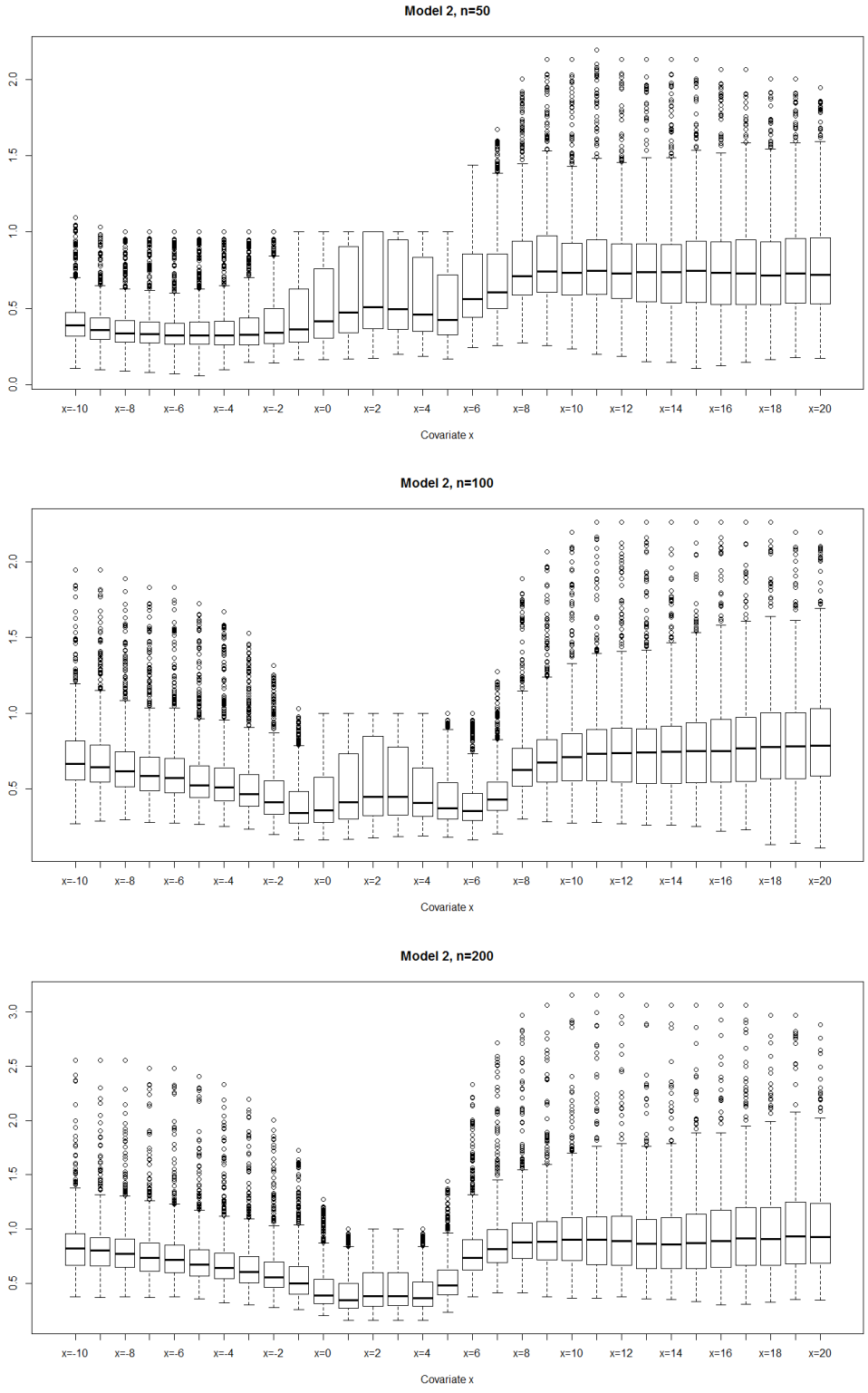

Figure 3.6: Boxplot of the ratio $b_x^*/b_{MISE,x}$ depending on the covariate, for Model 2 with sample sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom).

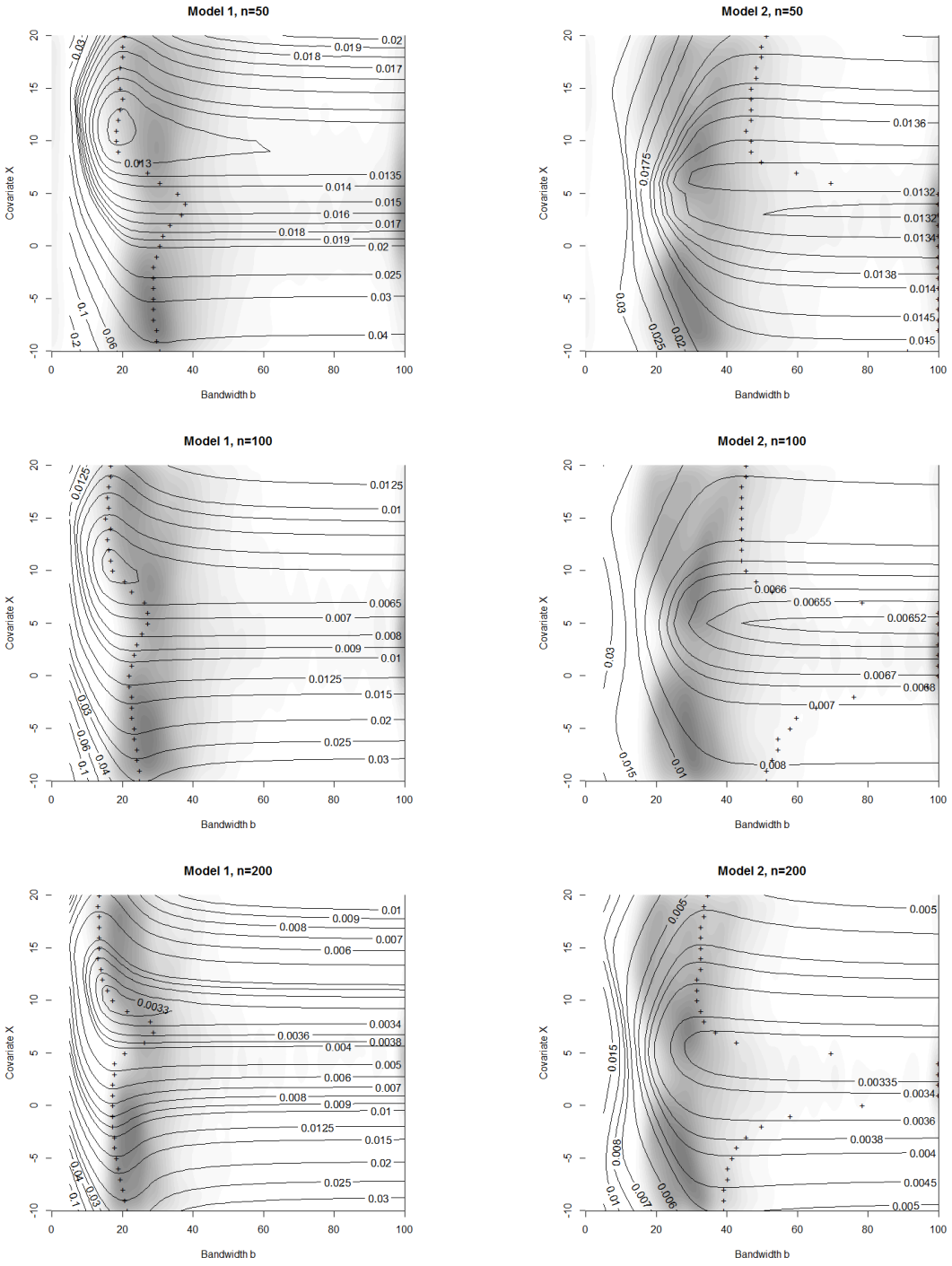

Figure 3.7: MISE contour plot depending on the bandwidth and on the covariate, for Model 1 (left) and Model 2 (right), with sample sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom). The density of the bootstrap bandwidth is displayed in grayscale and the $b_{MISE,x}$ bandwidth, for each covariate value, is represented with crosses.

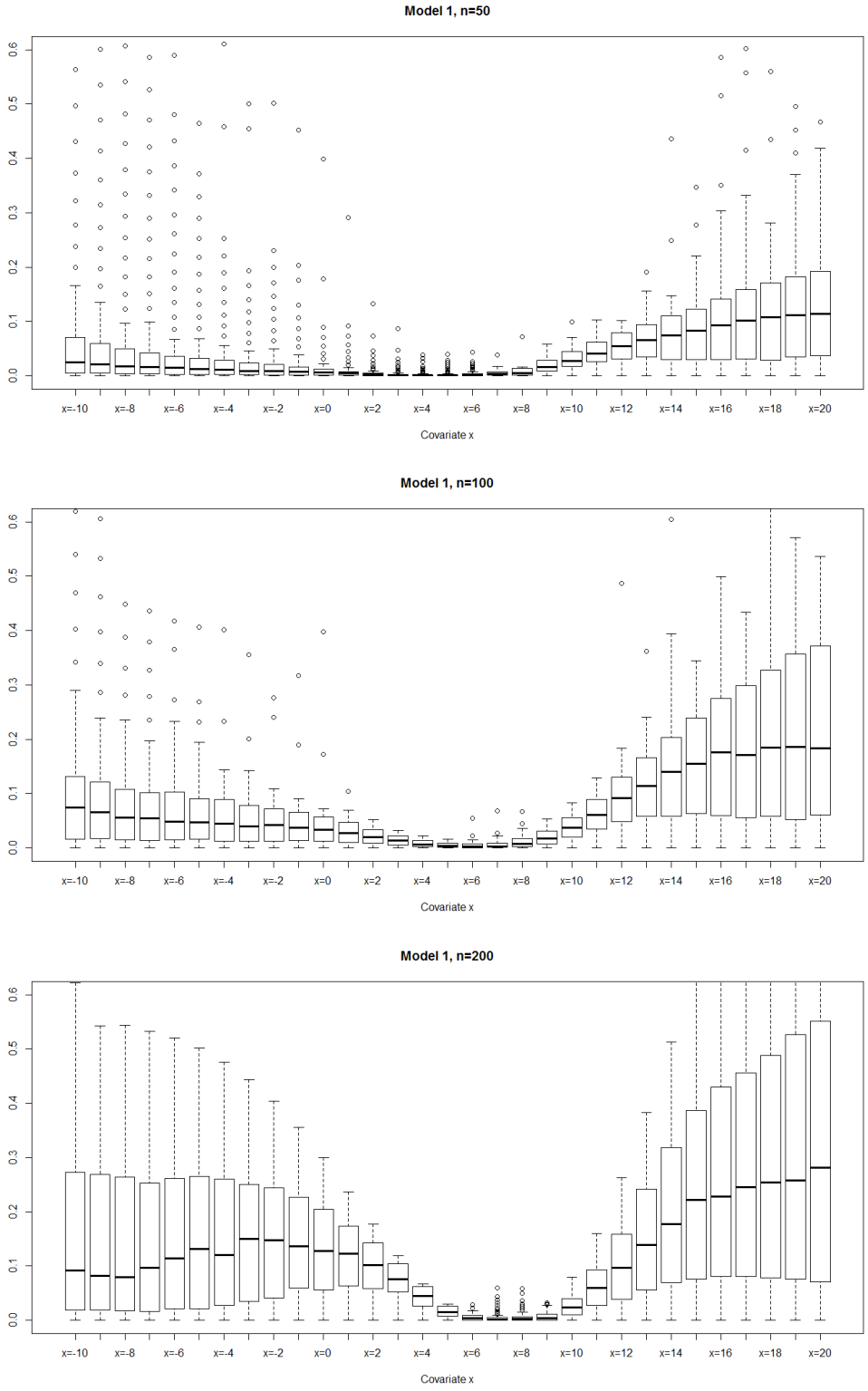

Figure 3.8: Boxplot of the ratio $(MISE(b_x^*) - MISE(b_{MISE,x}))/MISE(b_{MISE,x})$ depending on the covariate, for Model 1, with sample sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom).

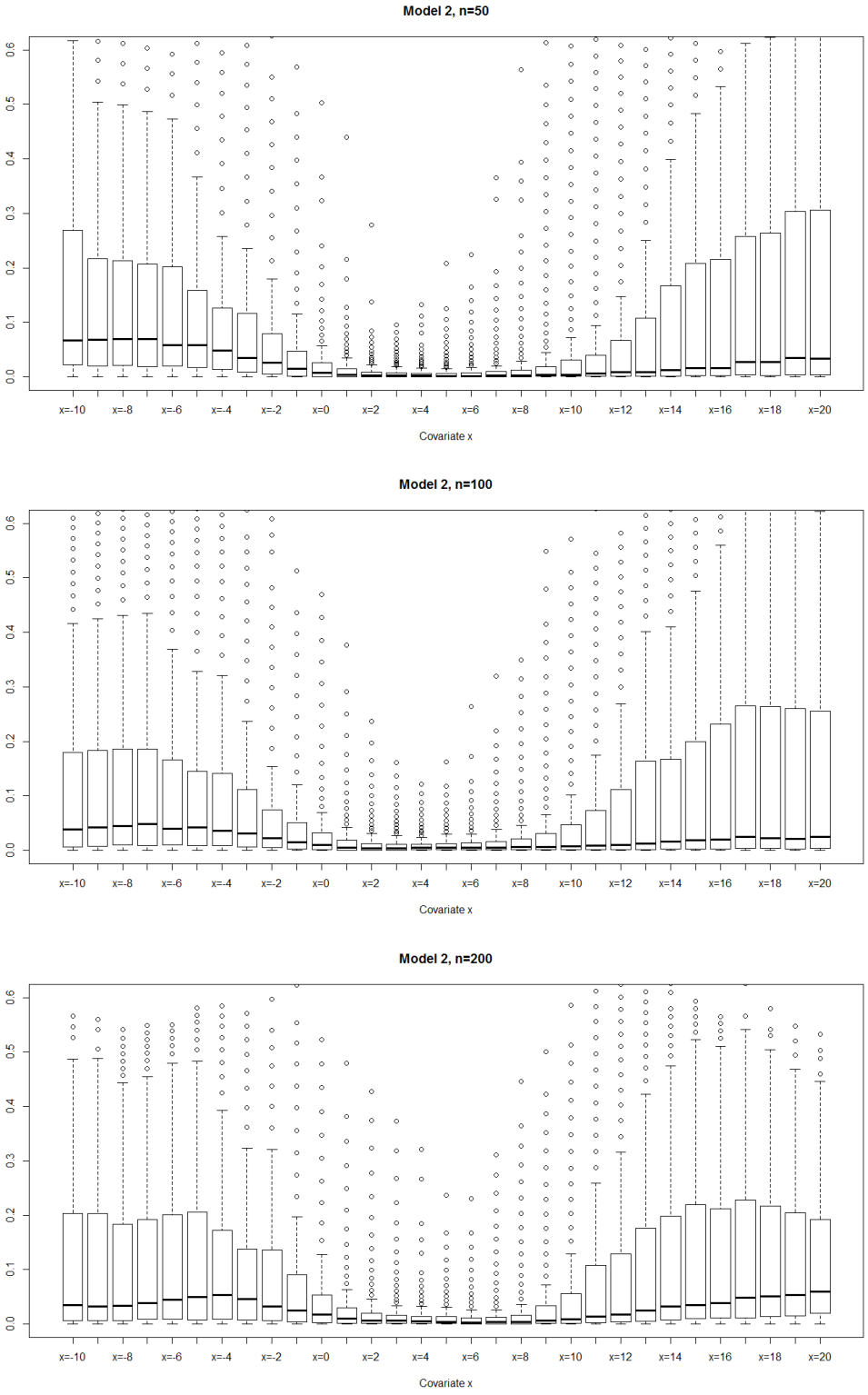

Figure 3.9: Boxplot of the ratio $(MISE(b_x^*) - MISE(b_{MISE,x}))/MISE(b_{MISE,x})$ depending on the covariate, for Model 1 (top) and Model 2 with sample sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom).

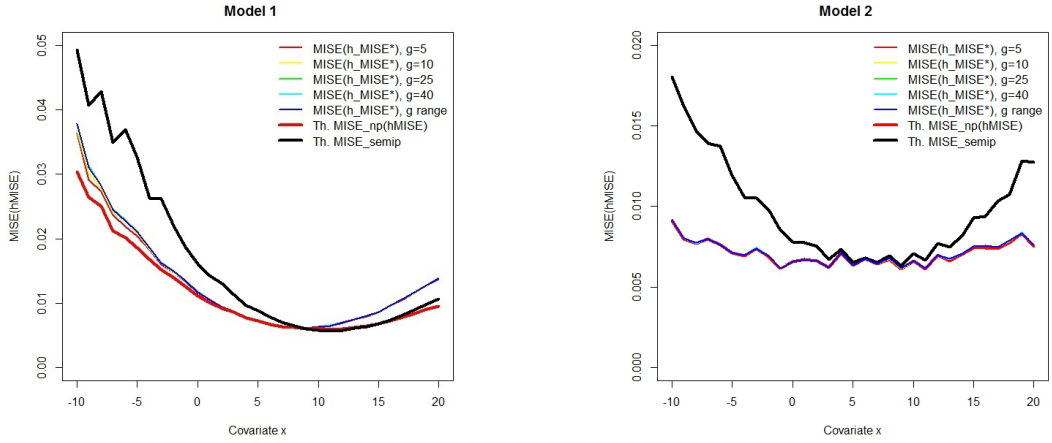

Figure 3.10: Theoretical optimal MISE for the nonparametric estimator (red), MISE for the semiparametric estimator (black), and MISE for the nonparametric estimator computed with the bootstrap bandwidth and using the different pilot bandwidths, for Model 1 (left) and Model 2 (right).

Since MISE for the nonparametric latency estimator computed with different pilot bandwidths is very similar in all cases, we decided to work with the following pilot bandwidth, $g = C(X_{(n)} - X_{(1)})n^{-1/9}$, with $C = 0.75$.

3.6 Application to real data

The proposed method was applied to the dataset used in Section 2.6, composed of 414 colorectal cancer patients from CHUAC.

In Figures 3.11 and 3.12 we show the latency estimation for Stages $1, 2, 3$ and 4 for two different ages, 45 and 76. The nonparametric estimator $\hat{S}_{0,b_x}$ is computed with five different constant bandwidths: $b = 10, 15, 20, 25$ and 30. It is noteworthy that in Stages 1 and 2 for 45 years, the bandwidth selection influences considerably latency estimation. This is due to the low density of the covariate around this age, as we can see in Figure 2.9.

Due to the small sample sizes in each stage, the results are presented in two groups: Stages 1-2 and Stages 3-4. First, the bootstrap bandwidth depending on the covariate age is studied. Then, the latency estimation computed with b_x^* for three different ages (35, 50 and 80) is shown. The number of bootstrap resamples are $B = 200$. Similarly to the simulation study in Section 3.5, we considered a grid of 35 bandwidths from $b_1 = 5$ to $b_{35} = 100$ equispaced on a logarithmic scale.

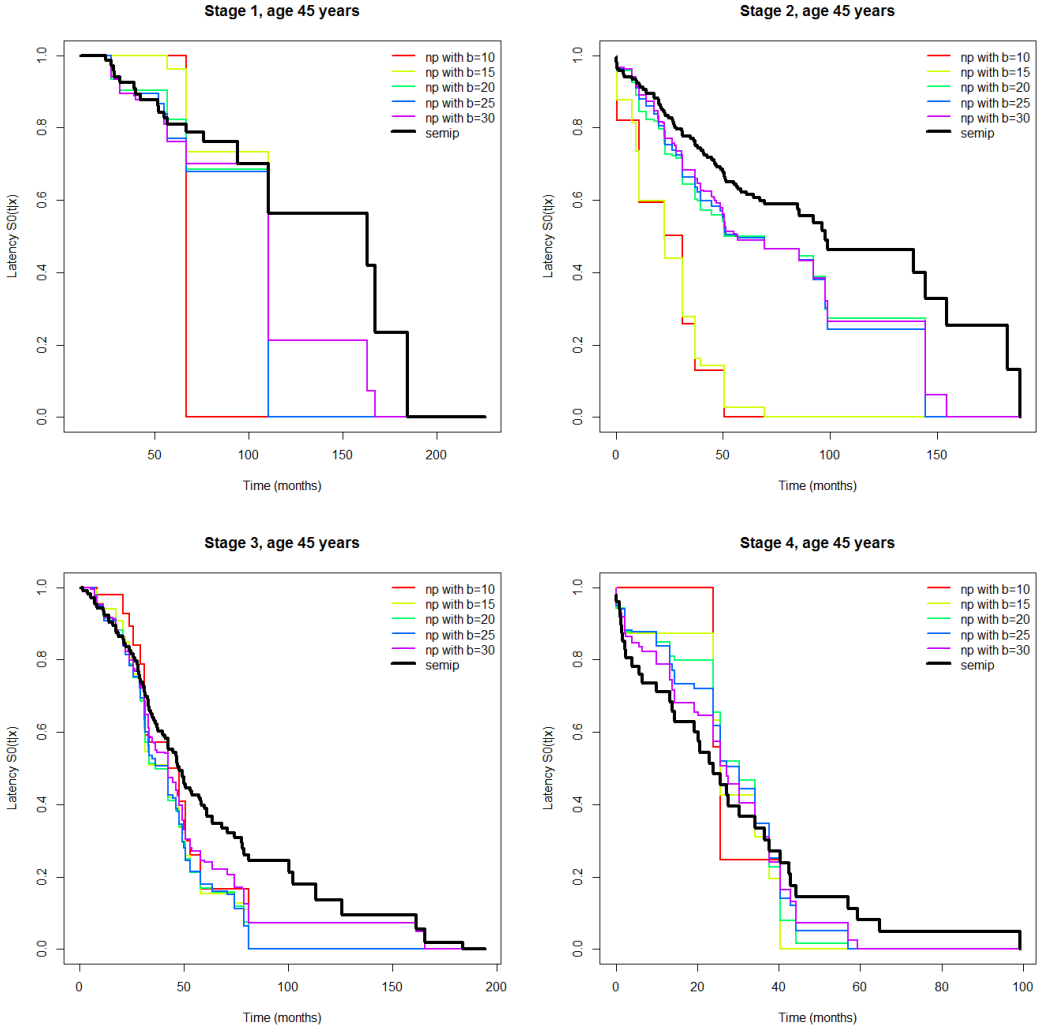

Figure 3.11: Estimated latency for patients of age 45 in Stages 1-4, using the semiparametric (black line) and nonparametric estimators with 5 equispaced bandwidths ranging from $b_0 = 10$ (red line) to $b_4 = 30$ (purple line).

Figure 3.13 shows the resulting bootstrap bandwidth, b_x^*, for Stages 1-2 and Stages 3-4. Note that since there is no enough data near the endpoints of the support of X, obtaining the bootstrap bandwidth for all the covariate values is not possible. It is remarkable that for patients younger than 65, the bandwidth b_x^* for Stages 1-2 is larger than the resulting bandwidth for Stages 3-4. However, for patients older than 65, both bandwidths are very similar. This can be explained because in Stages 1-2 the censoring is much higher and there are less young patients than in Stages 3-4, which leads to a large resulting bootstrap bandwidth.

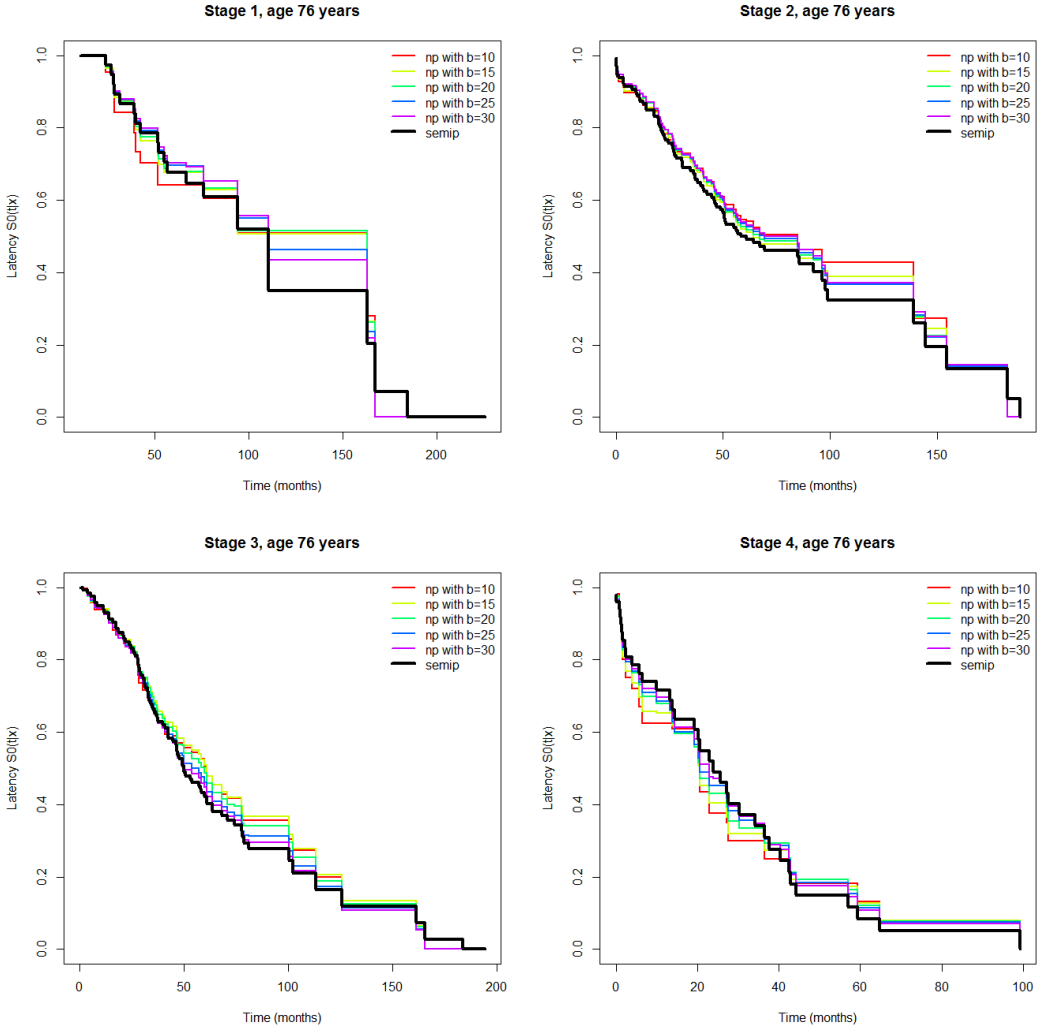

Figure 3.12: Estimated latency for patients of age 76 in Stages 1-4, using the semiparametric (black line) and nonparametric estimators with 5 equispaced bandwidths ranging from $b_0 = 10$ (red line) to $b_4 = 30$ (purple line).

The latency estimation computed with the bootstrap bandwidth, $\hat{S}_{0,b^*}(t|x)$, for different ages (35, 50 and 80) is shown in Figure 3.14. We can observe that for Stages 1-2 the covariate age does not seem to be determining for the latency estimation, since all the estimated latency functions are very similar for the whole grid of ages. On the contrary, for Stages 3-4, the latency estimation varies considerably depending on the age: the short-term survival is larger in young patients, whereas the long-term survival is larger in old individuals. For example, the probability that the follow-up time since the diagnostic until death is larger than 4.5 years (54 months)

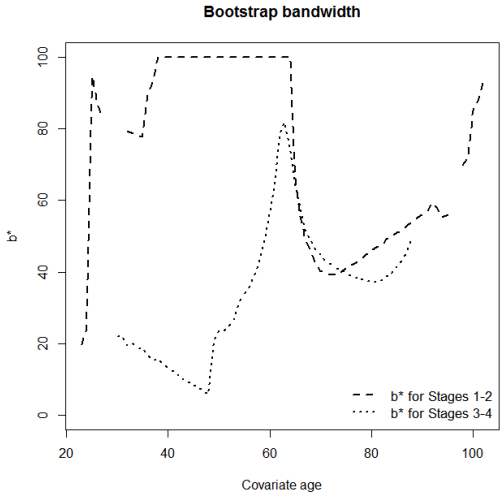

Figure 3.13: Bootstrap bandwidth, b_x^*, depending on the covariate, for patients in Stages 1-2 (dashed line) and 3-4 (dotted line).

is around 0.2 for patients with ages 35 and 50, whereas for 80 year old patients, that probability is larger than 0.4. The reason is that when a colorrectal cancer is diagnosed in a young patient, it is usually in an advanced stage and with worse prognosis, since the cancer cells are more active in young individuals.

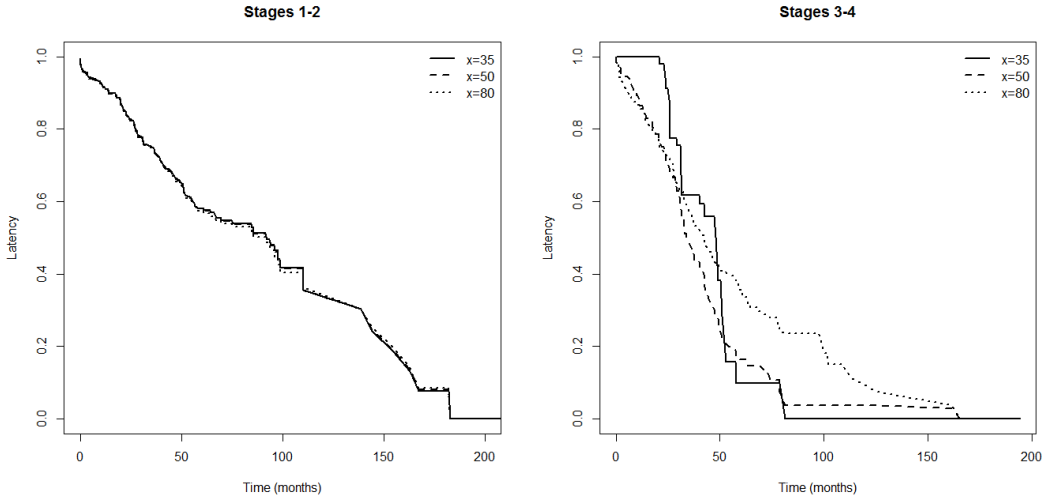

Figure 3.14: Latency estimation for patients in Stages 1-2 (left) and 3-4 (right) with ages 35 (solid line), 50 (dashed line) and 80 (dotted line), computed using the nonparametric estimator, $\hat{S}_{0,b}(t|x)$, with the bootstrap bandwidth, b_x^*.

Chapter 4

Covariate significance testing

4.1 Introduction

Significance testing is of primary importance in regression analysis, because the number of potential covariates to be included in the model can be large. In particular, in mixture cure models, variable selection is of outstanding interest, since the covariates having an effect on the survival of the uncured patients are not necessarily the same as those impacting the probability of cure. In this chapter, we propose a covariate significance test for the incidence based on the method by Delgado & González-Manteiga (2001), who introduced a test for selecting explanatory variables in nonparametric regression without censoring. The main advantage over other smoothed tests is that it only requires a smooth nonparametric estimator of the regression function depending on the explanatory variables which are significant under the null hypothesis. This feature is computationally convenient and solves, in part, the problem of the "curse of dimensionality" when selecting regressors in a nonparametric context.

Following Delgado & González-Manteiga (2001), let us denote $\mathcal{Y}_n = \{\mathcal{X}_i, i = 1, \ldots, n\}$, independent copies of $\mathcal{X} = (Y, \mathbf{W})$, which has probability space $(S, \mathcal{F}, P)$, where Y is unidimensional and $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ is $\mathbb{R}^q$-valued and $\mathbf{Z}$ is $\mathbb{R}^m$-valued. Consider the regression function $m(\cdot) = E(Y|\mathbf{X} = \cdot)$. The goal is to test $H_0 : E(Y|\mathbf{W}) = m(\mathbf{X})$, that is, if the conditional expectation of Y given $\mathbf{W}$ depends only on $\mathbf{X}$ but not on $\mathbf{Z}$. The test is based on:

$$T_n(\mathbf{w}) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \hat{f}_h(\mathbf{X}_i)(Y_i - \hat{m}_h(\mathbf{X}_i))I(\mathbf{W}_i \leq \mathbf{w}), \tag{4.1}$$

where $\hat{f}_h(\mathbf{X}_i)$ is the nonparametric estimator of the density function of $\mathbf{X}$, and

$\hat{m}_h(\mathbf{X}_i)$ is a nonparametric estimator of $m(\cdot) = E(Y|\mathbf{X} = \cdot)$.

The test statistic is a functional of $n^{1/2}T_n$, for instance, the Cramer-von Mises' statistic

$$C_n = \sum_{i=1}^{n} T_n(\mathbf{W}_i)^2,$$

or the Kolmogorov-Smirnov statistic

$$K_n = \sup_{\mathbf{w}} |n^{1/2}T_n(\mathbf{w})|.$$

In the particular case that X has one dimension, the bootstrap version corresponding to T_n, proposed by Delgado & González-Manteiga (2001), is

$$T_n^*(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_h(X_i)(Y_i^* - \hat{m}_h^*(X_i))I(\mathbf{W}_i \leq \mathbf{w}),$$

with $Y_i^* = \hat{m}_h(X_i) + \varepsilon_i^*$, where $\{\hat{\varepsilon}_i^* = V_i\hat{\varepsilon}_i, i = 1, \ldots, n\}$ is the bootstrap resample of the nonparametric residuals, with V_i obtained from a $N(0,1)$ and $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$, and considering $\hat{m}_h^*(X_i) = (nh^q\hat{f}_h(X_i))^{-1}\sum_{i=1}^{n} Y_i^*K_{ij}$, where $K_{ij} = K((X_i - X_j)/h)$. That is, T_n^* is the bootstrap version of T_n computed with the "wild resample" $\{(Y_i^*, X_i), i = 1, \ldots, n\}$. Furthermore, the bootstrap version of C_n and K_n are

$$C_n^* = \sum_{i=1}^{n} T_n^*(\mathbf{W}_i)^2 \text{ and } K_n^* = \sup_{\mathbf{w}} |n^{1/2}T_n^*(\mathbf{w})|.$$

When the number of simultaneous tests is large, like in genomics and other biology-related fields, the probability of getting a significant result simply due to chance is high. In order to deal with this problem, we consider the method by Benjamini & Hochberg (1995). It consists of controlling the expected proportion of falsely rejected hypotheses, that is, the false discovery rate (FDR). This error rate is equivalent to the familywise error rate (FWER) when all hypotheses are true, but is smaller otherwise. Therefore, in problems where the control of the FDR rather than that of the FWER is desired, there is potential for a gain in power.

Benjamini & Hochberg (1995) consider testing H_0^1, H_0^2, ..., H_0^m, based on the corresponding p-values: p_1, p_2, ..., p_m. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered p-values, and denote by $H_0^{(i)}$ the null hypothesis corresponding to $p_{(i)}$. The Bonferroni-type multiple-testing procedure consists of defining k as the largest i for which $p_{(i)} \leq \frac{i}{m}\alpha$, and then rejecting all $H_0^{(i)}$, for $i = 1, 2, \ldots, k$. If no such i exists,

reject no hypothesis. Note that this procedure controls the FDR at α, for independent test statistics and for any configuration of false null hypotheses.

When trying to use the FDR approach in practice, dependent test statistics are encountered more often than independent ones. Therefore, for the study of significance tests with high dimensional data, appart from the approach by Benjamini & Hochberg (1995), we also consider the conservative modification by Benjamini & Yekutieli (2001) to achieve the FWER control. When many of the tested hypotheses are rejected, indicating that many hypotheses are not true, the error from a single erroneous rejection is not always as crucial for drawing conclusions from the family tested, and the proportion of errors is controlled instead. The correction consists of comparing $p_{(i)}$ to $\frac{\alpha}{m-i+1}$ instead of $\frac{i}{m}\alpha$. Specifically, the method can be described as follows. Define:

$$k = \max\left\{ i : p_{(i)} \leq \frac{\alpha}{m-i+1} \right\}$$

and reject $H_0^{(1)}, \ldots, H_0^{(k)}$.

4.2 Significance tests for the incidence

In cancer studies it is interesting to test if a covariate has some influence on the cure rate or on the survival time of the susceptible patients. Müller & Van Keilegom (2018) propose a test statistic to assess whether the cure rate, $1-p$ (as a function of the covariates) satisfies a certain parametric model. However, to the best of our knowledge, no significance testing has been proposed yet for nonparametric cure models. We fill this important gap by proposing a covariate significance test for the incidence. The behavior of this method is assessed in some simulation studies.

Let us denote by $\mathbf{W} = (\mathbf{X}, \mathbf{Z}) = (X_1, \ldots, X_q, Z_1, \ldots, Z_m)$ the explanatory covariates. We would like to test if the cure probability, as a function of the covariate vector $\mathbf{W}$, only depends on $\mathbf{X}$, but not on $\mathbf{Z}$:

$$H_0 : E\left(\nu|\mathbf{X}, \mathbf{Z}\right) \equiv 1 - p\left(\mathbf{X}\right) \text{ vs. } H_1 : E\left(\nu|\mathbf{X}, \mathbf{Z}\right) \equiv 1 - p\left(\mathbf{X}, \mathbf{Z}\right),$$

where the function $p(\mathbf{X}, \mathbf{Z})$ depends on $\mathbf{Z}$ (i.e. is a function that depends not only on $\mathbf{X}$ but also on $\mathbf{Z}$).

Note that ν is not observed due to the censoring, since it is unknown if a censored individual will be eventually cured ($\nu = 1$) or not ($\nu = 0$). Let us define the variable

η as follows:

$$\eta = \frac{\nu(1 - I(\delta = 0, T \leq \tau(\mathbf{X})))}{1 - G\left(\tau(\mathbf{X})|\mathbf{X}\right)},$$

where $\tau(\mathbf{X})$ is a time beyond which a subject is considered cured and $\mathbf{X}$ is the covariate vector that influences the cure rate under H_0. The idea of the unknown variable, ν, is easy to understand, since it consists of identifying the cured individuals with weight equal to 1, and the uncured patients with weight equal to 0. Therefore, if we knew exactly which subjects are cured and which ones are not, we would use the corresponding weights. Note that the average of these weights is $E(\nu) = P(\nu = 1) = P(cure)$. Nevertheless, we only know that the individuals with $\delta = 1$ are not cured (then $\eta = 0$), and for the subset of patients with $\delta = 0$, there are both cured and uncured subjects. By definition, in the subset with $\delta = 0$, we set as cured those who, in addition, fulfill the condition $T > \tau(X)$ (then $\eta \neq 0$). But in this subset with $\delta = 0$ there are also some cured individuals with $T \leq \tau(X)$, and they are impossible to identify (we would assign, erroneously, weight $\eta = 0$). For this reason, we use weight $\eta > 1$ (specifically, $\eta = 1/(1 - G(\tau(X)))$ for the patients which we know that they are cured ($\delta = 0, T > \tau(X)$), in order to balance out the cured subjects in ($\delta = 0, T \leq \tau(X)$), with $\eta = 0$.

Furthermore, it is easy to check that $E(\eta|\mathbf{X}) = E(\nu|\mathbf{X})$ under the following assumption:

(A15) The distribution of $(C|\mathbf{X}, \nu = 0)$ equals that of $(C|\mathbf{X}, \nu = 1)$.

Specifically,

$$E(\eta|\mathbf{X}) \quad = \quad E(\eta|\mathbf{X}, \nu = 0)P(\nu = 0|\mathbf{X}) + E(\eta|\mathbf{X}, \nu = 1)P(\nu = 1|\mathbf{X}).$$

Since $\nu = 0$ implies $\eta = 0$, then $E(\eta|\mathbf{X}, \nu = 0) = 0$ and $E(\eta|\mathbf{X})$ reduces to

$$E(\eta|\mathbf{X}) \quad = \quad \frac{E(\nu(1 - I(\delta = 0, T \leq \tau(\mathbf{X})))|\mathbf{X}, \nu = 1)}{1 - G(\tau(\mathbf{X})|\mathbf{X})} P(\nu = 1|\mathbf{X}). \qquad (4.2)$$

Note that $\nu = 1$ implies $\delta = 0$, hence $T = C$ and the numerator in (4.2) is,

$$\begin{aligned} & E(\nu(1 - I(\delta = 0, T \leq \tau(\mathbf{X})))|\mathbf{X}, \nu = 1) \\ = \; & E(1 - I(C \leq \tau(\mathbf{X}))|\mathbf{X}, \nu = 1) = E(T > \tau(\mathbf{X})|\mathbf{X}, \nu = 1) \\ = \; & P(C > \tau(\mathbf{X})|\mathbf{X}, \nu = 1) = 1 - G(\tau(\mathbf{X})|\mathbf{X}, \nu = 1). \end{aligned}$$

Considering assumption (A15), C and ν are independent conditionally on $\mathbf{X}$, then

$$\begin{aligned} & 1 - G(\tau(\mathbf{X})|\mathbf{X}, \nu = 1) \\ = \; & P(C > \tau(\mathbf{X})|\mathbf{X}, \nu = 1)[P(\nu = 1|\mathbf{X}) + P(\nu = 0|\mathbf{X})] \\ = \; & P(C > \tau(\mathbf{X})|\mathbf{X}, \nu = 1)P(\nu = 1|\mathbf{X}) + P(C > \tau(\mathbf{X})|\mathbf{X}, \nu = 0)P(\nu = 0|\mathbf{X}) \\ = \; & P(C > \tau(\mathbf{X})|\mathbf{X}) = 1 - G(\tau(\mathbf{X})|\mathbf{X}). \end{aligned}$$

As a consequence, $E(\eta|\mathbf{X})$ in (4.2) is

$$E(\eta|\mathbf{X}) \;\; = \;\; \frac{1 - G(\tau(\mathbf{X})|\mathbf{X}, \nu = 1)}{1 - G(\tau(\mathbf{X})|\mathbf{X})} P(\nu = 1|\mathbf{X}) = P(\nu = 1|\mathbf{X}) = E(\nu|\mathbf{X}).$$

Note that if there is no covariate $\mathbf{X}$, in practice, we consider $\tau(\mathbf{x}_i) = T^1_{\max} = \max_{i:\delta_i=1} T_i$. In any other case, without loss of generality, we estimate $\tau(\mathbf{X})$ in the following way for a continuous univariate covariate X: using a bandwidth h_τ, we consider a subset of individuals j with $|x_j - x_i| < h_\tau$, and $\tau(x_i)$ will be estimated as the largest T_j with $\delta_j = 1$ in the subset. If there is no $\delta_j = 1$ in the subset, then $\tau(x_i)$ is equal to the available $\tau(x_l)$ for the nearest x_l to x_i. If there are several nearest values to determine x_l, then we estimate $\tau(x_i)$ as the mean of those. Preliminary studies suggested that a good bandwidth choice is:

$$h_\tau = \left(X_{(n)} - X_{(1)}\right)\ 0.25\, n^{-1/9}.$$

Moreover, it is assumed that C does not depend on the covariates, $\mathbf{X}$, and then $G\left(\tau(\mathbf{X}_i)|\mathbf{X}\right)$ is estimated by the product limit estimator, $\hat{G}(\tau(\mathbf{X}_i))$. This gives the following estimations for the η_i:

- $\delta_i = 1$ (failure) $\Rightarrow$ uncured ($\nu_i = 0$) $\Rightarrow \hat{\eta}_i = 0$.
- $\delta_i = 0$ (censored) and $T_i \leq \tau(\mathbf{X}_i) \Rightarrow \hat{\eta}_i = \frac{\nu_i(1-1)}{1-\hat{G}(\tau(\mathbf{X}_i))} = 0$.
- $\delta_i = 0$ (censored) and $T_i > \tau(\mathbf{X}_i) \Rightarrow$ cured ($\nu_i = 1$) $\Rightarrow \hat{\eta}_i = \frac{1}{1-\hat{G}(\tau(\mathbf{X}_i))}$.

For $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ the three cases in Table 4.1 can be considered.

Case	Dimension $\mathbf{W}$	Dimension $\mathbf{X}$	Dimension $\mathbf{Z}$
1: $H_0 : 1 - p(\mathbf{z}) = 1 - p$	m	0	m
2: $H_0 : 1 - p(x, \mathbf{z}) = 1 - p(x)$	$1 + m$	1	m
3: $H_0 : 1 - p(\mathbf{x}, \mathbf{z}) = 1 - p(\mathbf{x})$	$q + m$	q	m

Table 4.1: Different cases of covariate significant testing.

In Sections 4.3 and 4.4 we introduce Case 1 and Case 2, respectively. Moreover, in Section 4.5, Case 1 is extended to contexts with a large number of covariates. The behavior of these methods is assessed in a simulation study in Section 4.6. In Section 4.7, the test is applied to the colorectal cancer dataset and to the sarcomas dataset.

4.3 Case 1

First, we focus on the case with only one covariate, $\mathbf{W} = Z$. We study if the cure rate, as a function of Z, is a constant value versus if it depends on the covariate:

$$H_0 : E\left(\nu|Z\right) = 1 - p \;\text{ constant }\; \text{ vs } H_1 : E\left(\nu|Z\right) = 1 - p(Z), \tag{4.3}$$

where $p(Z)$ is not a constant function of Z. Our test will be based on the following observations: $\{(Z_i, \hat{\eta}_i), i = 1, \ldots, n\}$.

4.3.1 *Z* quantitative

Following Delgado & González-Manteiga (2001), the statistics we will propose when Z is a quantitative variable is based on the following process:

$$T_n(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j\right)\right) I\left(Z_i \leq z\right), \tag{4.4}$$

which is a weighted mean of the difference between the observations of η and the values of the conditional mean of η under the null hypothesis. Possible test statistics are the Cramér-von Mises (CvM) test, $C_n = \sum_{i=1}^{n} T_n^2(Z_i)$, or the Kolmogorov-Smirnov (KS) test, $K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(Z_i)|$. The test statistic null distribution is approximated by bootstrap, using an independent naive resampling. Specifically, the bootstrap procedure is the following:

1. For $i = 1, 2, \ldots, n$, obtain Z_i^* and $\hat{\eta}_i^*$ from $(Z_1, \ldots, Z_n)$ and $(\hat{\eta}_1, \ldots, \hat{\eta}_n)$ independently, by random resampling with replacement.

2. With the bootstrap resample, $\{(Z_i^*, \hat{\eta}_i^*), i = 1, \ldots, n\}$, obtain the bootstrap version of T_n:

$$T_n^*(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i^* - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j^*\right)\right) I(Z_i^* \leq z)$$

and the corresponding bootstrap version of the Cramér-von Mises and Kolmogorov-Smirnov statistics:

$$C_n^* = \sum_{i=1}^{n} T_n^*(Z_i^*)^2 \quad \text{and} \quad K_n^* = \max_{i=1,\ldots,n} |n^{1/2} T_n^*(Z_i^*)|.$$

3. Repeat B times Steps 1-2 in order to generate B values of C_n^* and K_n^*. Define the critical values d_C^* and d_K^* as the values which are in position $\lceil (1-\alpha)B \rceil$ in the sorted vectors.

4. Compare the value of the statistic, C_n (respectively, K_n), obtained with the original sample with d_C^* (respectively, d_K^*), and reject the null hypothesis if $C_n > d_C^*$ (respectively, $K_n > d_K^*$). In addition, the p-value can be calculated as the proportion of resamples for which the bootstrap statistic, C_n^* (K_n^*) is larger than the value of the statistic with the original sample, C_n (K_n).

Remark: Since Z_i^* and $\hat{\eta}_i^*$ are resampled independently in Step 1, the bootstrap resampling plan mimics H_0.

In the simulations in Section 4.6 we will repeat Steps 1-4 κ times. The power of the test is approximated as the proportion of rejections out of κ.

4.3.2 *Z* ordinal qualitative

The procedure is identical to when Z is quantitative. Specifically, departing from (4.3), the implementation of the algorithm is the same as for discrete Z.

4.3.3 *Z* non ordinal qualitative

In the case with only one qualitative non ordinal covariate, $\mathbf{W} = Z$, there is no natural way to order the values of Z from lowest to highest. This makes impossible to compute the indicator function in the test statistic (4.4). We propose to consider all the possible $k!$ combinations of the values of Z and compute $T_n(z)$ (and also C_n and K_n) for each "ordered" combination. For example, if Z can take 3 values, there will be $3! = 6$ possible values of $T_n(z)$ (from which we obtain C_n and K_n), each corresponding to every sorting of those values. Finally, we compute the maximum of C_n and K_n along all these possible permutations and compare it with the critical point obtained by bootstrap likewise.

A different approach consists of working with dummy variables. For example, considering a qualitative variable Z which can take $k = 3$ different values, we define D_1 (equal to 1 if the original qualitative variable is the first value of the covariate and 0 otherwise) and D_2 (equal to 1 if the original variable takes the second value and 0 otherwise). In general, it would be necessary to define $k-1$ dummy variables. Note that the vector of the dummy variables, (D_1, D_2), is equal to $(0, 0)$ if the value is the third one, $(0, 1)$ if it is the second value and $(1, 0)$ if it is the first value. The main advantage of this method is that we only need to compute $k - 1$ times the value of the statistic (one for each value of the covariate), whereas with the previous method, we have to compute the statistic $k!$ times (one per each permutation of the k values). Therefore, this approach is considerably less computationally expensive. Specifically, considering sample size $n = 100$, $\kappa = 5000$ trials and $B = 2000$ bootstrap resamples, the previous method takes 10050 seconds (see Table 4.13) and the approach using dummy variables takes 5657 seconds. On the other hand, by adressing the covariance testing using dummy variables, every new dummy variable have to be tested individually and it could be the case that the test leads to different conclusions for the dummy variables.

4.4 Case 2

In this case, $\mathbf{W} = (X, \mathbf{Z})$ has $m + 1$ dimension, with a one-dimensional covariate X and an m-dimensional covariate $\mathbf{Z}$. We study if the cure probability, as a function of $(X, \mathbf{Z})$, only depends on the covariate X, that is:

$$H_0 : E\left(\nu|X, \mathbf{Z}\right) = 1 - p(X), \quad \text{vs } H_1 : E\left(\nu|X, \mathbf{Z}\right) = 1 - p(X, \mathbf{Z}), \tag{4.5}$$

where $p(X, \mathbf{Z})$ depends on $\mathbf{Z}$ under the alternative hypothesis. To do this, we use the observations $\{(X_i, \mathbf{Z}_i, \hat{\eta}_i), i = 1, \ldots, n\}$. Note that in Case 2, we estimate $\tau(x_i)$ as mentioned in Section 4.2.

For the sake of simplicity, in this section we only considered a univariate Z and the most representative situations of Case 2, according to the distribution of X and Z. Nevertheless, in Section 4.6.2 we will show the results considering all the different possibilities: X and Z continuous, X continuous and Z discrete, X continuous and Z binary, X continuous and Z qualitative, X discrete and Z continuous, ..., X qualitative and Z qualitative.

4.4.1 $\boldsymbol{X}$ continuous

Based on the process T_n in (4.1), the statistic is defined as:

$$T_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_h(X_i)\left(\hat{\eta}_i - \hat{m}_h(X_i)\right) I\left(\mathbf{W}_i \leq \mathbf{w}\right), \tag{4.6}$$

where $\hat{f}_h$ is the estimated density function of the covariate X which depends on the bandwidth h, $\hat{m}_h$ is the nonparametric estimator of the regression function $m(x) = E\left(\hat{\eta}|X = x\right)$, obtained by the Nadaraya-Watson kernel method, and $\mathbf{W}_i \leq \mathbf{w}$ means that the inequality $\leq$ is checked for each component: $W_j \leq w_j,\ j = 1, \ldots, m+1$. Note that the process in (4.6) is a weighted mean of the difference between the observations of η and the values of the conditional mean of η under the null hypothesis.

Similarly to Case 1, we consider the Cramér-von Mises, $C_n = \sum_{i=1}^{n} T_n^2(\mathbf{W}_i)$ and the Kolmogorov-Smirnov, $K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(\mathbf{W}_i)|$ statistics. The test distribution under H_0 is approximated by bootstrap, considering the following procedure:

1. We fix the covariate $X_i^* = X_i$ and we obtain $\mathbf{Z}_i^*$ from $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ by random resampling with replacement, for $i = 1, 2, \ldots, n$. Furthermore, we compute $\hat{\eta}_{gi}^* = \hat{\hat{\eta}}_{gi} + v_i \hat{\varepsilon}_i$, where $\hat{\hat{\eta}}_{gi} = \hat{m}_g(X_i)$ is the Nadaraya-Watson kernel regression estimate computed with the original sample, v_i is obtained from a $N(0,1)$ and $\hat{\varepsilon}_i = \hat{\eta}_i - \hat{\hat{\eta}}_{gi}$ is the i-th residual. Note that the Nadaraya-Watson estimation of $m(X_i)$ is bounded between 0 and 1, i.e. $0 \leq \hat{\hat{\eta}}_{gi} \leq 1$, $i = 1, 2, \ldots, n$.

2. With the bootstrap resample, $\{(X_i, \mathbf{Z}_i^*, \hat{\eta}_i^*), i = 1, \ldots, n\}$, obtain the bootstrap version of T_n:

$$T_n^*(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_h(X_i)\left(\hat{\eta}_{gi}^* - \hat{m}_h(X_i)\right) I\left(\mathbf{W}_i^* \leq \mathbf{w}\right),$$

   and the corresponding bootstrap version of the Cramér-von Mises and Kolmogorov-Smirnov statistics:

$$C_n^* = \sum_{i=1}^{n} T_n^*(\mathbf{W}_i^*)^2 \text{ and } K_n^* = \max_{i=1,\ldots,n} |n^{1/2} T_n^*(\mathbf{W}_i^*)|.$$

3. Repeat Steps 1-2 B times in order to generate B values of C_n^* and K_n^*. Define the critical values d_C^* and d_K^* as the values which are in position $\lceil (1-\alpha)B \rceil$ in the sorted vectors.

4. Compare the value of the statistic, C_n (respectively, K_n) obtained with the original sample with d_C^* (respectively, d_K^*), and reject the null hypothesis if $C_n > d_C^*$ (respectively, $K_n > d_K^*$).

Steps 1-4 will be repeated κ times in the simulations in Section 4.6. The power of the test is approximated as the proportion of rejections out of κ.

Remark: To mimic H_0 in (4.5), the values $\hat{\eta}_{gi}^*$ defined in Step 1 do not depend on $\mathbf{Z}_i^*$ (just on X_i^*).

Note that in this procedure we need to select a bandwidth, h, and a pilot bandwidth, g. Although the results are quite insensitive to the value of the pilot bandwidth, preliminary studies showed that a good choice is $g = 2h$ (see related results below). Moreover, since we do not have a bandwidth selection method, we consider the values $h = Cn^{-1/3}$, where n is the sample size and $C = 10, 20, 40, 60$. In practice, we suggest to use any of the bandwidth selection methods for nonparametric tests proposed in the literature. There are two main approaches: Kulasekera & Wang (1997) focuses on power maximization under the alternative hypothesis, whereas Martínez-Camblor (2010); Martínez-Camblor & de Uña-Álvarez (2013) considers the idea of minimizing p-values. The two approaches are strongly related (see Martínez-Camblor & de Uña-Álvarez (2013)).

4.4.2 $\boldsymbol{X}$ categorical or discrete

For a categorical (dichotomous, qualitative) or discrete variable X, the estimated density, $\hat{f}_h(X_i)$, and the estimated regression function, $\hat{m}_h(X_i)$, in the test statistic in (4.6) are replaced by

$$\hat{\Pi}(X_i) = \frac{1}{n}\sum_{j=1}^{n} I(X_j = X_i)$$

and

$$\hat{m}(X_i) = \frac{\sum_{j=1}^{n} I(X_j = X_i)\hat{\eta}_j}{\sum_{j=1}^{n} I(X_j = X_i)} = \frac{\frac{1}{n}\sum_{j=1}^{n} I(X_j = X_i)\hat{\eta}_j}{\hat{\Pi}(X_i)},$$

respectively.

Similarly as in Case 1, for a qualitative variable in $\mathbf{W}$ with no intrinsic order in its values, the indicator function $I(\mathbf{W}_i \leq \mathbf{w})$ in the test statistic is computed for all the possible "ordered" permutations of the values of $\mathbf{W}$.

For example, for the case of $\mathbf{W} = (X, Z)$ qualitative, we consider that the values of the variables are $X_i \in \{a_1, \ldots, a_{l_x}\}$ and $Z_i \in \{b_1, \ldots, b_{l_z}\}$. The general test statistic is defined as:

$$T_n(x, z) = \frac{1}{n} \sum_{i=1}^{n} \hat{\Pi}(X_i) \left(\hat{\eta}_i - (1 - \hat{m}(X_i))\right) I\left(X_i \leq x\right) I\left(Z_i \leq z\right). \tag{4.7}$$

We will study the CvM, $C_n = \sum_{i=1}^{n} T_n^2(\mathbf{W}_i)$, and the KS, $K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(\mathbf{W}_i)|$ tests.

Since both variables are qualitative (which are neither numeric, nor ordinal) then there is no a natural order among them. Therefore, we obtain all the possible ordered values of $\mathbf{W}$ (for instance, if $m = 2$ and each variable can take 3 different values, then we will have $(3!)(3!) = 36$ different orders of these values), we sort them and we consider the maximum values of the tests C_n and K_n evaluated in $\mathbf{W}$. Finally, we compare them with their corresponding critical point obtained by bootstrap likewise.

Specifically, let us denote by $S_n(a_j, b_k)$:

$$S_n(a_j, b_k) \;=\; \frac{1}{n} \sum_{i=1}^{n} \hat{\Pi}(X_i) \left(\hat{\eta}_i - (1 - \hat{m}(X_i))\right) I\left(X_i = a_j\right) I\left(Z_i = b_k\right).$$

In order to compute the statistic $T_n(x, z)$ in (4.7), we take into account that $T_n(a_1, b_1) = S_n(a_1, b_1)$. In general:

$$T_n(a_j, b_k) \;=\; \sum_{r=1}^{j} \sum_{s=1}^{k} S_n(a_r, b_s),\ j = 1, \ldots, l_x;\ k = 1, \ldots, l_z,$$

for each "ordered" combination of the values of X and Z: $a_1 < a_2 < \cdots < a_{l_x}$; $b_1 < b_2 < \cdots < b_{l_z}$.

4.5 Case 1 with high dimensional covariate vector $\boldsymbol{Z}$

This procedure can be applied to Case 1 or Case 2, provided that $\mathbf{Z}$ is a m-dimensional covariate vector. In this thesis, only Case 1 is considered.

Departing from (4.3), we consider the case with an m-dimensional covariate $\mathbf{Z}$. The method consists of considering m hypotheses in Equation (4.3) to be tested independently, $H_0^1, \ldots, H_0^m$. Depending on the type (quantitative or qualitative) of each covariate, Z_i, we use its corresponding T_n defined in Sections 4.3.1-4.3.3.

In order to control the false discovery rate (FDR), we depart from the approach by Benjamini & Hochberg (1995) to problems of multiple significance testing. Furthermore, to achieve the family wise error rate (FWE) control, we consider the method by Benjamini & Yekutieli (2001). The procedure to test $H_0^1, \ldots, H_0^m$ with level α is the following:

1. Sort the p-values obtained when testing H_0^j, p_j, for $j = 1, 2, \ldots, m$: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Define: $k = \max\{i : p_{(i)} \leq \frac{\alpha}{m-i+1}\}$.
3. Reject $H_0^{(1)}, \ldots, H_0^{(k)}$, that is, the hypotheses which lead to the k smallest p-values.

Note that if there is no i which fulfills the condition $p_{(i)} \leq \frac{\alpha}{m-i+1}$ in Step 2, then no hypotheses will be rejected.

4.6 Simulation studies

The purpose of the simulation studies is to assess the practical behavior of the proposed significance tests. We work with three different sample sizes: $n = 50$, $n = 100$ and $n = 200$. A total of $\kappa = 5000$ trials and $B = 2000$ bootstrap resamples are drawn.

4.6.1 Case 1

Under the null hypothesis, $H_0 : \ E(\nu|Z) = 1 - p$, we consider four different constant values for the incidence: $1 - p = 0.7$, $1 - p = 0.5$, $1 - p = 0.3$ and $1 - p = 0.2$. Under the alternative hypothesis, $H_1 : \ E(\nu|Z) = 1 - p(Z)$, and when Z is a continuous covariate, we study the same two models as in Section 2.5, but replacing the covariate X by Z. If Z is discrete, binary or qualitative, we only consider Model 1 in order to generate the data.

Model 1 The incidence is $1 - p(z)$, where

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)}, \tag{4.8}$$

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$, and the latency is a function which almost fulfills the proportional hazard model and that it has been truncated

$$S_0(t|z) = \begin{cases} \frac{\exp(-\lambda(z)t)-\exp(-\lambda(z)\tau_0)}{1-\exp(-\lambda(z)\tau_0)} & \text{if } t \leq \tau_0 \\ 0 & \text{if } t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda(z) = \exp((z+20)/40)$. The percentage of censored data is 54% and of cured data is 47%.

Model 2 The probability of uncure is:

$$p(z) = \frac{\exp\left(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3\right)}{1 + \exp\left(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3\right)}, \tag{4.9}$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$ and $\beta_3 = 0.0020$ and the survival function of the susceptible population is

$$S_0(t|z) = \frac{1}{2}\left(\exp(-\alpha(z)t^5) + \exp(-100t^5)\right),$$

with

$$\alpha(z) = \frac{1}{5}\exp((z+20)/40).$$

The percentages of censored and cured data are 62% and 53%, respectively.

Preliminary studies for the selection of the bandwidths

For sample size $n = 100$, $B = 1000$ bootstrap resamples and $\kappa = 1000$ trials, we draw the significance tests for Case 2, with X and Z continuous. We work with the same Model 1 and Model 2 defined in Section 4.6.2. The significance level considered is $\alpha = 0.05$.

For these preliminary simulations, we considered 3 studies. In the first one, we set the pilot bandwidth, g, equal to the bandwidth h. The results are shown in Table 4.2. In the second study (see Table 4.3), we fix the bandwidth $h = 8.62$, since it seems to give good results, and we work with different pilot bandwidths. Finally, in Table 4.4, we use $g = 2h$, for a grid of bandwidths h.

According to the results in Table 4.3, it seems that the effect of the bandwidth g is negligible for a fixed bandwidth h. From Tables 4.2 and 4.4, we can appreciate that $g = 2h$ could be an appropriate choice, since the asymptotic pilot bandwidth

	H_0				H_1			
	Model 1		Model 2		Model 1		Model 2	
$g = h$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
4.31	0.039	0.044	0.040	0.048	0.428	0.457	0.665	0.571
8.62	0.033	0.041	0.054	0.048	0.427	0.448	0.651	0.579
12.93	0.039	0.066	0.078	0.083	0.390	0.412	0.632	0.576
17.24	0.070	0.124	0.097	0.102	0.335	0.369	0.614	0.571
34.47	0.712	0.903	0.125	0.147	0.165	0.213	0.615	0.586

Table 4.2: Size of the test under the null hypothesis (on the left) and power of the test under the alternative hypothesis (on the right), for Case 2 with X and Z continuous with distribution $U(-20, 20)$ for Model 1 and Model 2.

	H_0				H_1			
	Model 1		Model 2		Model 1		Model 2	
g	CvM	KS	CvM	KS	CvM	KS	CvM	KS
4.31	0.033	0.043	0.048	0.047	0.439	0.450	0.660	0.591
8.62	0.034	0.042	0.050	0.049	0.431	0.448	0.660	0.588
12.93	0.033	0.041	0.054	0.048	0.427	0.448	0.651	0.579
17.24	0.034	0.048	0.051	0.041	0.424	0.447	0.642	0.567
34.47	0.038	0.059	0.045	0.035	0.423	0.444	0.642	0.556

Table 4.3: Size of the test under the null hypothesis (on the left) and power of the test under the alternative hypothesis (on the right), for Case 2 with X and Z continuous with distribution $U(-20, 20)$ for Model 1 and Model 2, and considering a fixed value for the bandwidth, $h = 8.62$.

	H_0				H_1			
	Model 1		Model 2		Model 1		Model 2	
h	CvM	KS	CvM	KS	CvM	KS	CvM	KS
2.15	0.044	0.037	0.043	0.046	0.280	0.331	0.575	0.492
4.31	0.034	0.036	0.044	0.050	0.427	0.442	0.647	0.549
8.62	0.038	0.059	0.045	0.035	0.423	0.444	0.642	0.556
12.93	0.054	0.087	0.067	0.068	0.383	0.405	0.632	0.565
17.24	0.090	0.152	0.099	0.104	0.335	0.363	0.613	0.563

Table 4.4: Size of the test under the null hypothesis (on the left) and power of the test under the alternative hypothesis (on the right), for Case 2 with X and Z continuous with distribution $U(-20, 20)$ for Model 1 and Model 2, and considering $g = 2h$.

is larger than h. Regarding the value of h, we follow the approach by Delgado & González-Manteiga (2001). They choose a bandwidth of the form $h = Cn^{-1/3m}$, for different values of C, where m is the dimension of the covariate vector Z that is

being tested. Note that in our case, $m = 1$. Delgado & González-Manteiga (2001) explains that this bandwidth is compatible with assumptions $(A4)$ and $(A4')$ in their paper. Therefore, we decided to work with $h = Cn^{-1/3}$, where n is the sample size, and $C = 10, 20, 40, 60$.

Z continuous

Under H_1, we considered Models 1 and 2 for a continuous covariate $Z \sim U(-20, 20)$. Figure 4.1 shows the cure probability, $1 - p(z)$, under H_0 and H_1. The results in Table 4.5 (Table 4.6) were obtained under the null (alternative) hypothesis. It is noteworthy that, under H_0, the size of the test is very similar to the significance level, α=0.05, for the different constant values for p. Furthermore, under H_1, the power of the test is very close (or even equal) to 1.

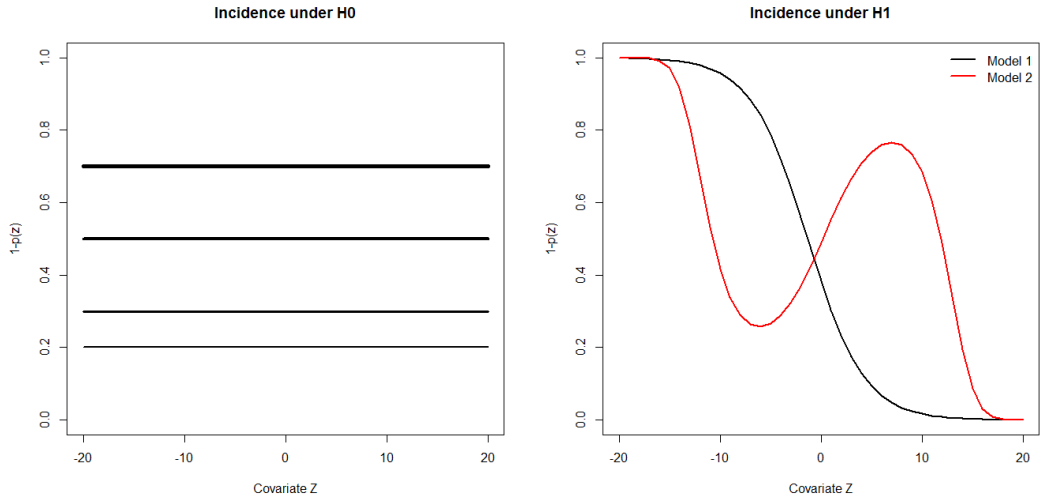

Figure 4.1: Cure probability, $1 - p(z)$, under the null (left) and the alternative (right) hypotheses, with Z continuous, for Case 1 in Model 1 (black) and Model 2 (red).

Z discrete

We work with a discrete covariate Z with 3 ordered values $b_1 < b_2 < b_3$. The values $p(b_i),\ i = 1, 2, 3$ under H_0 and H_1 are given in Tables 4.7 and 4.8, respectively. Under H_1, there are two scenarios, one with similar values for $p(b_i),\ i = 1, 2, 3$, and another one with more variable values of $p(\cdot)$ as a function of Z. We consider two situations according to the probability mass function of Z given by $\Pi_z(b_i) = P(Z = b_i)$: in the first one, $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (1/3, 1/3, 1/3)$ and in the second one, $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (3/5, 1/5, 1/5)$, where the value b_1, with the lowest value

		Model 1		Model 2	
n	p	CvM	KS	CvM	KS
50	0.3	0.0484	0.0600	0.0550	0.0636
	0.5	0.0500	0.0564	0.0490	0.0582
	0.7	0.0446	0.0472	0.0510	0.0504
	0.8	0.0418	0.0360	0.0400	0.0412
100	0.3	0.0608	0.0636	0.0544	0.0584
	0.5	0.0494	0.0562	0.0488	0.0586
	0.7	0.0444	0.0520	0.0474	0.0468
	0.8	0.0414	0.0396	0.0484	0.0470
200	0.3	0.0490	0.0530	0.0528	0.0534
	0.5	0.0552	0.0616	0.0540	0.0572
	0.7	0.0498	0.0508	0.0516	0.0514
	0.8	0.0446	0.0432	0.0466	0.0480

Table 4.5: Size of the test for Case 1 with Z continuous with distribution $U(-20, 20)$ under the null hypothesis.

	Model 1		Model 2	
n	CvM	KS	CvM	KS
50	0.9890	0.9862	0.4200	0.4148
100	0.9994	0.9992	0.7330	0.7402
200	1	1	0.9670	0.9746

Table 4.6: Power of the test for Case 1 with Z continuous with distribution $U(-20, 20)$ under the alternative hypothesis.

of $p(\cdot)$, is given the highest probability.

Even when Z is not a continuous covariate, the data are also generated from Model 1 in the following way: the values of Z, $\{b_1, b_2, b_3\}$, will be those that, in Model 1, yield fixed probabilities $p(b_1) = p(b_2) = p(b_3) \in \{0.2, 0.3, 0.5, 0.7\}$, under H_0 (see Table 4.7), and $p(b_1) \in \{0.1, 0.3\}$, $p(b_2) = 0.5$ and $p(b_3) \in \{0.7, 0.9\}$ (see Table 4.8) under H_1. This procedure is also applicable in Case 2 when X and/or Z are discrete, for the uncure probability $p(x, z)$ in (4.10).

The results in Table 4.7 (Table 4.8) were obtained under the null (alternative) hypothesis. We can see that, for the different values of p, the results under the null hypothesis are very similar to the significance level, $\alpha = 0.05$, regardless the value of $p(b_i)$, $i = 1, 2, 3$. Moreover, under the alternative hypothesis, the power of the test is very close to 1 for large sample sizes and when the values of the incidence, as

a function of z, are more distant, that is, $p(b_1) = 0.1$, $p(b_2) = 0.5$ and $p(b_3) = 0.9$ and the values of Z are equiprobable.

n	p	CvM	KS
50	0.2	0.0472	0.0530
	0.3	0.0472	0.0536
	0.5	0.0474	0.0534
	0.7	0.0488	0.0436
100	0.2	0.0502	0.0526
	0.3	0.0596	0.0610
	0.5	0.0490	0.0534
	0.7	0.0498	0.0494
200	0.2	0.0526	0.0510
	0.3	0.0504	0.0528
	0.5	0.0518	0.0512
	0.7	0.0514	0.0536

Table 4.7: Size of the test for Case 1 with Z discrete with probability mass function $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (1/3, 1/3, 1/3)$ under the null hypothesis, $H_0 : E(\nu|Z) = 1 - p$. Note: the distribution of Z does not have any influence on the results.

n	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	$p(b_1)$	$p(b_2)$	$p(b_3)$	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.3	0.5	0.7	0.2968	0.2764
	$(3/5, 1/5, 1/5)$	0.3	0.5	0.7	0.2690	0.2528
	$(1/3, 1/3, 1/3)$	0.1	0.5	0.9	0.8574	0.8214
	$(3/5, 1/5, 1/5)$	0.1	0.5	0.9	0.8076	0.7880
100	$(1/3, 1/3, 1/3)$	0.3	0.5	0.7	0.4800	0.4358
	$(3/5, 1/5, 1/5)$	0.3	0.5	0.7	0.4264	0.4090
	$(1/3, 1/3, 1/3)$	0.1	0.5	0.9	0.9804	0.9714
	$(3/5, 1/5, 1/5)$	0.1	0.5	0.9	0.9504	0.9440
200	$(1/3, 1/3, 1/3)$	0.3	0.5	0.7	0.6970	0.6556
	$(3/5, 1/5, 1/5)$	0.3	0.5	0.7	0.6376	0.6218
	$(1/3, 1/3, 1/3)$	0.1	0.5	0.9	0.9982	0.9966
	$(3/5, 1/5, 1/5)$	0.1	0.5	0.9	0.9964	0.9956

Table 4.8: Power of the test for Case 1 with Z discrete, with values $\{b_1, b_2, b_3\}$, under the alternative hypothesis.

Z binary

Let Z be a binary variable with values $b_1 = 0$, $b_2 = 1$. We consider 3 situations depending on the probability mass function, $(\Pi_z(b_1), \Pi_z(b_2))$. In the first one, $(9/10, 1/10)$, in the second one, $(7/10, 3/10)$ and in the third one, $(1/2, 1/2)$. The

binary covariate Z was generated from Model 1 as explained for a discrete covariate. Let $p(0) = p(b_1)$ and $p(1) = p(b_2)$. Under H_0, $p(0) = p(1) = p \in \{0.2, 0.3, 0.5, 0.7\}$, whereas under H_1, $p(0)$,$p(1) \in \{0.2, 0.3, 0.5, 0.7\}$, with $p(0) > p(1)$.

Table 4.9 shows the results obtained under the null hypothesis. Similarly as it happened in the case where Z is discrete, for the different values of p, the size of the test is very close to the significance level, $\alpha = 0.05$. Furthermore, Table 4.10 contains the results under the alternative hypothesis. As it was expected, the lowest power is obtained when the values of $p(0)$ and $p(1)$ are similar. It also happens when $p(0)$ and $p(1)$ are small, since in such a case the probability of cure is high, and therefore the censoring percentage is large (see, for instance, the case $p(0) = 0.3$ and $p(1) = 0.2$ in Table 4.10). On the contrary, if $p(0) = 0.7$ and $p(1) = 0.2$, then the power of the test is very close to 1 for sample sizes $n = 200$ and when $\Pi_z(1) = 3/10$ or $\Pi_z(1) = 1/2$.

p	n	CvM	KS
0.2	50	0.0536	0.0534
	100	0.0528	0.0536
	200	0.0512	0.0516
0.3	50	0.0538	0.0542
	100	0.0578	0.0582
	200	0.0514	0.0518
0.5	50	0.0530	0.0528
	100	0.0590	0.0588
	200	0.0510	0.0512
0.7	50	0.0506	0.0504
	100	0.0538	0.0544
	200	0.0536	0.0538

Table 4.9: Size of the test for Case 1 with Z binary under the null hypothesis.

Z qualitative

A qualitative covariate Z with three possible values $\{b_1, b_2, b_3\}$ was considered. Two situations, according to the probability mass function given by $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$, were studied. The observations were simulated from Model 1, finding the numerical values (b_1', b_2', b_3') that yield $p(b_1') = 0.5$, $p(b_2') = 0.2$, and $p(b_3') = 0.7$, that is, b_i', $i = 1, 2, 3$ denotes the value z at which $p(z)$ was evaluated to get $p(b_i)$.

Note that the power of the test will be higher when the values b_2 and b_3 are more frequent in the sample, since $p(b_2) = 0.2$ and $p(b_3) = 0.7$ correspond to

			CvM $\Pi_z(1)$			KS $\Pi_z(1)$		
$p(0)$	$p(1)$	n	1/10	3/10	1/2	1/10	3/10	1/2
0.7	0.5	50	0.1460	0.1722	0.1576	0.1418	0.1686	0.1586
		100	0.1760	0.2462	0.2458	0.1730	0.2440	0.2460
		200	0.2354	0.3658	0.3806	0.2342	0.3626	0.3816
0.7	0.3	50	0.2766	0.4350	0.4548	0.2682	0.4288	0.4586
		100	0.3964	0.6184	0.6658	0.3924	0.6164	0.6658
		200	0.5692	0.8260	0.8712	0.5676	0.8250	0.8718
0.7	0.2	50	0.3838	0.5830	0.6160	0.3762	0.5746	0.6174
		100	0.5158	0.7836	0.8316	0.5112	0.7828	0.8322
		200	0.6968	0.9332	0.9614	0.6946	0.9333	0.9616
0.5	0.3	50	0.1114	0.1478	0.1418	0.1068	0.1426	0.1408
		100	0.1302	0.1998	0.2096	0.1280	0.2002	0.2118
		200	0.1756	0.2910	0.3250	0.1740	0.2912	0.3258
0.5	0.2	50	0.1558	0.2432	0.2674	0.1508	0.2382	0.2664
		100	0.2076	0.3548	0.3936	0.2058	0.3522	0.3956
		200	0.3052	0.5308	0.5936	0.3032	0.5294	0.5954
0.3	0.2	50	0.0670	0.0806	0.0818	0.0626	0.0780	0.0814
		100	0.0722	0.0900	0.0886	0.0706	0.0900	0.0894
		200	0.0834	0.1018	0.1106	0.0828	0.1004	0.1114

Table 4.10: Power of the test for Case 1 with Z binary under the alternative hypothesis.

the extreme values of $p(\cdot)$ as a function of Z. On the contrary, the power will be lower if b_1 is more frequent than b_2 or b_3, because $p(b_1) = 0.5$ is an intermediate value between 0.2 and 0.7. Therefore, we expect to obtain low power when $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (3/5, 1/5, 1/5)$, whereas with probability mass function $(1/3, 1/3, 1/3)$ the power of the test will be higher.

The results in Table 4.11, which were obtained under the null hypothesis, are very similar to the significance level, $\alpha = 0.05$. Regarding the alternative hypothesis (see Table 4.12), the power of the test is higher for large sample sizes and when the probability mass function of Z is equal in probability, as expected.

Computational summary in Case 1

All the simulation studies were coded in R language. The procedures were drawn in the computers of the Department of Mathematics, at the Faculty of Computer Sciences in the University of A Coruña. The computational times taken for each study are shown in Table 4.13.

n	p	CvM	KS
50	0.2	0.0512	0.0526
	0.5	0.0494	0.0520
100	0.2	0.0544	0.0538
	0.5	0.0488	0.0532
200	0.2	0.0486	0.0500
	0.5	0.0456	0.0516

Table 4.11: Size of the test for Case 1 with Z qualitative under the null hypothesis with $\Pi_z(b_1) = \Pi_z(b_2) = \Pi_z(b_3) = 1/3$. Note: the probability mass function of Z does not have any influence on the results.

n	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	$p(b_1)$	$p(b_2)$	$p(b_3)$	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.3402	0.3408
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.1680	0.1606
100	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.5588	0.5600
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.2748	0.2690
200	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.8028	0.7994
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.4552	0.4448

Table 4.12: Power of the test for Case 1 with Z qualitative under the alternative hypothesis.

Type of Z	Scenario	Computational time
Continuous	Model 1	270610 sec (75.20 h)
	Model 2	269420 sec (74.84 h)
Discrete	Model 1	970 sec (16.17 min)
Binary	Model 1	3560 sec (59.33 min)
Qualitative	Model 1	10050 sec (2.79 h)

Table 4.13: Computational times for simulations in Case 1, considering sample size $n = 100$, $\kappa = 5000$ trials and $B = 2000$ bootstrap resamples.

4.6.2 Case 2

In this case, $\mathbf{W} = (X, \mathbf{Z})$ has dimension $m + 1$, with a one-dimensional covariate X and a m-dimensional covariate $\mathbf{Z}$. For the sake of simplicity, in this simulation study we suppose that Z is also one-dimensional.

We consider two different scenarios, Model 1 and Model 2, when X and Z are continuous. In any other case, we only consider Model 1.

Model 1 Under the null hypothesis, $H_0 : E(\nu|X, Z) = 1 - p(X)$, the probability of uncure $p(x)$ is that in (4.8), corresponding to Model 1 in Section 4.6.1. Under

the alternative, the incidence is $1 - p(x, z)$, where

$$p(x, z) = \frac{\exp(\beta_0 + \beta_1 x(1 + \beta_2 z))}{1 + \exp(\beta_0 + \beta_1 x(1 + \beta_2 z))}, \tag{4.10}$$

with $\beta_0 = 0.476$, $\beta_1 = 0.358$ and $\beta_2 = 0.225$, and the latency is:

$$S_0(t|x, z) = \begin{cases} \frac{\exp(-\lambda(x,z)t) - \exp(-\lambda(x,z)\tau_0)}{1 - \exp(-\lambda(x,z)\tau_0)} & \text{if } t \leq \tau_0 \\ 0 & \text{if } t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda(x, z) = \exp((x + z + 20)/40)$. Figure 4.2 (top) shows the incidence function in Model 1.

Model 2 The probability of uncure, $p(x)$, under $H_0 : E(\nu|X, Z) = 1 - p(X)$, is that in (4.9), corresponding to Model 2 in Section 4.6.1. Under the alternative, the probability of uncure is:

$$p(x, z) = \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3(1 + \beta_4 z)\right)}{1 + \exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3(1 + \beta_4 z)\right)},$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$, $\beta_3 = 0.0020$ and $\beta_4 = 0.5$, and the survival function of the susceptible population is

$$S_0(t|x, z) = \frac{1}{2}\left(\exp(-\alpha(x, z)t^5) + \exp(-100t^5)\right),$$

with

$$\alpha(x, z) = \frac{1}{5}\exp((x + z + 20)/40).$$

Figure 4.2 (bottom) shows the incidence function in Model 2.

Remark: Similarly as in Section 4.6.1, we study different situations depending on the type of X and Z. If both X and Z are continuous, the data are simulated from Models 1 and 2. If X and/or Z are not continuous, the data follow only Model 1, with functions $p(\cdot)$, $S_0(t|\cdot)$ and $G(\cdot)$ defined there. For binary and qualitative covariates, the computation of the probability of cure deserves special attention, since that probability can not be obtained directly evaluating $p(\cdot)$ in (4.10) in the values a_i (and/or b_j) of X (and/or Z), because they are not numerical. Therefore, let a'_i (and b'_j) be the numerical values associated to a_i for X (and to b_j for Z), in the sense that the distributions of Y and C conditioned on a_i (and/or b_j) are the same as the conditional distributions given a'_i (and/or b'_j). Under the alternative hypothesis, the probability of cure derives from evaluating the function $p(\cdot)$ in (4.10) not in a_i and/or b_j, but in the corresponding numerical values, a'_i and/or b'_j. Under the null

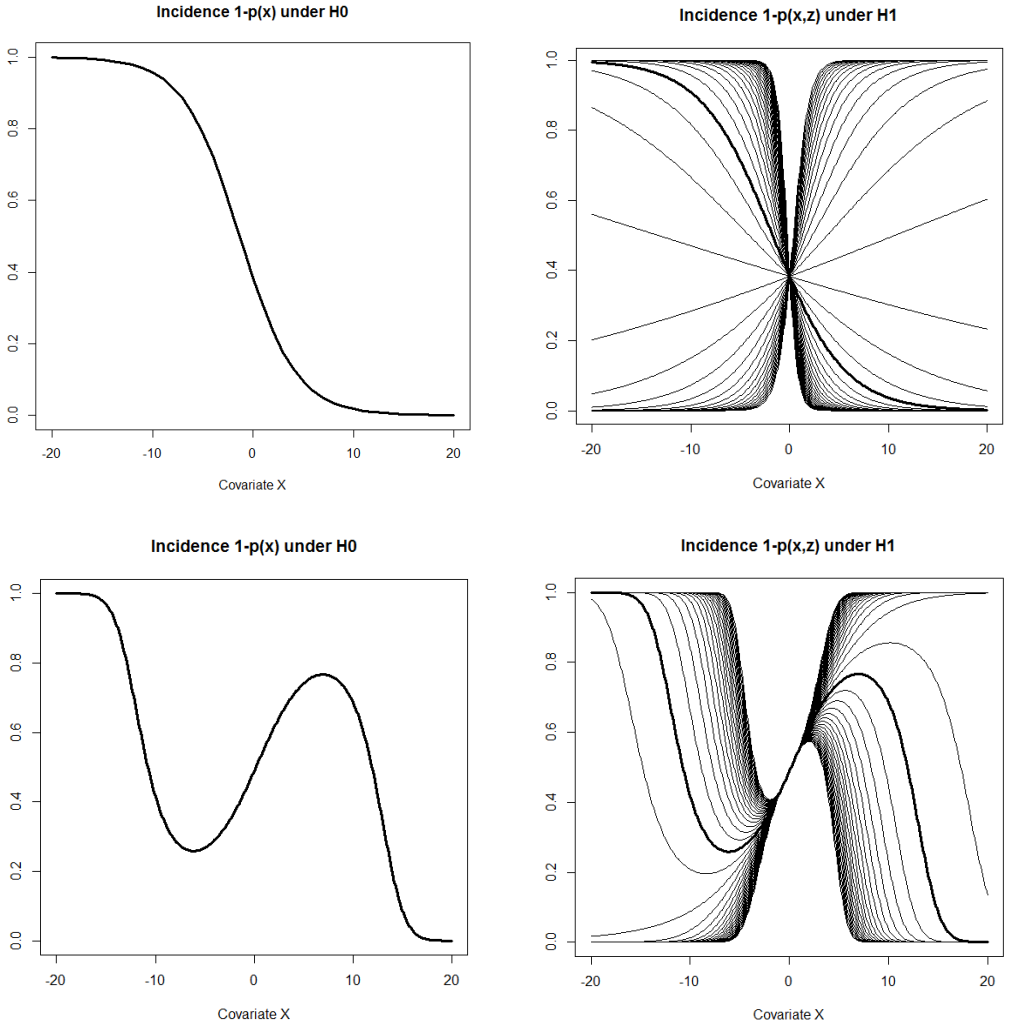

Figure 4.2: Probability of cure for Case 2 in Model 1 (top) and Model 2 (bottom), under the null (left) and under the alternative (right) hypothesis, for different values of z.

hypothesis, let x be the value of the covariate X, then $p(x, b_1) = p(x, b_2) = p(x, b_3)$ is given by function $p(\cdot)$ in (4.10) evaluated in $(x, \bar{b})$, with $\bar{b}$ suitably selected to get the desired constant probability of cure.

Furthermore, under H_0, if the covariates are not continuous, we define the distribution function of the variable Y:

$$F(y|x = 0) = F_1^0(y|x = a_1') \text{ and } F(y|x = 1) = F_1^0(y|x = a_2')$$

and the distribution function of the censoring variable,

$$G(y|x=0) = G_1^0(y|x=a_1') \text{ and } G(y|x=1) = G_1^0(y|x=a_2'),$$

where F_1^0 and G_1^0 are the conditional distribution functions for Model 1 under H_0. Analogously, under the alternative hypothesis:

$$F(y|x=0, z=b_i) = F_1^1(y|x=a_1', z=b_i), \text{ with } i=1,2,3$$

and

$$G(y|x=1, z=b_i) = G_1^1(y|x=a_2', z=b_i), \text{ with } i=1,2,3,$$

where F_1^1 and G_1^1 are the conditional distribution functions for Model 1 under H_1. Note that the distribution function of the variable Y, $F(y|x)$, and the distribution function of the censoring variable, $G(y|x)$, are the corresponding distribution functions of Model 1 considered in the simulation studies for Case 2.

$\boldsymbol{X}$ continuous, $\boldsymbol{Z}$ continuous

We consider two continuous covariates X, Z with distribution $U(-20, 20)$ and Models 1 and 2 (see Figure 4.2). Since we do not have a bandwidth selection method for h in (4.6), we consider the grid of bandwidths $h = Cn^{-1/3}$, where $C = 10, 20, 40$ and 60. Our numerical experience shows that $g = 2h$ is a good choice for the pilot bandwidth. Under H_0, the results are very similar to the significance level, $\alpha = 0.05$, except for very large bandwidths (see Table 4.14). Furthermore, under H_1, the power of the test is considerably high (see Table 4.15).

$\boldsymbol{X}$ continuous, $\boldsymbol{Z}$ discrete

We consider a continuous covariate X with distribution $U(-20, 20)$ and a discrete covariate Z with values $(b_1, b_2, b_3) = (-5.2019, -3.6964, -1.3296)$. The cure probability under H_0 is shown in Figure 4.3. Under the alternative hypothesis, we study two different scenarios: values $(b_1, b_2, b_3) = (-3.6964, -1.3296, 1.0371)$ and values $(b_1, b_2, b_3) = (-7.4671, -1.3296, 4.8079)$, with two probability mass functions for each scenario, $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$. Figure 4.4 shows the probability of cure under H_1, $1-p(x,z)$, for the first and the second situations with the different values of b_i. Note that higher power is expected in the second scenario, since the incidence for these specific values of b_i is more variable.

Table 4.16 shows the results under the null hypothesis, which are very similar to the significance level, $\alpha = 0.05$, except in the situations where the bandwidth is

		Model 1		Model 2	
n	h	CvM	KS	CvM	KS
50	2.71	0.0504	0.0524	0.0494	0.0500
	5.43	0.0428	0.0436	0.0464	0.0444
	10.86	0.0372	0.0486	0.0448	0.0438
	16.28	0.0442	0.0782	0.0692	0.0734
100	2.15	0.0494	0.0514	0.0526	0.0520
	4.31	0.0476	0.0556	0.0544	0.0542
	8.62	0.0580	0.0708	0.0514	0.0490
	12.93	0.0590	0.0988	0.0700	0.0782
200	1.71	0.0436	0.0412	0.0510	0.0496
	3.42	0.0420	0.0448	0.0558	0.0542
	6.84	0.0522	0.0606	0.0508	0.0494
	10.26	0.0590	0.0948	0.0614	0.0708

Table 4.14: Size of the test for Case 2 with X and Z continuous with distribution $U(-20, 20)$, under the null hypothesis.

		Model 1		Model 2	
n	h	CvM	KS	CvM	KS
50	2.71	0.2016	0.2182	0.2632	0.2466
	5.43	0.2596	0.2698	0.3244	0.3074
	10.86	0.2610	0.2696	0.3210	0.3150
	16.28	0.2278	0.2442	0.3084	0.3152
100	2.15	0.3906	0.4736	0.6228	0.5662
	4.31	0.5132	0.5556	0.6918	0.6128
	8.62	0.5160	0.5564	0.6864	0.6320
	12.93	0.4718	0.5206	0.6746	0.6438
200	1.71	0.7428	0.8554	0.9730	0.9364
	3.42	0.8492	0.9156	0.9832	0.9486
	6.84	0.8582	0.9114	0.9830	0.9568
	10.26	0.8358	0.8914	0.9852	0.9576

Table 4.15: Power of the test for Case 2 with X and Z continuous with distribution $U(-20, 20)$, under the alternative hypothesis.

very large. On the contrary as in Case 2, when X and Z are continuous covariates, the best choice of the bandwidth is $h = 5.43$ for $n = 50$, $h = 4.31$ for $n = 100$ and $h = 3.42$ for $n = 200$, that is, $h = Cn^{-1/3}$ with $C = 20$.

Regarding the alternative hypothesis, Table 4.17 contains the results for both considered scenarios. As expected, the power of the test is higher in the second

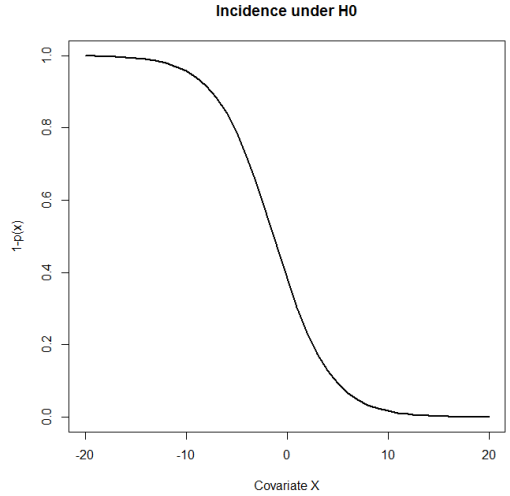

Figure 4.3: Probability of cure, $1 - p(x, b_1) = 1 - p(x, b_2) = 1 - p(x, b_3)$, under the null hypothesis, with X continuous and Z discrete, for Case 2 in Model 1.

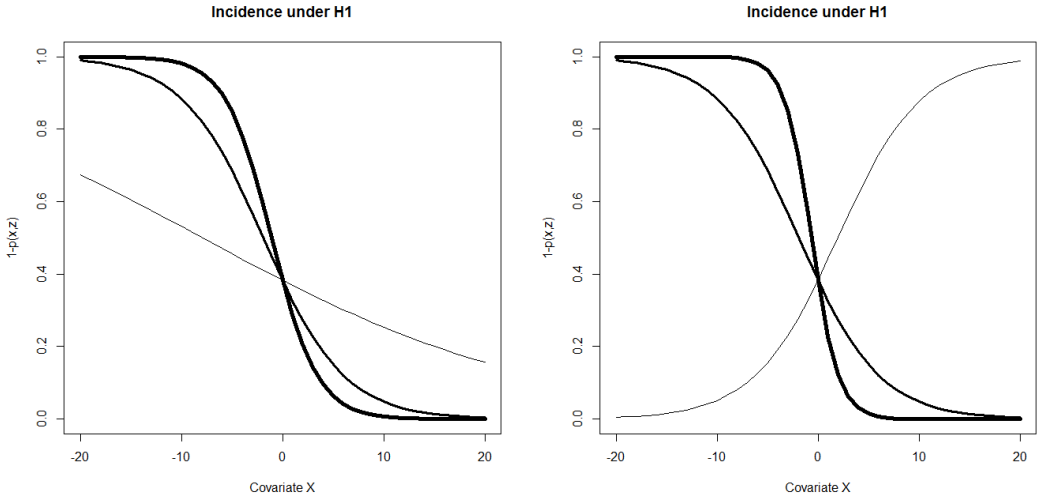

Figure 4.4: Probability of cure, $1 - p(x, b_i)$, under the alternative hypothesis, with X continuous and Z discrete, for Case 2 in Model 1. On the left, we consider $b_1 = -3.6964$ (thin line), $b_2 = -1.3296$ (medium line) and $b_3 = 1.0371$ (thick line). On the right, $b_1 = -7.4671$ (thin line), $b_2 = -1.3296$ (medium line) and $b_3 = 4.8079$ (thick line).

scenario, since the function $p(x, b_i)$ is more variable as a function of Z (see Figure 4.4). Note that in Scenario 1, the power is lower due to the similarity between the functions $p(x, b_i)$, for $i = 1, 2, 3$. In addition, the best results are obtained with bandwidth $h = 10.86$ for $n = 50$, $h = 8.62$ for $n = 100$ and $h = 6.84$ for $n = 200$, that is, $h = Cn^{-1/3}$ with $C = 40$.

n	h	CvM	KS
50	2.71	0.0588	0.0570
	5.43	0.0468	0.0562
	10.86	0.0466	0.0730
	16.28	0.0618	0.1508
100	2.15	0.0482	0.0494
	4.31	0.0478	0.0514
	8.62	0.0638	0.0856
	12.93	0.0672	0.1604
200	1.71	0.0426	0.0406
	3.42	0.0406	0.0430
	6.84	0.0566	0.0722
	10.26	0.0708	0.1454

Table 4.16: Size of the test for Case 2 and Model 1, with X continuous with distribution $U(-20, 20)$, and Z discrete with values $\{b_1, b_2, b_3\}$ and probability mass function $(1/3, 1/3, 1/3)$, under the null hypothesis (see Figure 4.3). Note that under H_0, the probability mass function does not have any influence on the results.

X continuous, Z binary

X is a continuous $U(-20, 20)$ variable, and Z is binary with values $\{0, 1\}$. The probabilities of uncure $p(x, 0)$ and $p(x, 1)$ have been computed as follows: under H_0, the function $p(x, 0) = p(x, 1)$ is $p(x)$ in (4.8), given in Model 1 in Section 4.6.1 (see Figure 4.5). Under H_1, $p(x, 0)$ and $p(x, 1)$ are given by $p(x, z)$ in (4.10) evaluated at $b_1' = -5.2019$ and $b_2' = 1.0371$, respectively, in Scenario 1, and in $b_1' = -3.6964$ and $b_2' = -1.3296$ in Scenario 2 (see Figure 4.6). For each situation, 3 probability mass functions, $(\Pi_z(0), \Pi_z(1)) = (1/10, 9/10), (7/10, 3/10)$ and $(1/2, 1/2)$, are considered.

In Table 4.18 we can see the results under the null hypothesis. For most of the bandwidths, except for the largest ones, the size of the test is very similar to the significance level, $\alpha = 0.05$. Likewise in Case 2 when X is continuous and Z is discrete, the best choice of the bandwidth is $h = 5.43$ for $n = 50$, $h = 4.31$ for $n = 100$ and $h = 3.42$ for $n = 200$. Note that these values of the bandwidth correspond to $h = Cn^{-1/3}$, with $C = 20$.

The results under H_1 are shown in Table 4.19. As expected, in the second situation we obtained low power, since $p(x, 0)$ and $p(x, 1)$ are very similar functions (see Figure 4.6). On the contrary, in the first situation the power is higher because the functions $p(x, 0)$ and $p(x, 1)$ present more separate shapes.

			Scenario 1		Scenario 2	
n	h	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	CvM	KS	CvM	KS
50	2.71	$(1/3, 1/3, 1/3)$	0.0496	0.0534	0.1686	0.2208
	5.43	$(1/3, 1/3, 1/3)$	0.0708	0.0686	0.2348	0.2690
	10.86	$(1/3, 1/3, 1/3)$	0.0726	0.0700	0.2390	0.2662
	16.28	$(1/3, 1/3, 1/3)$	0.0592	0.0674	0.1984	0.2206
	2.71	$(3/5, 1/5, 1/5)$	0.0590	0.0606	0.3118	0.3244
	5.43	$(3/5, 1/5, 1/5)$	0.0842	0.0798	0.3746	0.3888
	10.86	$(3/5, 1/5, 1/5)$	0.0820	0.0772	0.3820	0.4214
	16.28	$(3/5, 1/5, 1/5)$	0.0674	0.0648	0.3744	0.4226
100	2.15	$(1/3, 1/3, 1/3)$	0.0540	0.0772	0.3298	0.4282
	4.31	$(1/3, 1/3, 1/3)$	0.0970	0.1102	0.4422	0.5224
	8.62	$(1/3, 1/3, 1/3)$	0.1100	0.1140	0.4628	0.5276
	12.93	$(1/3, 1/3, 1/3)$	0.0878	0.0910	0.4118	0.4776
	2.15	$(3/5, 1/5, 1/5)$	0.0964	0.1046	0.6026	0.6068
	4.31	$(3/5, 1/5, 1/5)$	0.1372	0.1428	0.6630	0.6730
	8.62	$(3/5, 1/5, 1/5)$	0.1378	0.1432	0.6640	0.6818
	12.93	$(3/5, 1/5, 1/5)$	0.1130	0.1120	0.6572	0.6900
200	1.71	$(1/3, 1/3, 1/3)$	0.0884	0.1232	0.6638	0.8134
	3.42	$(1/3, 1/3, 1/3)$	0.1566	0.1868	0.7934	0.8828
	6.84	$(1/3, 1/3, 1/3)$	0.1866	0.1996	0.8116	0.8836
	10.26	$(1/3, 1/3, 1/3)$	0.1638	0.1628	0.7798	0.8532
	1.71	$(3/5, 1/5, 1/5)$	0.1632	0.1858	0.9104	0.9240
	3.42	$(3/5, 1/5, 1/5)$	0.2284	0.2540	0.9406	0.9444
	6.84	$(3/5, 1/5, 1/5)$	0.2310	0.2558	0.9380	0.9444
	10.26	$(3/5, 1/5, 1/5)$	0.1968	0.2102	0.9360	0.9390

Table 4.17: Power of the test for Case 2 and Model 1, with X continuous with distribution $U(-20, 20)$, and Z discrete under the alternative hypothesis (see Figure 4.4).

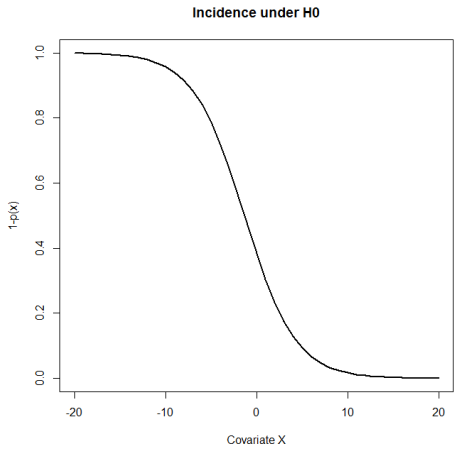

Figure 4.5: Probability of cure, $1 - p(x, 0) = 1 - p(x, 1)$, under the null hypothesis, with X continuous and Z binary, for Case 2 in Model 1.

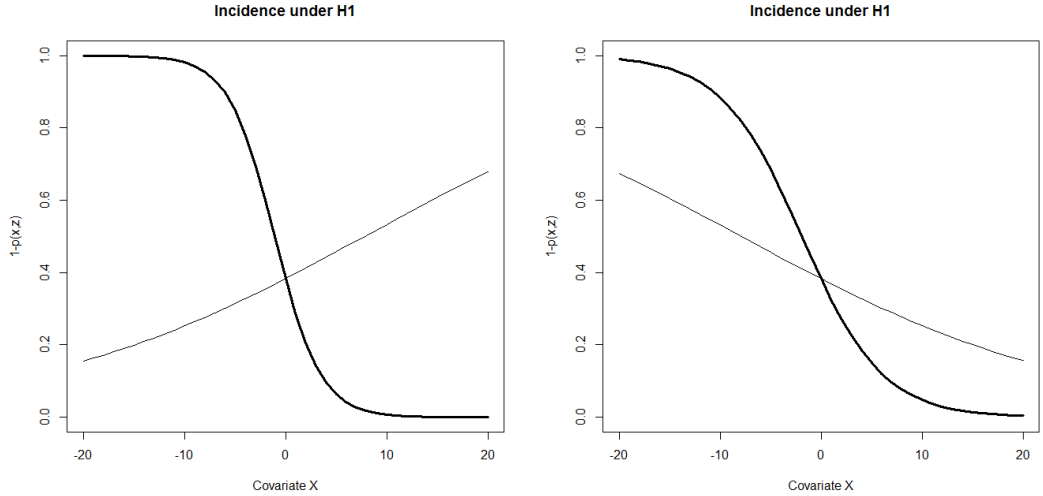

Figure 4.6: Probability of cure, $1 - p(x, 0)$ (thin line) and $1 - p(x, 1)$ (thick line), under the alternative hypothesis, with X continuous and Z binary, for Case 2 in Model 1, with $p(x, 0)$ and $p(x, 1)$ given by $p(x, z)$ in (4.10) evaluated at (x, b_1') and (x, b_2'), where $(b_1', b_2') = (-5.2019, 1.0371)$ (left) and $(b_1', b_2') = (-3.6964, -1.3296)$ (right).

n	h	CvM	KS
50	2.71	0.0626	0.0584
	5.43	0.0544	0.0602
	10.86	0.0540	0.0930
	16.28	0.0796	0.1996
100	2.15	0.0534	0.0518
	4.31	0.0510	0.0564
	8.62	0.0664	0.0956
	12.93	0.0782	0.1984
200	1.71	0.0410	0.0420
	3.42	0.0404	0.0450
	6.84	0.0554	0.0730
	10.26	0.0720	0.1696

Table 4.18: Size of the test for Case 2, with X continuous with distribution $U(-20, 20)$, and Z binary with values $\{0, 1\}$ and probability mass function $(1/2, 1/2)$, under the null hypothesis (see Figure 4.5).

			Scenario 1		Scenario 2	
n	h	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	2.71	$(9/10, 1/10)$	0.1382	0.1074	0.0294	0.0258
	5.43	$(9/10, 1/10)$	0.1838	0.1470	0.0554	0.0456
	10.86	$(9/10, 1/10)$	0.1916	0.1654	0.0622	0.0982
	16.28	$(9/10, 1/10)$	0.1996	0.1850	0.0422	0.0514
	2.71	$(7/10, 3/10)$	0.1910	0.1882	0.0540	0.0524
	5.43	$(7/10, 3/10)$	0.2404	0.2378	0.0838	0.0830
	10.86	$(7/10, 3/10)$	0.2378	0.2432	0.0454	0.0572
	16.28	$(7/10, 3/10)$	0.2230	0.2354	0.0596	0.0656
	2.71	$(1/2, 1/2)$	0.1778	0.2266	0.0484	0.0610
	5.43	$(1/2, 1/2)$	0.2370	0.2724	0.0904	0.1068
	10.86	$(1/2, 1/2)$	0.2472	0.2794	0.0534	0.0750
	16.28	$(1/2, 1/2)$	0.2276	0.2482	0.0638	0.0742
100	2.15	$(9/10, 1/10)$	0.2270	0.1876	0.0334	0.0324
	4.31	$(9/10, 1/10)$	0.2656	0.2310	0.0816	0.0746
	8.62	$(9/10, 1/10)$	0.2744	0.2576	0.0430	0.0586
	12.93	$(9/10, 1/10)$	0.2858	0.2760	0.0478	0.0536
	2.15	$(7/10, 3/10)$	0.3712	0.3826	0.0746	0.0810
	4.31	$(7/10, 3/10)$	0.4336	0.4368	0.1498	0.1594
	8.62	$(7/10, 3/10)$	0.4254	0.4438	0.0450	0.0548
	12.93	$(7/10, 3/10)$	0.4040	0.4284	0.0842	0.0822
	2.15	$(1/2, 1/2)$	0.3496	0.4320	0.0630	0.0900
	4.31	$(1/2, 1/2)$	0.4440	0.5068	0.1478	0.1920
	8.62	$(1/2, 1/2)$	0.4454	0.5034	0.0534	0.0694
	12.93	$(1/2, 1/2)$	0.4052	0.4624	0.0972	0.0998
200	1.71	$(9/10, 1/10)$	0.3784	0.3440	0.0644	0.0578
	3.42	$(9/10, 1/10)$	0.4078	0.3778	0.1356	0.1394
	6.84	$(9/10, 1/10)$	0.4112	0.3900	0.0458	0.0532
	10.26	$(9/10, 1/10)$	0.4252	0.4036	0.0778	0.0728
	1.71	$(7/10, 3/10)$	0.6498	0.6810	0.1144	0.1306
	3.42	$(7/10, 3/10)$	0.6872	0.7192	0.1568	0.1792
	6.84	$(7/10, 3/10)$	0.6764	0.7128	0.1574	0.1768
	10.26	$(7/10, 3/10)$	0.6572	0.6944	0.1288	0.1430
	1.71	$(1/2, 1/2)$	0.6546	0.7684	0.1136	0.1670
	3.42	$(1/2, 1/2)$	0.7434	0.8184	0.2652	0.3582
	6.84	$(1/2, 1/2)$	0.7432	0.8144	0.0592	0.0704
	10.26	$(1/2, 1/2)$	0.7156	0.7870	0.1616	0.1896

Table 4.19: Power of the test for Case 2 and Model 1, with X continuous with distribution $U(-20, 20)$, and Z binary with values $\{0, 1\}$, under the alternative hypothesis (see Figure 4.6).

X continuous, Z qualitative

As in the previous cases, X is $U(-20, 20)$, and Z is a categorical variable with values $\{b_1, b_2, b_3\}$, with probability mass functions $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$. The probabilities of uncure, $p(x, b_1) = p(x, b_2) = p(x, b_3)$, under H_0, are the function $p(x)$ in (4.8) in Model 1 of Section 4.6.1 (see Figure 4.7, left). Under H_1, the incidence functions, $p(x, b_i)$ with $i = 1, 2, 3$, are given by the function $p(x, z)$ in (4.10), evaluated at (x, b'_1), (x, b'_2) and (x, b'_3), where $(b'_1, b'_2, b'_3) = (-5.2019, -1.3296, 1.0371)$ (see Figure 4.7, right).

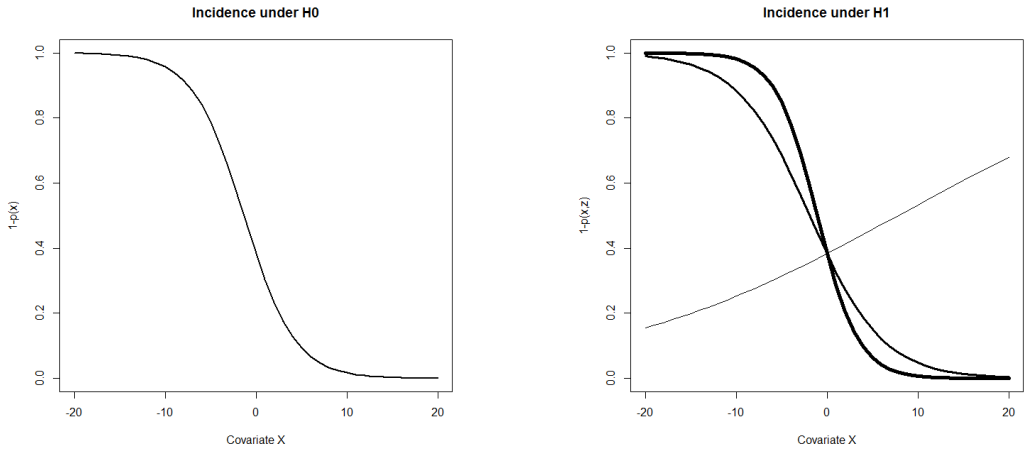

Figure 4.7: Probabilities of cure, $1 - p(x, z)$, for Case 2 with X continuous $U(-20, 20)$ and Z qualitative with values $\{b_1, b_2, b_3\}$. On the left, $1 - p(x, z)$ under the null hypothesis. On the right, $1 - p(x, b_1)$ (thin), $1 - p(x, b_2)$ (medium) and $1 - p(x, b_3)$ (thick line), given by $p(x, z)$ in (4.10) evaluated at (x, b'_1), (x, b'_2) and (x, b'_3), where $b'_1 = -5.2019$, $b'_2 = -1.3296$ and $b'_3 = 1.0371$.

The results under the null (alternative) hypothesis are shown in Table 4.20 (Table 4.21). Under the null hypothesis, the best choice of the bandwidth is $h = 5.43$ for $n = 50$, $h = 2.15$ for $n = 100$ and $h = 1.71$ for $n = 200$. Moreover, under H_1 we obtain higher power when using $h = 16.28$ for $n = 50$, $h = 8.62$ or 12.93 for $n = 100$ and $h = 6.84$ or 10.26 for $n = 200$, that is, $h = Cn^{-1/3}$ with $C = 60$.

X discrete, Z continuous

The variable X is discrete with values $\{a_1, a_2, a_3\}$, and Z is $U(-20, 20)$. We consider two scenarios depending on the values of X. In the first scenario, $a_1 = -3.6964$, $a_2 = -1.3296$ and $a_3 = 1.0371$; and in the second one, $a_1 = -7.4671$, $a_2 = -1.3296$ and $a_3 = 4.8079$. Furthermore, the probability mass functions considered are

n	h	CvM	KS
50	2.71	0.0520	0.0600
	5.43	0.0448	0.0554
	10.86	0.0450	0.0832
	16.28	0.0558	0.1546
100	2.15	0.0436	0.0530
	4.31	0.0414	0.0560
	8.62	0.0584	0.0886
	12.93	0.0656	0.1538
200	1.71	0.0406	0.0480
	3.42	0.0374	0.0480
	6.84	0.0584	0.0800
	10.26	0.0780	0.1462

Table 4.20: Size of the test for Case 2 with X continuous with distribution $U(-20, 20)$, and Z qualitative with values $\{b_1, b_2, b_3\}$, with probability mass function $(1/3, 1/3, 1/3)$, under the null hypothesis (see Figure 4.7, left).

$(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$ for each scenario. The incidence, $1 - p(a_i, z)$, reduces, under H_0, to $1 - p(a_i)$, with $i = 1, 2, 3$, given by the function $p(x)$ in (4.8) of Model 1 in Section 4.6.1, evaluated at the aforementioned values a_i of X (see Figure 4.8). Under H_1, the incidence, $1 - p(a_i, z)$, can be obtained from $p(x, z)$ in (4.10) of Model 1 in Section 4.6.2 (see Figure 4.9).

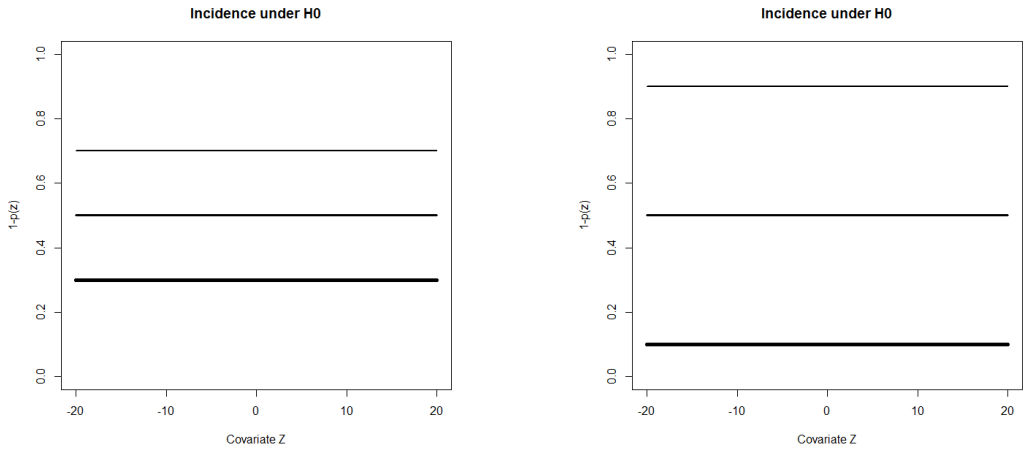

Figure 4.8: Probability of cure, $1 - p(a_i, z)$, with $i = 1, 2, 3$, under H_0 for Case 2, where X is discrete with values $\{a_1, a_2, a_3\}$, and Z is continuous with distribution $U(-20, 20)$. On the left, for Scenario 1, we consider $a_1 = -3.6964$ (thin line), $a_2 = -1.3296$ (medium line) and $a_3 = 1.0371$ (thick line). On the right, for Scenario 2, with $a_1 = -7.4671$ (thin line), $a_2 = -1.3296$ (medium line) and $a_3 = 4.8079$ (thick line).

n	h	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	CvM	KS
50	2.71	$(1/3, 1/3, 1/3)$	0.0976	0.1270
	5.43	$(1/3, 1/3, 1/3)$	0.1382	0.1598
	10.86	$(1/3, 1/3, 1/3)$	0.1562	0.1990
	16.28	$(1/3, 1/3, 1/3)$	0.1686	0.2448
	2.71	$(3/5, 1/5, 1/5)$	0.1592	0.1536
	5.43	$(3/5, 1/5, 1/5)$	0.2086	0.2018
	10.86	$(3/5, 1/5, 1/5)$	0.2090	0.2176
	16.28	$(3/5, 1/5, 1/5)$	0.1920	0.2200
100	2.15	$(1/3, 1/3, 1/3)$	0.1796	0.2524
	4.31	$(1/3, 1/3, 1/3)$	0.2452	0.3160
	8.62	$(1/3, 1/3, 1/3)$	0.2796	0.3730
	12.93	$(1/3, 1/3, 1/3)$	0.2858	0.4276
	2.15	$(3/5, 1/5, 1/5)$	0.3202	0.3336
	4.31	$(3/5, 1/5, 1/5)$	0.3848	0.3876
	8.62	$(3/5, 1/5, 1/5)$	0.3868	0.4038
	12.93	$(3/5, 1/5, 1/5)$	0.3550	0.4060
200	1.71	$(1/3, 1/3, 1/3)$	0.3666	0.5278
	3.42	$(1/3, 1/3, 1/3)$	0.4698	0.6044
	6.84	$(1/3, 1/3, 1/3)$	0.5112	0.6566
	10.26	$(1/3, 1/3, 1/3)$	0.5212	0.6988
	1.71	$(3/5, 1/5, 1/5)$	0.6048	0.6342
	3.42	$(3/5, 1/5, 1/5)$	0.6532	0.6804
	6.84	$(3/5, 1/5, 1/5)$	0.6410	0.6850
	10.26	$(3/5, 1/5, 1/5)$	0.6068	0.6772

Table 4.21: Power of the test for Case 2 with X continuous with distribution $U(-20, 20)$, and Z qualitative with values $\{b_1, b_2, b_3\}$ under H_1 (see Figure 4.7, right).

The results under the null hypothesis are shown in Table 4.22. For the different values of the probability mass functions of X, $\Pi_x(a_i)$, with $i = 1, 2, 3$, the results are very close to the significance level, $\alpha = 0.05$. Table 4.23 shows the results under the alternative hypothesis. Note that in Scenario 1, $p(a_1, b_1) = 0.3$, $p(a_2, b_2) = 0.5$ and $p(a_3, b_3) = 0.7$, whereas in Scenario 2, $p(a_1, b_1) = 0.1$, $p(a_2, b_2) = 0.5$ and $p(a_3, b_3) = 0.9$. The highest power is obtained when we consider that the probability mass function of X is $(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3)) = (3/5, 1/5, 1/5)$, that is, the power of the test is higher when the extreme values are more frequent (i.e. when $p(a_1)$ has more weight).

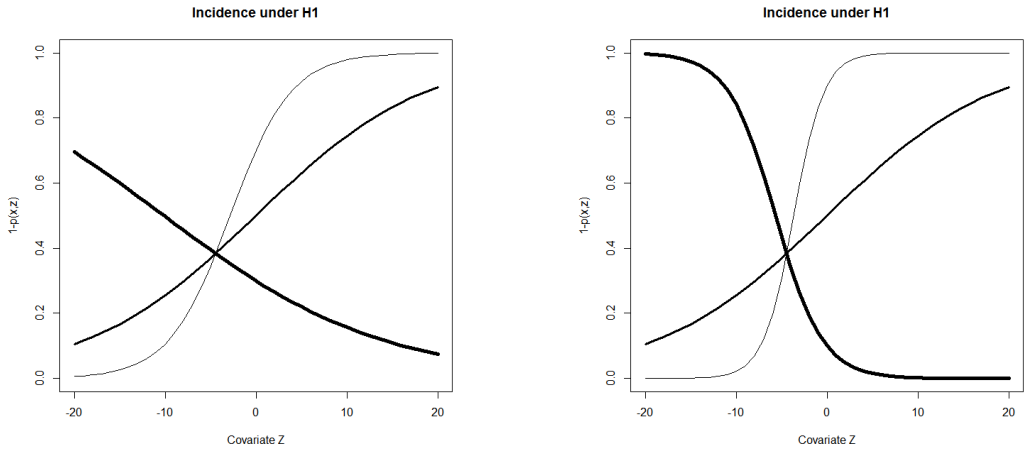

Figure 4.9: Probability of cure, $1 - p(a_i, z)$, with $i = 1, 2, 3$, under H_1 for Case 2, where X is discrete with values $\{a_1, a_2, a_3\}$, and Z is continuous with distribution $U(-20, 20)$. On the left, $p(a_i, z)$ in (4.10) for Scenario 1, considering $a_1 = -5.2019$ (thin line), $a_2 = -1.3296$ (medium line) and $a_3 = 1.0371$ (thick line). On the right, $p(a_i, z)$ in (4.10) for Scenario 2, with $a_1 = -7.4671$ (thin line), $a_2 = -1.3296$ (medium line) and $a_3 = 4.8079$ (thick line).

		Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.0498	0.0602	0.0540	0.0570
	$(3/5, 1/5, 1/5)$	0.0512	0.0612	0.0440	0.0558
100	$(1/3, 1/3, 1/3)$	0.0518	0.0584	0.0500	0.0560
	$(3/5, 1/5, 1/5)$	0.0454	0.0494	0.0420	0.0498
200	$(1/3, 1/3, 1/3)$	0.0494	0.0534	0.0512	0.0576
	$(3/5, 1/5, 1/5)$	0.0476	0.0504	0.0536	0.0578

Table 4.22: Size of the test for Case 2 with X discrete, with values $\{a_1, a_2, a_3\}$, and Z continuous with distribution $U(-20, 20)$, under the null hypothesis (see Figure 4.8).

		Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.4522	0.4150	0.4276	0.4372
	$(3/5, 1/5, 1/5)$	0.7306	0.7314	0.7478	0.7786
100	$(1/3, 1/3, 1/3)$	0.7248	0.6818	0.7168	0.7426
	$(3/5, 1/5, 1/5)$	0.9258	0.9260	0.9442	0.9548
200	$(1/3, 1/3, 1/3)$	0.9466	0.9262	0.9456	0.9540
	$(3/5, 1/5, 1/5)$	0.9944	0.9940	0.9976	0.9986

Table 4.23: Power of the test for Case 2 with X discrete, with values $\{a_1, a_2, a_3\}$, and Z continuous with distribution $U(-20, 20)$, under the alternative hypothesis (see Figure 4.9).

X discrete, Z discrete

The covariates X and Z are discrete variables with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. We work with two different situations depending on the corresponding probability mass functions: in the first one, both for X and Z are $(1/3, 1/3, 1/3)$, whereas in the second one, both are $(3/5, 1/5, 1/5)$.

Under H_1, the values of Z are $\{b_1, b_2, b_3\} = \{0.6157, -3.5434, -7.7026\}$, and those of X, $\{a_1, a_2, a_3\}$, are chosen so $p(a_i, b_j)$ has high dependency on b_j, with $j = 1, 2, 3$, only for $x = a_1$ (Scenario 1), for $x \in \{a_1, a_2\}$ (Scenario 2), and for $x \in \{a_1, a_2, a_3\}$ (Scenarios 3 and 4). The values $\{a_1, a_2, a_3\}$ in the four scenarios are given in Table 4.25. Note that in these last two scenarios more power is expected.

Under H_0, the values $\{a_1, a_2, a_3\}$ of X and $\{b_1, b_2, b_3\}$ of Z are the same as those under H_1. The probabilities $p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$, with $i = 1, 2, 3$, however, are given now by function p in (4.10) evaluated at $(a_i, \bar{b})$, with $\bar{b}$ suitable chosen. Two scenarios are considered, one when the probability mass functions of X and Z are $(1/3, 1/3, 1/3)$, and the other when the probability mass functions are $(3/5, 1/5, 1/5)$. See Table 4.24 for details.

Scenario 1	$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$		$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$	
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -1.1678$	$(\bar{b} = -3.4427)$	0.5943	$(\bar{b} = -1.7995)$	0.5566
$a_3 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -2.0064)$	0.6581
Scenario 2				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a_3 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -1.9950)$	0.6583
Scenario 3				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a_3 = 4.2229$	$(\bar{b} = -4.0960)$	0.6444	$(\bar{b} = -2.6671)$	0.7466
Scenario 4				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a_3 = -3.2466$	$(\bar{b} = -3.3932)$	0.5501	$(\bar{b} = -1.8585)$	0.4501

Table 4.24: Uncure probabilities considered under H_0, for Case 2 when X and Z are discrete with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. The probability mass functions for both X and Z are $(1/3, 1/3, 1/3)$ (first column), and $(3/5, 1/5, 1/5)$ (second column). See Remark in Section 4.6.2 for details.

	$b_1 = 0.6157$	$b_2 = -3.5434$	$b_3 = -7.7026$
Scenario 1 (Least favorable situation)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -1.1678$	0.5	0.5966	0.6862
$a_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 2 (Intermediate situation)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -4.5690$	0.2	0.5360	0.8423
$a_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 3 (Favorable situation a)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -4.5690$	0.2	0.5360	0.8422
$a_3 = 4.2229$	0.9	0.6862	0.3470
Scenario 4 (Favorable situation b)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -4.5690$	0.2	0.5360	0.8423
$a_3 = -3.2466$	0.3	0.5598	0.7905

Table 4.25: Uncure probabilities under H_1, for Case 2 when X and Z are discrete with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively.

Table 4.26 shows the results under the null hypothesis, which are very close to the significance level even for small sample sizes. The results under the alternative hypothesis for the 4 scenarios are given in Table 4.27. As expected, the power of the test is higher in Scenario 4. It is important to highlight that for large sample sizes ($n = 200$) and for the probability mass function $(3/5, 1/5, 1/5)$ for both X and Z, the power is also very close to 1 for the 4 scenarios.

X discrete, *Z* binary

The variable X is discrete, with values $\{a_1, a_2, a_3\}$ and probability mass functions $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$, and Z is binary with values $\{0, 1\}$ and probability mass functions $(1/2, 1/2)$ and $(7/10, 3/10)$.

The probabilities $p(a_i, 0)$ and $p(a_i, 1)$, $i = 1, 2, 3$, are obtained evaluating $p(x, z)$ in (4.10). The scenarios under H_0 are detailed in Table 4.28. Under H_1, we consider 4 scenarios according to the values of X: $\{a_1, a_2, a_3\}$, are chosen so there is a high dependence on the values $\{0, 1\}$ of Z in $p(a_1, \cdot)$, but not in $p(a_i, \cdot)$, $i = 2, 3$ (Scenario 1), in $p(a_1, \cdot)$ and $p(a_2, \cdot)$ but not in $p(a_3, \cdot)$ (Scenario 2). In Scenarios 3 and 4, there is a high dependence on the values $\{0, 1\}$ in $p(a_i, \cdot)$, $i = 1, 2, 3$ (see Table 4.29).

n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS
Scenario 1			
50	$(1/3, 1/3, 1/3)$	0.0490	0.0500
	$(3/5, 1/5, 1/5)$	0.0574	0.0556
100	$(1/3, 1/3, 1/3)$	0.0480	0.0508
	$(3/5, 1/5, 1/5)$	0.0530	0.0538
200	$(1/3, 1/3, 1/3)$	0.0506	0.0558
	$(3/5, 1/5, 1/5)$	0.0484	0.0504
Scenario 2			
50	$(1/3, 1/3, 1/3)$	0.0458	0.0522
	$(3/5, 1/5, 1/5)$	0.0550	0.0546
100	$(1/3, 1/3, 1/3)$	0.0514	0.0530
	$(3/5, 1/5, 1/5)$	0.0558	0.0552
200	$(1/3, 1/3, 1/3)$	0.0536	0.0506
	$(3/5, 1/5, 1/5)$	0.0528	0.0544
Scenario 3			
50	$(1/3, 1/3, 1/3)$	0.0556	0.0564
	$(3/5, 1/5, 1/5)$	0.0540	0.0526
100	$(1/3, 1/3, 1/3)$	0.0512	0.0568
	$(3/5, 1/5, 1/5)$	0.0544	0.0528
200	$(1/3, 1/3, 1/3)$	0.0514	0.0498
	$(3/5, 1/5, 1/5)$	0.0560	0.0502
Scenario 4			
50	$(1/3, 1/3, 1/3)$	0.0504	0.0526
	$(3/5, 1/5, 1/5)$	0.0594	0.0538
100	$(1/3, 1/3, 1/3)$	0.0504	0.0492
	$(3/5, 1/5, 1/5)$	0.0526	0.0512
200	$(1/3, 1/3, 1/3)$	0.0496	0.0526
	$(3/5, 1/5, 1/5)$	0.0464	0.0490

Table 4.26: Size of the test under H_0, for Case 2 with X and Z discrete, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$ (see Table 4.24). The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.2774	0.2172	0.4592	0.3868	0.3690	0.3286	0.6248	0.5830
	$(3/5, 1/5, 1/5)$	0.5328	0.5060	0.5544	0.5444	0.5556	0.5418	0.5948	0.5998
100	$(1/3, 1/3, 1/3)$	0.4586	0.3408	0.7262	0.6364	0.6268	0.5902	0.8820	0.8476
	$(3/5, 1/5, 1/5)$	0.7734	0.7580	0.7954	0.7924	0.7748	0.7766	0.8034	0.8284
200	$(1/3, 1/3, 1/3)$	0.7106	0.5674	0.9344	0.8874	0.8780	0.8696	0.9864	0.9764
	$(3/5, 1/5, 1/5)$	0.9396	0.9336	0.9476	0.9484	0.9438	0.9506	0.9492	0.9642

Table 4.27: Power of the test under H_1, for Case 2 with X and Z discrete, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$ (see Table 4.25). The probability mass function of Z equals that of X.

Scenario 1				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2 = -1.1678$	$(\bar{b} = -3.3900)$	0.5931	$(\bar{b} = -5.0579)$	0.6303
$a_3 = 0.9109$	$(\bar{b} = -3.7087)$	0.6295	$(\bar{b} = -5.3335)$	0.6013
Scenario 2				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2 = -4.5690$	$(\bar{b} = -3.3809)$	0.5211	$(\bar{b} = -4.8281)$	0.6496
$a_3 = 0.9109$	$(\bar{b} = -3.7087)$	0.6295	$(\bar{b} = -5.3335)$	0.6013
Scenario 3				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2 = -4.5690$	$(\bar{b} = -3.3809)$	0.5211	$(\bar{b} = -4.8281)$	0.6496
$a_3 = 4.2229$	$(\bar{b} = -4.3609)$	0.6295	$(\bar{b} = -5.6922)$	0.5129
Scenario 4				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2 = -4.5690$	$(\bar{b} = -3.3809)$	0.5211	$(\bar{b} = -4.8281)$	0.6496
$a_3 = -3.2466$	$(\bar{b} = -3.3184)$	0.5453	$(\bar{b} = -4.8803)$	0.6434

Table 4.28: Uncure probabilities considered under H_0, for Case 2 when X is discrete with values $\{a_1, a_2, a_3\}$, and Z is binary with values $\{0, 1\}$. The probability mass functions for Z are $(1/2, 1/2)$ (first column), and $(7/10, 3/10)$ (second column). See Remark in Section 4.6.2 for details.

	$p(a_i, 0)$	$p(a_i, 1)$
Scenario 1 (Least favorable situation)		
$a_1 = -6.5585$	0.1	0.9000
$a_2 = -1.1678$	0.5	0.6862
$a_3 = 0.9109$	0.7	0.5590
Scenario 2 (Intermediate situation)		
$a_1 = -6.5585$	0.1	0.9000
$a_2 = -4.5690$	0.2	0.8423
$a_3 = 0.9109$	0.7	0.5590
Scenario 3 (Favorable situation a)		
$a_1 = -6.5585$	0.1	0.9000
$a_2 = -4.5690$	0.2	0.8423
$a_3 = 4.2229$	0.9	0.3470
Scenario 4 (Favorable situation b)		
$a_1 = -6.5585$	0.1	0.9000
$a_2 = -4.5690$	0.2	0.8423
$a_3 = -3.2466$	0.3	0.7905

Table 4.29: Uncure probabilities considered under H_1, for Case 2 when X is discrete with values $\{a_1, a_2, a_3\}$, and Z is binary with values $\{0, 1\}$. See Remark in Section 4.6.2 for details.

n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS
Scenario 1				
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0460	0.0566
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0434	0.0488
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0574	0.0632
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0562	0.0578
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0468	0.0532
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0486	0.0502
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0512	0.0526
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0502	0.0504
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0464	0.0490
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0532	0.0548
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0578	0.0562
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0504	0.0500
Scenario 2				
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0420	0.0514
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0426	0.0496
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0584	0.0610
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0596	0.0616
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0482	0.0526
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0484	0.0514
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0558	0.0522
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0538	0.0540
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0512	0.0544
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0504	0.0544
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0462	0.0482
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0502	0.0502
Scenario 3				
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0424	0.0468
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0402	0.0466
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0568	0.0604
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0558	0.0566
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0492	0.0548
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0436	0.0462
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0556	0.0534
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0512	0.0528
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0504	0.0530
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0506	0.0512
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0498	0.0506
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0532	0.0516
Scenario 4				
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0464	0.0524
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0458	0.0510
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0512	0.0540
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0542	0.0544
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0464	0.0510
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0422	0.0498
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0544	0.0550
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0508	0.0522
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0512	0.0552
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0422	0.0452
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0500	0.0484
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0482	0.0466

Table 4.30: Size of the test under the null hypothesis for Case 2 with X discrete with values $\{a_1, a_2, a_3\}$, and Z binary with values $\{0, 1\}$ (see Table 4.28).

			Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.3652	0.3396	0.5976	0.5828
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.3970	0.3446	0.6574	0.6370
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.7212	0.7344	0.7482	0.7744
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.7896	0.7862	0.8154	0.8336
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.5850	0.5244	0.8216	0.8068
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.6414	0.5732	0.9002	0.8886
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9106	0.9106	0.9300	0.9408
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9470	0.9458	0.9616	0.9668
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.8304	0.7732	0.9682	0.9644
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.8804	0.8370	0.9888	0.9874
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9888	0.9892	0.9912	0.9924
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9936	0.9938	0.9968	0.9978
			Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.4922	0.5186	0.7780	0.7872
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.5538	0.5938	0.8232	0.8262
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.7308	0.7592	0.7818	0.8164
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.8074	0.8254	0.8290	0.8626
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.7472	0.7960	0.9488	0.9484
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.8206	0.8716	0.9722	0.9712
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9200	0.9334	0.9328	0.9520
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9564	0.9640	0.9622	0.9756
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.9426	0.9606	0.9972	0.9968
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.9788	0.9894	0.9988	0.9992
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9866	0.9896	0.9916	0.9956
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9970	0.9982	0.9976	0.9992

Table 4.31: Power of the test under H_1, for Case 2 with X discrete with values $\{a_1, a_2, a_3\}$, and Z binary with values $\{0, 1\}$ (see Table 4.29).

The results under the null hypothesis (the alternative hypothesis) are shown in Table 4.30 (4.31). For all the scenarios, determined by the probability mass functions of X and Z, the size of the test is close to $\alpha = 0.05$ for all the sample sizes. The power is higher when $\Pi_x = (3/5, 1/5, 1/5)$ and $\Pi_z = (1/2, 1/2)$. In Scenario 4 it is very close to 1 even for small sample sizes.

X discrete, Z qualitative

X is a discrete variable with values $\{a_1, a_2, a_3\}$, and Z is qualitative with values $\{b_1, b_2, b_3\}$. The cure probabilities, $1 - p(a_i, b_j)$, with $i, j = 1, 2, 3$, are computed from the function $p(x, z)$ in (4.10), evaluated at the numerical values $(a_i, \bar{b})$, with $i = 1, 2, 3$, given in Table 4.32 (under the null hypothesis), and at (a_i, b'_j), with $i, j = 1, 2, 3$, given in Table 4.33 (under the alternative hypothesis).

Two different situations are considered within each scenario, one when the probability mass functions of X and Z are $(1/3, 1/3, 1/3)$, and the other when the probability mass functions are $(3/5, 1/5, 1/5)$.

Scenario 1	$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$		$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$	
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -1.1678$	$(\bar{b} = -3.4427)$	0.5943	$(\bar{b} = -1.7995)$	0.5566
$a_3 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -2.0064)$	0.6581
Scenario 2				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a_3 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -1.9950)$	0.6583
Scenario 3				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a_3 = 4.2229$	$(\bar{b} = -4.0960)$	0.6444	$(\bar{b} = -2.6671)$	0.7466
Scenario 4				
$a_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a_3 = -3.2466$	$(\bar{b} = -3.3932)$	0.5501	$(\bar{b} = -1.8585)$	0.4501

Table 4.32: Uncure probabilities under H_0, for Case 2 when X is discrete with values $\{a_1, a_2, a_3\}$, and Z is qualitative with values $\{b_1, b_2, b_3\}$. The probability mass functions for both X and Z are $(1/3, 1/3, 1/3)$ (first column), and $(3/5, 1/5, 1/5)$ (second column). See Remark in Section 4.6.2 for details.

The results under H_0 (H_1) are shown in Table 4.34 (Table 4.35). In the 4 scenarios, the size of the test is very similar to the significance level, $\alpha = 0.05$. Similarly as it happened in Case 2 when X and Z are discrete, the highest powers are obtained if the distribution mass functions of X and Z are $(3/5, 1/5, 1/5)$.

	$b'_1 = 0.6157$	$b'_2 = -3.5434$	$b'_3 = -7.7026$
Scenario 1 (Least favorable situation)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -1.1678$	0.5	0.5966	0.6862
$a_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 2 (Intermediate situation)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -4.5690$	0.2	0.5360	0.8423
$a_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 3 (Favorable situation a)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -4.5690$	0.2	0.5360	0.8423
$a_3 = 4.2229$	0.9	0.6862	0.3470
Scenario 4 (Favorable situation b)			
$a_1 = -6.5585$	0.1	0.5000	0.9000
$a_2 = -4.5690$	0.2	0.5360	0.8423
$a_3 = -3.2466$	0.3	0.5598	0.7905

Table 4.33: Uncure probabilities considered under H_1, for Case 2 when X is discrete with values $\{a_1, a_2, a_3\}$, and Z is qualitative with values $\{b_1, b_2, b_3\}$. See Remark in Section 4.6.2 for details.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.0424	0.0488	0.0476	0.0552	0.0516	0.0576	0.0442	0.0514
	$(3/5, 1/5, 1/5)$	0.0508	0.0526	0.0494	0.0522	0.0516	0.0514	0.0524	0.0516
100	$(1/3, 1/3, 1/3)$	0.0470	0.0532	0.0508	0.0546	0.0484	0.0588	0.0494	0.0516
	$(3/5, 1/5, 1/5)$	0.0536	0.0518	0.0508	0.0524	0.0508	0.0536	0.0530	0.0522
200	$(1/3, 1/3, 1/3)$	0.0488	0.0534	0.0494	0.0534	0.0452	0.0490	0.0504	0.0526
	$(3/5, 1/5, 1/5)$	0.0466	0.0492	0.0540	0.0546	0.0526	0.0500	0.0486	0.0486

Table 4.34: Size of the test under H_0 for Case 2 with X discrete with values $\{a_1, a_2, a_3\}$, and Z qualitative with values $\{b_1, b_2, b_3\}$ (see Table 4.32). The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.2096	0.1866	0.3592	0.3396	0.2882	0.2882	0.5150	0.5294
	$(3/5, 1/5, 1/5)$	0.4952	0.4852	0.5100	0.5152	0.5106	0.5144	0.5384	0.5726
100	$(1/3, 1/3, 1/3)$	0.3626	0.2966	0.6242	0.5748	0.5144	0.5270	0.8142	0.8052
	$(3/5, 1/5, 1/5)$	0.7394	0.7302	0.7582	0.7698	0.7458	0.7584	0.7728	0.8088
200	$(1/3, 1/3, 1/3)$	0.7582	0.7698	0.8820	0.8488	0.7894	0.8266	0.9714	0.9678
	$(3/5, 1/5, 1/5)$	0.9262	0.9214	0.9364	0.9438	0.9378	0.9470	0.9400	0.9538

Table 4.35: Power of the test under H_1 for Case 2 with X discrete with values $\{a_1, a_2, a_3\}$, and Z qualitative with values $\{b_1, b_2, b_3\}$ (see Table 4.33). The probability mass function of Z equals that of X.

X binary, Z continuous

We consider X binary with values $\{0, 1\}$ and probability mass functions $(1/2, 1/2)$ and $(7/10, 3/10)$, and Z continuous with distribution $U(-20, 20)$. Two scenarios are simulated. The probabilities $p(0, z)$ and $p(1, z)$ are given by $p(x, z)$ in (4.10) evaluated at (a'_1, z) and (a'_2, z), respectively, in two scenarios: for the first one $a'_1 = -3.6964$ and $a'_2 = 1.0371$, and for the second one $a'_1 = -7.4671$ and $a'_2 = 4.8079$. Figures 4.10 and 4.11 show the corresponding cure probabilities under the null and the alternative hypothesis, respectively.

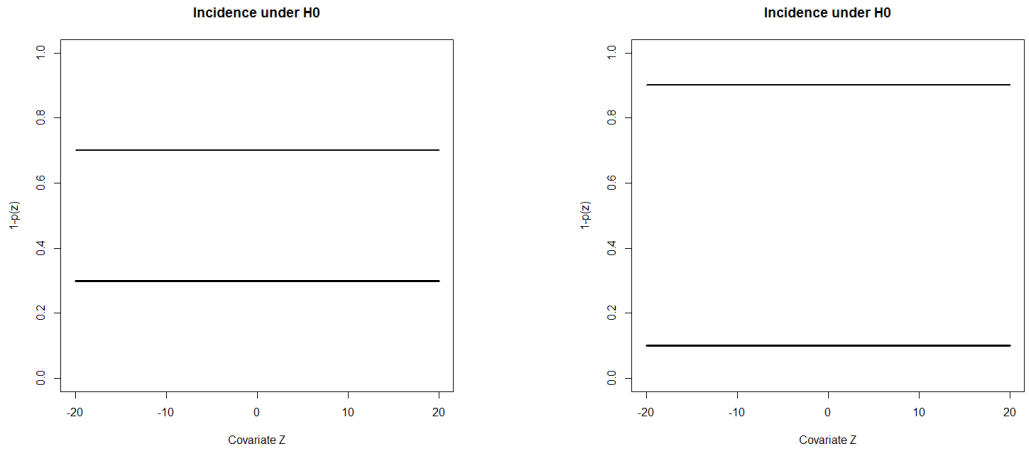

Figure 4.10: Probability of cure, $1 - p(0, z)$ (thin line) and $1 - p(1, z)$ (thick line), under H_0 for Case 2, where X is binary with values $\{0, 1\}$, and Z is continuous with distribution $U(-20, 20)$, for Scenario 1 (left) and Scenario 2 (right).

The results under the null hypothesis are shown in Table 4.36. Note that in Scenario 1, we obtain results very close to the significance level even for small sample sizes.

In Table 4.37 we can see the results under the alternative hypothesis. On the contrary as it happened under H_0, the best results are obtained in Scenario 2: the power is almost equal to 1 for $n = 200$ and when the probability mass function of X is $(7/10, 3/10)$.

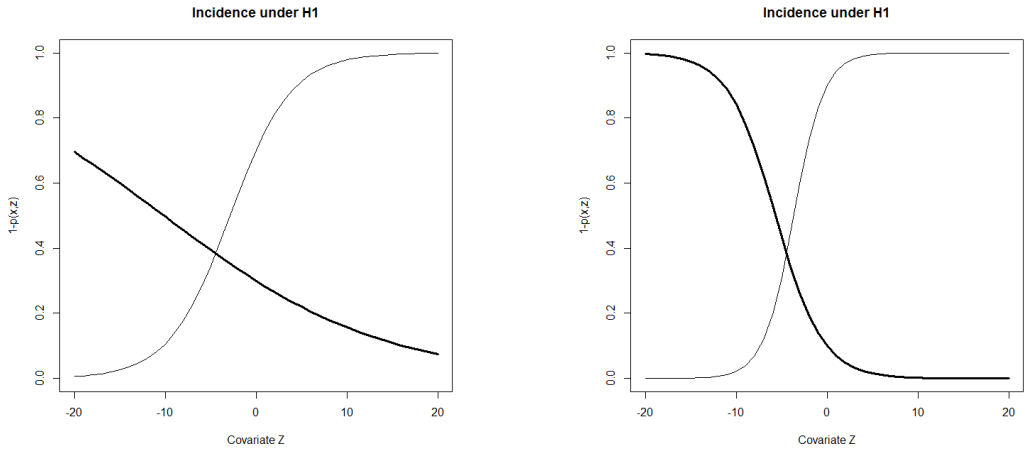

Figure 4.11: Probability of cure, $1-p(0,z)$ (thin line) and $1-p(1,z)$ (thick line), under H_1 for Case 2, where X is binary with values $\{0,1\}$, and Z is continuous with distribution $U(-20,20)$, for Scenario 1 (left) and Scenario 2 (right).

		Scenario 1		Scenario 2	
n	$(\Pi_x(0),\Pi_x(1))$	CvM	KS	CvM	KS
50	$(1/2,1/2)$	0.0502	0.0548	0.0480	0.0488
	$(7/10,3/10)$	0.0422	0.0518	0.0298	0.0308
100	$(1/2,1/2)$	0.0492	0.0538	0.0492	0.0494
	$(7/10,3/10)$	0.0428	0.0602	0.0368	0.0380
200	$(1/2,1/2)$	0.0526	0.0536	0.0474	0.0524
	$(7/10,3/10)$	0.0496	0.0588	0.0416	0.0452

Table 4.36: Size of the test for Case 2 with X binary with values $\{0,1\}$, and Z continuous with distribution $U(-20,20)$, under the null hypothesis (see Figure 4.10).

		Scenario 1		Scenario 2	
n	$(\Pi_x(0),\Pi_x(1))$	CvM	KS	CvM	KS
50	$(1/2,1/2)$	0.3482	0.3414	0.3590	0.4476
	$(7/10,3/10)$	0.1620	0.1694	0.7422	0.7420
100	$(1/2,1/2)$	0.5724	0.5698	0.6282	0.7156
	$(7/10,3/10)$	0.2690	0.2514	0.9354	0.9126
200	$(1/2,1/2)$	0.8606	0.8674	0.9228	0.9600
	$(7/10,3/10)$	0.4460	0.3802	0.9970	0.9908

Table 4.37: Power of the test for Case 2 with X binary with values $\{0,1\}$ and Z continuous with distribution $U(-20,20)$, under the alternative hypothesis (see Figure 4.11).

X binary, Z discrete

We consider a binary covariate X with values $\{0, 1\}$ and probability mass functions $(7/10, 3/10)$ and $(1/2, 1/2)$; and a discrete covariate Z with values $\{b_1, b_2, b_3\}$ and probability mass functions $(3/5, 1/5, 1/5)$ and $(1/3, 1/3, 1/3)$. The probabilities of uncure, $p(0, b_j)$ and $p(1, b_j)$, with $j = 1, 2, 3$, are computed from $p(x, z)$ in (4.10) evaluated at $(a_i', \bar{b})$, with $i = 1, 2$, given in Table 4.38 under H_0, and at (a_1', b_j) and (a_2', b_j), given in Table 4.39 under H_1.

Scenario 1	$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$		$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$	
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2' = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -2.0064)$	0.6581
Scenario 2				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2' = 4.2229$	$(\bar{b} = -4.0960)$	0.6444	$(\bar{b} = -2.6671)$	0.7466
Scenario 3				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a_2' = -3.2466$	$(\bar{b} = -3.3932)$	0.5501	$(\bar{b} = -1.8585)$	0.4501

Table 4.38: Uncure probabilities under H_0, for Case 2 when X is binary with values $\{0, 1\}$ and Z is discrete with values $\{b_1, b_2, b_3\}$. The probability mass functions of Z are $(1/3, 1/3, 1/3)$ (first column), and $(3/5, 1/5, 1/5)$ (second column). See Remark in Section 4.6.2 for details.

Under the alternative hypothesis, we consider 3 situations (see Table 4.39). In the first one, $p(x, z)$ has high dependency on z only for $x = 0$. In the second and third situations, in which more power is expected, $p(x, z)$ has strong dependency on z for both $x = 0$ and $x = 1$.

	$b_1 = 0.6157$	$b_2 = -3.5434$	$b_3 = -7.7026$
Scenario 1 (Least favorable situation)			
$a_1' = -6.5585$	0.1	0.5000	0.9000
$a_2' = 0.9109$	0.7	0.6323	0.5590
Scenario 2 (Favorable situation a)			
$a_1' = -6.5585$	0.1	0.5000	0.9000
$a_2' = 4.2229$	0.9	0.6862	0.3470
Scenario 3 (Favorable situation b)			
$a_1' = -6.5585$	0.1	0.5000	0.9000
$a_2' = -3.2466$	0.3	0.5598	0.7905

Table 4.39: Uncure probabilities considered under H_1, for Case 2 when X is binary with values $\{0, 1\}$, and Z is discrete with values $\{b_1, b_2, b_3\}$. See Remark in Section 4.6.2 for details.

In Table 4.40 we can see the results under H_0. Note that even with small sample size ($n = 50$), the results are very close to the significance level, $\alpha = 0.05$.

			Scenario 1		Scenario 2		Scenario 3	
n	$(\Pi_x(0), \Pi_x(1))$	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	CvM	KS	CvM	KS	CvM	KS
50	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.0554	0.0580	0.0476	0.0516	0.0486	0.0512
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0472	0.0490	0.0442	0.0460	0.0432	0.0492
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.0552	0.0540	0.0550	0.0532	0.0562	0.0568
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0530	0.0490	0.0520	0.0510	0.0566	0.0538
100	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.0510	0.0534	0.0536	0.0528	0.0530	0.0542
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0478	0.0496	0.0478	0.0462	0.0530	0.0538
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.0518	0.0490	0.0614	0.0580	0.0564	0.0516
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0552	0.0544	0.0488	0.0464	0.0470	0.0484
200	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.0566	0.0510	0.0502	0.0508	0.0508	0.0532
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0488	0.0490	0.0502	0.0502	0.0452	0.0480
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.0490	0.0490	0.0490	0.0462	0.0534	0.0496
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0552	0.0510	0.0474	0.0480	0.0510	0.0474

Table 4.40: Size of the test under H_0 for Case 2 with X binary with values $\{0, 1\}$, and Z discrete with values $\{b_1, b_2, b_3\}$ (see Table 4.38).

The results under H_1 are shown in Table 4.41. In the least favorable situation, the power is only high when we consider $\Pi_x(a_1) = \Pi_x(a_2) = 1/2$, regardless of the distribution of Z. In the second and third situations, the results are much better. For example, in Scenario 3 we can see that the power is very close to 1 for $n = 200$, and it is also high if the probability mass function of X is $(1/2, 1/2)$.

			Scenario 1		Scenario 2		Scenario 3	
n	$(\Pi_x(0), \Pi_x(1))$	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	CvM	KS	CvM	KS	CvM	KS
50	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.3244	0.2572	0.2288	0.2100	0.6302	0.5846
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0652	0.0582	0.2734	0.2544	0.4466	0.4052
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.3212	0.2656	0.2374	0.2490	0.5918	0.5542
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0716	0.0614	0.2918	0.2946	0.4366	0.4036
100	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.5530	0.4414	0.4054	0.4002	0.8686	0.8202
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0730	0.0596	0.4464	0.3970	0.6826	0.6254
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.5126	0.4420	0.3878	0.4336	0.8226	0.7896
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0740	0.0664	0.4516	0.4488	0.6404	0.6020
200	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.8074	0.7172	0.6680	0.6920	0.9812	0.9680
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0678	0.0558	0.6922	0.6126	0.8972	0.8584
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.7670	0.7140	0.6286	0.7210	0.9688	0.9632
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0762	0.0612	0.6822	0.6606	0.8598	0.8366

Table 4.41: Power of the test under H_1 for Case 2 with X binary with values $\{0, 1\}$ and Z discrete with values $\{b_1, b_2, b_3\}$ (see Table 4.39).

X binary, Z binary

The covariates X and Z are binary variables, with values $\{0, 1\}$ both of them. We work with two different situations depending on the corresponding probability mass functions: in the first one, both for X and Z are $(1/2, 1/2)$, whereas in the second one, both are $(7/10, 3/10)$.

The four probabilities, $p(0,0), p(0,1), p(1,0)$ and $p(1,1)$, are computed evaluating the function $p(x,z)$ in (4.10) as follows: under H_1, at the points (a_1', b_1'), (a_1', b_2'), (a_2', b_1') and (a_2', b_2') respectively, given in Table 4.43; under H_0, $p(0,0) = p(0,1)$ is given by point $(a_1', \bar{b}_1)$ and $p(1,0) = p(1,1)$ by point $(a_2', \bar{b}_2)$ (see Table 4.42).

Under H_1, we consider 3 situations: in the first one, $p(x,z)$ has high dependency on z only for $x = 0$, whereas in the second and third situations (in which more power is expected), $p(x,z)$ has high dependency on z for both $x = 0$ and $x = 1$. Under the null hypothesis, the $\{a_1', a_2'\}$ associated to the values $\{0, 1\}$ of X are the same as those in the scenarios under H_1.

Scenario 1	$p(0,0) = p(0,1)$		$p(1,0) = p(1,1)$	
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = 0.9109$	$(\bar{b} = -3.7087)$	0.6295	$(\bar{b} = -5.3335)$	0.6013
Scenario 2				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = 4.2229$	$(\bar{b} = -4.3609)$	0.6295	$(\bar{b} = -5.6922)$	0.5129
Scenario 3				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = -3.2466$	$(\bar{b} = -3.3184)$	0.5452	$(\bar{b} = -4.8803)$	0.6434

Table 4.42: Uncure probabilities under H_0, for Case 2 when X and Z are binary with values $\{0, 1\}$ both of them, when the probability mass functions for both X and Z are $(1/2, 1/2)$ (first column), and $(7/10, 3/10)$ (second column). See Remark in Section 4.6.2 for details.

Table 4.44 shows the results under H_0, which are very close to the significance level, $\alpha = 0.05$, in the 3 situations and for the different sample sizes. The results under the alternative hypothesis are shown in Table 4.45. The highest powers are obtained when the values $\{0, 1\}$ are equiprobable for both X and Z.

	$b_1' = 0.6157$	$b_2' = -7.7027$
Scenario 1 (Least favorable situation)		
$a_1' = -6.5585$	0.1	0.9000
$a_2' = 0.9109$	0.7	0.5590
Scenario 2 (Favorable situation a)		
$a_1' = -6.5585$	0.1	0.9000
$a_2' = 4.2229$	0.9	0.3470
Scenario 3 (Favorable situation b)		
$a_1' = -6.5585$	0.1	0.9000
$a_2' = -3.2466$	0.3	0.7905

Table 4.43: Uncure probabilities considered under H_1, for Case 2 when X and Z are binary with values $\{0, 1\}$ both of them. See Remark in Section 4.6.2 for details.

		Scenario 1		Scenario 2		Scenario 3	
n	$(\Pi_x(0), \Pi_x(1))$	CvM	KS	CvM	KS	CvM	KS
50	$(1/2, 1/2)$	0.0616	0.0626	0.0546	0.0562	0.0544	0.0602
	$(7/10, 3/10)$	0.0394	0.0450	0.0486	0.0544	0.0434	0.0484
100	$(1/2, 1/2)$	0.0518	0.0502	0.0542	0.0544	0.0514	0.0530
	$(7/10, 3/10)$	0.0486	0.0510	0.0520	0.0542	0.0446	0.0476
200	$(1/2, 1/2)$	0.0544	0.0536	0.0496	0.0508	0.0540	0.0552
	$(7/10, 3/10)$	0.0458	0.0470	0.0472	0.0502	0.0506	0.0504

Table 4.44: Size of the test under H_0 for Case 2 with X and Z binary with values $\{0, 1\}$ both of them (see Table 4.42). The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3	
n	$(\Pi_x(0), \Pi_x(1))$	CvM	KS	CvM	KS	CvM	KS
50	$(1/2, 1/2)$	0.4896	0.4558	0.3480	0.4178	0.8330	0.8208
	$(7/10, 3/10)$	0.0682	0.0650	0.2726	0.2918	0.5704	0.5840
100	$(1/2, 1/2)$	0.7480	0.7342	0.5922	0.7130	0.9724	0.9658
	$(7/10, 3/10)$	0.066	0.0604	0.4754	0.4812	0.8046	0.8064
200	$(1/2, 1/2)$	0.9362	0.9416	0.8982	0.9520	0.9976	0.9978
	$(7/10, 3/10)$	0.0744	0.0706	0.7626	0.7334	0.9494	0.9458

Table 4.45: Power of the test under H_1 for Case 2 with X and Z binary with values $\{0, 1\}$ both of them (see Table 4.43). The probability mass function of Z equals that of X.

X binary, Z qualitative

X is a binary variable with values $\{0, 1\}$, and Z is qualitative with values $\{b_1, b_2, b_3\}$. The uncure probabilities, $p(0, b_j)$ and $p(1, b_j)$, with $j = 1, 2, 3$, are computed from the function $p(x, z)$ in (4.10), replacing x by a'_1 and a'_2 respectively, and z by the values $\bar{b}$ in Table 4.46 (under H_0) and b'_j in Table 4.47 (under H_1).

Scenario 1	$p(0,b_1) = p(0,b_2) = p(0,b_3)$		$p(1,b_1) = p(1,b_2) = p(1,b_3)$	
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -2.0064)$	0.6581
Scenario 2				
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = 4.2229$	$(\bar{b} = -4.0960)$	0.6444	$(\bar{b} = -2.6671)$	0.7466
Scenario 3				
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = -3.2466$	$(\bar{b} = -3.3932)$	0.5501	$(\bar{b} = -1.8585)$	0.4501

Table 4.46: Uncure probabilities under H_0, for Case 2 when X is binary with values $\{0, 1\}$, and Z is qualitative with values $\{b_1, b_2, b_3\}$. The probability mass functions of Z are $(1/3, 1/3, 1/3)$ (first column), and $(3/5, 1/5, 1/5)$ (second column). See Remark in Section 4.6.2 for details.

	$b'_1 = 0.6157$	$b'_2 = -3.5434$	$b'_3 = -7.7027$
Scenario 1 (Least favorable situation)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = 0.9109$	0.7	0.6323	0.5590
Scenario 2 (Favorable situation a)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = 4.2229$	0.9	0.6862	0.3470
Scenario 3 (Favorable situation b)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -3.2466$	0.3	0.5598	0.7905

Table 4.47: Uncure probabilities under H_1, for Case 2 when X is binary with values $\{0, 1\}$, and Z is qualitative with values $\{b_1, b_2, b_3\}$. See Remark in Section 4.6.2 for details.

The results under H_0 are shown in Table 4.48. For the 3 situations, the size of the test is very close to the significance level, $\alpha = 0.05$. The results under H_1 are shown in Table 4.49. In the least favorable and in the most favorable situations, the best results are obtained when we consider that the probability mass function of X is $(1/2, 1/2)$. Regarding the second situation, the powers are slightly higher for $(7/10, 3/10)$.

			Scenario 1		Scenario 2		Scenario 3	
n	$(\Pi_x(0), \Pi_x(1))$	$\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)$	CvM	KS	CvM	KS	CvM	KS
50	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.0494	0.0530	0.0450	0.0528	0.0468	0.0544
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0472	0.0516	0.0412	0.0484	0.0440	0.0476
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.0528	0.0528	0.0526	0.0504	0.0522	0.0530
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0444	0.0452	0.0494	0.0508	0.0548	0.0524
100	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.0508	0.0552	0.0484	0.0528	0.0510	0.0534
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0458	0.0478	0.0418	0.0484	0.0520	0.0550
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.0490	0.0504	0.0558	0.0542	0.0550	0.0502
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0526	0.0514	0.0464	0.0436	0.0470	0.0456
200	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.0472	0.0510	0.0482	0.0496	0.0460	0.0522
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0472	0.0474	0.0492	0.0520	0.0476	0.0476
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.0478	0.0470	0.0478	0.0452	0.0512	0.0500
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0526	0.0518	0.0468	0.0474	0.0482	0.0472

Table 4.48: Size of the test under H_0, for Case 2 with X binary with values $\{0, 1\}$, and Z qualitative with values $\{b_1, b_2, b_3\}$ (see Table 4.46).

			Scenario 1		Scenario 2		Scenario 3	
n	$(\Pi_x(0), \Pi_x(1))$	$\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)$	CvM	KS	CvM	KS	CvM	KS
50	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.2488	0.2212	0.1700	0.1790	0.5156	0.5248
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0570	0.0544	0.2066	0.2186	0.3566	0.3510
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.2882	0.2426	0.2122	0.2266	0.5456	0.5222
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0636	0.0582	0.2810	0.2780	0.3932	0.3772
100	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.4432	0.3770	0.2864	0.3292	0.8024	0.7768
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0592	0.0566	0.3350	0.3394	0.5820	0.5692
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.4644	0.4078	0.3372	0.4042	0.8020	0.7828
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0730	0.0640	0.4434	0.4368	0.6070	0.5848
200	$(1/2, 1/2)$	$(1/3, 1/3, 1/3)$	0.7066	0.6472	0.5244	0.6182	0.9602	0.9526
	$(7/10, 3/10)$	$(1/3, 1/3, 1/3)$	0.0604	0.0556	0.5800	0.5626	0.8346	0.8212
	$(1/2, 1/2)$	$(3/5, 1/5, 1/5)$	0.7360	0.6962	0.5640	0.6952	0.9622	0.9576
	$(7/10, 3/10)$	$(3/5, 1/5, 1/5)$	0.0676	0.0578	0.6564	0.6304	0.8368	0.8232

Table 4.49: Power of the test under H_1, for Case 2 with X binary with values $\{0, 1\}$, and Z qualitative with values $\{b_1, b_2, b_3\}$ (see Table 4.47).

X qualitative, *Z* continuous

The variable X is qualitative with values $\{a_1, a_2, a_3\}$, and Z is $U(-20, 20)$. Let a_i' be the numerical value x at which $p(x, z)$ in (4.10) is evaluated to get $p(a_i, z)$. We considered two scenarios depending on the values (a_1', a_2', a_3'): $(-3.6964, -1.3296, 1.0371)$ in the first scenario, and $(-7.4671, -1.3296, 4.8079)$ in the second one, with the following probability mass functions for each scenario: $(1/3, 1/3, 1/3)$ and $(3/5, 1/5, 1/5)$ (see Figure 4.9 for the probabilities of cure under H_1). The incidence, $1-p(a_i, z)$, $i = 1, 2, 3$, reduces, under H_0, to $1 - p(a_i')$, with $i = 1, 2, 3$, with $p(x)$ in (4.8), evaluated at the aforementioned values (a_1', a_2', a_3') (see Figure 4.8).

Table 4.50 shows the results under the null hypothesis, which are close to the significance level, $\alpha = 0.05$, for both scenarios. The results under the alternative hypothesis are given in Table 4.51. The power is higher if we consider that the probability mass function of X is $(3/5, 1/5, 1/5)$.

		Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.0502	0.0658	0.0498	0.0624
	$(3/5, 1/5, 1/5)$	0.0524	0.0696	0.0458	0.0616
100	$(1/3, 1/3, 1/3)$	0.0494	0.0632	0.0446	0.0574
	$(3/5, 1/5, 1/5)$	0.0468	0.0558	0.0392	0.0560
200	$(1/3, 1/3, 1/3)$	0.0496	0.0572	0.0486	0.0614
	$(3/5, 1/5, 1/5)$	0.0512	0.0550	0.0538	0.0612

Table 4.50: Size of the test for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$, and Z continuous with distribution $U(-20, 20)$, under the null hypothesis (see Figure 4.8).

		Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.3888	0.4148	0.4136	0.4866
	$(3/5, 1/5, 1/5)$	0.7196	0.7418	0.7436	0.7906
100	$(1/3, 1/3, 1/3)$	0.6552	0.6834	0.7058	0.7958
	$(3/5, 1/5, 1/5)$	0.9290	0.9348	0.9380	0.9554
200	$(1/3, 1/3, 1/3)$	0.9126	0.9280	0.9460	0.9742
	$(3/5, 1/5, 1/5)$	0.9938	0.9940	0.9984	0.9992

Table 4.51: Power of the test for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$, and Z continuous with distribution $U(-20, 20)$, under the alternative hypothesis (see Figure 4.9).

X qualitative, *Z* discrete

X is a qualitative variable with values $\{a_1, a_2, a_3\}$, and Z is discrete with values $\{b_1, b_2, b_3\}$. The cure probabilities, $1 - p(a_i, b_j)$, with $i, j = 1, 2, 3$, are computed from the function $p(x, z)$ in (4.10), replacing (a_i, b_j) with the numerical values $(a_i', \bar{b})$ given in Table 4.52 (under H_0), and with (a_i', b_j) given in Table 4.53 (under H_1). We work with two different situations depending on the corresponding probability mass functions: in the first one, both for X and Z are $(1/3, 1/3, 1/3)$, whereas in the second one, both are $(3/5, 1/5, 1/5)$.

Under H_1, the values of Z are $\{b_1, b_2, b_3\} = \{0.6157, -3.5434, -7.7026\}$, and $\{a_1', a_2', a_3'\}$ are chosen so $p(a_i, b_j)$, with $i, j = 1, 2, 3$, in (4.10), has high dependency

on b_j, with $j = 1, 2, 3$ only for $x = a_1$ (Scenario 1), for $x \in \{a_1, a_2\}$ (Scenario 2), and for $x \in \{a_1, a_2, a_3\}$ (Scenarios 3 and 4). The specific values of $\{a'_1, a'_2, a'_3\}$ in the four scenarios are given in Table 4.53. Note that in these last two scenarios more power is expected.

Under H_0, the values $\{a'_1, a'_2, a'_3\}$ in the four scenarios are the same as those under H_1. The probabilities of cure, $1 - p(a_i, b_1) = 1 - p(a_i, b_2) = 1 - p(a_i, b_3)$ are now given by $p(a'_i, \bar{b})$, $i = 1, 2, 3$ with p in (4.10) (see Table 4.52 for details). Two situations are considered within each scenario, one when the probability mass functions of X and Z are $(1/3, 1/3, 1/3)$, and the other when the probability mass functions are $(3/5, 1/5, 1/5)$.

Scenario 1	$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$		$p(a_i, b_1) = p(a_i, b_2) = p(a_i, b_3)$	
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = -1.1678$	$(\bar{b} = -3.4427)$	0.5943	$(\bar{b} = -1.7995)$	0.5566
$a'_3 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -2.0064)$	0.6581
Scenario 2				
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a'_3 = 0.9109$	$(\bar{b} = -3.6537)$	0.6304	$(\bar{b} = -1.9950)$	0.6583
Scenario 3				
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a'_3 = 4.2229$	$(\bar{b} = -4.0960)$	0.6444	$(\bar{b} = -2.6671)$	0.7466
Scenario 4				
$a'_1 = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -2.2879)$	0.3400
$a'_2 = -4.5690$	$(\bar{b} = -3.4350)$	0.5261	$(\bar{b} = -2.0006)$	0.3957
$a'_3 = -3.2466$	$(\bar{b} = -3.3932)$	0.5501	$(\bar{b} = -1.8585)$	0.4501

Table 4.52: Uncure probabilities considered under H_0, for Case 2 when X is qualitative with values $\{a_1, a_2, a_3\}$, and Z is discrete with values $\{b_1, b_2, b_3\}$. The probability mass functions for both X and Z are $(1/3, 1/3, 1/3)$ (first column) and $(3/5, 1/5, 1/5)$ (second column). See Remark in Section 4.6.2 for details.

The results under the null hypothesis are shown in Table 4.54. In the 4 scenarios, the size of the test is very close to the significance level, $\alpha = 0.05$. The results under the alternative hypothesis are shown in Table 4.55. Note that even when the probability mass function of X and Z is $(3/5, 1/5, 1/5)$ (the least favorable situation), the power of the test is considerably high. In the rest of the situations, the sample size has much influence, since the results are significantly higher when we work with large sample sizes.

	$b_1 = 0.6157$	$b_2 = -3.5434$	$b_3 = -7.7026$
Scenario 1 (Least favorable situation)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -1.1678$	0.5	0.5966	0.6862
$a'_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 2 (Intermediate situation)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.2	0.5360	0.8423
$a'_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 3 (Favorable situation a)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.2	0.5360	0.8423
$a'_3 = 4.2229$	0.9	0.6862	0.3470
Scenario 4 (Favorable situation b)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.2	0.5360	0.8423
$a'_3 = -3.2466$	0.3	0.5598	0.7905

Table 4.53: Uncure probabilities considered under H_1, for Case 2 when X is qualitative with values $\{a_1, a_2, a_3\}$, and Z is discrete with values $\{b_1, b_2, b_3\}$. See Remark in Section 4.6.2 for details.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.0468	0.0494	0.0454	0.0512	0.0544	0.0544	0.0486	0.0508
	$(3/5, 1/5, 1/5)$	0.0558	0.0554	0.0550	0.0552	0.0548	0.0536	0.0598	0.0526
100	$(1/3, 1/3, 1/3)$	0.0440	0.0504	0.0508	0.0536	0.0528	0.0564	0.0498	0.0510
	$(3/5, 1/5, 1/5)$	0.0514	0.0528	0.0520	0.0510	0.0530	0.0540	0.0530	0.0510
200	$(1/3, 1/3, 1/3)$	0.0570	0.0588	0.0534	0.0536	0.0528	0.0502	0.0488	0.0550
	$(3/5, 1/5, 1/5)$	0.0504	0.0528	0.0540	0.0542	0.0572	0.0518	0.0490	0.0504

Table 4.54: Size of the test under H_0 for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$, and Z discrete with values $\{b_1, b_2, b_3\}$ (see Table 4.52). The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.2258	0.2072	0.3818	0.3694	0.3220	0.3250	0.5812	0.5668
	$(3/5, 1/5, 1/5)$	0.5280	0.5062	0.5524	0.5394	0.5504	0.5362	0.6026	0.6024
100	$(1/3, 1/3, 1/3)$	0.3836	0.3314	0.6462	0.6090	0.5650	0.5818	0.8458	0.8322
	$(3/5, 1/5, 1/5)$	0.7690	0.7490	0.7916	0.7844	0.7820	0.7824	0.8070	0.8218
200	$(1/3, 1/3, 1/3)$	0.6236	0.5474	0.8926	0.8740	0.8378	0.8646	0.9804	0.9746
	$(3/5, 1/5, 1/5)$	0.9348	0.9252	0.9450	0.9466	0.9446	0.9484	0.9502	0.9636

Table 4.55: Power of the test under H_1 for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$ and Z discrete with values $\{b_1, b_2, b_3\}$ (see Table 4.53). The probability mass function of Z equals that of X.

X qualitative, Z binary

X is a qualitative variable with values $\{a_1, a_2, a_3\}$, and Z is binary with values $\{0, 1\}$. The uncure probabilities, $p(a_i, 0)$ and $p(a_i, 1)$, with $i = 1, 2, 3$, are computed from the function $p(x, z)$ in (4.10) evaluated at the numerical values $(a_i', \bar{b}_i)$, $i = 1, 2, 3$ under H_0 (see Table 4.56) and (a_i', b_1') and (a_i', b_2'), $i = 1, 2, 3$ under H_1 (see Table 4.57).

Scenario 1	$p(a_i,0) = p(a_i,1)$		$p(a_i,0) = p(a_i,1)$	
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = -1.1678$	$(\bar{b} = -3.3900)$	0.5931	$(\bar{b} = -5.0579)$	0.6303
$a_3' = 0.9109$	$(\bar{b} = -3.7087)$	0.6295	$(\bar{b} = -5.3335)$	0.6013
Scenario 2				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = -4.5690$	$(\bar{b} = -3.3809)$	0.5211	$(\bar{b} = -4.8281)$	0.6496
$a_3' = 0.9109$	$(\bar{b} = -3.7087)$	0.6295	$(\bar{b} = -5.3335)$	0.6013
Scenario 3				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = -4.5690$	$(\bar{b} = -3.3809)$	0.5211	$(\bar{b} = -4.8281)$	0.6496
$a_3' = 4.2229$	$(\bar{b} = -4.3609)$	0.6295	$(\bar{b} = -5.6922)$	0.5129
Scenario 4				
$a_1' = -6.5585$	$(\bar{b} = -3.5434)$	0.5000	$(\bar{b} = -4.7990)$	0.6600
$a_2' = -4.5690$	$(\bar{b} = -3.3809)$	0.5211	$(\bar{b} = -4.8281)$	0.6496
$a_3' = -3.2466$	$(\bar{b} = -3.3184)$	0.5452	$(\bar{b} = -4.8803)$	0.6434

Table 4.56: Uncure probabilities considered under H_0, for Case 2 when X is qualitative with values $\{a_1, a_2, a_3\}$, and Z is binary with values $\{0, 1\}$. The probability mass functions for Z are $(1/2, 1/2)$ (first column) and $(7/10, 3/10)$ (second column). See Remark in Section 4.6.2 for details.

The results under the null hypothesis are shown in Table 4.58. Regardless the sample size and the different values of $p(a_i, 0) = p(a_i, 1)$, with $i = 1, 2, 3$, the results are very close to the significance level, $\alpha = 0.05$.

Table 4.59 shows the results under the alternative hypothesis. In all situations except in the most favorable one (Scenario 4), the highest power is obtained when the probability mass fuction of X is $(3/5, 1/5, 1/5)$. Furthermore, with sample sizes $n = 100$ and $n = 200$, the power of the test is very close to 1 in the 4 scenarios and regardless the probability mass function of Z.

	$b'_1 = 0.6157$	$b'_2 = -7.7027$
Scenario 1 (Least favorable situation)		
$a'_1 = -6.5585$	0.1	0.9000
$a'_2 = -1.1678$	0.5	0.6862
$a'_3 = 0.9109$	0.7	0.5590
Scenario 2 (Intermediate situation)		
$a'_1 = -6.5585$	0.1	0.9000
$a'_2 = -4.5690$	0.2	0.8423
$a'_3 = 0.9109$	0.7	0.5590
Scenario 3 (Favorable situation a)		
$a'_1 = -6.5585$	0.1	0.9000
$a'_2 = -4.5690$	0.2	0.8423
$a'_3 = 4.2229$	0.9	0.3470
Scenario 4 (Favorable situation b)		
$a'_1 = -6.5585$	0.1	0.9000
$a'_2 = -4.5690$	0.2	0.8423
$a'_3 = -3.2466$	0.3	0.7905

Table 4.57: Uncure probabilities considered under H_1, for Case 2 when X is qualitative with values $\{a_1, a_2, a_3\}$ and Z is binary with values $\{0, 1\}$. See Remark in Section 4.6.2 for details.

X qualitative, *Z* qualitative

Let both X and Z be qualitative variables with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. The cure probabilities, $1 - p(a_i, b_j)$, with $i, j = 1, 2, 3$, are computed from the function $p(x, z)$ in (4.10) evaluated at the numerical values $(a'_i, \bar{b}_i)$, $i = 1, 2, 3$, given in Table 4.60 (under H_0), and (a'_i, b'_j), $i, j = 1, 2, 3$, given in Table 4.61 (under H_1). We work with two different situations depending on the corresponding probability mass functions: in the first one, both for X and Z are $(1/3, 1/3, 1/3)$, whereas in the second one, both are $(3/5, 1/5, 1/5)$.

Table 4.62 shows the results under the null hypothesis. In Scenario 1, the results are close to the significance level ($\alpha = 0.05$), except for the CvM test, with $n = 50, 100$ and when the probability mass function of X is $(1/3, 1/3, 1/3)$. In the other 3 scenarios, the results are very competent regardless the probability mass function of X.

The results under the alternative hypothesis are shown in Table 4.63. Note that in the 4 scenarios, the power is higher if we consider that the probability mass function of X is $(3/5, 1/5, 1/5)$.

			Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0444	0.0556	0.0410	0.0510
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0420	0.0492	0.0424	0.0494
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0574	0.0632	0.0522	0.0598
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0572	0.0588	0.0588	0.0606
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0428	0.0520	0.0446	0.0528
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0472	0.0498	0.0464	0.0504
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0502	0.0512	0.0522	0.0528
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0496	0.0522	0.0540	0.0552
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0450	0.0490	0.0464	0.0524
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0560	0.0576	0.0514	0.0548
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0578	0.0570	0.0468	0.0462
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0508	0.0492	0.0490	0.0520
			Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0382	0.0480	0.0424	0.0518
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0406	0.0474	0.0462	0.0512
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0534	0.0598	0.0498	0.0544
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0568	0.0576	0.0540	0.0542
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0446	0.0536	0.0428	0.0510
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0452	0.0472	0.0424	0.0494
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0498	0.0514	0.0546	0.0560
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0522	0.0540	0.0530	0.0540
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.0490	0.0530	0.0486	0.0528
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.0512	0.0532	0.0414	0.0456
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.0498	0.0506	0.0484	0.0494
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.0526	0.0510	0.0484	0.0466

Table 4.58: Size of the test under H_0, for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$, and Z binary with values $\{0, 1\}$ (see Table 4.56).

			Scenario 1		Scenario 2	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.2944	0.3290	0.5144	0.5604
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.3236	0.3318	0.5814	0.6144
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.7180	0.7340	0.7440	0.7694
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.7838	0.7836	0.8176	0.8336
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.4954	0.5136	0.7582	0.7942
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.5574	0.5506	0.8430	0.8694
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9024	0.9058	0.9270	0.9382
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9488	0.9474	0.9608	0.9652
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.7450	0.7560	0.9466	0.9608
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.8216	0.8212	0.9764	0.9834
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9886	0.9884	0.9900	0.9910
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9936	0.9944	0.9966	0.9974
			Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	$(\Pi_z(0), \Pi_z(1))$	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.4272	0.5228	0.7416	0.7774
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.5088	0.5976	0.7914	0.8134
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.7320	0.7644	0.7828	0.8112
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.8036	0.8206	0.8340	0.8564
100	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.6760	0.7922	0.9340	0.9440
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.7644	0.8744	0.9646	0.9704
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9164	0.9288	0.9360	0.9524
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9520	0.9604	0.9640	0.9750
200	$(1/3, 1/3, 1/3)$	$(7/10, 3/10)$	0.9202	0.9672	0.9944	0.9968
	$(1/3, 1/3, 1/3)$	$(1/2, 1/2)$	0.9654	0.9922	0.9988	0.9990
	$(3/5, 1/5, 1/5)$	$(7/10, 3/10)$	0.9920	0.9936	0.9932	0.9962
	$(3/5, 1/5, 1/5)$	$(1/2, 1/2)$	0.9974	0.9982	0.9982	0.9992

Table 4.59: Power of the test under H_1 for Case 2 with X qualitative with values $\{a_1, a_2, a_3\}$ and Z binary with values $\{0, 1\}$ (see Table 4.57).

Scenario 1	$p(a_i,b_1)=p(a_i,b_2)=p(a_i,b_3)$		$p(a_i,b_1)=p(a_i,b_2)=p(a_i,b_3)$	
$a'_1=-6.5585$	$(\bar{b}=-3.5434)$	0.5000	$(\bar{b}=-2.2879)$	0.3400
$a'_2=-1.1678$	$(\bar{b}=-3.4427)$	0.5943	$(\bar{b}=-1.7995)$	0.5566
$a'_3=0.9109$	$(\bar{b}=-3.6537)$	0.6304	$(\bar{b}=-2.0064)$	0.6581
Scenario 2				
$a'_1=-6.5585$	$(\bar{b}=-3.5434)$	0.5000	$(\bar{b}=-2.2879)$	0.3400
$a'_2=-4.5690$	$(\bar{b}=-3.4350)$	0.5261	$(\bar{b}=-2.0006)$	0.3957
$a'_3=0.9109$	$(\bar{b}=-3.6537)$	0.6304	$(\bar{b}=-1.995)$	0.6583
Scenario 3				
$a'_1=-6.5585$	$(\bar{b}=-3.5434)$	0.5000	$(\bar{b}=-2.2879)$	0.3400
$a'_2=-4.5690$	$(\bar{b}=-3.4350)$	0.5261	$(\bar{b}=-2.0006)$	0.3957
$a'_3=4.2229$	$(\bar{b}=-4.0960)$	0.6444	$(\bar{b}=-2.6671)$	0.7466
Scenario 4				
$a'_1=-6.5585$	$(\bar{b}=-3.5434)$	0.5000	$(\bar{b}=-2.2879)$	0.3400
$a'_2=-4.5690$	$(\bar{b}=-3.4350)$	0.5261	$(\bar{b}=-2.0006)$	0.3957
$a'_3=-3.2466$	$(\bar{b}=-3.3932)$	0.5501	$(\bar{b}=-1.8585)$	0.4501

Table 4.60: Uncure probabilities considered under H_0, for Case 2 when X and Z are qualitative with values $\{a_1,a_2,a_3\}$ and $\{b_1,b_2,b_3\}$, respectively. The probability mass functions for both X and Z are $(1/3,1/3,1/3)$ (first column), and $(3/5,1/5,1/5)$ (second column). See Remark in Section 4.6.2 for details.

	$b'_1=0.6157$	$b'_2=-3.5434$	$b'_3=-7.7026$
Scenario 1 (Least favorable situation)			
$a'_1=-6.5585$	0.1	0.5000	0.9000
$a'_2=-1.1678$	0.5	0.5966	0.6862
$a'_3=0.9109$	0.7	0.6323	0.5590
Scenario 2 (Intermediate situation)			
$a'_1=-6.5585$	0.1	0.5000	0.9000
$a'_2=-4.5690$	0.2	0.5360	0.8423
$a'_3=0.9109$	0.7	0.6323	0.5590
Scenario 3 (Favorable situation a)			
$a'_1=-6.5585$	0.1	0.5000	0.9000
$a'_2=-4.5690$	0.2	0.5360	0.8423
$a'_3=4.2229$	0.9	0.6862	0.3470
Scenario 4 (Favorable situation b)			
$a'_1=-6.5585$	0.1	0.5000	0.9000
$a'_2=-4.5690$	0.2	0.5360	0.8423
$a'_3=-3.2466$	0.3	0.5598	0.7905

Table 4.61: Uncure probabilities considered under H_1, for Case 2 when X and Z are qualitative with values $\{a_1,a_2,a_3\}$ and $\{b_1,b_2,b_3\}$, respectively. See Remark in Section 4.6.2 for details.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.0392	0.0490	0.0452	0.0548	0.0480	0.0558	0.0432	0.0526
	$(3/5, 1/5, 1/5)$	0.0496	0.0528	0.0490	0.0510	0.0524	0.0524	0.0530	0.0512
100	$(1/3, 1/3, 1/3)$	0.0392	0.0512	0.0500	0.0552	0.0472	0.0562	0.0452	0.0546
	$(3/5, 1/5, 1/5)$	0.0524	0.0502	0.0454	0.0496	0.0520	0.0546	0.0514	0.0512
200	$(1/3, 1/3, 1/3)$	0.0478	0.0552	0.0496	0.0588	0.0470	0.0496	0.0486	0.0526
	$(3/5, 1/5, 1/5)$	0.0496	0.0520	0.0540	0.0544	0.0532	0.0514	0.0506	0.0490

Table 4.62: Size of the test under H_0 for Case 2 with X and Z qualitative with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively (see Table 4.60). The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.1658	0.1802	0.2968	0.3224	0.2410	0.2780	0.4586	0.5164
	$(3/5, 1/5, 1/5)$	0.4874	0.4818	0.5054	0.5104	0.5056	0.5126	0.5482	0.5740
100	$(1/3, 1/3, 1/3)$	0.2932	0.2888	0.5358	0.5478	0.4556	0.5206	0.7584	0.7896
	$(3/5, 1/5, 1/5)$	0.7324	0.7220	0.7536	0.7612	0.7500	0.7616	0.7774	0.8060
200	$(1/3, 1/3, 1/3)$	0.5128	0.4854	0.8232	0.8324	0.7350	0.8266	0.9578	0.9640
	$(3/5, 1/5, 1/5)$	0.9218	0.9142	0.9340	0.9382	0.9354	0.9408	0.9412	0.9516

Table 4.63: Power of the test under H_1 for Case 2 with X and Z qualitative with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively (see Table 4.61). The probability mass function of Z equals that of X.

Computational summary in Case 2

Table 4.64 shows a complete summary of the simulation studies, together with the computational times, considered for Case 2. Note that the computational times neither vary with the distributions of X and Z, nor with the different scenarios. All the procedures were coded in R language and run in the computers of the Department of Mathematics, at UDC.

X	Z	Scenario	Computational time
Continuous	Continuous	Model 1	597463 sec (165.96 h)
		Model 2	605182 sec (168.11 h)
	Discrete	Scenario 1	471430 sec (130.95 h)
		Scenario 2	465230 sec (129.23 h)
	Binary	Scenario 1	459990 sec (127.78 h)
		Scenario 2	466280 sec (129.52 h)
	Qualitative		425990 sec (118.33 h)
Discrete	Continuous	Scenario 1	18040 sec (5.01 h)
		Scenario 2	18230 sec (5.06 h)
	Discrete	Scenario 1, 2, 3, 4	11300 sec (3.14 h)
	Binary	Scenario 1, 2, 3, 4	9096 sec (2.53 h)
	Qualitative	Scenario 1, 2, 3, 4	43696 sec (12.14 h)
Binary	Continuous	Scenario 1	13661 sec (3.80 h)
		Scenario 2	13489 sec (3.75 h)
	Discrete	Scenario 1, 2, 3	7496 sec (2.08 h)
	Binary	Scenario 1, 2, 3	4945 sec (1.37 h)
	qualitative	Scenario 1, 2, 3	26633 (7.40 h)
Qualitative	Continuous	Scenario 1	156413 sec (43.45 h)
		Scenario 2	154693 sec (42.97 h)
	Discrete	Scenario 1, 2, 3, 4	11694 sec (3.25 h)
	Binary	Scenario 1, 2, 3, 4	9786 sec (2.72 h)
	Qualitative	Scenario 1, 2, 3, 4	30780 sec (8.55 h)

Table 4.64: Computational times for simulations in Case 2, considering sample size $n = 100$, $\kappa = 5000$ trials and $B = 2000$ bootstrap resamples. Note that in cases where a continuous covariate is involved, the times are given for only one bandwidth.

4.6.3 Case 1 with high dimensional covariate vector $\boldsymbol{Z}$

In the case of an m-dimensional set of covariates $\mathbf{W} = \mathbf{Z}$, we test each of them separately to obtain m p-values associated to each $H_0^1, \ldots, H_0^m$ in (4.3). For the sake of simplicity, all the covariates are binary with two specific values between -20 and 20, given in columns z_1 and z_2 in Tables 4.65-4.68. The censoring variable follows an exponential distribution with mean 1/0.3. Moreover, we work with different

sample sizes: $n = 50$, $n = 100$ and $n = 200$, considering $\kappa = 1000$ trials, $B = 1000$ bootstrap resamples and significance level $\alpha = 0.05$.

Two different mixture cure models are considered, both with the form:

$$S(t|w_1, \ldots, w_m) = 1 - p(w_1, \ldots, w_m) + p(w_1, \ldots, w_m) \cdot S_0(t|w_1, \ldots, w_m).$$

In Model 1, under the alternative hypothesis, only the covariates $W_{s_1}, \ldots, W_{s_j}$, with $s_1, \ldots, s_j \in \{1, \ldots, m\}$ are influencing the cure rate. We also assume that the uncure probability is of multiplicative form:

$$p(w_1, \ldots, w_m) = p(w_{s_1}) \times \cdots \times p(w_{s_j}),$$

where

$$p(w_{s_i}) = \frac{\exp\left(0.476 + 0.358 w_{s_i}\right)}{1 + \exp\left(0.476 + 0.358 w_{s_i}\right)}.$$

Furthermore,

$$S_0(t|w_1, \ldots, w_m) = \begin{cases} \frac{\exp(-\lambda(w_{s_1}, \ldots, w_{s_j})t) - \exp(-\lambda(w_{s_1}, \ldots, w_{s_j})4.605)}{1 - \exp(-\lambda(w_{s_1}, \ldots, w_{s_j})4.605)}, & \text{if } t \leq 4.605 \\ 0, & \text{if } t > 4.605 \end{cases},$$

where $\lambda\left(w_{s_1}, \ldots, w_{s_j}\right) = \exp\left(\frac{w_{s_1}+20}{40} \times \cdots \times \frac{w_{s_j}+20}{40}\right)$.

In Model 2, the uncure probability is defined under H_1 by:

$$p(w_1, \ldots, w_m) = \frac{\exp(\beta_0 + \beta_1 w_{s_1} + \beta_2 w_{s_2} + \beta_{12} w_{s_1} w_{s_2})}{1 + \exp(\beta_0 + \beta_1 w_{s_1} + \beta_2 w_{s_2} + \beta_{12} w_{s_1} w_{s_2})},$$

where $\beta_0 = 0.476$, $\beta_1 = 0.358$, $\beta_2 = 0.225$, $\beta_{12} = 0.195$. Furthermore,

$$S_0(t|w_1, \ldots, w_m) = \begin{cases} \frac{\exp(-\lambda(w_{s_1}, w_{s_2})t) - \exp(-\lambda(w_{s_1}, w_{s_2})4.605)}{1 - \exp(-\lambda(w_{s_1}, w_{s_2})4.605)}, & \text{if } t \leq 4.605 \\ 0, & \text{if } t > 4.605 \end{cases},$$

with $\lambda\left(w_{s_1}, w_{s_2}\right) = \exp\left(\frac{w_{s_1}+20}{40} \times \frac{w_{s_2}+20}{40}\right)$.

We consider 4 simulation studies and in each one of them, we work with different scenarios. In simulation studies 1, 2 and 3, we consider a covariate vector $\mathbf{W}$ with dimension 10 and we work with Model 1. In simulation study 4, we consider that $\mathbf{W}$ has dimension 100 and we use Model 2. Note that all the covariates are equally distributed, binary with values z_1 and z_2 and corresponding $p(z_1)$ and $p(z_2)$

given in Tables 4.65-4.68 for the different scenarios. For the scenario with influential variables $Z_{s_1}, \ldots, Z_{s_j}$, the probability of cure associated to each scenario can be computed as $1 - p(z_{s_1}, \ldots, z_{s_j}) = 1 - p(z_{s_1}) \times \cdots \times p(z_{s_j})$ (not reported in the Tables 4.65-4.67). For example, in Scenario a we should consider the probabilities: $p(-1.3, -1.3, -1.3) = 0.5027 \times 0.5027 \times 0.5027 = 0.1270$, $p(20, -1.3, -1.3) = 0.9995 \times 0.5027 \times 0.5027 = 0.2525$, $\ldots$, $p(20, 20, 20) = 0.9995 \times 0.9995 \times 0.9995 = 0.9986$.

Scenario	Influential variables	z_1	z_2	$(\Pi_z(z_1), \Pi_z(z_2))$	$p(z_1)$	$p(z_2)$	% cens.	% cure
a	$Z_1,\ Z_3,\ Z_5$	-1.3	20	$(1/2, 1/2)$	0.5027	0.9995	63.8	57.6
b	$Z_5, \ldots,\ Z_9$	-1.3	20	$(1/2, 1/2)$	0.5027	0.9995	80.3	76.2
c	$Z_1,\ Z_3,\ Z_5$	-7.4	20	$(1/2, 1/2)$	0.1022	0.9995	85.5	83.5
d	$Z_5, \ldots,\ Z_9$	-7.4	20	$(1/2, 1/2)$	0.1022	0.9995	95.5	94.9

Table 4.65: Scenarios a-d, considering Model 1, for simulation study 1.

Scenario	Influential variables	z_1	z_2	$(\Pi_z(z_1), \Pi_z(z_2))$	$p(z_1)$	$p(z_2)$	% cens.	% cure
e	$Z_1,\ Z_3,\ Z_5$	-1.3	20	$(1/4, 3/4)$	0.5027	0.9995	41.2	32.9
f	$Z_5, \ldots,\ Z_9$	-1.3	20	$(1/4, 3/4)$	0.5027	0.9995	55.4	48.4
g	$Z_1,\ Z_3,\ Z_5$	-7.4	20	$(1/4, 3/4)$	0.1022	0.9995	58.4	53.4
h	$Z_5, \ldots,\ Z_9$	-7.4	20	$(1/4, 3/4)$	0.1022	0.9995	75.1	72.0

Table 4.66: Scenarios e-h, considering Model 1, for simulation study 2.

Scenario	Influential variables	z_1	z_2	$(\Pi_z(z_1), \Pi_z(z_2))$	$p(z_1)$	$p(z_2)$	% cens.	% cure
i	Z_1	-1.3	20	$(1/4, 3/4)$	0.5027	0.9995	22.0	12.5
j	Z_1, Z_3	-1.3	20	$(1/4, 3/4)$	0.5027	0.9995	32.2	23.3
k	Z_1	-7.4	20	$(1/4, 3/4)$	0.1022	0.9995	30.4	22.5
l	Z_1, Z_3	-7.4	20	$(1/4, 3/4)$	0.1022	0.9995	46.2	39.8

Table 4.67: Scenarios i-l, considering Model 1, for simulation study 3.

Scenario	Influential variables	z_1	z_2	$(\Pi_z(z_1), \Pi_z(z_2))$	$p(z_1)$	$p(z_2)$	% cens.	% cure
m	Z_1, Z_2	-7	-3	$(1/4, 3/4)$	0.1161	0.3548	42.3	27.7
n	Z_1, Z_2	-7	-3	$(1/2, 1/2)$	0.1161	0.3548	34.4	17.7

Table 4.68: Scenarios m and n, considering Model 2, for simulation study 4.

Simulation study 1:
In Scenarios a and c, there are 3 influential covariates: Z_1, Z_3 and Z_5; whereas in Scenarios b and d, there are 5 influential covariates: Z_5, Z_6, Z_7, Z_8 and Z_9. Table 4.65 shows the values of z_1 and z_2 and the probability mass function for each scenario.

Tables 4.69, 4.70, 4.71 and 4.72 show the results obtained in Scenarios a, b, c and d, respectively. The behavior of the tests is satisfactory in Scenarios a (Table 4.69) and c (Table 4.71). For example, for sample size $n = 200$, the rejection percentage of the influential covariates is around 25% and 35%, respectively. On the contrary, in Scenarios b and d lower power is obtained.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	5.5	6.1	0.2946	0.2951
50	Z_2	0.8	1.0	0.4848	0.4863
50	Z_3	6.9	7.2	0.3089	0.3098
50	Z_4	0.8	1.0	0.5047	0.5065
50	Z_5	5.5	5.7	0.3043	0.3052
50	Z_6	0.7	0.6	0.4879	0.4893
50	Z_7	0.4	0.5	0.5066	0.5079
50	Z_8	0.2	0.2	0.4982	0.4997
50	Z_9	0.6	0.6	0.4975	0.4989
50	Z_{10}	0.6	0.9	0.4899	0.4913
100	Z_1	13.3	13.3	0.2149	0.2153
100	Z_2	0.4	0.5	0.4852	0.4861
100	Z_3	12.9	13.2	0.2273	0.2279
100	Z_4	0.2	0.2	0.5077	0.5086
100	Z_5	13.1	13.0	0.2336	0.2340
100	Z_6	0.8	1.0	0.4808	0.4815
100	Z_7	0.7	0.6	0.4942	0.4950
100	Z_8	0.2	0.1	0.4992	0.5000
100	Z_9	0.5	0.4	0.5084	0.5092
100	Z_{10}	0.6	0.7	0.5030	0.5039
200	Z_1	25.6	26.1	0.1364	0.1364
200	Z_2	0.3	0.4	0.5124	0.5127
200	Z_3	28.3	28.9	0.1399	0.1398
200	Z_4	0.4	0.4	0.5074	0.5079
200	Z_5	25.2	24.8	0.1254	0.1256
200	Z_6	0.2	0.2	0.4973	0.4978
200	Z_7	0.3	0.3	0.4983	0.4987
200	Z_8	0.7	0.7	0.5049	0.5052
200	Z_9	0.3	0.4	0.5091	0.5095
200	Z_{10}	0.5	0.5	0.4848	0.4853

Table 4.69: Results obtained in Scenario *a*, where Z_1, Z_3 and Z_5 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	0.3	0.7	0.4881	0.4892
50	Z_2	0.7	0.7	0.5014	0.5030
50	Z_3	0.6	0.6	0.5010	0.5025
50	Z_4	0.6	0.6	0.4921	0.4932
50	Z_5	2.7	2.7	0.4255	0.4268
50	Z_6	2.2	2.5	0.4364	0.4376
50	Z_7	1.5	1.7	0.4169	0.4181
50	Z_8	2.3	2.6	0.4324	0.4336
50	Z_9	1.9	2.2	0.4191	0.4208
50	Z_{10}	1.0	1.1	0.4952	0.4966
100	Z_1	1.3	1.2	0.4815	0.4825
100	Z_2	0.5	0.6	0.4957	0.4965
100	Z_3	0.6	0.8	0.4892	0.4897
100	Z_4	0.5	0.6	0.5071	0.5078
100	Z_5	2.5	2.8	0.3829	0.3838
100	Z_6	3.0	3.0	0.3972	0.3981
100	Z_7	3.3	3.0	0.3868	0.3876
100	Z_8	2.7	2.6	0.4011	0.4019
100	Z_9	4.2	4.1	0.3890	0.3897
100	Z_{10}	0.7	0.9	0.4932	0.4940
200	Z_1	0.4	0.4	0.4917	0.4922
200	Z_2	0.2	0.2	0.4930	0.4934
200	Z_3	0.6	0.6	0.4857	0.4860
200	Z_4	0.2	0.2	0.5001	0.5005
200	Z_5	6.3	6.5	0.3444	0.3447
200	Z_6	5.5	5.4	0.3314	0.3316
200	Z_7	5.1	5.0	0.3516	0.3518
200	Z_8	4.1	4.5	0.3422	0.3425
200	Z_9	4.5	5.0	0.3302	0.3303
200	Z_{10}	0.5	0.5	0.5052	0.5055

Table 4.70: Results obtained in Scenario b, where Z_5, Z_6, Z_7, Z_8 and Z_9 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	10.5	10.8	0.2462	0.2468
50	Z_2	0.5	0.6	0.4737	0.4753
50	Z_3	9.0	9.4	0.2603	0.2615
50	Z_4	0.8	0.8	0.4959	0.4974
50	Z_5	10.3	11.2	0.2390	0.2397
50	Z_6	0.6	0.8	0.4771	0.4786
50	Z_7	0.8	0.9	0.4790	0.4804
50	Z_8	0.8	0.8	0.5017	0.5037
50	Z_9	0.7	0.8	0.4861	0.4877
50	Z_{10}	1.0	1.0	0.4928	0.4943
100	Z_1	21.3	21.3	0.1713	0.1717
100	Z_2	0.2	0.3	0.4964	0.4971
100	Z_3	17.2	18.3	0.1852	0.1857
100	Z_4	0.4	0.4	0.5061	0.5069
100	Z_5	19.7	19.9	0.1710	0.1716
100	Z_6	0.3	0.4	0.5019	0.5028
100	Z_7	0.9	0.9	0.4975	0.4984
100	Z_8	0.7	0.8	0.5085	0.5094
100	Z_9	1.0	0.9	0.5045	0.5053
100	Z_{10}	0.7	0.7	0.4698	0.4708
200	Z_1	35.4	35.0	0.1025	0.1025
200	Z_2	0.7	0.7	0.5111	0.5116
200	Z_3	34.6	35.4	0.0959	0.0960
200	Z_4	0.6	0.7	0.5064	0.5067
200	Z_5	37.4	37.3	0.0903	0.0903
200	Z_6	0.8	0.9	0.5002	0.5005
200	Z_7	0.2	0.1	0.4806	0.4807
200	Z_8	0.9	1.2	0.4997	0.5000
200	Z_9	0.6	0.4	0.5003	0.5005
200	Z_{10}	0.7	0.7	0.5072	0.5078

Table 4.71: Results obtained in Scenario c, where Z_1, Z_3 and Z_5 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	0.5	0.9	0.4799	0.4812
50	Z_2	0.5	0.7	0.4837	0.4854
50	Z_3	0.1	0.0	0.4769	0.4783
50	Z_4	0.4	0.5	0.4635	0.4649
50	Z_5	1.1	1.3	0.4269	0.4283
50	Z_6	1.1	1.3	0.4264	0.4282
50	Z_7	1.1	1.3	0.4175	0.4185
50	Z_8	0.8	0.9	0.4235	0.4248
50	Z_9	1.1	1.3	0.4192	0.4208
50	Z_{10}	0.6	0.8	0.4593	0.4613
100	Z_1	0.6	0.6	0.4922	0.4928
100	Z_2	0.3	0.4	0.4884	0.4890
100	Z_3	0.7	0.7	0.4976	0.4982
100	Z_4	0.9	0.7	0.4811	0.4818
100	Z_5	1.5	1.6	0.4166	0.4174
100	Z_6	2.7	2.8	0.4120	0.4128
100	Z_7	1.7	1.9	0.4187	0.4197
100	Z_8	1.9	1.9	0.4149	0.4156
100	Z_9	1.7	2.0	0.4165	0.4171
100	Z_{10}	0.7	0.9	0.4909	0.4915
200	Z_1	0.9	1.0	0.4860	0.4865
200	Z_2	0.2	0.2	0.5034	0.5038
200	Z_3	0.7	0.7	0.4955	0.4958
200	Z_4	0.9	1.0	0.4865	0.4868
200	Z_5	4.2	4.2	0.3764	0.3768
200	Z_6	2.9	2.9	0.3910	0.3912
200	Z_7	2.8	2.6	0.4095	0.4100
200	Z_8	2.5	2.6	0.3767	0.3770
200	Z_9	3.5	3.7	0.3984	0.3988
200	Z_{10}	0.6	0.6	0.5070	0.5073

Table 4.72: Results obtained in Scenario d, where Z_5, Z_6, Z_7, Z_8 and Z_9 are influential.

Simulation study 2:
Similarly as in simulation study 1, in Scenarios e and g, there are 3 influential covariates: Z_1, Z_3 and Z_5; whereas in Scenarios f and h, there are 5 influential covariates: Z_5, Z_6, Z_7, Z_8 and Z_9. Table 4.66 shows the values of z_1 and z_2, and the probability mass function for each scenario. Note that in this case, since the probability mass function of Z is $(1/4, 3/4)$, there are fewer censored data than in simulation study 1, where $(\Pi_z(z_1), \Pi_z(z_2)) = (1/2, 1/2)$.

Tables 4.73, 4.74, 4.75 and 4.76 show the results obtained in Scenarios e, f, g and h, respectively. Note that the percentages in Scenarios f (Table 4.74) and h (Table 4.76) are very similar to the ones in Scenarios a and c, where there are two less influential covariates.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	14.2	19.7	0.1974	0.1961
50	Z_2	0.3	0.6	0.4855	0.4894
50	Z_3	13.7	18.8	0.1995	0.1978
50	Z_4	0.2	0.5	0.4948	0.4989
50	Z_5	14.0	17.9	0.2024	0.2012
50	Z_6	0.0	0.2	0.4773	0.4808
50	Z_7	0.5	1.0	0.4761	0.4798
50	Z_8	0.2	0.6	0.4786	0.4822
50	Z_9	0.2	0.2	0.4971	0.5007
50	Z_{10}	0.3	0.7	0.4834	0.4872
100	Z_1	31.6	36.2	0.1075	0.1063
100	Z_2	0.1	0.4	0.4823	0.4844
100	Z_3	31.7	35.4	0.1174	0.1164
100	Z_4	0.1	0.2	0.4860	0.4882
100	Z_5	31.2	35.6	0.1305	0.1295
100	Z_6	0.7	0.8	0.4990	0.5014
100	Z_7	0.4	0.7	0.5006	0.5026
100	Z_8	0.5	0.5	0.5004	0.5030
100	Z_9	0.5	0.6	0.4873	0.4895
100	Z_{10}	0.3	0.6	0.4968	0.4990
200	Z_1	59.4	61.8	0.0420	0.0413
200	Z_2	0.4	0.5	0.4995	0.5006
200	Z_3	59.7	61.4	0.0418	0.0413
200	Z_4	0.6	0.5	0.4766	0.4778
200	Z_5	58.7	61.0	0.0432	0.0427
200	Z_6	0.3	0.4	0.5112	0.5124
200	Z_7	0.5	0.8	0.4998	0.5009
200	Z_8	0.5	0.5	0.5016	0.5029
200	Z_9	0.8	1.0	0.4981	0.4994
200	Z_{10}	0.4	0.4	0.5180	0.5193

Table 4.73: Results obtained in Scenario e, where Z_1, Z_3 and Z_5 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	0.3	0.4	0.4852	0.4897
50	Z_2	0.2	0.8	0.4883	0.4921
50	Z_3	0.3	0.6	0.4831	0.4870
50	Z_4	0.4	0.5	0.4887	0.4923
50	Z_5	3.8	5.6	0.3273	0.3279
50	Z_6	4.3	7.0	0.3198	0.3210
50	Z_7	5.5	8.5	0.3066	0.3074
50	Z_8	5.2	7.9	0.3222	0.3235
50	Z_9	4.1	6.2	0.3165	0.3172
50	Z_{10}	0.5	0.6	0.4708	0.4745
100	Z_1	0.7	1.2	0.5015	0.5036
100	Z_2	0.3	0.5	0.5102	0.5127
100	Z_3	0.5	0.6	0.4820	0.4842
100	Z_4	0.2	0.4	0.4872	0.4895
100	Z_5	11.0	13.0	0.2395	0.2396
100	Z_6	11.6	14.3	0.2364	0.2369
100	Z_7	11.1	14.1	0.2561	0.2564
100	Z_8	10.9	12.7	0.2361	0.2364
100	Z_9	11.7	13.2	0.2340	0.2341
100	Z_{10}	0.3	0.4	0.4863	0.4884
200	Z_1	0.3	0.3	0.4996	0.5007
200	Z_2	0.1	0.3	0.4919	0.4931
200	Z_3	0.4	0.6	0.4975	0.4986
200	Z_4	0.5	0.7	0.4946	0.4957
200	Z_5	24.7	27.0	0.1664	0.1660
200	Z_6	25.7	28.2	0.1627	0.1624
200	Z_7	21.7	24.4	0.1540	0.1537
200	Z_8	24.7	25.9	0.1470	0.1465
200	Z_9	24.6	25.5	0.1454	0.1449
200	Z_{10}	0.5	0.7	0.4856	0.4869

Table 4.74: Results obtained in Scenario *f*, where Z_5, Z_6, Z_7, Z_8 and Z_9 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	26.9	35.9	0.1035	0.1004
50	Z_2	0.1	0.3	0.4880	0.4929
50	Z_3	26.8	35.4	0.1029	0.0999
50	Z_4	0.2	0.7	0.4926	0.4972
50	Z_5	27.6	35.5	0.1043	0.1012
50	Z_6	0.2	0.4	0.4849	0.4893
50	Z_7	0.4	0.8	0.4857	0.4908
50	Z_8	0.2	0.6	0.4702	0.4746
50	Z_9	0.3	0.8	0.4825	0.4870
50	Z_{10}	0.1	0.7	0.4833	0.4877
100	Z_1	58.8	64.8	0.0386	0.0373
100	Z_2	0.4	0.6	0.4935	0.4957
100	Z_3	56.8	60.6	0.0470	0.0457
100	Z_4	0.6	1.0	0.4839	0.4864
100	Z_5	57.7	60.7	0.0404	0.0392
100	Z_6	0.6	1.0	0.5018	0.5041
100	Z_7	0.6	1.0	0.4802	0.4824
100	Z_8	0.4	0.6	0.4897	0.4921
100	Z_9	0.3	0.3	0.5174	0.5198
100	Z_{10}	0.6	0.9	0.5040	0.5063
200	Z_1	88.7	90.1	0.0062	0.0059
200	Z_2	0.8	1.0	0.4976	0.4990
200	Z_3	86.0	87.3	0.0090	0.0087
200	Z_4	0.5	0.5	0.5043	0.5056
200	Z_5	88.7	89.8	0.0084	0.0081
200	Z_6	0.5	0.7	0.5031	0.5043
200	Z_7	0.8	0.8	0.4981	0.4996
200	Z_8	0.6	0.6	0.4879	0.4889
200	Z_9	0.6	0.6	0.4813	0.4823
200	Z_{10}	0.9	1.0	0.4967	0.4977

Table 4.75: Results obtained in Scenario g, where Z_1, Z_3 and Z_5 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	0.4	0.6	0.4762	0.4806
50	Z_2	0.0	0.2	0.4803	0.4843
50	Z_3	0.0	0.4	0.4889	0.4938
50	Z_4	0.3	0.5	0.4858	0.4900
50	Z_5	4.3	7.9	0.2575	0.2569
50	Z_6	5.1	7.9	0.2470	0.2462
50	Z_7	6.1	10.1	0.2426	0.2424
50	Z_8	3.9	6.2	0.2534	0.2534
50	Z_9	4.8	7.2	0.2618	0.2621
50	Z_{10}	0.6	0.9	0.4851	0.4899
100	Z_1	0.8	0.9	0.5028	0.5054
100	Z_2	0.4	0.7	0.4911	0.4932
100	Z_3	0.5	0.6	0.4856	0.4877
100	Z_4	0.9	1.0	0.4814	0.4836
100	Z_5	15.8	18.5	0.1759	0.1754
100	Z_6	16.6	19.5	0.1802	0.1799
100	Z_7	16.3	19.3	0.1799	0.1793
100	Z_8	14.3	17.7	0.1712	0.1703
100	Z_9	13.8	17.6	0.1838	0.1833
100	Z_{10}	0.8	0.8	0.4825	0.4847
200	Z_1	0.3	0.4	0.5050	0.5061
200	Z_2	0.4	0.4	0.4962	0.4973
200	Z_3	0.4	0.4	0.4908	0.4921
200	Z_4	0.8	1.0	0.4892	0.4901
200	Z_5	35.6	37.5	0.0940	0.0935
200	Z_6	35.7	37.9	0.1084	0.1078
200	Z_7	34.8	38.3	0.1029	0.1023
200	Z_8	35.0	37.6	0.0942	0.0936
200	Z_9	35.7	38.5	0.0999	0.0994
200	Z_{10}	0.6	0.6	0.5029	0.5041

Table 4.76: Results obtained in Scenario h, where Z_5, Z_6, Z_7, Z_8 and Z_9 are influential.

Simulation study 3:
In Scenarios i and k, there is only 1 influential covariate, Z_1, whereas in Scenarios j and l, there are 2 influential covariates, Z_1 and Z_3. Table 4.67 shows the values z_1 and z_2 of those variables, and the probability mass for each scenario. Note that, since the probability mass function of Z is $(1/4, 3/4)$, there is a lower censoring percentage than in simulation study 1, where $(\Pi_z(z_1), \Pi_z(z_2)) = (1/2, 1/2)$.

Tables 4.77, 4.78, 4.79 and 4.80 show the results obtained in Scenarios i, j, k and l, respectively. In general, the behavior of the tests is satisfactory, increasing the power when the sample size becomes larger and the distance with H_0 increases. It is important to highlight that, for the four Scenarios, i, j, k, l, the rejection percentage for influential covariates is very close (or even equal) to 100% for large sample sizes.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	59.6	66.5	0.0139	0.0111
50	Z_2	2.2	2.2	0.4301	0.4332
50	Z_3	2.2	2.2	0.4409	0.4437
50	Z_4	2.1	2.3	0.4324	0.4350
50	Z_5	2.3	2.4	0.4451	0.4483
50	Z_6	2.1	2.1	0.4362	0.4393
50	Z_7	2.2	2.5	0.4469	0.4499
50	Z_8	2.2	2.4	0.4461	0.4497
50	Z_9	2.0	2.1	0.4509	0.4543
50	Z_{10}	2.2	2.3	0.4417	0.4449
100	Z_1	93.8	94.5	0.0020	0.0016
100	Z_2	0.3	0.4	0.4768	0.4792
100	Z_3	0.2	0.4	0.4751	0.4772
100	Z_4	0.1	0.1	0.4672	0.4695
100	Z_5	0.6	0.7	0.4837	0.4853
100	Z_6	0.7	0.7	0.4760	0.4782
100	Z_7	0.4	0.4	0.4906	0.4929
100	Z_8	0.6	0.7	0.4828	0.4852
100	Z_9	0.4	0.4	0.4934	0.4963
100	Z_{10}	0.3	0.5	0.4888	0.4908
200	Z_1	99.7	99.7	0.0001	0.0001
200	Z_2	0.2	0.4	0.4972	0.4984
200	Z_3	0.5	0.7	0.4945	0.4956
200	Z_4	0.4	0.6	0.4845	0.4854
200	Z_5	0.3	0.3	0.4993	0.5004
200	Z_6	0.3	0.3	0.4984	0.4995
200	Z_7	0.5	0.5	0.4856	0.4868
200	Z_8	0.2	0.4	0.5031	0.5043
200	Z_9	0.7	0.7	0.5028	0.5042
200	Z_{10}	0.6	0.6	0.5007	0.5024

Table 4.77: Results obtained in Scenario i, where Z_1 is influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	27.7	34.0	0.1216	0.1191
50	Z_2	0.7	0.9	0.4706	0.4737
50	Z_3	25.2	31.6	0.1155	0.1128
50	Z_4	0.2	0.4	0.4804	0.4839
50	Z_5	0.3	0.6	0.4644	0.4677
50	Z_6	0.2	0.3	0.4792	0.4818
50	Z_7	0.2	0.4	0.4806	0.4843
50	Z_8	0.2	0.3	0.4850	0.4879
50	Z_9	0.4	0.7	0.4794	0.4826
50	Z_{10}	0.5	0.9	0.4814	0.4854
100	Z_1	56.3	60.1	0.0508	0.0497
100	Z_2	0.2	0.3	0.4749	0.4771
100	Z_3	55.7	59.9	0.0424	0.0414
100	Z_4	0.3	0.7	0.4893	0.4912
100	Z_5	0.3	0.4	0.5044	0.5065
100	Z_6	0.5	0.8	0.4818	0.4839
100	Z_7	0.5	0.6	0.5022	0.5046
100	Z_8	0.2	0.2	0.4937	0.4959
100	Z_9	0.3	0.4	0.4918	0.4943
100	Z_{10}	0.3	0.5	0.4903	0.4926
200	Z_1	82.2	83.1	0.0124	0.0120
200	Z_2	0.3	0.4	0.4910	0.4921
200	Z_3	84.0	85.6	0.0099	0.0095
200	Z_4	0.4	0.8	0.4964	0.4974
200	Z_5	0.6	0.7	0.4886	0.4897
200	Z_6	0.4	0.8	0.5044	0.5057
200	Z_7	0.2	0.2	0.4963	0.4975
200	Z_8	0.5	0.6	0.4885	0.4899
200	Z_9	0.4	0.7	0.4916	0.4929
200	Z_{10}	0.5	0.7	0.4923	0.4935

Table 4.78: Results obtained in Scenario j, where Z_1 and Z_3 are influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	96.1	97.2	0.0012	0.0009
50	Z_2	0.4	0.9	0.4671	0.4705
50	Z_3	0.6	1.0	0.4782	0.4816
50	Z_4	0.4	0.7	0.4779	0.4811
50	Z_5	0.3	0.6	0.4805	0.4836
50	Z_6	0.6	0.9	0.4800	0.4829
50	Z_7	0.5	0.7	0.4703	0.4737
50	Z_8	0.5	0.7	0.4846	0.4878
50	Z_9	0.3	0.7	0.4812	0.4842
50	Z_{10}	0.4	0.7	0.4619	0.4653
100	Z_1	100.0	100.0	0.0000	0.0000
100	Z_2	0.5	0.6	0.4856	0.4880
100	Z_3	0.3	0.4	0.4845	0.4869
100	Z_4	0.1	0.1	0.5065	0.5089
100	Z_5	0.0	0.0	0.5038	0.5057
100	Z_6	0.5	0.5	0.4858	0.4881
100	Z_7	0.3	0.5	0.4948	0.4973
100	Z_8	0.2	0.3	0.4940	0.4963
100	Z_9	0.1	0.5	0.5081	0.5105
100	Z_{10}	0.3	0.4	0.4745	0.4766
200	Z_1	100.0	100.0	0.0000	0.0000
200	Z_2	0.3	0.4	0.5017	0.5030
200	Z_3	0.6	0.6	0.4869	0.4883
200	Z_4	0.1	0.1	0.5179	0.5191
200	Z_5	0.5	0.4	0.5003	0.5014
200	Z_6	0.5	0.7	0.4891	0.4904
200	Z_7	0.2	0.3	0.4940	0.4949
200	Z_8	0.3	0.3	0.4960	0.4972
200	Z_9	0.2	0.4	0.5057	0.5068
200	Z_{10}	0.0	0.0	0.4882	0.4890

Table 4.79: Results obtained in Scenario k, where Z_1 is influential.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	58.7	66.4	0.0349	0.0324
50	Z_2	0.2	0.4	0.4916	0.4957
50	Z_3	57.5	66.9	0.0356	0.0329
50	Z_4	0.5	0.8	0.4761	0.4806
50	Z_5	0.1	0.2	0.4923	0.4969
50	Z_6	0.5	0.8	0.4778	0.4822
50	Z_7	0.0	0.2	0.4927	0.4971
50	Z_8	0.3	0.7	0.4735	0.4775
50	Z_9	0.4	1.0	0.4858	0.4900
50	Z_{10}	0.2	0.3	0.4995	0.5041
100	Z_1	89.2	90.8	0.0054	0.0050
100	Z_2	0.6	0.7	0.5030	0.5056
100	Z_3	88.2	90.5	0.0061	0.0056
100	Z_4	0.7	0.8	0.4978	0.5000
100	Z_5	0.5	0.6	0.4977	0.4999
100	Z_6	0.5	0.7	0.4990	0.5013
100	Z_7	0.1	0.2	0.4978	0.5005
100	Z_8	0.5	1.0	0.5086	0.5108
100	Z_9	0.4	0.9	0.4954	0.4973
100	Z_{10}	0.2	0.4	0.4721	0.4740
200	Z_1	98.6	98.8	0.0005	0.0004
200	Z_2	0.6	0.9	0.4934	0.4947
200	Z_3	99.3	99.3	0.0004	0.0003
200	Z_4	0.6	0.8	0.5164	0.5176
200	Z_5	0.5	0.4	0.4921	0.4934
200	Z_6	0.5	0.6	0.4860	0.4871
200	Z_7	0.4	0.5	0.4991	0.5003
200	Z_8	0.6	0.6	0.5064	0.5076
200	Z_9	0.6	0.7	0.4963	0.4976
200	Z_{10}	0.4	0.4	0.4964	0.4976

Table 4.80: Results obtained in Scenario l, where Z_1 and Z_3 are influential.

Simulation study 4:
In both Scenarios m and n, there are 2 influential covariates: Z_1 and Z_2. Table 4.68 shows the values z_1 and z_2 of both variables, and the probability mass function for each scenario. Note that in Scenario m there will be more censored data.

Tables 4.81 and 4.82 show the results obtained in Scenarios m and n, respectively. In both scenarios, low power is obtained due to the large number of covariates involved.

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	0.2	0.2	0.4030	0.4044
50	Z_2	0.2	0.2	0.3406	0.3409
50	Z_3	0.2	0.2	0.4565	0.4600
50	Z_4	0.2	0.2	0.4637	0.4673
50	Z_5	0.3	0.3	0.4708	0.4750
50	Z_6	0.2	0.2	0.4738	0.4772
⋮	⋮	⋮	⋮	⋮	⋮
50	Z_{98}	0.2	0.2	0.4759	0.4789
50	Z_{99}	0.2	0.3	0.4763	0.4798
50	Z_{100}	0.2	0.2	0.4705	0.4740
100	Z_1	0.0	0.0	0.3680	0.3688
100	Z_2	0.2	0.2	0.2657	0.2657
100	Z_3	0.1	0.1	0.4921	0.4943
100	Z_4	0.0	0.1	0.4905	0.4926
100	Z_5	0.0	0.1	0.4731	0.4750
100	Z_6	0.0	0.0	0.4815	0.4839
⋮	⋮	⋮	⋮	⋮	⋮
100	Z_{98}	0.0	0.0	0.5003	0.5025
100	Z_{99}	0.0	0.0	0.5014	0.5036
100	Z_{100}	0.1	0.1	0.4926	0.4952
200	Z_1	0.3	0.2	0.2816	0.2816
200	Z_2	0.3	0.3	0.1516	0.1509
200	Z_3	0.0	0.0	0.5051	0.5067
200	Z_4	0.2	0.2	0.4906	0.4919
200	Z_5	0.0	0.0	0.4983	0.4997
200	Z_6	0.1	0.1	0.5017	0.5028
⋮	⋮	⋮	⋮	⋮	⋮
200	Z_{98}	0.0	0.1	0.5030	0.5042
200	Z_{99}	0.0	0.1	0.5025	0.5038
200	Z_{100}	0.0	0.0	0.4942	0.4955

Table 4.81: Results considering Scenario m, where $(\Pi_z(z_1), \Pi_z(z_2)) = (1/4, 3/4)$ (least favorable).

n	Variable	% rejections CvM	% rejections KS	mean p-value CvM	mean p-value KS
50	Z_1	2.1	2.1	0.3460	0.3467
50	Z_2	2.3	2.3	0.2765	0.2770
50	Z_3	2.1	2.1	0.4439	0.4448
50	Z_4	2.1	2.1	0.4567	0.4580
50	Z_5	2.1	2.1	0.4627	0.4636
50	Z_6	2.1	2.1	0.4468	0.4481
⋮	⋮	⋮	⋮	⋮	⋮
50	Z_{98}	2.1	2.1	0.4623	0.4634
50	Z_{99}	2.1	2.1	0.4535	0.4542
50	Z_{100}	2.1	2.1	0.4705	0.4716
100	Z_1	0.4	0.8	0.3208	0.3211
100	Z_2	2.3	2.3	0.1758	0.1759
100	Z_3	0.1	0.1	0.4944	0.4950
100	Z_4	0.3	0.3	0.4791	0.4796
100	Z_5	0.1	0.1	0.4870	0.4876
100	Z_6	0.1	0.1	0.4973	0.4981
⋮	⋮	⋮	⋮	⋮	⋮
100	Z_{98}	0.1	0.1	0.4933	0.4939
100	Z_{99}	0.1	0.1	0.4887	0.4894
100	Z_{100}	0.1	0.1	0.4924	0.4934
200	Z_1	1.8	2.0	0.2180	0.2180
200	Z_2	8.1	8.4	0.0889	0.0889
200	Z_3	0.0	0.0	0.5132	0.5137
200	Z_4	0.0	0.0	0.5194	0.5197
200	Z_5	0.1	0.1	0.4869	0.4873
200	Z_6	0.0	0.0	0.4955	0.4957
⋮	⋮	⋮	⋮	⋮	⋮
200	Z_{98}	0.1	0.1	0.4991	0.4995
200	Z_{99}	0.0	0.0	0.4759	0.4764
200	Z_{100}	0.0	0.0	0.4948	0.4952

Table 4.82: Results considering Scenario n, where $(\Pi_z(z_1), \Pi_z(z_2)) = (1/2, 3/2)$ (favorable).

Computational summary in Case 1 with FDR

The computational times taken for each study are shown in Table 4.83.

Type of Z	Scenario	Computational time
Z binary with FDR	Scenario a^*	2910 sec (48.5 min)
	Scenario m^{**}	26080 sec (7.24 h)

Table 4.83: Computational times for simulations with FDR in Case 1 considering Z binary. Sample sizes $n = 100$, $\kappa = 1000$ trials and $B = 1000$ bootstrap resamples are considered. (*Similar computational times for Scenarios b, c, d, e, f, g, h, i, j, k and l. **Similar computational times for Scenario n).

4.7 Application to real data

We applied the proposed methodology to the dataset used in Sections 2.6 and 3.6. Table 2.1 shows a summary of the data.

4.7.1 Colorectal cancer data (Case 1)

We consider three studies: in the first one, we split the data into four groups according to the categorical covariate stage. Then, we study if the covariate age (continuous and denoted by Z) has a significant effect on the cure rate, that is:

$$H_0 : E\left(\nu|Z\right) = 1 - p \ \text{ constant } \ \text{vs } H_1 : E\left(\nu|Z\right) = 1 - p(Z),$$

where $p(Z)$ is not a constant value under the alternative hypothesis. In the second (third) study, we test if the covariate age (stage) has some influence on the cure rate, regardless of the covariate stage (age). Note that in studies 1 and 2, Z is a continuous covariate, whereas in study 3, we work with Z ordinal. In the three studies we consider $B = 1000$ bootstrap resamples.

We start with study 1. Figure 4.12 shows the estimated values of $\tau(z)$, defined in Section 4.1, for Stages 1-4. We can see that the time beyond which a subject is considered cured varies with the stage, but in general it increases with the age. Furthermore, in Figure 2.9 of Section 2.6, we can appreciate how the age influences on the incidence for each stage. Considering only this figure, it seems reasonable to suppose that in Stages 1 and 4 the cure rate could be a constant value, as a function of the age, whereas in Stages 2 and 3 the age may have some influence. The corresponding p-values for each stage are shown in Table 4.84.

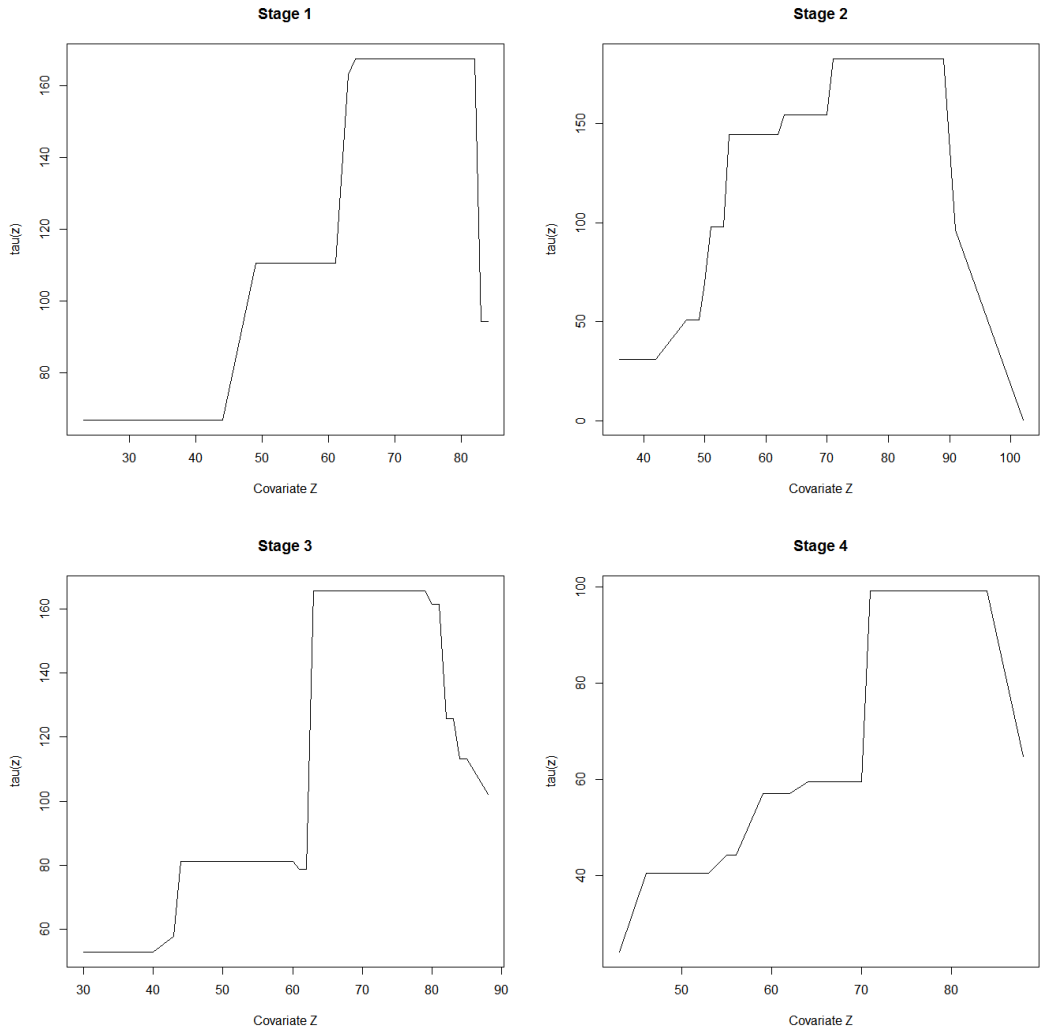

Figure 4.12: Values of $\tau(z)$ for Stages 1, 2, 3 and 4.

Stage	CvM	KS
1	0.396	0.257
2	0.082	0.067
3	0.002	0
4	0.587	0.551

Table 4.84: p-values when testing the effect of the age (continuous) on the probability of cure in stages 1, 2, 3 and 4.

Therefore, after applying the test and considering a significance level $\alpha = 0.05$, the test is only significant in Stage 3 (see Table 4.84). In Stages 1, 2 and 4 there is not enough evidence to reject the null hypothesis, so it is possible that the covariate age in colorectal cancer patients in these stages does not have any influence on the cure rate. Note that if we work with $\alpha = 0.10$, then the test would be significant in both Stages 2 and 3.

Regarding the second and third studies, Tables 4.85 and 4.86 show the p-values for Case 1 considering the covariate age and stage, respectively. In these cases, there is not enough evidence to reject the null hypothesis and therefore, similarly as in the first study, the covariate Z does not have any influence on the cure rate.

CvM	KS
0.142	0.146

Table 4.85: p-values when testing the effect of the age (continuous) on the probability of cure.

CvM	KS
0.581	0.483

Table 4.86: p-values when testing the effect of the stage (discrete) on the probability of cure.

4.7.2 Colorectal cancer data (Case 2)

We consider that $\mathbf{W} = (X, Z)$ is the vector of the covariate age (continuous) and Stage (discrete). The covariate significance test is the following:

$$H_0 : E\left(\nu|X, Z\right) = 1 - p(X), \quad \text{vs } H_1 : E\left(\nu|X, Z\right) = 1 - p(X, Z),$$

where $p(X, Z)$ depends on Z under the alternative hypothesis. We show the results of two studies, one considering $X = Age$ and $Z = Stage$, for a wide range of bandwidths: $h = Cn^{-1/3}$, with $C = 10, 20, 40, 60, 120, 240, 300$ and 375; and the other with $X = Stage$ and $Z = Age$. A total of $B = 1000$ bootstrap resamples are drawn. The results in both studies show that there is not enough evidence to reject H_0 (see Tables 4.87 and 4.88). In other terms, the cure rate can be explained with just any of the two covariates: age or stage.

h	CvM	KS
1.342	0.721	0.815
2.683	0.655	0.720
5.367	0.611	0.647
8.050	0.597	0.647
16.101	0.537	0.517
32.202	0.233	0.218
40.252	0.168	0.189
50.315	0.137	0.153

Table 4.87: p-values considering that X is the covariate age and Z is the covariate stage, using different bandwidths.

CvM	KS
0.370	0.267

Table 4.88: p-values considering that X is the covariate stage and Z is the covariate age.

4.7.3 Sarcomas data (Case 1 with FDR)

We study a dataset related to patients with sarcomas, provided by Angel Díaz-Lagares, postdoc researcher in Cancer Epigenomics from Translational Medical Oncology (OMT) group, Health Research Institute of Santiago (IDIS) and the University Hospital of Santiago (CHUS).

Sarcomas are an uncommon and histologically heterogeneous group of neoplasias, which stand for 1% of the diagnosed tumors in adults and around 20% in children and teens (Burningham et al., 2012). Sarcomas proceed from embryonal cells with mesodermal origin, except the peripheral nerve sheath tumors, which are ectodermal. Since these cells can be differentiated from adipose, muscular, fribous, cartilaginous and bone tissues, they can be generally classified in two tumor locations: soft tissue and hard tissue. Nevertheless, the wide range of histologic entities which form the sarcomas can also be found in the different systems or viscus, such as the digestive system.

The universality of locations in the human body and the large number of different histologies have complicated the epidemiological analysis of sarcomas. The population-based cancer registries (RCBP) analize and describe the epidemiology in all the tumor locations. The purpose is to register all the cases of neoplasias from a specific population in a precise and systematic way, obtaining data from different sources which vary depending on the health system of each country. Specifically,

in Spain, the most of the RCBP are compilled, essentially, from the information provided by the pathology services and the hospital centers in each area, and by the mortality registers in each community (see Muir & Percy, 1991).

The epidemiological analysis is based on the tumor location. The characteristics of the tumor, the average age of diagnosis and the distribution by stages or sublocations are analized, together with some parameters such as the number of new cases in a population group from a specific time, the mortality (deaths caused by this pathology in a population group from a specific time), the prevalence (proportion of current, previous and new cases in a population group from a specific time) and the survival, which reflects the time until a new vital event (usually, the reappearance of the tumor or the death of the patient). In epidemiological studies, the observed and the relative survival are considered. The latter one is obtained as a ratio between the observed and the expected survival as a function of the variables age and sex, in the absence of cancer. Therefore, the mortality originated by different causes other than cancer in each group of age can be fit. The acquisition of genetic changes is essential in the development of neoplasias. Such alterations are modifications in DNA sequence, like mutations, translocations, delections (mutations in which a part of a chromosome or a sequence of DNA is lost during DNA replication), and amplifications (production of multiple copies of a sequence of DNA). The epigenetic changes that represent reversible modifications which affect the gene expression but not the DNA sequence, are also a milestone in the origin of cancer (Esteller, 2008).

Epigenetics is the study of heritable changes in gene expression (active versus inactive genes) that do not involve changes to the underlying DNA sequence - a change in the phenotype without a change in the genotype - which in turn affects how cells read the genes. The epigenetic regulation allows the plasticity of the genome for the adaptation to the environment. A CpG island is a region of DNA with high concentration of cytosine and guanine nucleotides. The hypermethylation (an increase in the epigenetic methylation of cytosine and adenosine residues in DNA) of the CpG islands and other epigenetic mechanisms constitute alterations of the gene expression associated to the human cancer (Kang, 2012). Therefore, the hypermethylation can be associated to the age and the development of cancer. A high grade methylation is related to gene silencing. Usually, the methylation appears with more frequency in the CpG islands.

Cancer cells are less methylated than the normal ones, and their promoters are

hypermethylated. This is considered for the tumor diagnosis and the identification of therapeutic alternatives. The abnormal DNA methylation is known as an early event of the tumorigenesis (see Suzuki et al., 2004; Ushijima, 2005, between others), and therefore, variations in DNA methylation patterns identified between the normal and the tumor cells can help doctors detect tumor cells in biopsy samples or in body fluids (Tsuchiya et al., 2000).

In some tumor subtypes it is possible to identify epigenetic profiles which are used to determine the aggressiveness of the tumor, and to determine if they will react to a specific treatment. For example, in brain tumors, the gene MGMT is analized. Its effect is repressed when it is hypermethylated. Therefore, chemotherapies which work methylating will only be effective when the gene is not methylated (Bennani-Baiti, 2011; Esteller et al., 2000). Furthermore, there are some epigenetically alterated oncogenes and tumor suppressor genes in sarcomas, such as APC, CDKN1A, CDKN2A, CDKN2B, Ezrin, FGFR1, GADD45A, MGMT, STK3, STK4, PTEN, RASSF1A, WIF1; and some epigenetic deregulated indicators in bone sarcomas, as PCR1, BMI1 and LSD1. The objective consists of analizying if the epigenetic expression is correlated with the probability of death in a group of soft part sarcomas.

The database consists of 261 observations with 372452 covariates. Specifically, there are 372420 covariates with information about DNA methylations and 32 covariates with clinical data. The methylation covariates are continuous, with values between 0 and 1. The covariates with clinical data are introduced in Table 4.89. The event of interest is the death due to the appearance of sarcomas in different parts of the body. A total of 195 observations are censored, which corresponds to 74.71% of the data.

Test by Maller & Zhou (1992)

Similarly as in Section 2.6.1, we apply the test by Maller & Zhou (1992) to the sarcomas dataset. In this case, the largest uncensored failure time is $T^1_{\max} = 2575$; and the largest time (censored or uncensored), is $T_{(n)} = 5324$. Therefore, the difference is given by $\varsigma = T_{(n)} - T^1_{\max} = 2749$, and the interval to study is $(2575 - 2749, 2575] = (-174, 2575]$. Note that since it includes the value 0, then the interval to consider is $(0, 2575]$, which has 76 uncensored observations. Finally, we obtain the value of α_n:

$$\alpha_n = \left(1 - \frac{76}{261}\right)^{261} = 9.742013 \cdot 10^{-40}.$$

Since $\alpha_n < \alpha = 0.05$, we reject the null hypothesis and then condition (2.8) holds.

Name	Type	Values	n without NA
Days to birth	Continuous	Between 7530 (20.6 years) and 32873 (90.1 years)	260
Gender	Binary	male ($n = 119$), female ($n = 142$)	261
Race	Qualitative	asian ($n = 6$), black or african american ($n = 18$), white ($n = 228$)	252
Ethnicity	Binary	hispanic or latino ($n = 5$), not hispanic or latino ($n = 223$)	228
Other diagnosis	Binary	yes ($n = 42$), no ($n = 219$)	261
Radiation therapy	Binary	yes ($n = 73$), no ($n = 142$)	215
Pharmaceutical adjuvant postop.	Binary	yes ($n = 37$), no ($n = 178$)	215
Histological type	Qualitative	dedifferentiated liposarcoma ($n = 59$), leiomyosarcoma (LMS) ($n = 105$), undifferentiated pleomorphic sarcoma (UPS) ($n = 51$), synovial sarcoma ($n = 10$), malignant peripheral nerve sheath tumors (MPNST) ($n = 9$), myxofibrosarcoma ($n = 25$), desmoid tumor ($n = 2$)	261
Leiomyosarcoma histologic subtype	Qualitative	poorly differentiated ($n = 34$), conventional ($n = 66$), well-differentiated ($n = 4$)	104
Mpnst neurofibromatosis heredity	Binary	familial ($n = 4$), sporadic ($n = 2$)	6
Tumor depth	Binary	deep ($n = 187$), superficial ($n = 21$)	208
Year of initial pathologic diagnosis	Continuous	Between 1994 and 2013	257
Age at initial pathologic diagnosis	Continuous	Between 20 and 90	261
Margin status	Binary	negative ($n = 137$), positive ($n = 74$)	211
Residual tumor	Qualitative	R0 ($n = 155$), R1 ($n = 70$), R2 ($n = 9$), RX ($n = 26$)	260
Tumor total necrosis percent	Qualitative	no necrosis ($n = 70$), focal ($n = 39$), extensive ($n = 12$), moderate ($n = 61$)	182
Mitotic count	Continuous	Between 0 and 102	90
Tumor multifocal	Binary	yes ($n = 40$), no ($n = 199$)	239
Discontiguous lesion count	Discrete	0 ($n = 8$), 1 ($n = 16$), 2 ($n = 10$), 3 ($n = 8$), 4 ($n = 3$), 6 ($n = 1$), 7 ($n = 1$), 9 ($n = 1$)	48
Radiologic tumor burden	Continuous	Between 2 and 800	48
Pathologic tumor burden	Continuous	Between 1.6 and 800	75
Local disease recurrence	Binary	yes ($n = 29$), no ($n = 144$)	173
Metastatic diagnosis	Binary	yes ($n = 56$), no ($n = 121$)	177
Metastatic site at diagnosis	Qualitative	bone ($n = 2$), liver ($n = 2$), lung ($n = 40$), lung and liver ($n = 1$), other ($n = 10$)	55
New tumor after initial treatment	Binary	yes ($n = 89$), no ($n = 129$)	218
Radiologic tumor length	Continuous	Between 1.8 and 30	108
Radiologic tumor width	Continuous	Between 1.1 and 27.8	103
Radiologic tumor depth	Continuous	Between 0.9 and 25	76
Pathologic tumor length	Continuous	Between 1.2 and 39.5	227
Pathologic tumor width	Continuous	Between 0 and 30	188
Pathologic tumor depth	Continuous	Between 0 and 18	184
Location	Qualitative	viscera ($n = 6$), retroperitoneum ($n = 101$), gynecological ($n = 32$), upper extremity ($n = 9$), lower extremity ($n = 68$), head and neck ($n = 6$) trunck ($n = 37$)	259

Table 4.89: Covariates with clinical information in the sarcomas dataset.

Significance tests

In order to perform significant tests for the 372452 covariates, we use the FDR approach by Benjamini & Hochberg (1995), and the conservative alternative by Benjamini & Yekutieli (2001), both together with the bootstrap method. Since the p-values estimated by the bootstrap need to be compared to eventually small numbers:

$$\frac{\alpha}{m-i+1} = \frac{0.05}{372452} = 1.3 \cdot 10^{-7}, \text{ for } i = 1 \text{ and } \alpha = 0.05,$$

the number of bootstrap replications are required to be large when the estimated p-values are close (or equal) to zero. But performing tens or hundreds of millions bootstrap replications for nearly 400000 covariates is a very time consuming process. Consequently, we designed an incremental mechanism in order to use a small number of bootstrap replications and only when the estimated p-value is not conclusive, we increase B by multiplying it by 10 at every step.

First, we consider $B = 10$ bootstrap resamples, we draw the test and we sort the 372452 p-values obtained. We define $q' = \frac{i\alpha}{m}$ for the non-conservative approach, and $q' = \frac{\alpha}{m-i+1}$ for the conservative alternative. Since p is estimated by Monte Carlo (using the bootstrap) classical hypothesis tests for proportions can be used to accept, reject or keep increasing the value B when testing if the theoretical p-value is larger, smaller or equal to q'. For each p-value:

a) If $p > q' + \sqrt{\frac{q'(1-q')}{B}}2.32$, then the covariate is not significant.

b) If $p < q' - \sqrt{\frac{q'(1-q')}{B}}2.32$, then the covariate is significant.

c) If $p \in \left(q' - \sqrt{\frac{q'(1-q')}{B}}2.32, q' + \sqrt{\frac{q'(1-q')}{B}}2.32\right)$, then the results are not conclusive.

Note that for covariates in cases a) and b), it is not necessary to continue with the procedure. For the non-conclusive covariates (case c)), we consider $B = 10^2$ bootstrap resamples and we draw the test again. We continue with this algorithm, multiplying B by 10 in each step, until one of the following conditions is fulfilled: either all of the covariates are conclusive or the number of bootstrap resamples is larger than $B = 10^9$.

Regarding the conservative alternative, after 5 days of CPU time, the results show that only one covariate is significant for the cure rate: "Year of initial pathologic diagnosis". The p-value is equal to 0 with the CvM statistic, and equal to 10^{-5}

with the KS statistic. The histogram of the p-values (obtained with CvM and KS) for the 372452 covariates is shown in Figure 4.13.

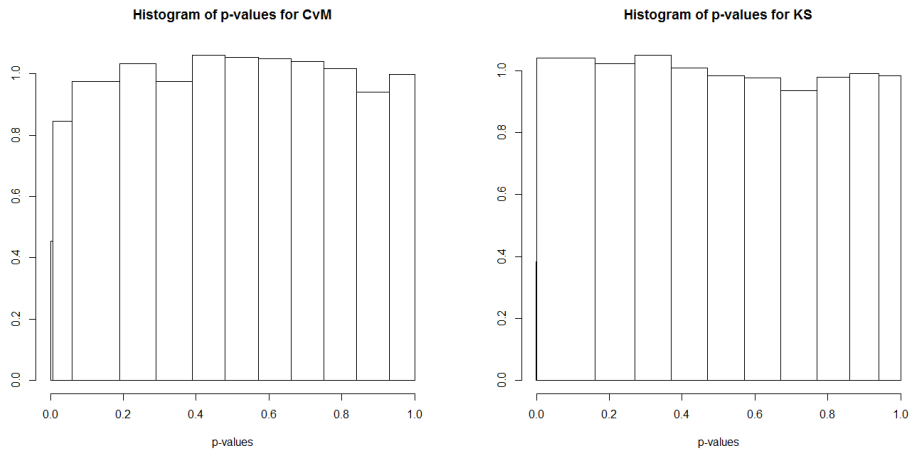

Figure 4.13: p-values obtained with CvM (left) and KS (right) statistics.

In Figure 4.14 we can appreciate the estimated probability of cure obtained with the nonparametric incidence estimator using the bootstrap bandwidth for the covariate "Year of initial pathologic diagnosis". The density of the covariate is also shown (green line). Similarly as in Section 2.6, alongside the bootstrap bandwidth, we have also used a smoothed bootstrap bandwidth for the incidence estimator, following Cao et al. (2001). The bootstrap, the smoothed and the pilot bandwidths are shown in Figure 4.15.

With respect to the non-conservative method, after 10 days of CPU time, the results for $B = 1000$ bootstrap resamples show that for the CvM statistic, there are 4179 significant covariates and 6924 non conclusive covariates, which need to be considered again in the next iteration of the process. For the KS statistic, there are 3457 significant covariates, and 6263 non conclusive covariates. At the moment of the deposit of this thesis, the program is still running for $B = 10000$ bootstrap resamples.

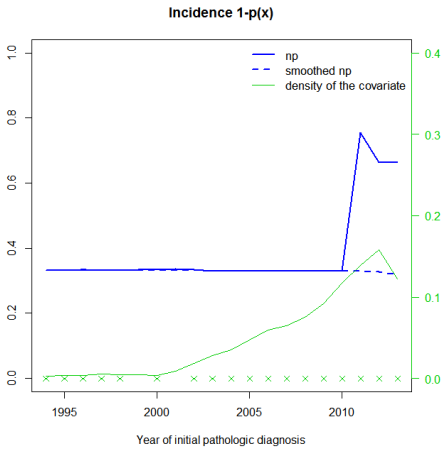

Figure 4.14: Incidence estimator computed with the bootstrap bandwidth (solid blue line) and a smoothed version (dashed blue line), for the covariate "Year of initial pathologic diagnosis". The density of the covariate is also shown (green line).

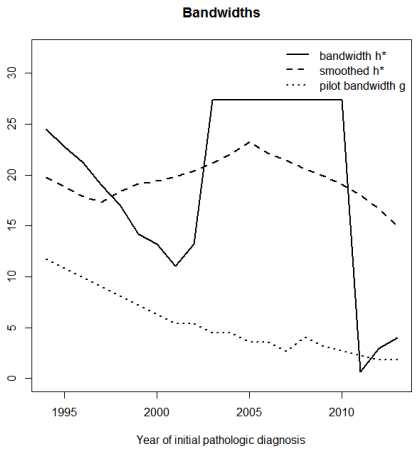

Figure 4.15: Bootstrap bandwidth h_x^* (solid line), smoothed bootstrap bandwidth $h_x^{*smoothed}$ (dashed line) and local pilot bandwidth g_x (dotted line) used for the nonparametric incidence estimator for the covariate "Year of initial pathologic diagnosis".

Chapter 5

Future work

Interesting challenges remain as open problems to be dealt in the future:

- The proposed methodology will be applied to high dimensional datasets, including analysis of images, related to cancer for medical diagnosis. The idea is to extract a numeric vector from these images and applying covariate significance tests.

- A bandwidth selection method will be proposed for the significance tests in Case 2 when the covariate X is continuous. There are two approaches for selecting the bandwidths that differ with respect to the optimality measure used. The first approach focuses on power maximization under the alternative hypothesis, proposed by Kulasekera & Wang (1997) and subsequently developed in Cao & Van Keilegom (2006), Gao & Gijbels (2008) and Martínez-Camblor & de Uña-Álvarez (2013). On the other hand, the second procedure considers the idea of minimizing p-values (see Martínez-Camblor, 2010; Martínez-Camblor & de Uña-Álvarez, 2013). As Martínez-Camblor & de Uña-Álvarez (2013) point out, the two approaches are strongly related.
  These procedures are computationally expensive, since for each value of the bandwidth in a grid of bandwidths, a double bootstrap is needed: one to approximate the power (or the p-value); and another to approximate the critical value of the test distribution, for each one of the resamples.
  We therefore propose that the bandwidth can be selected by means of the following procedure, which estimates the bandwidth that maximizes the power (or minimizes the p-value) by means of a double-bootstrap procedure:

  1. For each bandwidth in the grid of bandwidths, generate the following resample: (X^*, T^*, δ^*), and compute the bootstrap version of test statistic,

T_n, in (4.6).

2. For each resample, determine the critical value by generating (second level) bootstrap resamples under the null hypothesis.
3. Finally, for each bandwidth, compute the power (or the p-value), and select the optimal bandwidth.

- Covariate significance tests for the latency function will be proposed, based on the same ideas in which the covariate significance tests for the incidence are based (see Delgado & González-Manteiga, 2001).

- The estimators of $p(\mathbf{x})$ and $S_0(t|\mathbf{x})$ will be redefined, so that they only depend on $(\mathbf{x}_I, \mathbf{x}_C)$ and $(\mathbf{x}_L, \mathbf{x}_C)$, respectively, with $\mathbf{x} = (\mathbf{x}_I, \mathbf{x}_L, \mathbf{x}_C)$ and where:

  – $\mathbf{x}_I$ are the covariates that influence the incidence but not the latency.

  – $\mathbf{x}_L$ are the covariates that influence the latency but not the incidence.

  – $\mathbf{x}_C$ are the common covariates, that influence both the incidence and the latency.

  Therefore, the mixture cure model will be written as:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}_I, \mathbf{x}_C) + p(\mathbf{x}_I, \mathbf{x}_C) \cdot S_0(t|\mathbf{x}_L, \mathbf{x}_C).$$

- An R package will be developed with all the techniques studied, including the implementation of the nonparametric incidence and latency estimators, as well as the covariate significance tests for the different types of data: continuous, discrete, binary and qualitative. This R package will be uploaded in CRAN and any R user will be able to use the methodology developed in this thesis.

- The presmoothed estimator of the incidence will be considered. The purpose is to estimate properly the tail of the survival function, so that the estimation of the cure probability is more efficient. Presmoothing (see Jácome & Cao, 2004) consists of the nonparametric estimation of the conditional probability of no censoring, which is a regression function. The name of the presmoothed estimation stems from the fact that the smoothing is used only for obtaining a smoothed version of the Kaplan-Meier weigths, but the resulting estimator of the distribution function is not smoothed.

- The proposed methods will be extended to cases with truncated and interval censored data.

- Single-index models have been recently used in survival analysis for censored data (see Strzalkowska-Kominiak & Cao, 2013). Extensions for the incidence and the latency in cure models will be considered to handle large number of covariates.
- The consistency of the bootstrap methods in Chapters 2 and 3 will be proven, since the bootstrap mean squared error is a plug-in estimation of the original mean squared error.
- The limit behavior of the bootstrap version of the test statistic in Chapter 4 will be considered.
- For the covariate significance testing a new bootstrap approach will be proposed, disturbing the main terms of the process similarly to a wild bootstrap. This would make the whole computational process much less time consuming.

Appendix A

Appendix

We need to consider the following assumptions, to be used in the asymptotic results for the incidence and the latency estimators:

(A1) X, Y and C are absolutely continuous random variables.

(A2) Condition (2.8) holds.

(A3) (a) Let $I = [x_1, x_2]$ be an interval contained in the support of m, and $I_\delta = [x_1 - \delta, x_2 + \delta]$ for some $\delta > 0$ such that

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

and $0 < \delta\Gamma < 1$. For all $x \in I_\delta$ the random variables Y and C are conditionally independent given $X = x$.

(b) There exist $a, b \in \mathbb{R}$, with $a < b$ satisfying $1 - H(t|x) \geq \theta > 0$ for $(t, x) \in [a, b] \times I_\delta$.

(A4) The first derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the first derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A5) The second derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the second derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A6) The first derivatives with respect to t of the functions $G(t|x)$, $H(t|x)$, $H^1(t|x)$ and $S_0(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A7) The second derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A8) The second partial derivatives with respect to t and x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded for $(t, x) \in [0, \infty) \times D$.

(A9) Let us define $H_{c,1}(t) = P(T < t|\delta = 1)$. The first and second derivatives of the distribution and subdistribution functions $H(t)$ and $H_{c,1}(t)$ are bounded and bounded away from zero in $[a, b]$. Moreover, $H'_{c,1}(\tau_0) > 0$.

(A10) The functions $H(t|x)$, $S_0(t|x)$ and $G(t|x)$ have bounded second-order derivatives with respect to x for any given value of t.

(A11) The kernel function, K, is a symmetric density vanishing outside $(-1, 1)$ and the total variation of K is less than some $\lambda < \infty$.

(A12) The density function of T, f_T, is bounded away from 0 in $[a, b]$.

(A13) $\displaystyle\int_0^\infty \frac{dH^1(t|x)}{(1 - H(t|x))^2} < \infty \ \forall x \in I.$

(A14) $b_i \to 0$, $\frac{\ln n}{nb_i} \to 0$, $\frac{nb_i^5}{\ln n} = O(1)$, $\frac{(\ln \ln n)^4}{(\ln n)^3} \frac{b_i}{nb_j^2} = O(1)$ and $\frac{(\ln \ln n)^2}{(\ln n)^3} \frac{nb_i^{11}}{b_j^2} = O(1)$, for $i, j = 1, 2,\ i \neq j$.

(A15) The distribution of $(C|\mathbf{X}, \nu = 0)$ equals that of $(C|\mathbf{X}, \nu = 1)$.

A.1 Proofs of the results in Chapter 2

Lemma 2.2.1. *Let D be the support of X. Model (2.1), with $p(x)$ and $S_0(t|x)$ unspecified, is identifiable if $S_0(t|x)$ is a proper survival function for $x \in D$.*

Proof of Lemma 2.2.1. Suppose we have two formulations of model (2.1):

$$S(t|x) = 1 - p(x) + p(x)S_0(t|x) \text{ and } S^*(t|x) = 1 - p^*(x) + p^*(x)S_0^*(t|x).$$

We need to show that $S(t|x) = S^*(t|x)$ if and only if $p(x) = p^*(x)$ and $S_0(t|x) = S_0^*(t|x)$, for all $x \in D$. The "if"part is clearly true in all cases, so we concentrate on "only if": suppose that $S(t|x) = S^*(t|x)$, then, rearranging Equation (2.1) gives the ratio:

$$\frac{p(x)}{p^*(x)} = \frac{1 - S_0^*(t|x)}{1 - S_0(t|x)} = c(x) \text{ for all } x \in D. \tag{A.1}$$

In particular,

$$S_0^*(t|x) = 1 - c(x)(1 - S_0(t|x)) \text{ for all } x \in D.$$

Since $S_0(t|x)$ and $S_0^*(t|x)$ are proper survival functions, that is, $S_0(+\infty|x) = 0$ and $S_0^*(+\infty|x) = 0$, then

$$0 = S_0^*(+\infty|x) = 1 - c(x)(1 - S_0(+\infty|x)) = 1 - c(x) \text{ for all } x \in D.$$

Hence, $c(x)$ is constant and equal to one for all x and thus, from (A.1), $p(x) = p^*(x)$ and $S_0(t|x) = S_0^*(t|x)$, so $S(t|x)$ is uniquely represented by $1 - p(x) + p(x)S_0(t|x)$. □

Theorem 2.3.1. *The estimator $1 - \hat{p}_h(x)$ given in (2.4) is the local maximum likelihood estimator of $1 - p(x)$ for the mixture cure model (2.1), for any $x \in D$.*

Theorem 3.3.1. *The estimator $\hat{S}_{0,b}(t|x)$, given in (3.2) is the local maximum likelihood estimator of $S_0(t|x)$ for the mixture cure model (2.1), for any $x \in D$ and $t \geq 0$.*

Proof of Theorems 2.3.1 and 3.3.1. The idea is to estimate $p(x)$ locally, maximizing the observed local likelihood function around x. It can be proven that the maximum likelihood estimator of the survival function $S_0(t|x) = 1 - F_0(t|x)$ has jumps only at the observations $(X_i, T_i, \delta_i), i = 1, \ldots, n$ with jumps

$$q_i(x) = S_0(T_{(i)}^-|x) - S_0(T_{(i)}|x).$$

Maximizing the likelihood of the observations for the cure model is equivalent to maximizing L given by

$$L\left(p\left(x\right),S_0\left(\cdot|x\right)\right) = \prod_{i=1}^{n}\left\{\left[p\left(x\right)q_i\left(x\right)\right]^{B_{h(i)}(x)\delta_{(i)}}\left[1-p\left(x\right)+p\left(x\right)\right.\right.$$
$$\left.\left.\times\left(1-\textstyle\sum_{j=1}^{i-1}q_j\left(x\right)\right)\right]^{\left(1-\delta_{(i)}\right)B_{h(i)}(x)}\right\}.$$

Let $D_i\left(x\right)=B_{h(i)}\left(x\right)\delta_{(i)}$ and $P_i\left(x\right)=p\left(x\right)q_i\left(x\right)$, then

$$L\left(p\left(x\right),S_0\left(\cdot|x\right)\right)=\prod_{i=1}^{n}\left\{P_i\left(x\right)^{D_i(x)}\left(1-\sum_{j=1}^{i-1}P_j\left(x\right)\right)^{B_{h(i)}(x)-D_i(x)}\right\}.$$

Consider now the functions $\lambda_i\left(x\right)=P_i\left(x\right)/\left(1-\sum_{j=1}^{i-1}P_j\left(x\right)\right)$ satisfying

$$1-\sum_{j=1}^{k}P_j\left(x\right)=\prod_{j=1}^{k}(1-\lambda_j\left(x\right)). \tag{A.2}$$

Since $\lambda_1(x)=P_1(x)/(1-\sum_{j=1}^{0}P_j(x))=P_1(x)$, then Equation (A.2) holds for $k=1$. To prove (A.2) for a general k we proceed by induction. Let us assume that (A.2) holds for k and let us prove it for $k+1$:

$$1-\sum_{j=1}^{k+1}P_j\left(x\right)=1-\sum_{j=1}^{k}P_j\left(x\right)-P_{k+1}(x)=1-\sum_{j=1}^{k}P_j\left(x\right)-\lambda_{k+1}(x)\left(1-\sum_{j=1}^{k}P_j(x)\right)$$
$$=\left(1-\sum_{j=1}^{k}P_j\left(x\right)\right)(1-\lambda_{k+1}(x))=\left[\prod_{j=1}^{k}(1-\lambda_j(x))\right](1-\lambda_{k+1}(x))=\prod_{j=1}^{k+1}(1-\lambda_j(x)).$$

Now straightforward calculations yield

$$\begin{aligned}L\left(\lambda_1(x),\ldots,\lambda_n(x)\right) &= \prod_{i=1}^{n}\lambda_i\left(x\right)^{D_i(x)}\left(\prod_{j=1}^{i-1}\left(1-\lambda_j\left(x\right)\right)\right)^{B_{h(i)}(x)}\\ &= \left(\prod_{i=1}^{n}\lambda_i\left(x\right)^{D_i(x)}\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{i-1}\left(1-\lambda_j\left(x\right)\right)^{B_{h(i)}(x)}\right)\\ &= \left(\prod_{i=1}^{n}\lambda_i\left(x\right)^{D_i(x)}\right)\left(\prod_{\substack{i,j=1\\ j<i}}^{n}\left(1-\lambda_j\left(x\right)\right)^{B_{h(i)}(x)}\right)\end{aligned}$$

$$= \left(\prod_{i=1}^{n} \lambda_i(x)^{D_i(x)}\right)\left(\prod_{j=1}^{n}(1-\lambda_j(x))^{\sum_{r=j+1}^{n} B_{h(r)}(x)}\right)$$
$$= \prod_{i=1}^{n} \lambda_i(x)^{D_i(x)}(1-\lambda_i(x))^{\sum_{r=i+1}^{n} B_{h(r)}(x)}.$$

Maximizing the likelihood of the observations for the cure model is equivalent to maximizing

$$\max_{\lambda_i \geq 0; i=1,\ldots,n} \Psi(\lambda_1, \ldots, \lambda_n),$$

where Ψ is the local loglikelihood:

$$\Psi(\lambda_1(x), \ldots, \lambda_n(x)) = \sum_{i=1}^{n}\left[D_i(x)\ln\lambda_i(x) + \left(\sum_{r=i+1}^{n} B_{h(r)}(x)\right)\ln(1-\lambda_i(x))\right],$$

subject to

$$\prod_{i=1}^{n}(1-\lambda_i(x)) = 1 - \sum_{j=1}^{n} P_j(x) = 1 - \sum_{j=1}^{n} p(x)q_j(x) = 1 - p(x). \tag{A.3}$$

Using standard maximization techniques, we have

$$\frac{\partial\Psi(\lambda_1(x), \ldots, \lambda_n(x))}{\partial\lambda_i(x)} = \frac{D_i(x)}{\lambda_i(x)} - \frac{\sum_{r=i+1}^{n} B_{h(r)}(x)}{1-\lambda_i(x)} = 0, \ \forall i = 1, 2, \ldots, n,$$

so $D_i(x)(1-\hat{\lambda}_i(x)) = \hat{\lambda}_i(x)\sum_{r=i+1}^{n} B_{h(r)}(x)$ and thus

$$\widehat{\lambda}_i(x) = \frac{D_i(x)}{\sum_{r=i+1}^{n} B_{h(r)}(x) + D_i(x)} = \frac{\delta_{(i)}B_{h(i)}(x)}{\sum_{r=i+1}^{n} B_{h(r)}(x) + \delta_{(i)}B_{h(i)}(x)}.$$

Replacing λ_i in (A.3) by $\widehat{\lambda}_i(x)$, we obtain the estimator of $1-p(x)$ given in (2.4).

With respect to the distribution of the uncured subjects, note that

$$F_0\left(T_{(i)}|x\right) = \sum_{j=1}^{i} q_j(x).$$

Since the jumps satisfy $P_i(x) = p(x)\,q_i(x)$ and using (A.2), we find that the local maximum likelihood estimator is given by

$$\hat{F}_0\left(T_{(i)}|x\right) = \frac{1}{\hat{p}_h(x)}\left[1 - \prod_{j=1}^{i}\left(1-\hat{\lambda}_j(x)\right)\right] = \frac{\hat{F}_h\left(T_{(i)}|x\right)}{\hat{p}_h(x)},$$

with $\hat{F}_h\left(T_{(i)}|x\right)$ the Beran estimator of $F = 1-S$ computed at time $T_{(i)}$. □

The following auxiliary results are necessary to prove Theorem 2.3.2.

Lemma A.1.1. *(Xu & Peng (2014)) Under assumption* $(A10)$*,*

$$T^1_{\max} = \max_{i:\delta_i=1}(T_i) \to \tau_0 \text{ in probability as } n \to \infty.$$

Lemma A.1.2. *Under assumption* $(A9)$*, we have that*

$$n^\alpha(\tau_0 - T^1_{\max}) \to 0 \ a.s.$$

for any $\alpha \in (0,1)$*. In particular, for a sequence of bandwidths satisfying* $nh^5(\ln n)^{-1} = O(1)$*, we have*

$$\tau_0 - T^1_{\max} = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s. \tag{A.4}$$

Proof of Lemma A.1.2. Using the Borel-Cantelli lemma, it is sufficient to prove that

$$\sum_{n=1}^{\infty} P\left(|a_n(\tau_0 - T^1_{\max})| > \epsilon\right) < \infty, \text{ for all } \epsilon > 0, \tag{A.5}$$

where $a_n = n^\alpha$. Let us fix $\epsilon > 0$ and consider:

$$\begin{aligned}
&P(|a_n(\tau_0 - T^1_{\max})| > \epsilon)\\
=&P\left(T^1_{\max} < \tau_0 - \frac{\epsilon}{a_n}\right)\\
=&P\left(T_i < \tau_0 - \frac{\epsilon}{a_n}, \text{ for all } i = 1, 2, \dots n \text{ where } \delta_i = 1\right)\\
=&E\left[P\left(T_i < \tau_0 - \frac{\epsilon}{a_n}, \text{ for all } i = 1, 2, \dots n \text{ where } \delta_i = 1 \middle| \delta_1, \delta_2, \dots, \delta_n\right)\right]\\
=&E\left[\prod_{i=1}^{n} P\left(T_i < \tau_0 - \frac{\epsilon}{a_n} \middle| \delta_i = 1\right)^{\delta_i}\right] = E\left[P\left(T_1 < \tau_0 - \frac{\epsilon}{a_n} \middle| \delta_1 = 1\right)^{\sum_{i=1}^n \delta_i}\right]\\
=&E\left[\left(H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right)\right)^{\sum_{i=1}^n \delta_i}\right],
\end{aligned}$$

where

$$H_{c,1}(t) = P\left(T < t \middle| \delta = 1\right) = \frac{P(T < t, \delta = 1)}{P(\delta = 1)} = \frac{H_1(t)}{\rho},$$

with $\rho = P(\delta = 1) = E(\delta)$ and $H_1(t) = P(T < t, \delta = 1)$. Consequently, since

$\sum_{i=1}^{n}\delta_i \overset{\mathrm{d}}{=} B(n,\rho)$, we get:

$$\begin{aligned} & P(|a_n(\tau_0 - T^1_{\max})| > \epsilon) \qquad (A.6)\\ = \; & E\left[H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right)^{\sum_{i=1}^{n}\delta_i}\right] = \sum_{j=0}^{n}\binom{n}{j}\rho^j(1-\rho)^{n-j}H_{c,1}\left(\tau_0-\frac{\epsilon}{a_n}\right)^j \\ = \; & \sum_{j=0}^{n}\binom{n}{j}\left[\rho H_{c,1}\left(\tau_0-\frac{\epsilon}{a_n}\right)\right]^j(1-\rho)^{n-j} = \left[\rho H_{c,1}\left(\tau_0-\frac{\epsilon}{a_n}\right)+1-\rho\right]^n \\ = \; & \left[\rho\left(H_{c,1}(\tau_0) - \frac{\epsilon}{a_n}H^{'}_{c,1}(\tau_0) + \frac{\epsilon^2}{2a_n^2}H^{''}_{c,1}(\xi_n)\right)+1-\rho\right]^n \\ = \; & \left[\rho - \rho\frac{\epsilon}{a_n}H^{'}_{c,1}(\tau_0) + \rho\frac{\epsilon^2}{2a_n^2}H^{''}_{c,1}(\xi_n) + 1 - \rho\right]^n \\ = \; & \left(1 - \rho\frac{\epsilon}{a_n}H^{'}_{c,1}(\tau_0) + \rho\frac{\epsilon^2}{2a_n^2}H^{''}_{c,1}(\xi_n)\right)^n, \end{aligned}$$

for some $\xi_n \in \left[\tau_0 - \frac{\epsilon}{a_n}, \tau_0\right]$, since $H_{c,1}(\tau_0) = 1$.

Using assumption $(A9)$, $\sup_{t\geq 0}|H^{''}_{c,1}(t)| = C < \infty$. As a consequence, since $\epsilon/a_n \to 0$ as $n \to \infty$, then there exists some $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$:

$$\left|\rho\frac{\epsilon^2}{2a_n^2}H^{''}_{c,1}(\xi_n)\right| \leq \frac{\rho\epsilon^2}{2a_n^2}C \leq \rho\frac{\epsilon}{2a_n}H^{'}_{c,1}(\tau_0). \qquad (A.7)$$

From (A.6) and (A.7), we have that:

$$P(|a_n(\tau_0 - T^1_{\max})| > \epsilon) \leq \left(1 - \rho\frac{\epsilon}{2a_n}H^{'}_{c,1}(\tau_0)\right)^n = \left(1 - \frac{\epsilon}{2a_n}H^{'}_{1}(\tau_0)\right)^n = b_n^{n/a_n}, \qquad (A.8)$$

where

$$b_n = \left(1 - \frac{\epsilon}{2a_n}H^{'}_{1}(\tau_0)\right)^{a_n} \xrightarrow[n\to\infty]{} r, \qquad (A.9)$$

with $r = \exp\left(-\frac{\epsilon H^{'}_1(\tau_0)}{2}\right) < 1$.

Using (A.8) and (A.9), to prove (A.5) it suffices to show that $\sum_{n=1}^{\infty} r^{n/a_n} < \infty$. For that purpose, we will prove that

$$r^{n/a_n} < n^{-2}, \text{ for } n \text{ large enough} \qquad (A.10)$$

and, since the hyperharmonic series $\sum_{n=1}^{\infty} n^{-2}$ is convergent, the series $\sum_{n=1}^{\infty} r^{n/a_n}$ will also be convergent.

Note that inequality (A.10) can be written as

$$2\log_R n < \frac{n}{a_n}, \qquad (A.11)$$

with $R = r^{-1} \in (1, \infty)$. Recall that $a_n = n^\alpha$ for some $\alpha \in (0, 1)$. Now condition (A.11) becomes

$$2 \log_R n < n^{1-\alpha},$$

which is true for n large enough, since $n^{-(1-\alpha)} 2 \log_R n \to 0$. As a consequence, $n^\alpha(\tau_0 - T^1_{\max}) \to 0$ *a.s.* for any $\alpha \in (0, 1)$. On the other hand, note that:

$$\frac{n^{-\alpha}}{\left(\frac{\ln n}{nh}\right)^{3/4}} = \left[\frac{nh^5}{\ln n} \frac{n^{4-20\alpha/3}}{(\ln n)^4}\right]^{3/20} \xrightarrow[n\to\infty]{} 0,$$

for $\alpha \geq 3/5$ and a sequence of bandwidths satisfying $(\ln n)^{-1} nh^5 = O(1)$. Therefore, the result in (A.4) holds. This completes the proof. □

In the next three lemmas, we use existing results in the literature for a fixed t such that $1 - H(t|x) \geq \theta > 0$ in $(t, x) \in [a, b] \times I_\delta$, and apply them to the random value $t = T^1_{\max}$. Note that if $\tau_0 < \tau_G(x) = \tau_H(x)$ for all $x \in I_\delta$, then from Lemma A.1.1, under assumption $(A10)$, we have that:

$$T^1_{\max} = \max_{i:\delta_i=1}(T_i) \to \tau_0 < \tau_H(x) \text{ in probability as } n \to \infty.$$

Therefore, for n large enough, $T^1_{\max} \leq \tau_0 < \tau_H(x)$ for all $x \in I_\delta$ and taking $b = \tau_0$ we can apply the results considering $t = T^1_{\max}$.

Lemma A.1.3. *Under assumptions* $(A1)$-$(A5)$, $(A10)$ *and* $(A12)$, *and if* $nh^5/\ln n = O(1)$ *and* $\ln n/(nh) \to 0$, *then the incidence estimator satisfies:*

$$1 - \hat{p}_h(x) = \exp\left(-\widehat{\Lambda}_h(T^1_{\max}|x)\right) + R_n(x), \text{ for all } x \in I,$$

where $\widehat{\Lambda}_h(t|x)$ *is the estimator of the conditional cumulative hazard function:*

$$\widehat{\Lambda}_h(t|x) = \sum_{i=1}^{n} \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)} I(T_{(i)} \leq t) = \int_0^t \frac{d\hat{H}^1_h(v|x)}{1 - \hat{H}_h(v^-|x)},$$

with $\hat{H}_h(t|x)$ *in (2.5),* $\hat{H}^1_h(t|x)$ *in (2.6), and*

$$\sup_{x\in I} |R_n(x)| = O\left((nh)^{-1}\right) a.s.$$

Proof of Lemma A.1.3. The incidence estimator is equal to:

$$1 - \hat{p}_h(x) = 1 - \hat{F}_h(T^1_{\max}|x),$$

where $\hat{F}_h(t|x) = 1 - \hat{S}_h(t|x)$ is the Beran estimator in (2.2). The result derives directly for $\hat{F}_h(t|x)$ from the so-called property 3) in the proof of part c) of Theorem 2

in Iglesias-Pérez & González-Manteiga (1999), when the data are subject to random left truncation and right censorship, for which assumptions $(A1)$,$(A3)$-$(A5)$ and $(A12)$ are required. Assumptions $(A2)$ and $(A10)$ allow to use the aforementioned property when $t = T^1_{\max}$. González-Manteiga & Cadarso-Suárez (1994) proved a similar result under right random censoring with fixed design on the covariate. □

Lemma A.1.4. *Under assumptions* $(A1)$-$(A11)$ *and* $(A13)$ *for* $x \in I$ *and if* $nh^5/\ln n = O(1)$, $\ln n/(nh) \to 0$, *then*

$$\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x) = \sum_{i=1}^{n} \tilde{B}_{hi}(x)\xi\left(T_i, \delta_i, \infty, x\right) + \tilde{R}_n\left(x\right),$$

with $\tilde{B}_{hi}$ *in (2.9),* ξ *in (2.11) and*

$$\sup_{x\in I}\left|\tilde{R}_n\left(x\right)\right| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

Proof of Lemma A.1.4. Under assumptions $(A1)$-$(A8)$, $(A10)$ and $(A11)$, we apply Theorem 2(b) of Iglesias-Pérez & González-Manteiga (1999) (similarly Theorem 2.2 of González-Manteiga & Cadarso-Suárez (1994) with fixed design using GM weights) to $t = T^1_{\max}$:

$$\begin{aligned}
&\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x) \qquad (A.12)\\
&= \sum_{i=1}^{n} \tilde{B}_{hi}(x)\xi\left(T_i, \delta_i, \infty, x\right) + \sum_{i=1}^{n}(B_{hi}(x) - \tilde{B}_{hi}(x))\xi\left(T_i, \delta_i, \infty, x\right)\\
&+ \sum_{i=1}^{n} B_{hi}(x)\left(\tilde{\xi}\left(T_i, \delta_i, \infty, x\right) - \xi\left(T_i, \delta_i, \infty, x\right)\right) + \tilde{\tilde{R}}_n\left(x\right),
\end{aligned}$$

with ξ in (2.11),

$$\tilde{\xi}\left(T_i, \delta_i, \infty, x\right) = \frac{\delta_i}{1 - H(T_i|x)} - \int_0^{T^1_{\max}} \frac{I(t < T_i)}{(1 - H(t|x))^2} dH^1(t|x)$$

and

$$\sup_{x\in I}\left|\tilde{\tilde{R}}_n\left(x\right)\right| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

Note that

$$\left|\tilde{\xi}\left(T_i, \delta_i, \infty, x\right) - \xi\left(T_i, \delta_i, \infty, x\right)\right| \leq \int_{T^1_{\max}}^{\tau_0} \frac{dH^1(t|x)}{(1 - H(t^-|x))^2} \quad \text{for all } i = 1, \ldots, n.$$

Then, under assumption $(A9)$ we apply Lemma A.1.2, and assuming $(A13)$, it is easy to prove that for a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$, the

third term in (A.12) is,

$$\sup_{x\in I}\left|\sum_{i=1}^{n} B_{hi}(x)\left(\tilde{\xi}\left(T_i,\delta_i,\infty,x\right)-\xi\left(T_i,\delta_i,\infty,x\right)\right)\right| = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

For the second term in (A.12), it is important to note that:

$$\begin{aligned}&\sum_{i=1}^{n}(B_{hi}(x)-\tilde{B}_{hi}(x))\xi(T_i,\delta_i,\infty,x)\\ &=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)\frac{m(x)-\hat{m}_h(x)}{\hat{m}_h(x)m(x)},\end{aligned}$$

with $\hat{m}_h(x)$ the Parzen-Rosenblatt estimator of $m(x)$. Using Theorem 3.3 of Arcones (1997), standard bias and variance calculations and Taylor expansions lead to

$$\sup_{x\in I}\left|\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)\right| = O\left(h^2+\sqrt{\frac{\ln\ln n}{nh}}\right) a.s.$$

Using again Theorem 3.3 of Arcones (1997), it is easy to prove that:

$$\sup_{x\in I}\left|\frac{m(x)-\hat{m}_h(x)}{\hat{m}_h(x)m(x)}\right| = O\left(h^2+\sqrt{\frac{\ln\ln n}{nh}}\right) a.s.$$

Therefore,

$$\sup_{x\in I}\left|\sum_{i=1}^{n}(B_{hi}(x)-\tilde{B}_{hi}(x))\xi(T_i,\delta_i,\infty,x)\right| = O\left(\left(h^2+\sqrt{\frac{\ln\ln n}{nh}}\right)^2\right) a.s.$$

For a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$ and $\frac{\ln n}{nh}\to 0$, it is immediate to prove that

$$\sup_{x\in I}\left|\sum_{i=1}^{n}(B_{hi}(x)-\tilde{B}_{hi}(x))\xi(T_i,\delta_i,\infty,x)\right| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

This completes the proof. □

Lemma A.1.5. *Under assumptions* (A1)-(A8) *and* (A10)-(A12), *and if* $nh^5/\ln n = O(1)$, $\ln n/(nh)\to 0$, *then*

$$\sup_{x\in I}\left|\widehat{\Lambda}_h\left(T^1_{\max}|x\right)-\Lambda\left(T^1_{\max}|x\right)\right| = O\left(\left(\frac{\ln n}{nh}\right)^{1/2}\right) a.s.$$

Proof of Lemma A.1.5. The equivalent result for a fixed $t \in [a, b]$ is within property 2 in the proof of part c) of Theorem 2 in Iglesias-Pérez & González-Manteiga (1999), for which assumptions $(A1)$, $(A3)$-$(A8)$, $(A11)$ and $(A12)$ are required. Assumptions $(A2)$ and $(A10)$ are needed to apply that result to $t = T^1_{\max}$. For the uniform strong consistency of the Beran estimator $\hat{F}_h(t|x)$, see also Dabrowska (1989). □

Theorem 2.3.2. *Under assumptions* $(A1)$-$(A13)$*, for any sequence of bandwidths satisfying* $nh^5(\ln n)^{-1} = O(1)$ *and* $\ln n/(nh) \to 0$*, then*

$$(1 - \hat{p}_h(x)) - (1 - p(x)) = (1 - p(x)) \sum_{i=1}^{n} \tilde{B}_{hi}(x) \xi(T_i, \delta_i, \infty, x) + R_n(x),$$

where

$$\tilde{B}_{hi}(x) = \frac{\frac{1}{nh} K\left(\frac{x - X_i}{h}\right)}{m(x)}, \tag{2.9}$$

$$\xi(T_i, \delta_i, t, x) = \frac{I(T_i \le t, \delta_i = 1)}{1 - H(T_i|x)} - \int_0^t \frac{I(u \le T_i) dH^1(u|x)}{(1 - H(u|x))^2} \tag{2.10}$$

and

$$\sup_{x \in I} |R_n(x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

Proof of Theorem 2.3.2. The incidence estimator can be split into the following terms:

$$\begin{aligned}
&(1 - \hat{p}_h(x)) - (1 - p(x)) \\
&= \hat{S}_h(T^1_{\max}|x) - (1 - p(x)) \\
&= \exp\left[-\widehat{\Lambda}_h(T^1_{\max}|x)\right] - \exp\left[-\Lambda(T^1_{\max}|x)\right] + R_2(x) + R_3(x),
\end{aligned} \tag{A.13}$$

with

$$R_2(x) = \hat{S}_h(T^1_{\max}|x) - \exp\left[-\widehat{\Lambda}_h(T^1_{\max}|x)\right]$$

and

$$R_3(x) = S(T^1_{\max}|x) - (1 - p(x)).$$

For the first term of (A.13) we apply a Taylor expansion of the function $\exp(y)$ around the value $y = -\Lambda(T^1_{\max}|x)$:

$$\begin{aligned}
&\exp\left[-\widehat{\Lambda}_h(T^1_{\max}|x)\right] - \exp\left[-\Lambda(T^1_{\max}|x)\right] \\
&= -\exp\left[-\Lambda(T^1_{\max}|x)\right] \left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right) + R_1(x),
\end{aligned} \tag{A.14}$$

with

$$R_1(x) = \frac{1}{2}\exp\left[-\Lambda^*(T^1_{\max}|x)\right]\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right)^2$$

and $\Lambda^*(T^1_{\max}|x) = \eta_n(x)$ a value between $\widehat{\Lambda}_h(T^1_{\max}|x)$ and $\Lambda(T^1_{\max}|x)$. Now, adding and substracting $1-p(x)$, and bearing in mind that $S(T^1_{\max}|x) = \exp[-\Lambda(T^1_{\max})|x]$,

$$\begin{aligned}&\exp\left[-\widehat{\Lambda}_h(T^1_{\max}x)\right] - \exp\left[-\Lambda(T^1_{\max}|x)\right]\\ &= -\left(1-p\left(x\right)\right)\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right) + R_1(x) - R_4(x),\end{aligned}$$

where

$$R_4(x) = \left[S\left(T^1_{\max}|x\right) - \left(1-p\left(x\right)\right)\right]\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right).$$

Now, inserting (A.14) in (A.13), we have:

$$\begin{aligned}&(1-\hat{p}_h(x)) - (1-p\left(x\right)) \qquad (A.15)\\ &= -\left(1-p\left(x\right)\right)\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right) + R_1(x) + R_2(x) + R_3(x) - R_4(x).\end{aligned}$$

The i.i.d. representation of $1-\hat{p}_h(x)$ now follows, assuming $(A1)$-$(A11)$ and $(A13)$, from Lemma A.1.4.

Let us study the remainder terms in (A.15), starting with $R_1(x)$. Taking into account that $\exp\left[-\Lambda^*(T^1_{\max}|x)\right]$ is bounded for all $x \in I$, and applying Lemma A.1.5, under the assumptions $(A1)$-$(A8)$ and $(A10)$-$(A12)$, we have

$$\sup_{x\in I}|R_1(x)| = O\left(\frac{\ln n}{nh}\right) a.s.$$

Regarding $R_2(x)$, under the assumptions $(A1)$, $(A3)$-$(A5)$, $(A10)$ and $(A12)$, directly from Lemma A.1.3 and using $\ln n/(nh) \to 0$ we obtain:

$$\sup_{x\in I}|R_2(x)| = O\left((nh)^{-1}\right) = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

Focusing on $R_3(x)$, note that it can be bounded as follows:

$$\begin{aligned}\sup_{x\in I}\left|R_3(x)\right| &= \sup_{x\in I}\left|S(T^1_{\max}|x) - (1-p(x))\,\right|\\ &= \sup_{x\in I}\left|\,\left[(1-p(x)) + p(x)S_0(T^1_{\max}|x)\right] - (1-p(x))\,\right|\\ &= \sup_{x\in I}\left|p(x)S_0(T^1_{\max}|x)\right| \le \sup_{x\in I}\left|S_0(T^1_{\max}|x)\right|\\ &= \sup_{x\in I}\left|S_0(T^1_{\max}|x) - S_0(\tau_0|x)\right|\\ &\le \sup_{x\in I}|(T^1_{\max} - \tau_0)S_0'(\tau_n|x)|, \qquad (A.16)\end{aligned}$$

with $\tau_n \in [T^1_{\max}, \tau_0]$. From condition $(A6)$, that implies that there exists some $\lambda > 0$ such that $\sup_{(t,x)\in[a,b]\times I} |S_0'(t|x)| \leq \lambda$, and using (A.4) and (A.16) for a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$, we have that:

$$\sup_{x\in I} |R_3(x)| = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

Finally, from Lemma A.1.5, the term R_4 is negligible with respect to R_3, and therefore:

$$\sup_{x\in I} |R_4(x)| = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

This completes the proof. □

The following Lemmas A.1.6, A.1.7 and A.1.8 are necessary to prove Corollary 2.3.1.

Lemma A.1.6. *Recall $m(x)$, the density function of X, and let us define $\Phi(y,t,x)$ as:*

$$\Phi(y,t,x) = E[\xi(T,\delta,t,x)|X = y], \tag{A.17}$$

with $\xi(T,\delta,t,x)$ in (2.10). If the kernel K is a symmetric density function, and assuming that the second derivative with respect to y of the function $\Phi(y,t,x)$ exists, then:

$$\begin{aligned} & E\left[K\left(\frac{x-X}{b}\right)\xi(T,\delta,t,x)\right] \\ = \;& \Phi(x,t,x)m(x)b + \frac{1}{2}b^3 d_K \frac{\partial^2}{\partial y^2}\left[\Phi(y,t,x)m(y)\right]|_{y=x} + o(b^3), \end{aligned}$$

where d_K was defined in (1.9).

Proof of Lemma A.1.6. Note that

$$\begin{aligned} E\left[K\left(\frac{x-X}{b}\right)\xi(T,\delta,t,x)\right] &= E\left[E\left(K\left(\frac{x-X}{b}\right)\xi(T,\delta,t,x)|X\right)\right] \\ &= E\left[K\left(\frac{x-X}{b}\right)E\left(\xi(T,\delta,t,x)|X\right)\right] \\ &= \int_{-\infty}^{\infty} K\left(\frac{x-y}{b}\right)\Phi(y,t,x)m(y)dy. \end{aligned}$$

Applying a change of variable and a Taylor expansion,

$$\begin{aligned}
&\int_{-\infty}^{\infty} K(u)\Phi(x-bu,t,x)m(x-bu)bdu \\
=&\int_{-\infty}^{\infty} K(u)\left[\Phi(x,t,x)m(x) - bu\frac{\partial}{\partial y}\left[\Phi(y,t,x)m(y)\right]|_{y=x}\right. \\
&\left.+\frac{1}{2}b^2u^2\frac{\partial^2}{\partial y^2}\left[\Phi(y,t,x)m(y)\right]|_{y=x}\right]bdu + o(b^3) \\
=&\Phi(x,t,x)m(x)b + \frac{1}{2}b^3 d_K \frac{\partial^2}{\partial y^2}\left[\Phi(y,t,x)m(y)\right]|_{y=x} + o(b^3).
\end{aligned}$$

This concludes the proof. □

Lemma A.1.7. *Let $\Phi(y,t,x) = E[\xi(T,\delta,t,x)|X=y]$, with $\xi(T,\delta,t,x)$ in (2.10). The following equality holds:*

$$\Phi(y,t,x) = \int_0^t \frac{dH^1(v|y)}{1-H(v|x)} - \int_0^t (1-H(v|y))\frac{dH^1(v|x)}{(1-H(v|x))^2}$$

and then

$$\Phi(x,t,x) = 0 \ \forall t \geq \infty. \tag{A.18}$$

Proof of Lemma A.1.7. Considering

$$\begin{aligned}
\Phi(y,t,x) &= E\left[\frac{I(T\leq t,\delta=1)}{1-H(T|x)}\middle|X=y\right] - E\left[\int_0^t \frac{I(v\leq T)dH^1(v|x)}{(1-H(v|x))^2}\middle|X=y\right] \\
&= A' - A''.
\end{aligned}$$

We start with A':

$$\begin{aligned}
A' &= E\left[\frac{I(T\leq t,\delta=1)}{1-H(T|x)}\middle|X=y\right] = E\left[E\left(\frac{I(T\leq t,\delta=1)}{1-H(T|x)}\middle|T,X=y\right)\right] \\
&= E\left[\frac{I(T\leq t)}{1-H(T|x)}E\left(\delta|T,X=y\right)\middle|X=y\right] = E\left[\frac{I(T\leq t)}{1-H(T|x)}q(T,y)\middle|X=y\right] \\
&= \int_0^t \frac{q(v,y)dH(v|y)}{1-H(v|x)} = \int_0^t \frac{dH^1(v|y)}{1-H(v|x)},
\end{aligned}$$

where

$$q(t,y) = E\left(\delta|T=t,X=y\right) \text{ and } H_1(t|y) = P\left(T\leq t,\delta=1|X=y\right).$$

We continue with A'':

$$\begin{aligned}
A'' &= E\left[\int_0^t \frac{I(v\leq T)dH^1(v|x)}{(1-H(v|x))^2}\middle|X=y\right] = \int_0^t E\left[I(v\leq T)|X=y\right]\frac{dH^1(v|x)}{(1-H(v|x))^2} \\
&= \int_0^t (1-H(v|y))\frac{dH^1(v|x)}{(1-H(v|x))^2}.
\end{aligned}$$

Therefore,

$$\Phi\left(y,t,x\right)=\int_0^t\frac{dH^1\left(v|y\right)}{1-H(v|x)}-\int_0^t(1-H(v|y))\frac{dH^1(v|x)}{\left(1-H(v|x)\right)^2}$$

and then $\Phi(x,t,x)=0$. □

Lemma A.1.8. *Let $g(x,y)$ and $m(y)$ be two functions such that both $g''(x,y)=\partial^2 g(x,y)/\partial y^2$ and the second derivative of $m(y)$ exist, and considering that the kernel function K is a symmetric density, then:*

$$\int K^2\left(\frac{x-y}{b}\right)g(x,y)m(y)dy=bg(x,x)m(x)c_K$$
$$+b^3e_K\left(\frac{g(x,x)m''(x)}{2}+\frac{m(x)}{2}g''(x,x)+m'(x)g'(x,x)\right)+o(b^3),$$

where c_K was defined in (1.10) and $e_K=\int v^2K^2(v)dv$.

Proof of Lemma A.1.8. We apply a change of variable:

$$\int K^2\left(\frac{x-y}{b}\right)g(x,y)m(y)dy=b\int K^2(u)\;g(x,x-bu)\;m(x-bu)\;du.$$

Now using a Taylor expansion, the result is directly derived. □

Corollary 2.3.1. *An asymptotic expression of the mean squared error of the incidence estimator is given by:*

$$AMSE_x(h)=\frac{1}{nh}(1-p(x))^2c_K\sigma^2(x)+\left[h^2\frac{1}{2}d_K(1-p(x))\mu(x)\right]^2,\qquad(2.13)$$

where the first term corresponds to the asymptotic variance and the second one to the asymptotic squared bias, with d_K in (1.9) and c_K in (1.10) and, following a notation similar to that in Dabrowska (1992):

$$\sigma^2(x)=\frac{1}{m(x)}\int_0^\infty\frac{dH^1(t|x)}{(1-H(t|x))^2}$$

and

$$\mu(x)=\frac{2\Phi'(x,\infty,x)m'(x)+\Phi''(x,\infty,x)m(x)}{m(x)},\qquad(2.14)$$

where

$$\Phi\left(y,t,x\right)=\int_0^t\frac{dH^1\left(v|y\right)}{1-H(v|x)}-\int_0^t(1-H(v|y))\frac{dH^1(v|x)}{\left(1-H(v|x)\right)^2},\qquad(2.15)$$

with $\Phi'(y,t,x)=\partial\Phi(y,t,x)/\partial y$ and $\Phi''(y,t,x)=\partial^2\Phi(y,t,x)/\partial y^2$. Note that the AMSE denotes the MSE of the dominant part of the almost sure representation

of the incidence estimator. If the censoring distribution does not depend on the covariate, then $\mu(x)$ can also be written as follows:

$$\mu(x) = \frac{1}{m(x)}\left([p(x)m(x)]'' - p(x)m''(x)\right)\left(1 - \int_0^\infty \frac{S(dv|x)}{S(v|x)^2}\right). \quad (2.16)$$

Proof of Corollary 2.3.1. The dominant part of the bias and the variance of $\hat{p}_h(x) - p(x)$ is the same as that of the i.i.d. representation given in Theorem 2.3.2:

$$(1-p(x))\sum_{i=1}^n \tilde{B}_{hi}(x)\xi(T_i,\delta_i,\infty,x) = -\frac{1-p(x)}{m(x)}\sum_{i=1}^n \frac{1}{nh}K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x).$$

Let us define

$$I = \sum_{i=1}^n \frac{1}{nh}\left[K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x) - E\left(K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)\right)\right] \text{ and}$$

$$II = \sum_{i=1}^n \frac{1}{nh}E\left(K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)\right).$$

Then,

$$\hat{p}_h(x) - p(x) = \frac{1-p(x)}{m(x)}(I+II) + O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) a.s.$$

The expected value of the estimator of $p(x)$ is asymptotically equal to that of

$$p(x) + \frac{1-p(x)}{m(x)}(I+II).$$

Note that the expected value of I is $E(I) = 0$. We continue with the term II:

$$\begin{aligned} II &= \sum_{i=1}^n \frac{1}{nh}E\left[K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)\right] \\ &= \frac{1}{h}E\left[K\left(\frac{x-X_1}{h}\right)\xi(T_1,\delta_1,\infty,x)\right]. \end{aligned}$$

It follows directly from Lemma A.1.6 that

$$\left[E\left(K\left(\frac{x-X_1}{h}\right)\xi(T_1,\delta_1,t,x)\right)\right]^2 = \left[\Phi(x,t,x)m(x)h + O(b_2^3)\right]^2.$$

Lemma A.1.7 implies that $E[\xi(T,\delta,t,x)|X=x] = \Phi(x,t,x) = 0$ and, as a consequence,

$$\left[E\left(K\left(\frac{x-X_1}{h}\right)\xi(T_1,\delta_1,\infty,x)\right)\right] = O(b_2^3). \quad (A.19)$$

If the censoring distribution does not depend on the covariate, $\Phi(y,t,x)$ can be expressed as:

$$\begin{aligned}\Phi(y,t,x) &= \int_0^t \frac{F(dv|y)}{S(v|x)} - \int_0^t S(v|y)\frac{F(dv|x)}{S(v|x)^2} \\ &= -\int_0^t \frac{p(y)\,S_0(dv|y)}{S(v|x)} + \int_0^t \left[(1-p(y)) + p(y)\,S_0(v|y)\right]\frac{S(dv|x)}{S(v|x)^2}. \quad \text{(A.20)}\end{aligned}$$

We obtain the first and second derivatives of $\Phi(y,t,x)$ in (2.15). Let us consider the notation $S_0'(t|x)$ for the corresponding derivative with respect to x. Furthermore, note that $F(v|x) = p(x)(1-S_0(v|x))$ and $F(dv|x) = -S(dv|x) = -p(x)\,S_0(dv|x)$. The derivative of $\Phi(y,t,x)$ in (A.20) is:

$$\begin{aligned}&\frac{\partial}{\partial y}\Phi(y,t,x)\,|_{y=x,t=\infty} \\ =\ & -\int_0^\infty \frac{\frac{\partial}{\partial y}\left[p(y)\,S_0(dv|y)\right]|_{y=x}}{S(v|x)} \\ +\ & \int_0^\infty \frac{\partial}{\partial y}\left[(1-p(y)) + p(y)\,S_0(v|y)\right]|_{y=x}\frac{S(dv|x)}{S(v|x)^2} \\ =\ & -\int_0^\infty \frac{p'(x)\,S_0(dv|x)}{S(v|x)} - \int_0^\infty \frac{p(x)\,S_0'(dv|x)}{S(v|x)} - p'(x)\int_0^\infty \frac{p(x)\,S_0(dv|x)}{S(v|x)^2} \\ +\ & \int_0^\infty p'(x)\,S_0(v|x)\frac{p(x)\,S_0(dv|x)}{S(v|x)^2} + \int_0^\infty p(x)\,S_0'(v|x)\frac{p(x)\,S_0(dv|x)}{S(v|x)^2}.\end{aligned}$$

Adding and substracting the same terms, suitably chosen, the derivative of $\Phi(y,t,x)$ equals:

$$\begin{aligned}&\frac{\partial}{\partial y}\Phi(y,t,x)\,|_{y=x,t=\infty} \\ =\ & -p'(x)\int_0^\infty \frac{S_0(dv|x)}{S(v|x)} - p'(x)\int_0^\infty p(x)\left(1-S_0(v|x)\right)\frac{S_0(dv|x)}{S(v|x)^2} \\ -\ & p(x)\int_0^\infty \frac{S_0'(dv|x)}{S(v|x)} \pm p'(x)\int_0^\infty \frac{S_0(dv|x)}{S(v|x)} \\ +\ & p^2(x)\int_0^\infty S_0'(v|x)\frac{S_0(dv|x)}{S(v|x)^2} \pm p(x)\int_0^\infty p'(x)\,S_0(v|x)\frac{S_0(dv|x)}{S(v|x)^2} \\ =\ & -p'(x)\int_0^\infty \frac{S_0(dv|x)}{S(v|x)} - p'(x)\int_0^\infty \left(1-S(v|x)\right)\frac{S_0(dv|x)}{S(v|x)^2} \\ -\ & \int_0^\infty \left[p(x)\,S_0(dv|x)\right]'\frac{1}{S(v|x)} + p'(x)\int_0^\infty \frac{S_0(dv|x)}{S(v|x)} \\ +\ & p(x)\int_0^\infty \left[p(x)\,S_0(v|x)\right]'\frac{S_0(dv|x)}{S(v|x)^2} - p(x)\int_0^\infty p'(x)\,S_0(v|x)\frac{S_0(dv|x)}{S(v|x)^2}.\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\partial}{\partial y}\Phi\left(y,t,x\right)|_{y=x,t=\infty}\\
=\;& -p'\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S^2\left(v|x\right)} + p'\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)}\\
-\;& p'\left(x\right)\int_0^\infty p\left(x\right)S_0\left(v|x\right)\frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2} - \int_0^\infty \left[p\left(x\right)S_0\left(dv|x\right)\right]'\frac{1}{S\left(v|x\right)}\\
+\;& p\left(x\right)\int_0^\infty \left[p\left(x\right)S_0\left(v|x\right)\right]'\frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2}.
\end{aligned}$$

Thus, considering that $S_0\left(\infty|x\right) = 0$, $S\left(\infty|x\right) = 1 - p\left(x\right)$ and $S_0\left(0|x\right) = S\left(0|x\right) = 1$, then

$$\begin{aligned}
&\frac{\partial}{\partial y}\Phi\left(y,t,x\right)|_{y=x,t=\infty} = -p'\left(x\right)p\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2} + p'\left(x\right)\\
=\;& p'\left(x\right)\left(1 - \int_0^\infty \frac{p\left(x\right)S_0\left(dv|x\right)}{S\left(v|x\right)^2}\right) = p'\left(x\right)\left(1 - \int_0^\infty \frac{S\left(dv|x\right)}{S\left(v|x\right)^2}\right). \quad \text{(A.21)}
\end{aligned}$$

Regarding the second derivative of $\Phi\left(y,t,x\right)$ in (2.15), we follow the same ideas:

$$\begin{aligned}
&\frac{\partial^2}{\partial y^2}\Phi\left(y,t,x\right)|_{y=x,t=\infty}\\
=\;& \int_0^\infty \frac{\frac{\partial^2}{\partial y^2}F\left(dv|y\right)|_{y=x}}{S\left(v|x\right)} - \int_0^\infty \frac{\partial^2}{\partial y^2}S\left(v|y\right)|_{y=x}\frac{F\left(dv|x\right)}{S\left(v|x\right)^2}\\
=\;& -\int_0^\infty \frac{\frac{\partial^2}{\partial y^2}\left[p\left(y\right)S_0\left(dv|y\right)\right]}{S\left(v|x\right)}\\
+\;& \int_0^\infty \frac{\partial^2}{\partial y^2}\left[\left(1 - p\left(y\right)\right) + p\left(y\right)S_0\left(v|y\right)\right]|_{y=x}\frac{p\left(x\right)S_0\left(dv|x\right)}{S\left(v|x\right)^2}.
\end{aligned}$$

Choosing a suitable term and then adding and substracting it, the second derivative of $\Phi\left(y,t,x\right)$ can be expressed as:

$$\begin{aligned}
&\frac{\partial^2}{\partial y^2}\Phi\left(y,t,x\right)|_{y=x,t=\infty}\\
=\;& -p''\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2}\\
-\;& 2p'\left(x\right)\int_0^\infty \frac{S_0'\left(dv|x\right)}{S\left(v|x\right)} - p\left(x\right)\int_0^\infty \frac{S_0''\left(dv|x\right)}{S\left(v|x\right)} \pm p''\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)}\\
+\;& p\left(x\right)\int_0^\infty \left(2p'\left(x\right)S_0'\left(v|x\right) + p\left(x\right)S_0''\left(v|x\right) \pm p''\left(x\right)S_0\left(v|x\right)\right)\frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2}.
\end{aligned}$$

Then,

$$\begin{aligned}
&\frac{\partial^2}{\partial y^2}\Phi\left(y,t,x\right)|_{y=x,t=\infty}\\
=\;& -p''\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2} + p''\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)}\\
-\;& p\left(x\right)p''\left(x\right)\int_0^\infty S_0\left(v|x\right)\frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2}\\
+\;& p\left(x\right)\int_0^\infty \left[p\left(x\right)S_0\left(v|x\right)\right]''\frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2} - \int_0^\infty \frac{1}{S\left(v|x\right)}\left[p\left(x\right)S_0\left(dv|x\right)\right]''
\end{aligned} \tag{A.22}$$

Note that the first three terms in (A.22) are equal to

$$\begin{aligned}
&p''\left(x\right)\int_0^\infty \left(-\frac{1}{S\left(v|x\right)^2} + \frac{1}{S\left(v|x\right)} - p\left(x\right)S_0\left(v|x\right)\frac{1}{S\left(v|x\right)^2}\right)S_0\left(dv|x\right)\\
=\;& p''\left(x\right)\int_0^\infty \left(-1 + S\left(v|x\right) - p\left(x\right)S_0\left(v|x\right)\right)\frac{1}{S\left(v|x\right)^2}S_0\left(dv|x\right)\\
=\;& -p''\left(x\right)p\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2}.
\end{aligned} \tag{A.23}$$

Regarding the last two terms in (A.22),

$$\begin{aligned}
&p\left(x\right)\int_0^\infty \left[p\left(x\right)S_0\left(v|x\right)\right]''\frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2} - \int_0^\infty \frac{1}{S\left(v|x\right)}\left[p\left(x\right)S_0\left(dv|x\right)\right]''\\
=\;& -\int_0^\infty \frac{1}{S^2\left(v|x\right)}\left(\left[p\left(x\right)S_0\left(dv|x\right)\right]''S\left(v|x\right) - \left[p\left(x\right)S_0\left(v|x\right)\right]''S\left(dv|x\right)\right)\\
=\;& -\int_0^\infty \frac{\partial}{\partial v}\left[\frac{\left[p\left(x\right)S_0\left(v|x\right)\right]''}{S\left(v|x\right)}\right]dv\\
=\;& -\left(\frac{\left[p\left(x\right)S_0\left(\infty|x\right)\right]''}{S\left(\infty|x\right)} - \frac{\left[p\left(x\right)S_0\left(0|x\right)\right]''}{S\left(0|x\right)}\right) = p''\left(x\right).
\end{aligned} \tag{A.24}$$

Considering that $S_0\left(\infty|x\right)=0$, $S\left(\infty|x\right)=1-p\left(x\right)$ and $S_0\left(0|x\right)=S\left(0|x\right)=1$, thus, using (A.23) and (A.24) in (A.22), and taking into account that $F(dv|x) = -S(dv|x) = -p(x)S_0(dv|x)$, the second derivative of $\Phi\left(y,t,x\right)$ reduces to

$$\begin{aligned}
&\frac{\partial^2}{\partial y^2}\Phi\left(y,t,x\right)|_{y=x,t=\infty}\\
=\;& -p''\left(x\right)p\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2} + p''\left(x\right) = p''\left(x\right)\left(1 - p\left(x\right)\int_0^\infty \frac{S_0\left(dv|x\right)}{S\left(v|x\right)^2}\right)\\
=\;& p''\left(x\right)\left(1 - \int_0^\infty \frac{S\left(dv|x\right)}{S\left(v|x\right)^2}\right).
\end{aligned} \tag{A.25}$$

Finally, let us recall $\Phi'(x,\infty,x)$ in (A.21) and $\Phi''(x,\infty,x)$ in (A.25). Therefore, the function $\mu(x)$ can be expressed as:

$$\begin{aligned}\mu(x) &= \frac{2\Phi'(x,\infty,x)\,m'(x)+\Phi''(x,\infty,x)\,m(x)}{m(x)}\\ &= \frac{1}{m(x)}\left[2p'(x)\left(1-\int_0^\infty \frac{S(dv|x)}{S(v|x)^2}\right)m'(x)\right.\\ &+ \left. p''(x)\left(1-\int_0^\infty \frac{S(dv|x)}{S(v|x)^2}\right)m(x)\right]\\ &= \frac{1}{m(x)}\left([p(x)\,m(x)]''-p(x)\,m''(x)\right)\left(1-\int_0^\infty \frac{S(dv|x)}{S(v|x)^2}\right).\end{aligned}$$

The variance of $\hat{p}_h$ is asymptotically equal to

$$\frac{(1-p(x))^2}{m^2(x)}Var(I),$$

since II is not a random term and therefore, its variance is zero. The variance of I is given by:

$$\begin{aligned}&Var(I)\\ =\ &\sum_{i=1}^{n} Var\left(\frac{1}{nh}\left[K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)-E\left(K\left(\frac{x-X_i}{h}\right)\xi(T_i,\delta_i,\infty,x)\right)\right]\right)\\ =\ &\frac{1}{n}Var\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\xi(T_1,\delta_1,\infty,x)\right).\end{aligned}$$

Note that

$$\begin{aligned}&Var\left(K\left(\frac{x-X_1}{h}\right)\xi(T_1,\delta_1,t,x)\right) \qquad (A.26)\\ =\ &E\left(K^2\left(\frac{x-X_1}{h}\right)\xi^2(T_1,\delta_1,t,x)\right)-\left[E\left(K\left(\frac{x-X_1}{h}\right)\xi(T_1,\delta_1,t,x)\right)\right]^2\end{aligned}$$

and we start with the first term in (A.26):

$$\begin{aligned}&E\left[K^2\left(\frac{x-X_1}{h}\right)\xi^2(T_1,\delta_1,t,x)\right]=E\left[K^2\left(\frac{x-X_1}{h}\right)E\left(\xi^2(T_1,\delta_1,t,x)|X_1\right)\right]\\ &=\int K^2\left(\frac{x-y}{h}\right)E\left(\xi^2(T_1,\delta_1,t,x)|X_1=y\right)m(y)dy.\end{aligned}$$

Let us define

$$\Phi_1(y,t,x)=E(\xi^2(T,\delta,t,x)|X=y), \qquad (A.27)$$

with ξ in (2.10). From Lemma A.1.8, considering $g(x,y)=\Phi_1(y,t,x)$ for a fixed t, then the first term in (A.26) is

$$E\left[K^2\left(\frac{x-X_1}{b_2}\right)\xi^2(T_1,\delta_1,t,x)\right]=b_2\Phi_1(x,t,x)m(x)c_K+O(b_2^3). \qquad (A.28)$$

Then, the variance is given by

$$\begin{aligned} Var(\hat{p}_h(x)) &= \frac{1}{n}\frac{(1-p(x))^2}{m(x)}\left(\frac{1}{h}\Phi_1(x,\infty,x)c_K + O(h)\right) \\ &= \frac{1}{nh}\frac{(1-p(x))^2}{m(x)}c_K\int_0^\infty \frac{dH^1(t|x)}{(1-H(t^-|x))^2} + o(nh^{-1}), \end{aligned}$$

with $\Phi_1(x,\infty,x) = \int_0^\infty \frac{dH^1(t|x)}{(1-H(t^-|x))^2}$.

□

A.2 Proofs of the results in Chapter 3

Lemma A.2.1. *Let $g(x,y)$ and $m(y)$ be two functions such that both $g''(x,y) = \partial^2 g(x,y)/\partial y^2$ and the second derivative of the function $m(y)$ exist, and considering that the kernel function K is a symmetric density, then:*

$$\begin{aligned}&\int K\left(\frac{x-y}{b_1}\right) K\left(\frac{x-y}{b_2}\right) g(x,y)m(y)dy \\ &= b_1 g(x,x)m(x)\int K(u)K\left(\frac{b_1}{b_2}u\right)du + O(b_1^2).\end{aligned}$$

Proof of Lemma A.2.1. We apply a change of variable and a Taylor expansion:

$$\begin{aligned}&\int K\left(\frac{x-y}{b_1}\right) K\left(\frac{x-y}{b_2}\right) g(x,y)m(y)dy \\ &= b_1\int K(u)\,K\left(\frac{b_1}{b_2}u\right)[g(x,x)m(x) \\ &+b_1 u g'(x,x)m(x) + b_1 u m'(x)g(x,x) + O(b_1^2)]\,du \\ &= b_1 g(x,x)m(x)\int K(u)\,K\left(\frac{b_1}{b_2}u\right)\,du + O(b_1^2),\end{aligned}$$

where $g'(x,y)$ is the derivative with respect to y. This concludes the proof. □

Theorem 3.2.1. *Under assumptions (A1)-(A13), and for two sequences of bandwidths satisfying (A14), then the i.i.d. representation of the nonparametric latency estimator in (3.1) is:*

$$\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) = \sum_{i=1}^{n}\eta_{b_1,b_2}(T_i,\delta_i,X_i,t,x) + O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4} + \left(\frac{\ln n}{nb_2}\right)^{3/4}\right) a.s.,$$

where

$$\begin{aligned}\eta_{b_1,b_2}(T_i,\delta_i,X_i,t,x) &= -\frac{S(t|x)}{p(x)}\tilde{B}_{b_2 i}(x)\xi(T_i,\delta_i,t,x) \\ &- \frac{(1-p(x))(1-S(t|x))}{p(x)^2}\tilde{B}_{b_1 i}(x)\xi(T_i,\delta_i,\infty,x),\end{aligned} \tag{3.3}$$

with $\xi(T_i,\delta_i,t,x)$ defined in (2.10) and $\tilde{B}_{b_j i}(x)$, $j=1,2$ in (2.9).

Proof of Theorem 3.2.1. Departing from (3.1), then:

$$\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) = A_{11} + A_{21} + A_{31} + A_{12} + A_{22} + A_{32},$$

where

$$\begin{aligned}
A_{11} &= \frac{\hat{S}_{b_2}(t|x) - S(t|x)}{p(x)}, \\
A_{21} &= \frac{\hat{p}_{b_1}(x) - p(x)}{p(x)}, \\
A_{31} &= -\frac{(S(t|x) - (1-p(x)))(\hat{p}_{b_1}(x) - p(x))}{p(x)^2}, \\
A_{12} &= \frac{(\hat{S}_{b_2}(t|x) - S(t|x))(p(x) - \hat{p}_{b_1}(x))}{\hat{p}_{b_1}(x)p(x)}, \\
A_{22} &= -\frac{(\hat{p}_{b_1}(x) - p(x))^2}{\hat{p}_{b_1}(x)p(x)}
\end{aligned}$$

and

$$A_{32} = \frac{(S(t|x) - (1-p(x)))(p(x) - \hat{p}_{b_1}(x))^2}{\hat{p}_{b_1}(x)p(x)^2}.$$

Note that

$$A_{21} + A_{31} = \frac{1 - S(t|x)}{p(x)^2}(\hat{p}_{b_1}(x) - p(x))$$

and

$$A_{22} + A_{32} = \frac{S(t|x) - 1}{p(x)^2}\frac{(\hat{p}_{b_1}(x) - p(x))^2}{\hat{p}_{b_1}(x)}.$$

The i.i.d. representation derives from the terms A_{11} and $A_{21} + A_{31}$. From Theorem 2 in Iglesias-Pérez & González-Manteiga (1999), we obtain the i.i.d. representation of the term A_{11}:

$$A_{11} = \frac{1}{p(x)}\left(-S(t|x)\sum_{i=1}^{n}\tilde{B}_{b_2 i}(x)\xi(T_i, \delta_i, t, x) + O\left(\left(\frac{\ln n}{nb_2}\right)^{3/4}\right)\right) a.s.$$

We continue studying the terms $A_{21} + A_{31}$. From Theorem 2.3.2 in Chapter 2 we obtain the following i.i.d. representation:

$$\begin{aligned}
&A_{21} + A_{31} \\
&= \frac{(1 - S(t|x))}{p(x)^2}\left(-(1-p(x))\sum_{i=1}^{n}\tilde{B}_{b_1 i}(x)\xi(T_i, \delta_i, \infty, x) + O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4}\right)\right) a.s.
\end{aligned}$$

The remainder terms A_{12}, A_{22} and A_{32} are negligible. We will study them separately: on the one hand, A_{12} and on the other hand, $A_{22} + A_{32}$.

We start proving the negligibility of A_{12}. We depart from Lemma 5 in Iglesias-Pérez & González-Manteiga (1999),

$$\hat{S}_{b_2}(t|x) - S(t|x) = O\left(\sqrt{\frac{\ln\ln n}{nb_2}} + b_2^2\right) a.s. \tag{A.29}$$

and from Theorem 3.3 in Arcones (1997) and the Strong Law of Large Numbers (SLLN), then

$$p(x) - \hat{p}_{b_1}(x) = O\left(\sqrt{\frac{\ln\ln n}{nb_1}} + b_1^2\right) a.s. \tag{A.30}$$

To prove the negligibility of A_{12}, from (A.29) and (A.30), we have that:

$$(\hat{S}_{b_2}(t|x) - S(t|x))(p(x) - \hat{p}_{b_1}(x)) = O\left(\left(\sqrt{\frac{\ln\ln n}{nb_1}} + b_1^2\right)\left(\sqrt{\frac{\ln\ln n}{nb_2}} + b_2^2\right)\right) \ a.s.$$

Note that:

$$\begin{aligned}&\left(\sqrt{\frac{\ln\ln n}{nb_1}} + b_1^2\right)\left(\sqrt{\frac{\ln\ln n}{nb_2}} + b_2^2\right)\\ =\ & \frac{\ln\ln n}{n\sqrt{b_1}\sqrt{b_2}} + b_2^2\sqrt{\frac{\ln\ln n}{nb_1}} + b_1^2\sqrt{\frac{\ln\ln n}{nb_2}} + b_1^2 b_2^2. \end{aligned} \tag{A.31}$$

To prove that (A.31) is $O\left(\frac{\ln n}{nb_1}\right)^{3/4}$, it suffices to prove the following equalities:

$$\frac{\frac{\ln\ln n}{n\sqrt{b_1}\sqrt{b_2}}}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1),\ \frac{b_2^2\sqrt{\frac{\ln\ln n}{nb_1}}}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1),\ \frac{b_1^2\sqrt{\frac{\ln\ln n}{nb_2}}}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1) \text{ and } \frac{b_1^2 b_2^2}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1),$$

which, respectively, lead to the resulting conditions:

$$\frac{(\ln\ln n)^4 b_1}{nb_2^2(\ln n)^3} = O(1), \frac{(\ln\ln n)^2 nb_1 b_2^8}{(\ln n)^3} = O(1), \frac{(\ln\ln n)^2 nb_1^{11}}{(\ln n)^3 b_2^2} = O(1) \text{ and } \frac{n^3 b_1^{11} b_2^8}{(\ln n)^3} = O(1).$$

To prove that (A.31) is $O\left(\frac{\ln n}{nb_2}\right)^{3/4}$, a similar argument just changing b_1 by b_2 leads to the following conditions:

$$\frac{(\ln\ln n)^4 b_2}{nb_1^2(\ln n)^3} = O(1),\ \frac{(\ln\ln n)^2 nb_2^{11}}{(\ln n)^3 b_1^2} = O(1),\ \frac{(\ln\ln n)^2 nb_1^8 b_2}{(\ln n)^3} = O(1) \text{ and } \frac{n^3 b_1^8 b_2^{11}}{(\ln n)^3} = O(1).$$

It is straightforward to check that if the bandwidths satisfy $nb_i^5/(\ln n) = O(1)$, with $i = 1, 2$, then $(\ln\ln n)^2 nb_i b_j^8/(\ln n)^3 = O(1)$ and $n^3 b_i^{11} b_j^8/(\ln n)^3 = O(1)$, for $i, j = 1, 2$, with $i \neq j$. Then, under assumption $(A14)$ for the bandwidths,

$$(\hat{S}_{b_2}(t|x) - S(t|x))(p(x) - \hat{p}_{b_1}(x)) = O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4} + \left(\frac{\ln n}{nb_2}\right)^{3/4}\right) \ a.s.$$

For a sequence b_1 satisfying $b_1 \to 0$ and $nb_1 \to \infty$, from (A.30), the convergence $\hat{p}_{b_1}(x) \to p(x)$ *a.s.* is proven. It follows that

$$A_{12} = O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4} + \left(\frac{\ln n}{nb_2}\right)^{3/4}\right) \ a.s.$$

With respect to $A_{22} + A_{32}$, we will prove that

$$A_{22} + A_{32} = O\left(\left(\sqrt{\frac{\ln\ln n}{nb_1}} + b_1^2\right)^2\right) \ a.s.$$

Note that from (A.30)

$$\hat{p}_{b_1}(x) \to p(x) \ a.s. \qquad \text{and} \qquad (\hat{p}_{b_1}(x) - p(x))^2 = O\left(\left(\sqrt{\tfrac{\ln\ln n}{nb_1}} + b_1^2\right)^2\right) \ a.s.$$

As a consequence, $(\hat{p}_{b_1}(x) - p(x))^2 = O((\ln n/nb_1)^{3/4})$ $a.s.$ if

$$\frac{\frac{\ln\ln n}{nb_1} + b_1^4 + b_1^2\sqrt{\frac{\ln\ln n}{nb_1}}}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} \quad = \quad O(1). \tag{A.32}$$

In order to show (A.32), it suffices to prove that

$$\frac{\frac{\ln\ln n}{nb_1}}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1), \ \frac{b_1^4}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1) \text{ and } \frac{b_1^2\sqrt{\frac{\ln\ln n}{nb_1}}}{\left(\frac{\ln n}{nb_1}\right)^{3/4}} = O(1),$$

which lead to the resulting conditions, respectively:

$$\frac{(\ln\ln n)^4}{nb_1(\ln n)^3} = O(1), \ \frac{n^3 b_1^{19}}{(\ln n)^3} = O(1) \text{ and } \frac{nb_1^9(\ln\ln n)^2}{(\ln n)^3} = O(1). \tag{A.33}$$

Since $nb_1^5/(\ln n) = O(1)$ implies the second condition in (A.33), and this leads to the third condition in (A.33), then assumption $(A14)$ on the bandwidths implies (A.32), and therefore

$$(\hat{p}_{b_1}(x) - p(x))^2 = O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4}\right) \ a.s.$$

Finally, using that $\hat{p}_{b_1}(x) \to p(x)$ $a.s.$

$$A_{22} + A_{32} = O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4} + \left(\frac{\ln n}{nb_2}\right)^{3/4}\right) \ a.s.$$

This concludes the proof. □

The following Lemmas A.2.2 and A.2.3 are necessary to prove Theorem 3.2.2.

Lemma A.2.2. *Let us denote:*

$$\Phi_1(y,t,x) = E(\xi^2(T_1,\delta_1,t,x)|X = y), \tag{A.34}$$

with ξ in (2.10). The term $\Phi_1(y,t,x)$ satisfies, for any $t \in [a,b]$,

$$\Phi_1(x,t,x) = \int_0^t \frac{dH^1(v|x)}{(1 - H(v|x))^2}. \tag{A.35}$$

Proof of Lemma A.2.2. From $\Phi_1(y,t,x)$ in (A.34), with ξ in (2.10), then:

$$\begin{aligned}\Phi_1(y,t,x) &= E\left[\frac{I(T\leq t,\delta=1)}{(1-H(T|x))^2}\Big|X=y\right]\\ &+ E\left[\int_0^t\int_0^t \frac{I(u\leq T)I(v\leq T)}{(1-H(u|x))^2(1-H(v|x))^2}dH^1(u|x)dH^1(v|x)\Big|X=y\right]\\ &- 2E\left[\frac{I(T\leq t,\delta=1)}{1-H(T|x)}\int_0^t\frac{I(u\leq T)\,dH^1(u|x)}{(1-H(u|x))^2}\Big|X=y\right]\\ &= A+B-2C.\end{aligned}$$

The first term in the decomposition of $\Phi_1(y,t,x)$ is

$$A=\int_0^t\frac{q(v,y)}{(1-H(v|x))^2}dH(v|y)=\int_0^t\frac{dH^1(v|y)}{(1-H(v|x))^2},$$

where $q(t,y)=E(\delta|T=t,X=y)$.

The second term is

$$B=\int_0^t\int_0^t\frac{1-H(\max(w,v)|y)}{(1-H(v|x))^2(1-H(w|x))^2}dH^1(v|x)dH^1(w|x).$$

Integrating in the supports $\{(v,w)\in[0,t]\times[0,t]\,/v\leq w\}$ and $\{(v,w)\in[0,t]\times[0,t]\,/w<v\}$, the term B is

$$B=2\int_0^t\frac{1}{(1-H(v|x))^2}\left(\int_v^t\frac{1-H(w|y)}{(1-H(w|x))^2}dH^1(w|x)\right)dH^1(v|x).$$

Finally, the third term in the decomposition of $\Phi_1(y,t,x)$ is

$$C=\int_0^t\frac{1}{(1-H(u|x))^2}\left(\int_u^t\frac{dH^1(v|y)}{1-H(v|x)}\right)dH^1(u|x).$$

Note that, for $y=x$, we have that $B=2C$. This completes the proof.

□

Lemma A.2.3. *Let us denote:*

$$\Phi_2(y,t,x)=E(\xi(T,\delta,t,x)\xi(T,\delta,\infty,x)|X=y),\tag{A.36}$$

with ξ in (2.10). The expression for the term $\Phi_2(x,t,x)$, for any $t\in[a,b]$, is the following:

$$\Phi_2(x,t,x)=\int_0^t\frac{dH^1(v|x)}{(1-H(v|x))^2}.$$

Proof of Lemma A.2.3. Recall $\Phi_2(y,t,x)$ in (A.36) with $\xi(T,\delta,t,x)$ in (2.10). Then:

$$\begin{aligned}
& \Phi_2(y,t,x) \\
= \; & E\left[\frac{I\left(T\leq t,\delta=1\right)}{\left(1-H(T|x)\right)^2}\middle|X=y\right] \\
- \; & E\left[\frac{I\left(\delta=1\right)}{1-H(T|x)}\int_0^\infty \frac{I\left(u\leq T\leq t\right)}{\left(1-H(u|x)\right)^2}dH^1(u|x)\middle|X=y\right] \\
- \; & E\left[\frac{I\left(\delta=1\right)}{1-H(T|x)}\int_0^t \frac{I\left(v\leq T\right)}{\left(1-H(v|x)\right)^2}dH^1(v|x)\middle|X=y\right] \\
+ \; & E\left[\int_0^t \frac{I\left(v\leq T\right)dH^1(v|x)}{\left(1-H(v|x)\right)^2}\int_0^\infty \frac{I\left(u\leq T\right)dH^1(u|x)}{\left(1-H(u|x)\right)^2}\middle|X=y\right] \\
= \; & A-B-C+D.
\end{aligned}$$

Straightforward calculations yield:

$$\begin{aligned}
A &= \int_0^t \frac{dH^1\left(v|y\right)}{\left(1-H(v|x)\right)^2}, \\
B &= \int_0^\infty \left(\int_u^t \frac{dH^1\left(v|y\right)}{1-H(v|x)}\right)\frac{dH^1(u|x)}{\left(1-H(u|x)\right)^2}, \\
C &= \int_0^t \left(\int_v^\infty \frac{dH^1\left(u|y\right)}{1-H(u|x)}\right)\frac{dH^1(v|x)}{\left(1-H(v|x)\right)^2} \text{ and} \\
D &= \int_0^t \frac{1}{\left(1-H(v|x)\right)^2}\left(\int_0^\infty \frac{1-H\left(\max\left(u,v\right)|y\right)}{\left(1-H(u|x)\right)^2}dH^1(u|x)\right)dH^1(v|x).
\end{aligned}$$

Integrating in the supports $\{(u,v)\in[0,\infty)\times[0,t]\,/v\leq u\}$ and $\{(u,v)\in[0,\infty)\times[0,t]\,/u<v\}=\{(u,v)\in[0,t]\times[0,t]\,/u<v\}$, the term D is

$$\begin{aligned}
D &= \int_0^t \left(\int_v^\infty \frac{1-H\left(u|y\right)}{\left(1-H(u|x)\right)^2}dH^1(u|x)\right)\frac{dH^1(v|x)}{\left(1-H(v|x)\right)^2} \\
&+ \int_0^\infty \left(\int_u^t \frac{1-H\left(v|y\right)}{\left(1-H(v|x)\right)^2}dH^1(v|x)\right)\frac{dH^1(u|x)}{\left(1-H(u|x)\right)^2}.
\end{aligned}$$

When $y=x$, then $D=C+B$, which concludes the proof. □

Theorem 3.2.2. *Under assumptions (A1)-(A13), and for two sequences of bandwidths satisfying (A14), then the mean squared error of the latency estimator satisfies*

$$\begin{aligned}
MSE(\hat{S}_{0,b_1,b_2}(t|x)) &= AMSE(\hat{S}_{0,b_1,b_2}(t|x)) \\
&+ o(b_2^4)+o(b_1^4)+o\left(b_1^2b_2^2\right)+O\left(\frac{b_2}{n}\right)+O\left(\frac{b_1}{nb_2}\right),
\end{aligned}$$

where

$$\begin{aligned} AMSE(\hat{S}_{0,b_1,b_2}(t|x)) &= \left(\frac{b_2^2}{2} d_K B_1(t,x) + \frac{b_1^2}{2} d_K B_2(t,x)\right)^2 + \frac{1}{nb_2} V_1(t,x)\, c_K \\ &+ \frac{1}{nb_1} V_2(t,x)\, c_K + 2\frac{1}{nb_1} V_3(t,x) \int K(u)\, K\left(\frac{b_2}{b_1} u\right) du, \end{aligned}$$

and

$$\begin{aligned} B_1(t,x) &= \frac{S(t|x)}{p(x)m(x)} \left(\Phi''(x,t,x)\, m(x) + 2\Phi'(x,t,x)\, m'(x)\right), && (3.4) \\ B_2(t,x) &= \frac{(1-p(x))(1-S(t|x))}{p^2(x)m(x)} \\ &\times \left(\Phi''(x,\infty,x)\, m(x) + 2\Phi'(x,\infty,x)\, m'(x)\right), && (3.5) \\ \Phi(y,t,x) &= \int_0^t \frac{dH^1(v|y)}{1-H(v|x)} - \int_0^t (1-H(v|y)) \frac{dH^1(v|x)}{(1-H(v|x))^2}, \end{aligned}$$

where Φ' and Φ'' are the partial derivatives of $\Phi(y,t,x)$ with respect to y. Furthermore,

$$\begin{aligned} V_1(t,x) &= \left(\frac{S(t|x)}{p(x)}\right)^2 \frac{1}{m(x)} \int_0^t \frac{dH^1(v|x)}{(1-H(v|x))^2}, && (3.6) \\ V_2(t,x) &= \left(\frac{(1-p(x))(1-S(t|x))}{p^2(x)}\right)^2 \frac{1}{m(x)} \int_0^\infty \frac{dH^1(v|x)}{(1-H(v|x))^2} \text{ and} && (3.7) \\ V_3(t,x) &= \frac{(1-p(x))S(t|x)(1-S(t|x))}{p^3(x)m(x)} \int_0^t \frac{dH^1(v|x)}{(1-H(v|x))^2}, && (3.8) \end{aligned}$$

with d_K in (1.9) and c_K in (1.10).

Proof of Theorem 3.2.2. We define

$$C_1 = -\frac{S(t|x)}{p(x)} \sum_{i=1}^n \tilde{B}_{b_2 i}(x) \xi(T_i, \delta_i, t, x)$$

and

$$C_2 = -\frac{(1-p(x))(1-S(t|x))}{p(x)^2} \sum_{i=1}^n \tilde{B}_{b_1 i}(x) \xi(T_i, \delta_i, \infty, x),$$

with $\tilde{B}_{b_1 i}(x)$ and $\tilde{B}_{b_2 i}(x)$ in (2.9). Then, from Theorem 3.2.1, the i.i.d. representation of the nonparametric latency estimator is

$$\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) = C_1 + C_2 + O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4} + \left(\frac{\ln n}{nb_2}\right)^{3/4}\right) \; a.s.$$

and the AMSE is

$$AMSE(\hat{S}_{0,b_1,b_2}(t|x)) = E\left[(C_1+C_2)^2\right] = E(C_1^2)+E(C_2^2)+2E(C_1C_2). \quad \text{(A.37)}$$

We start with the first term in (A.37). Note that

$$E(C_1^2) = Var(C_1) + E(C_1)^2, \quad \text{(A.38)}$$

where

$$\begin{aligned} Var(C_1) &= \left(\frac{S(t|x)}{p(x)}\right)^2 nVar\left(\tilde{B}_{b_2 1}(x)\xi(T_1,\delta_1,t,x)\right) \\ &= \frac{1}{nb_2^2}\left(\frac{S(t|x)}{p(x)}\right)^2 \frac{1}{m^2(x)} Var\left(K\left(\frac{x-X_1}{b_2}\right)\xi(T_1,\delta_1,t,x)\right) \quad \text{(A.39)} \end{aligned}$$

From (A.39), (A.26), (A.28) and (A.19), then

$$\begin{aligned} Var(C_1) &= \frac{1}{nb_2^2}\left(\frac{S(t|x)}{p(x)}\right)^2 \frac{1}{m^2(x)}\left(b_2\Phi_1(x,t,x)m(x)c_K + O(b_2^3)\right) \\ &= \frac{1}{nb_2}\left(\frac{S(t|x)}{p(x)}\right)^2 \frac{1}{m(x)}\Phi_1(x,t,x)\,c_K + O\left(\frac{b_2}{n}\right), \end{aligned}$$

with $\Phi_1(x,t,x)$ in (A.35).

Continuing with the second term in the right hand side of (A.38):

$$\begin{aligned} E(C_1) &= -\frac{S(t|x)}{p(x)} nE\left[\tilde{B}_{b_2 1}(x)\xi(T_1,\delta_1,t,x)\right] \\ &= -\frac{S(t|x)}{p(x)}\frac{1}{b_2}\frac{1}{m(x)}E\left[K\left(\frac{x-X_1}{b_2}\right)\xi(T_1,\delta_1,t,x)\right]. \end{aligned}$$

From Lemma A.1.6, then

$$E(C_1) = -\frac{1}{2}b_2^2\frac{S(t|x)}{p(x)}\frac{1}{m(x)}d_K\frac{\partial^2}{\partial y^2}\left[\Phi(y,t,x)m(y)\right]\Big|_{y=x} + o(b_2^2),$$

and using (A.18)

$$[E(C_1)]^2 = \left[-\frac{1}{2}b_2^2\frac{S(t|x)}{p(x)}\frac{1}{m(x)}\left(\Phi''(x,t,x)m(x)+\Phi'(x,t,x)m'(x)\right)d_K + o(b_2^2)\right]^2.$$

So the first term in (A.37) is

$$\begin{aligned} E(C_1^2) &= \frac{1}{nb_2}\left(\frac{S(t|x)}{p(x)}\right)^2 \frac{1}{m(x)}\Phi_1(x,t,x)\,c_K \quad \text{(A.40)} \\ &+ \left[-\frac{1}{2}b_2^2\frac{S(t|x)}{p(x)}\frac{1}{m(x)}\left(\Phi''(x,t,x)m(x)+\Phi'(x,t,x)m'(x)\right)d_K + o(b_2^2)\right]^2 \\ &+ O\left(\frac{b_2}{n}\right). \end{aligned}$$

Following the same ideas as those for C_1, we obtain that

$$\begin{aligned} E(C_2^2) &= \frac{1}{nb_1}\left(\frac{(1-S(t|x))(1-p(x))}{p(x)^2}\right)^2 \frac{1}{m(x)}\Phi_1(x,\infty,x)\,c_K \\ &+ \left[-\frac{1}{2}b_1^2\frac{(1-S(t|x))(1-p(x))}{p(x)^2}\frac{1}{m(x)}\left(\Phi''(x,\infty,x)m(x)\right.\right. \\ &+ \left.\left.\Phi'(x,\infty,x)m'(x)\right)d_K + o(b_1^2)\right]^2 + O\left(\frac{b_1}{n}\right). \end{aligned} \tag{A.41}$$

We continue studying the third term in (A.37):

$$\begin{aligned} E\left(C_1C_2\right) &= \frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3} \\ &\times E\left[\left(\sum_{i=1}^n \tilde{B}_{b_2i}(x)\xi(T_i,\delta_i,t,x)\right)\left(\sum_{j=1}^n \tilde{B}_{b_1j}(x)\xi(T_j,\delta_j,\infty,x)\right)\right] \\ &= \frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3} \\ &\times \sum_{i,j=1}^n E\left[\tilde{B}_{b_2i}(x)\tilde{B}_{b_1j}(x)\xi(T_i,\delta_i,t,x)\xi(T_j,\delta_j,\infty,x)\right] \\ &= \frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3}\left[n(n-1)\alpha\beta+n\gamma\right], \end{aligned}$$

where

$$\begin{aligned} \alpha &= E\left[\tilde{B}_{b_21}(x)\xi(T_1,\delta_1,t,x)\right], \\ \beta &= E\left[\tilde{B}_{b_12}(x)\xi(T_2,\delta_2,\infty,x)\right] \text{ and} \\ \gamma &= E\left[\tilde{B}_{b_11}(x)\tilde{B}_{b_21}(x)\xi(T_1,\delta_1,t,x)\xi(T_1,\delta_1,\infty,x)\right]. \end{aligned}$$

We start with

$$\alpha = \frac{1}{nb_2}\frac{1}{m(x)}E\left[K\left(\frac{x-X_1}{b_2}\right)\xi(T_1,\delta_1,t,x)\right].$$

From Lemma A.1.6 and using (A.18), then:

$$\alpha = \frac{b_2^2}{n}\frac{1}{m(x)}\frac{1}{2}\left(\Phi''(x,t,x)m(x)+\Phi'(x,t,x)m'(x)\right)d_K + o\left(\frac{b_2^2}{n}\right). \tag{A.42}$$

The term β can be analyzed in a similar way. From Lemma A.1.6 and from (A.18), then:

$$\begin{aligned} \beta &= \frac{1}{nb_1}\frac{1}{m(x)}E\left[K\left(\frac{x-X_2}{b_1}\right)\xi(T_2,\delta_2,\infty,x)\right] \\ &= \frac{b_1^2}{n}\frac{1}{m(x)}\frac{1}{2}\left(\Phi''(x,\infty,x)m(x)+\Phi'(x,\infty,x)m'(x)\right)d_K + o\left(\frac{b_1^2}{n}\right). \end{aligned} \tag{A.43}$$

We continue with the third term, γ. Recall $\Phi_2(y,t,x)$ from (A.36). Then,

$$\begin{aligned}
\gamma &= \frac{1}{nb_1}\frac{1}{m(x)}\frac{1}{nb_2}\frac{1}{m(x)} \\
&\times E\left[K\left(\frac{x-X_1}{b_1}\right)K\left(\frac{x-X_1}{b_2}\right)\xi(T_1,\delta_1,t,x)\xi(T_1,\delta_1,\infty,x)\right] \\
&= \frac{1}{n^2b_1b_2}\frac{1}{m(x)^2}E\left[K\left(\frac{x-X_1}{b_1}\right)K\left(\frac{x-X_1}{b_2}\right)\right. \\
&\times \left.E\left(\xi(T_1,\delta_1,t,x)\xi(T_1,\delta_1,\infty,x)|X_1\right)\right] \\
&= \frac{1}{n^2b_1b_2}\frac{1}{m(x)^2}\int K\left(\frac{x-y}{b_1}\right)K\left(\frac{x-y}{b_2}\right)\Phi_2(y,t,x)m(y)dy.
\end{aligned}$$

Applying Lemma A.2.1 it follows that

$$\begin{aligned}
\gamma &= \frac{1}{n^2b_1b_2}\frac{1}{m(x)^2}\left(b_2\Phi_2(x,t,x)m(x)\int K(u)K\left(\frac{b_2}{b_1}udu\right)+O(b_2^2)\right) \\
&= \frac{1}{n^2b_2}\frac{1}{m(x)}\Phi_2(x,t,x)\int K(u)K\left(\frac{b_1}{b_2}udu\right)+O\left(\frac{b_1}{n^2b_2}\right), \qquad \text{(A.44)}
\end{aligned}$$

where $\Phi_2(y,t,x)$ was defined in (A.36).

From (A.42), (A.43) and (A.44),

$$\begin{aligned}
E\left(C_1C_2\right) &= \frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3}\left[n(n-1)\alpha\beta+n\gamma\right] \qquad \text{(A.45)} \\
&= \frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3}\left[b_1^2b_2^2\frac{1}{m(x)^2}\frac{1}{4}d_K^2\Phi_1(x,t,x)\Phi_1(x,\infty,x)\right. \\
&+ \left.\frac{1}{nb_1}\frac{1}{m(x)}\Phi_2(x,t,x)\int K(u)K\left(\frac{b_2}{b_1}u\right)du+o\left(b_1^2b_2^2\right)+O\left(\frac{b_2}{nb_1}\right)\right].
\end{aligned}$$

Finally, plugging the expression of $E(C_1^2)$ in (A.40), $E(C_2^2)$ in (A.41) and $E(C_1C_2)$ in (A.45) into the Equation (A.37), the asymptotic expression for the MSE is derived. This concludes the proof. □

Theorem 3.2.3. *The bandwidths which minimize the asymptotic expression of* $MSE(\hat{S}_{0,b_1,b_2}(t|x))$ *are*

$$\widehat{b}_1(t,x)=\left(\frac{\frac{c_K}{L_n(t,x)}V_1\left(t,x\right)+c_KV_2\left(t,x\right)+2V_3\left(t,x\right)\int K(u)K(L_n(t,x)u)du}{d_K^2\left(L_n^2(t,x)B_1\left(t,x\right)+B_2\left(t,x\right)\right)^2}\right)^{1/5}n^{-1/5}$$

and

$$\widehat{b}_2(t,x)=L_n(t,x)\widehat{b}_1(t,x),$$

where

$$L_n(t,x)=\arg\min_{L>0}\psi\left(t,x,L\right)$$

and

$$\begin{aligned}\psi(t,x,L) &= \left(L^2 B_1(t,x) + B_2(t,x)\right) \\ &\times \left(\frac{c_K}{L} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x) \int K(u)K(Lu)du\right)^2,\end{aligned} \tag{3.9}$$

with $B_1(t,x)$ *in (3.4),* $B_2(t,x)$ *in (3.5),* $V_1(t,x)$ *in (3.6),* $V_2(t,x)$ *in (3.7) and* $V_3(t,x)$ *in (3.8).*

Proof of Theorem 3.2.3. We first prove that both bandwidths must be of the same order. If $b_2/b_1 \to 0$ (that is, $b_2 = L_n b_1$ with $L_n = L_n(t,x) \to 0$), then the expression of b_1 which minimizes the dominant part of $MSE(\hat{S}_{0,b_1,b_2}(t|x)) \equiv MSE_{t,x}(b_1,b_2)$ is:

$$\widehat{b}_1(t,x) = \left(\frac{\frac{c_K}{L_n(t,x)} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\int K(u)K(L_n(t,x)u)du}{d_K^2\left(L_n^2(t,x)B_1(t,x) + B_2(t,x)\right)^2}\right)^{1/5} n^{-1/5}.$$

We evaluate $MSE_{t,x}(b_1,b_2)$ in $\widehat{b}_1$ and $\widehat{b}_2 = L_n\widehat{b}_1$:

$$\begin{aligned}MSE_{t,x}(\widehat{b}_1, L_n\widehat{b}_1) &= n^{-4/5}\frac{5}{4} d_K^{2/5}\left(L_n^2(t,x)B_1(t,x) + B_2(t,x)\right)^{2/5} \\ &\times \left(\frac{c_K}{L_n(t,x)} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\int K(u)K(L_n(t,x)u)du\right)^{4/5} \\ &+ o(n^{-4/5}).\end{aligned}$$

Then, the optimal bandwidths are

$$\widehat{b}_1(t,x) = \left(\frac{\frac{c_K}{L_n(t,x)} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\int K(u)K(L_n(t,x)u)du}{d_K^2\left(L_n^2(t,x)B_1(t,x) + B_2(t,x)\right)^2}\right)^{1/5} n^{-1/5}$$

and

$$\widehat{b}_2(t,x) = L_n(t,x)\widehat{b}_1(t,x),$$

where

$$\begin{aligned}L_n(t,x) &= \arg\min_{L>0}\left(L^2 B_1(t,x) + B_2(t,x)\right)^{2/5} \\ &\times \left(\frac{c_K}{L} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\int K(u)K(L_n(t,x)u)du\right)^{4/5},\end{aligned}$$

which contradicts the initial assumption $b_2/b_1 \to 0$.

The same argument applies in the opposite case, if $b_2/b_1 \to \infty$. Therefore, both optimal bandwidths are necessarily of the same order.

Since both b_1 and b_2 have the same order, then b_2 is asymptotically equal to Lb_1 for some $L > 0$. Let us rewrite $b_2 = L_n b_1$ for $L_n \to L$. Then the expression of b_1 which minimizes the dominant part of $MSE_{t,x}(b_1, L_n b_1)$ is:

$$\widehat{b}_1(t,x) = \left(\frac{\frac{c_K}{L_n} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\,\Upsilon(L_n)}{d_K^2 \left(L_n^2 B_1(t,x) + B_2(t,x)\right)^2} \right)^{1/5} n^{-1/5},$$

with

$$\Upsilon(L_n) = \int K(u)\, K(L_n u)\, du.$$

Therefore, $MSE_{t,x}(b_1, L_n b_1)$ evaluated in $\widehat{b}_1(t,x)$ is:

$$\begin{aligned} MSE_{t,x}(\widehat{b}_1, L_n \widehat{b}_1) &= n^{-4/5} \frac{5}{4} d_K^{2/5} \left(L_n^2 B_1(t,x) + B_2(t,x)\right)^{2/5} \\ &\times \left(\frac{c_K}{L_n} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\,\Upsilon(L_n) \right)^{4/5} + o(n^{-4/5}). \end{aligned}$$

Then, the optimal bandwidths are

$$\widehat{b}_1(t,x) = \left(\frac{\frac{c_K}{L_n(t,x)} V_1(t,x) + c_K V_2(t,x) + 2V_3(t,x)\,\Upsilon(L_n(t,x))}{d_K^2 \left(L_n^2(t,x) B_1(t,x) + B_2(t,x)\right)^2} \right)^{1/5} n^{-1/5}$$

and

$$\widehat{b}_2(t,x) = L_n(t,x)\widehat{b}_1(t,x),$$

where $L_n(t,x) = \arg\min_{L>0} \psi(t,x,L)$ with $\psi(t,x,L)$ defined in (3.9). This concludes the proof. □

Theorem 3.2.4. *Under assumptions (A1)-(A13), if $b_i \to 0$ for $i = 1,2$ and $((\ln n)^3 / n b_i) \times (b_j/(b_1+b_2))^2 \to 0$ for $i, j = 1,2$ with $i \neq j$, it follows that*

a) If $n b_i^5 \frac{b_j}{b_1+b_2} \to 0$ for $i, j = 1,2$ and $i \neq j$, then

$$\sqrt{n b_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) \right) \xrightarrow{d} N\left(0, \sigma_a^2(t,x)\right),$$

where

$$\sigma_a^2(t,x) = \begin{cases} V_2(t,x)\, c_K, & \text{if } b_1/b_2 \to 0 \\ V_1(t,x)\, c_K, & \text{if } b_2/b_1 \to 0 \\ \frac{C_1}{C_1+C_2} \left(V_1(t,x)\, c_K + 2V_3(t,x) \int K(u) K\left(\frac{C_1}{C_2} u\right) du \right) + \frac{C_2}{C_1+C_2} V_2(t,x)\, c_K, & \\ \qquad \text{if } b_1 = C_1 n^{-\alpha} + o(n^{-\alpha}),\ b_2 = C_2 n^{-\alpha} + o(n^{-\alpha}),\ \text{with } \alpha > \frac{1}{5} \end{cases}$$

with $V_1(t,x)$ in (3.6), $V_2(t,x)$ in (3.7) and $V_3(t,x)$ in (3.8).

b) If $nb_1^5 \to 0$ *and* $nb_2^5 \to C_2^5 > 0$, *then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) \right) \xrightarrow{d} N\left(0, \sigma_b^2(t,x)\right),$$

with $\sigma_b^2(t,x) = V_2(t,x)\, c_K$.

c) If $nb_1^5 \to C_1^5 > 0$ *and* $nb_2^5 \to 0$, *then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) \right) \xrightarrow{d} N\left(0, \sigma_c^2(t,x)\right),$$

with $\sigma_c^2(t,x) = V_1(t,x)\, c_K$.

d) If $nb_1^5 \to C_1^5 > 0$ *and* $nb_2^5 \to C_2^5 > 0$, *then*

$$\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \left(\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) \right) \xrightarrow{d} N\left(b(t,x), \sigma_d^2(t,x)\right),$$

where

$$b(t,x) = \frac{1}{2} d_K \left(\frac{C_1 C_2}{C_1+C_2} \right)^{1/2} \left(C_2^2 B_1(t,x) + C_1^2 B_2(t,x) \right)$$

and

$$\begin{aligned} \sigma_d^2(t,x) &= \frac{C_1}{C_1+C_2} \left(V_1(t,x)\, c_K + 2V_3(t,x) \int K(u) K\left(\frac{C_1}{C_2} u \right) du \right) \\ &+ \frac{C_2}{C_1+C_2} V_2(t,x)\, c_K. \end{aligned}$$

Proof of Theorem 3.2.4. The asymptotic normality is proven by following Theorem 2.3 in Iglesias-Pérez & González-Manteiga (1999). Under the assumptions $(A1)$-$(A13)$, we can apply Theorem 3.2.1, and then

$$\hat{S}_{0,b_1,b_2}(t|x) - S_0(t|x) = \sum_{i=1}^{n} \eta_{b_1,b_2}(T_i, \delta_i, X_i, t, x) + R_n(t,x),$$

with

$$\sup_{x \in I, t \geq 0} R_n(t,x) = O\left(\left(\frac{\ln n}{nb_1} \right)^{3/4} + \left(\frac{\ln n}{nb_2} \right)^{3/4} \right) \ a.s.,$$

where η_{b_1,b_2} is defined in (3.3).

The asymptotic distribution of $\sqrt{\frac{nb_1b_2}{b_1+b_2}}\left(\hat{S}_{0,b_1,b_2}(t|x)-S_0(t|x)\right)$ is the same as that of

$$\begin{aligned}
&\sqrt{\frac{nb_1b_2}{b_1+b_2}}\sum_{i=1}^{n}\eta_{b_1,b_2}(T_i,\delta_i,X_i,t,x)\\
=\;&-\sqrt{nb_1\frac{b_2}{b_1+b_2}}\frac{1}{nb_2}\frac{S(t|x)}{p(x)m(x)}\sum_{i=1}^{n}K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\\
-\;&\sqrt{nb_1\frac{b_2}{b_1+b_2}}\frac{1}{nb_1}\frac{(1-p(x))(1-S(t|x))}{p(x)^2m\,(x)}\sum_{i=1}^{n}K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x),
\end{aligned}$$

since the conditions $\left[(\ln n)^3/nb_i\right]\cdot\left[(b_j/(b_1+b_2))^2\right]\to 0, i,j=1,2$ and $i\neq j$ implies that

$$\sqrt{\frac{nb_1b_2}{b_1+b_2}}O\left(\left(\frac{\ln n}{nb_1}\right)^{3/4}+\left(\frac{\ln n}{nb_2}\right)^{3/4}\right)\to 0. \tag{A.46}$$

If we focus on the first part in (A.46) then,

$$\sqrt{\frac{nb_1b_2}{b_1+b_2}}\left(\frac{\ln n}{nb_1}\right)^{3/4}=\left(\frac{(\ln n)^3}{nb_1}\left(\frac{b_2}{b_1+b_2}\right)^2\right)^{1/4}\to 0 \text{ if } \frac{(\ln n)^3}{nb_1}\left(\frac{b_2}{b_1+b_2}\right)^2\to 0.$$

The same argument applies for the second condition in (A.46).

We have to study the limiting distribution of

$$\begin{aligned}
&\sqrt{\frac{nb_1b_2}{b_1+b_2}}\frac{1}{nb_2}\frac{S(t|x)}{p(x)m(x)}\sum_{i=1}^{n}K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\\
+\;&\sqrt{\frac{nb_1b_2}{b_1+b_2}}\frac{1}{nb_1}\frac{(1-p(x))(1-S(t|x))}{p(x)^2m\,(x)}\sum_{i=1}^{n}K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\\
=\;&I+II+III+IV,
\end{aligned}$$

where

$$\begin{aligned}
I\;=\;&\sqrt{\frac{nb_1b_2}{b_1+b_2}}\frac{1}{nb_2}\frac{S(t|x)}{p(x)m(x)}\\
\times\;&\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)-E\left(K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\right)\right],\\
II\;=\;&\sqrt{\frac{nb_1b_2}{b_1+b_2}}\frac{1}{nb_1}\frac{(1-p(x))(1-S(t|x))}{p(x)^2m\,(x)}\\
\times\;&\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)-E\left(K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\right)\right],
\end{aligned}$$

$$III = \sqrt{\frac{nb_1b_2}{b_1+b_2}}\frac{1}{nb_2}\frac{S(t|x)}{p(x)m(x)}\sum_{i=1}^{n}E\left(K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\right) \text{ and}$$

$$IV = \sqrt{\frac{nb_1b_2}{b_1+b_2}}\frac{1}{nb_1}\frac{(1-p(x))(1-S(t|x))}{p(x)^2m\,(x)}\sum_{i=1}^{n}E\left(K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\right).$$

The deterministic part $b(t,x)$ comes from $III+IV$. Recall the definition of $\Phi(y,t,x)$ in (A.17). Using (A.18), and from Lemma A.1.6, then

$$E\left(K\left(\frac{x-X}{b_2}\right)\xi(T,\delta,t,x)\right) = \frac{1}{2}b_2^3d_K\left(\Phi''(x,t,x)m(x)+2\Phi'(x,t,x)m'(x)\right)+o(b_2^3).$$

Therefore,

$$III = \sqrt{nb_2^5\frac{b_1}{b_1+b_2}}\frac{S(t|x)}{p(x)m(x)}\frac{1}{2}d_K\left(\Phi''(x,t,x)m(x)+2\Phi'(x,t,x)m'(x)\right)(1+o\,(1))$$

and

$$\begin{aligned} IV &= \sqrt{nb_1^5\frac{b_2}{b_1+b_2}}\frac{(1-p(x))(1-S(t|x))}{p(x)^2m\,(x)}\frac{1}{2}d_K \\ &\times \left(\Phi''(x,\infty,x)m(x)+2\Phi'(x,\infty,x)m'(x)\right)(1+o(1))\,. \end{aligned}$$

Focusing on III, then

$$\sqrt{nb_2^5\frac{b_1}{b_1+b_2}} = \left(nb_2^5\frac{b_1}{b_1+b_2}\right)^{1/2} \to 0 \text{ if } nb_2^5\frac{b_1}{b_1+b_2}.$$

A similar argument can be applied for IV. Therefore, under the hypothesis

a) $nb_i^5\frac{b_j}{b_1+b_2}\to 0, i,j=1,2, i\neq j$, then $III+IV=o\,(1)$ and $b\,(t,x)=0$.

b) If $nb_1^5\to 0$ and $nb_2^5\to C_2^5>0$, then $b_2/b_1\to\infty$ and thus

$$\begin{aligned} III &= \sqrt{nb_2^5\frac{1}{1+\frac{b_2}{b_1}}}\frac{S(t|x)}{p(x)m(x)}\frac{1}{2}d_K\left(\Phi''(x,t,x)m(x)+2\Phi'(x,t,x)m'(x)\right) \\ &\times (1+o\,(1))=o\,(1) \end{aligned}$$

and

$$\begin{aligned} IV &= \sqrt{nb_1^5\frac{1}{\frac{b_1}{b_2}+1}}\frac{(1-p(x))(1-S(t|x))}{p(x)^2m\,(x)}\frac{1}{2}d_K\left(\Phi''(x,\infty,x)m(x)\right. \\ &\left.+2\Phi'(x,\infty,x)m'(x)\right)(1+o(1))=o\,(1)\,. \end{aligned}$$

Thus, $III+IV=o\,(1)$ and $b\,(t,x)=0$.

c) If $nb_1^5 = C_1^5$ and $nb_2^5 \to 0$ then $b_2/b_1 \to 0$ and

$$\begin{aligned} III &= \sqrt{nb_2^5 \frac{1}{1+\frac{b_2}{b_1}}} \frac{S(t|x)}{p(x)m(x)} \frac{1}{2} d_K \left(\Phi''(x,t,x)m(x) + 2\Phi'(x,t,x)m'(x)\right) \\ &\times (1+o(1)) = o(1) \end{aligned}$$

and

$$\begin{aligned} IV &= \sqrt{nb_1^5 \frac{1}{\frac{b_1}{b_2}+1}} \frac{(1-p(x))(1-S(t|x))}{p(x)^2 m(x)} \frac{1}{2} d_K \left(\Phi''(x,\infty,x)m(x)\right. \\ &\left. +2\Phi'(x,\infty,x)m'(x)\right)(1+o(1)) = o(1). \end{aligned}$$

Consequently, $III + IV = o(1)$ and $b(t,x) = 0$.

d) If $nb_1^5 \to C_1^5$ and $nb_2^5 \to C_2^5$, then $b_2/b_1 \to C_2/C_1$ and then

$$\begin{aligned} & b(t,x) \\ = \; & C_2^{5/2} \left(\frac{C_1}{C_1+C_2}\right)^{1/2} \frac{S(t|x)}{p(x)m(x)} \frac{1}{2} d_K \left(\Phi''(x,t,x)m(x) + 2\Phi'(x,t,x)m'(x)\right) \\ + \; & C_1^{5/2} \left(\frac{C_2}{C_1+C_2}\right)^{1/2} \frac{(1-p(x))(1-S(t|x))}{p(x)^2 m(x)} \frac{1}{2} d_K \left(\Phi''(x,\infty,x)m(x)\right. \\ + \; & \left. 2\Phi'(x,\infty,x)m'(x)\right) \\ = \; & C_2^{5/2} \left(\frac{C_1}{C_1+C_2}\right)^{1/2} \frac{1}{2} d_K B_1(t,x) + C_1^{5/2} \left(\frac{C_2}{C_1+C_2}\right)^{1/2} \frac{1}{2} d_K B_2(t,x) \\ = \; & \frac{1}{2} d_K \left(\frac{C_1 C_2}{C_1+C_2}\right)^{1/2} \left[\frac{1}{C_2^{1/2}} C_2^{5/2} B_1(t,x) + \frac{1}{C_1^{1/2}} C_1^{5/2} B_2(t,x)\right] \\ = \; & \frac{1}{2} d_K \left(\frac{C_1 C_2}{C_1+C_2}\right)^{1/2} \left[C_2^2 B_1(t,x) + C_1^2 B_2(t,x)\right]. \end{aligned}$$

Now, we focus on the asymptotic distribution of $I + II$. It is immediate that:

$$I + II = \sum_{i=1}^{n} \left(\gamma_{i,n}(x,t) + \Gamma_{i,n}(x,t)\right),$$

where

$$\begin{aligned} \gamma_{i,n}(x,t) &= \sqrt{\frac{1}{nb_2} \frac{b_1}{b_1+b_2}} \frac{S(t|x)}{p(x)m(x)} \\ &\times \left[K\left(\frac{x-X_i}{b_2}\right) \xi(T_i,\delta_i,t,x) - E\left(K\left(\frac{x-X_i}{b_2}\right) \xi(T_i,\delta_i,t,x)\right)\right], \\ \Gamma_{i,n}(x,t) &= \sqrt{\frac{1}{nb_1} \frac{b_2}{b_1+b_2}} \frac{(1-p(x))(1-S(t|x))}{p(x)^2 m(x)} \\ &\times \left[K\left(\frac{x-X_i}{b_1}\right) \xi(T_i,\delta_i,\infty,x) - E\left(K\left(\frac{x-X_i}{b_1}\right) \xi(T_i,\delta_i,\infty,x)\right)\right], \end{aligned}$$

are n independent variables with mean 0. We only have to prove the asymptotic normality of $I + II$. Therefore, if $\sigma_{i,n}^2(x,t) = Var(\gamma_{i,n}(x,t) + \Gamma_{i,n}(x,t)) < \infty$, $\sigma_n^2(x,t) = \sum_{i=1}^n \sigma_{i,n}^2(x,t)$ is positive and if the Lindeberg condition is satisfied, we can apply Lindeberg's theorem for triangular arrays (Theorem 7.2 in Billingsley (1968), p.42) to obtain

$$\frac{\sum_{i=1}^n (\gamma_{i,n}(x,t) + \Gamma_{i,n}(x,t))}{\sigma_n(x,t)} \to N(0,1)$$

and, consequently,

$$\frac{\sqrt{nb_1 \frac{b_2}{b_1+b_2}} \sum_{i=1}^n (\eta_{b_1,b_2}(T_i,\delta_i,X_i,t,x) - E(\eta_{b_1,b_2}(T_i,\delta_i,X_i,t,x)))}{\sigma_n(x,t)} \to N(0,1).$$

We will start proving that the variance is finite:

$$\begin{aligned}
\sigma_{i,n}^2(x,t) &= Var(\gamma_{i,n}(x,t)) + Var(\Gamma_{i,n}(x,t)) + 2Cov(\gamma_{i,n}(x,t), \Gamma_{i,n}(x,t)) \\
&= \frac{1}{nb_2}\frac{b_1}{b_1+b_2}\left(\frac{S(t|x)}{p(x)m(x)}\right)^2 Var\left(K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\right) \\
&+ \frac{1}{nb_1}\frac{b_2}{b_1+b_2}\left(\frac{(1-p(x))(1-S(t|x))}{p(x)^2 m(x)}\right)^2 Var\left(K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\right) \\
&+ \sqrt{\frac{1}{n^2}\frac{1}{(b_1+b_2)^2}}\frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3 m(x)^2} \\
&\times \left[E\left(K\left(\frac{x-X_i}{b_2}\right)K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\xi(T_i,\delta_i,t,x)\right)\right. \\
&\left. - E\left[K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\right]E\left[K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\right]\right].
\end{aligned} \tag{A.47}$$

We first focus on $Var\left(K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\right)$. The first term in (A.47) is

$$Var(\gamma_{i,n}(x,t)) = \frac{1}{n}\frac{b_1}{b_1+b_2}\left(\frac{S(t|x)}{p(x)}\right)^2 \frac{\Phi_1(x,t,x)}{m(x)}c_K + O\left(\frac{1}{n}\frac{b_1 b_2^2}{b_1+b_2}\right). \tag{A.48}$$

In a similar way, the second term in (A.47) is

$$Var(\Gamma_{i,n}(x,t)) = \frac{1}{n}\frac{b_2}{b_1+b_2}\left(\frac{(1-p(x))(1-S(t|x))}{p(x)^2}\right)^2 \frac{\Phi_1(x,\infty,x)}{m(x)}c_K + O\left(\frac{1}{n}\frac{b_1^2 b_2}{b_1+b_2}\right). \tag{A.49}$$

Finally, for the third term in (A.47), recall $\Phi_2(y,t,x)$ in (A.36). Applying Lemma

A.2.1, then

$$\begin{aligned} & E\left(K\left(\frac{x-X_i}{b_2}\right)K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\xi(T_i,\delta_i,t,x)\right) \\ = & \int K\left(\frac{x-y}{b_2}\right)K\left(\frac{x-y}{b_1}\right)\Phi_2(y,t,x)m\left(y\right)dy \\ = & \ b_1\Phi_2(x,t,x)m(x)\int K(u)K\left(\frac{b_1}{b_2}udu\right)+O(b_1^2). \end{aligned}$$

On the other hand, using Lemmas A.1.6 and A.1.7, we get:

$$E\left[K\left(\frac{x-X_i}{b_2}\right)\xi(T_i,\delta_i,t,x)\right]E\left[K\left(\frac{x-X_i}{b_1}\right)\xi(T_i,\delta_i,\infty,x)\right]=O(b_1^3b_2^3).$$

Therefore, the third term in (A.47) is

$$\begin{aligned} & Cov\left(\gamma_{i,n}(x,t),\Gamma_{i,n}(x,t)\right) \qquad (A.50) \\ = & \frac{1}{n}\frac{b_1}{b_1+b_2}\frac{(1-p(x))S(t|x)(1-S(t|x))}{p(x)^3m\left(x\right)}\Phi_2(y,t,x)\int K(u)K\left(\frac{b_1}{b_2}udu\right) \\ + & O\left(\frac{1}{n}\frac{b_1^2}{b_1+b_2}\right). \end{aligned}$$

The results (A.48), (A.49) and (A.50) lead to

$$\begin{aligned} \sigma_{i,n}^2\left(x,t\right) = & \frac{1}{n}\left(\frac{b_1}{b_1+b_2}V_1\left(t,x\right)c_K+\frac{b_2}{b_1+b_2}V_2\left(t,x\right)c_K+2\frac{b_1}{b_1+b_2}V_3\left(t,x\right)\right. \\ \times & \left.\int K(u)K\left(\frac{b_1}{b_2}u\right)du+O\left(\frac{b_1b_2^2}{b_1+b_2}\right)+O\left(\frac{b_1^2}{b_1+b_2}\right)\right), \end{aligned}$$

where $V_1\left(t,x\right)$, $V_2\left(t,x\right)$ and $V_3\left(t,x\right)$, defined in (3.6)-(3.8), are finite, and as a consequence, $\sigma_{i,n}^2\left(x,t\right)<\infty$. The finiteness of the variance $\sigma_n^2\left(x,t\right)$ is also proven, since

$$\begin{aligned} \sigma_n^2\left(x,t\right) = & \sum_{i=1}^{n}\sigma_{i,n}^2\left(x,t\right) \\ = & \frac{b_1}{b_1+b_2}V_1\left(t,x\right)c_K+\frac{b_2}{b_1+b_2}V_2\left(t,x\right)c_K+2\frac{b_1}{b_1+b_2}V_3\left(t,x\right) \\ \times & \int K(u)K\left(\frac{b_1}{b_2}udu\right)+O\left(\frac{b_1b_2^2}{b_1+b_2}\right)+O\left(\frac{b_1^2}{b_1+b_2}\right) \\ < & V_1\left(t,x\right)c_K+V_2\left(t,x\right)c_K+2V_3\left(t,x\right)\int K(u)K\left(\frac{b_1}{b_2}u\right)du+o\left(1\right)<+\infty, \end{aligned}$$

for $b_1>0$ and $b_2>0$, since $b_i/\left(b_1+b_2\right)<1, i=1,2$.

a) Under the hypothesis $nb_i^5 \frac{b_j}{b_1+b_2} \to 0, i, j = 1, 2, i \neq j$, if $b_1/b_2 \to 0$ then

$$\begin{aligned}\sigma_n^2(x,t) &= \frac{b_1/b_2}{b_1/b_2+1} V_1(t,x)\, c_K + \frac{1}{b_1/b_2+1} V_2(t,x)\, c_K \\ &+ 2\frac{b_1/b_2}{b_1/b_2+1} V_3(t,x) \int K(u) K\left(\frac{b_1}{b_2}u\right) du \to V_2(t,x)\, c_K.\end{aligned}$$

If $b_2/b_1 \to 0$ then

$$\begin{aligned}\sigma_n^2(x,t) &= \frac{1}{1+b_2/b_1} V_1(t,x)\, c_K + \frac{b_2/b_1}{1+b_2/b_1} V_2(t,x)\, c_K \\ &+ 2\frac{1}{1+b_2/b_1} V_3(t,x) \int K(u) K\left(\frac{b_1}{b_2}u\right) du \to V_1(t,x)\, c_K,\end{aligned}$$

since K is a compact support kernel (using (A11)) and then $K(v) \to 0$ when $u \to \pm\infty$.

If $\frac{b_1}{b_2} \to \frac{C_1}{C_2} > 0$, then

$$\begin{aligned}\sigma_n^2(x,t) &\to \frac{C_1}{C_1+C_2}\left(V_1(t,x)c_K + 2V_3(t,x)\int K(u)K\left(\frac{C_1}{C_2}u\right)du\right) \\ &+ \frac{C_2}{C_1+C_2} V_2(t,x) c_K. \qquad \text{(A.51)}\end{aligned}$$

b) If $nb_1^5 \to 0$ and $nb_2^5 \to C_2^5 > 0$ then $\sigma_n^2(x,t) \to V_2(t,x)\, c_K$.
c) If $nb_1^5 \to C_1^5$ and $nb_2^5 \to 0$, then $\sigma_n^2(x,t) \to V_1(t,x)\, c_K$.
d) Finally, if $nb_1^5 \to C_1^5$ and $nb_2^5 \to C_2^5$, then we consider (A.51).
We continue studying the Lindeberg's condition:

$$\frac{1}{\sigma_n^2(x,t)} \sum_{i=1}^{n} \int_{\{|\gamma_{i,n}(x,t)+\Gamma_{i,n}(x,t)|>\epsilon\sigma_n(x,t)\}} (\gamma_{i,n}(x,t)+\Gamma_{i,n}(x,t))^2 dP \to 0, \forall \epsilon > 0. \qquad \text{(A.52)}$$

Let us define the indicator function:

$$\begin{aligned}I_{i,n}(x,t) &= I\left(|\gamma_{i,n}(x,t)+\Gamma_{i,n}(x,t)| > \epsilon\sigma_n(x,t)\right) \\ &= I\left(\left(\gamma_{i,n}(x,t)+\Gamma_{i,n}(x,t)\right)^2 > \epsilon^2\sigma_n^2(x,t)\right).\end{aligned}$$

Then (A.52) can be expressed as

$$\frac{1}{\sigma_n^2(x,t)} E\left[\sum_{i=1}^{n} (\gamma_{i,n}(x,t)+\Gamma_{i,n}(x,t))^2 I_{i,n}(x,t)\right] = \frac{1}{\sigma_n^2(x,t)} E\left(\eta_n(x,t)\right),$$

with

$$\eta_n(x,t) = \sum_{i=1}^{n} (\gamma_{i,n}(x,t)+\Gamma_{i,n}(x,t))^2 I_{i,n}(x,t).$$

Since $\frac{1}{nb_2}\frac{b_1}{b_1+b_2} \to 0$, $\frac{1}{nb_1}\frac{b_2}{b_1+b_2} \to 0$ and the functions K and ξ are bounded, then:

$$\begin{aligned} & \exists n_0 \in \mathbb{N}/n \geq n_0 \Rightarrow I_{i,n}(w) = 0, \forall w \in \Omega \text{ and } \forall i \in \{1, 2, \ldots, n\} \\ \Leftrightarrow \quad & \exists n_0 \in \mathbb{N}/n \geq n_0 \Rightarrow \eta_n(w) = 0, \forall w \in \Omega. \end{aligned}$$

Since $\eta_n(x,t)$ is bounded, then the previous condition implies:

$$\exists n_0 \in \mathbb{N}/n \geq n_0 \Rightarrow E(\eta_n(x,t)) = 0$$

and then

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} E(\eta_n(x,t)) = 0.$$

Therefore, Lindeberg's condition is proven and all these previous arguments lead to the proof of Theorem 3.2.4. □

Theorem 3.3.2. *Suppose that conditions (A1)-(A13) hold. If* $\frac{\ln n}{nb} \to 0$ *and* $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$*, then we have an i.i.d. representation for the nonparametric latency estimator for any* $t \in [a,b]$*:*

$$\hat{S}_{0,b}(t|x) - S_0(t|x) = \sum_{i=1}^{n} \eta_b(T_i, \delta_i, X_i, t, x) + O\left(\left(\frac{\ln n}{nb}\right)^{3/4}\right) a.s.,$$

with

$$\begin{aligned} \eta_b(T_i, \delta_i, X_i, t, x) &= -\frac{S(t|x)}{p(x)} \tilde{B}_{bi}(x) \xi(T_i, \delta_i, t, x) \\ &- \frac{(1-p(x))(1-S(t|x))}{p^2(x)} \tilde{B}_{bi}(x) \xi(T_i, \delta_i, \infty, x), \end{aligned}$$

where $\xi(T_i, \delta_i, t, x)$ *has been defined (2.10) and* $\tilde{B}_{bi}(x)$ *in (2.9).*

Proof of Theorem 3.3.2. Taking $b_1 = b_2 = b$ with $\frac{\ln n}{nb} \to 0$ and $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$ it is straightforward to prove that the conditions required for the bandwidths in Theorem 3.2.1 are fulfilled. So Theorem 3.2.1 can be applied and Theorem 3.3.2 is proven. □

Theorem 3.3.3. *Suppose that conditions (A1)-(A13) hold. If* $\frac{\ln n}{nb} \to 0$ *and* $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$*, then the mean squared error of the latency estimator is*

$$MSE(\hat{S}_{0,b}(t|x)) = \frac{b^4}{4} d_K^2 B^2(t,x) + \frac{c_K}{nb} V(t,x) + o(b^4) + O\left(\frac{1}{n}\right),$$

where d_K *and* c_K *have been defined in (1.9) and (1.10), respectively, and*

$$B(t,x) = B_1(t,x) + B_2(t,x), \tag{3.10}$$

$$V(t,x) = V_1(t,x) + V_2(t,x) + 2V_3(t,x), \tag{3.11}$$

with $t \in [a,b]$*,* $B_1(t,x)$*,* $B_2(t,x)$*,* $V_1(t,x)$*,* $V_2(t,x)$ *and* $V_3(t,x)$ *in (3.4)-(3.8).*

Proof of Theorem 3.3.3. Choosing again $b_1 = b_2 = b$, conditions $\frac{\ln n}{nb} \to 0$ and $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$ imply the conditions for the bandwidths in Theorem 3.2.2. As a consequence, Theorem 3.3.3 is proven. □

Theorem 3.3.4. *Suppose that conditions* $(A1)$-$(A13)$ *hold. If* $b \to 0$ *and* $\frac{(\ln n)^3}{nb} \to 0$, *it follows that, for any* $t \in [a, b]$*:*

a) If $nb^5 \to 0$, *then*

$$\sqrt{nb}\left(\hat{S}_{0,b}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(0, V\left(t, x\right) c_K\right).$$

b) If $nb^5 \to C^5 > 0$, *then*

$$\sqrt{nb}\left(\hat{S}_{0,b}(t|x) - S_0(t|x)\right) \xrightarrow{d} N\left(B\left(t, x\right) C^{5/2} d_K, V\left(t, x\right) c_K\right).$$

Proof of Theorem 3.3.4. Taking $b_1 = b_2 = b$, together with $\frac{\ln n}{nb} \to 0$ and $b = O\left(\left(\frac{\ln n}{n}\right)^{1/5}\right)$ imply the conditions for the bandwidths in Theorem 3.2.4. Thus, Theorem 3.3.4 is proven. □

Appendix B

Resumo en galego

Este traballo pretende resumir os estudos desenvolvidos ao longo do proceso de doutoramento. Principalmente, está centrado na estimación e nos contrastes de significación de covariables para modelos de curación de tipo mixtura non paramétricos. Especificamente, a metodoloxía é aplicada a dúas bases de datos médicas: unha relacionada con doentes de cancro colorrectal do Complexo Hospitalario Universitario da Coruña (CHUAC), e outra relacionada con doentes de sarcomas do Complexo Hospitalario Universitario de Santiago (CHUS).

Capítulo 1: Estado da arte

O primeiro capítulo da tese está dedicado a introducir ao lector ao contexto no que foi desenvolvido o traballo: os modelos de curación. A Sección 1.1 comeza cunha presentación da análise de supervivencia e dos diferentes tipos de censura. Como o traballo é levado a cabo nun contexto non paramétrico, na Sección 1.2 inclúese unha pequena introdución á estimación non paramétrica de curvas, incorporando algunhas definicións e unha pequena revisión dos conceptos básicos, consistente nunha reseña da estimación da función de distribución (incluíndo o estudo dos estimadores de Kaplan-Meier e de Beran, así como algunhas propiedades asintóticas de ambos), da función de densidade e da función de regresión. Na Sección 1.3 preséntase unha descrición detallada do método bootstrap, considerando tamén o caso con datos censurados e unha explicación do método de selección da ventá.

Os modelos de curación preséntanse na Sección 1.4. Nas dúas últimas décadas houbo un importante progreso nos tratamentos de cancro, que deu lugar

a unha supervivencia mais longa e á mellora da calidade de vida dos doentes desta enfermidade. Desta forma, os datos pertencentes a estudos de cancro tipicamente presentan unha forte censura pola dereita (debido á supervivencia tan longa) ao final do periodo de seguemento, polo que un modelo de supervivencia estándar non é adecuado. Esa proporción de individuos curados ou non susceptibles pódese incorporar explicitamente nos modelos de supervivencia, e como consecuencia, xorden os modelos de curación. Estes modelos son unha ferramenta moi útil para analizar e describir datos de supervivencia de cancro, xa que expresan e predín a prognose dun doente considerando, como novidade, a posibilidade real de que o suxeito nunca experimente o evento de interese. Dentro dos modelos de curación distínguense dous grandes tipos: os modelos de mixtura e os modelos de non mixtura. Os modelos de curación de tipo mixtura, sobre os que se centra esta tese, consideran a función de supervivencia como unha mixtura de dous tipos de doentes: os susceptibles e os non susceptibles ou curados. Mais especificamente, estes modelos permiten estimar a probabilidade de cura, tamén coñecida como *incidencia*, e a función de supervivencia dos individuos non curados, denominada *latencia*. Unha vantaxe dos modelos de curación de tipo mixtura é que permiten ás covariables ter diferente influencia nos doentes curados e non curados. A Sección 1.4 tamén inclúe unha revisión bibliográfica sobre métodos paramétricos e non paramétricos de estimación en modelos de curación, tanto de tipo mixtura como noutros modelos de curación, e presenta a necesidade de técnicas non paramétricas neste contexto.

Capítulo 2: Estimador non paramétrico da incidencia

Neste capítulo introdúcense os principais resultados para o estimador non paramétrico da probabilidade de cura (2.4), proposto por Xu & Peng (2014), que se corresponde co estimador de Beran avaliado no maior tempo de fallo non censurado. Xu & Peng (2014) probaron a consistencia e a normalidade asintótica de dito estimador. A Sección 2.2 presenta a notación que será usada ao longo da tese, e aborda o problema da identificabilidade, necesaria para obter estimacións únicas das funcións do modelo. Nun modelo de curación, todos os tempos de vida observados non censurados corresponden necesariamente con suxeitos non curados; pero é imposible distinguir se un suxeito cun tempo censurado pertence ao grupo de susceptibles ou ao grupo de non susceptibles, xa que algúns individuos censurados poden experimentar fallos despois

do periodo de estudo. Isto da lugar a dificultades en distinguir entre modelos con incidencia alta e colas largas na función de distribución da latencia, e modelos con incidencia baixa e colas curtas na función de distribución da latencia. Para afrontar este problema da identificabilidade, próbase que se a función da latencia dun modelo é unha función de supervivencia propia, entón dito modelo é identificable. Na Sección 2.3 estúdase con maior profundidade o estimador non paramétrico da incidencia proposto por Xu & Peng (2014). Especificamente, demóstrase que é o máximo verosímil local, e obtense a representación i.i.d., así como a expresión do erro cadrático medio asintótico para dito estimador. Na Sección 2.4 introdúcese o problema de selección da ventá. En estudos preliminares traballouse cun selector plug-in, que presenta termos moi complicados de estimar, o que non da lugar á obtención de bos resultados. Por outra parte, para a selección da ventá tamén se considerou un método de validación cruzada, pero os resultados obtidos non son competentes, xa que a ventá resultante é moi variable e presenta tendencia a subestimar. Polo tanto, para a selección da ventá decidiuse utilizar o método bootstrap, detallado na Sección 2.4.1.

Na Sección 2.5 lévase a cabo un estudo de simulación para poder avaliar o comportamento do estimador non paramétrico, así como o método de selección da ventá. O estimador proposto compárase co estimador semiparamétrico da incidencia de Peng & Dear (2000), que asume unha función loxística, e está implementado no paquete *smcure* de R. Este estudo de simulación ten dous obxectivos: primeiro, avaliar o grao de dependencia que ten o estimador non paramétrico do parámetro de suavización, considerando unha reixa de ventás, e comparar os resultados cos obtidos utilizando o estimador semiparamétrico; e segundo, avaliar o comportamento do selector bootstrap do parámetro de suavizado. Para ambos casos, considéranse dous modelos para xerar os datos: por motivos de comparación, co primeiro modelo dase vantaxe ao estimador semiparamétrico, xa que os datos son xerados a partir dunha función loxística, e co segundo modelo, trátase de mostrar a ganancia obtida ao traballar co estimador non paramétrico proposto, que non require de condicións paramétricas ou semiparamétricas. Os resultados obtidos respecto á eficiencia do estimador non paramétrico mostran que este é moi competente no primeiro modelo, e supera con moita diferenza ao estimador semiparamétrico no segundo modelo. Nos resultados dos estudos relativos á eficiencia do selector bootstrap da ventá,

podemos apreciar que a ventá bootstrap aproxímase á ventá óptima teórica, adaptándose adecuadamente á forma para os diferentes tamaños muestrais estudados.

A continuación, na Sección 2.6, a metodoloxía é aplicada a unha base de datos de doentes de cancro colorrectal do CHUAC. Este conxunto de datos consiste en oito variables tomadas en 414 individuos: o indicador de censura, o tempo de vida observado, a localización (colon ou recto), a idade (de 23 a 102 anos), e o estadio TNM, que é o principal determinante na prognose dos doentes. Dito estadio ten tres compoñentes: T, que describe o tamaño do tumor e indica se invadiu algún tecido cercano; N, que mide os nódulos linfáticos involucrados; e M, que avalía a presenza (ou non) de metástase. A información destes tres aspectos pode ser combinada, o que nos permite clasificar a cada doente nun único estadio numérico, dando lugar a unha nova variable, denominada *estadio*, que toma valores do 1 ao 4. Nesta base de datos, ao redor dun 50 % das observacións son censuradas, cunha porcentaxe de censura variando desde o 30 % ata o 70 %, dependendo do estadio. Para aplicar a metodoloxía proposta, os datos foron divididos en 4 grupos de acordo coa variable categórica estadio. Os resultados mostran que o efecto da covariable idade na probabilidade de cura cambia co estadio. Ademais, a incidencia é maior nos estadios 1 e 2 que nos estadios 3 e 4. O motivo é que, nos estadios iniciais, a maioría das ciruxías teñen finalidades de curación, mentres que en estadios avanzados, as ciruxías son, normalmente, tratamentos paliativos e, polo tanto, a supervivencia dos doentes é mais baixa. Por exemplo, co estimador non paramétrico da incidencia no estadio 1, os doentes teñen unha probabilidade de cura de entre un 25 % e un 65 %, dependendo da idade; mentres que no estadio 3, para individuos maiores de 60 anos, nun período de 10 anos dita probabilidade decrece considerablemente dende un 40 % ata case un 0 %. Respecto ao estadio 4, un total de 11 dos 12 maiores tempos de vida, incluíndo o maior tempo de vida, son non censurados, e polo tanto, non curados. Isto da lugar a que o estimador non paramétrico da probabilidade de cura sexa igual a 0. Aínda que non se pode afirmar que para un doente de cancro colorrectal en estadio 4 sexa imposible sobrevivir, esta estimación ratifica a afirmación de que a supervivencia a longo prazo de doentes en estadio 4 de cancro colorrectal non é común. Este feito, lonxe de ser unha debilidade do modelo non paramétrico, é unha vantaxe importante, xa que permite detectar situacións nas que introducir a posibilidade

de cura non contribúe a mellorar o modelo. Por outra parte, é importante destacar a diferencia entre as curvas semiparamétricas e non paramétricas, o que parece indicar que o modelo loxístico non é válido para os datos. Os resultados correspondentes a este capítulo da tese están publicados en López-Cheda et al. (2017a).

Capítulo 3: Estimador non paramétrico da latencia

O estimador non paramétrico da latencia estúdase detalladamente neste capítulo. Na Sección 3.1 comézase presentando o estimador xenérico (3.1), que utiliza dúas ventás diferentes (unha para o estimador da incidencia, e outra para o estimador de Beran da función de supervivencia), xa que a ventá óptima para a probabilidade de cura non ten por que ser a mesma que a ventá óptima para o estimador de Beran. Agora ben, o estimador da latencia en (3.1) non da lugar necesariamente a unha función de distribución propia e de feito, non está garantido que sexa non negativo. Ademais, na Sección 3.5.1 móstrase que os valores óptimos de ambas ventás no estimador (3.1) son moi similares. Polo tanto, o traballo presentado na tese céntrase, principalmente, no estimador non paramétrico da latencia que utiliza unha ventá (3.2).

Na Sección 3.2 inclúense os resultados asintóticos para o estimador non paramétrico da latencia considerando dúas ventás diferentes (3.1). Obtense a representación i.i.d. e a expresión asintótica do erro cadrático medio. Ademais, próbase a normalidade asintótica. Na Sección 3.3 móstranse os mesmos resultados relativos ao estimador non paramétrico da latencia utilizando unha ventá (3.2), que se obteñen directamente dos correspondentes do estimador (3.1), considerando que ambas ventás son iguais. De forma similar que para o estimador da incidencia, para o estimador da latencia proposto (3.2), introdúcese un método de selección da ventá tipo bootstrap na Sección 3.4.

Os resultados do estudo de simulación preséntanse na Sección 3.5. Estes consisten en tres partes: na primeira, móstrase que se perde pouca eficiencia cando se considera unha soa ventá no estimador non paramétrico da latencia; na segunda, avalíase o bo comportamento do estimador non paramétrico proposto; e na terceira, compróbase o bo comportamento do selector da ventá. Para este estudo completo de simulación considéranse os mesmos dous modelos

de xeración de datos que para o caso da incidencia. Con respecto ao primeiro estudo, os resultados mostran que para a maioría de valores da covariable, ambas ventás óptimas do estimador non paramétrico da latencia (3.1) son moi similares ou incluso iguais. Nos resultados do segundo estudo compárase o estimador non paramétrico da latencia proposto co estimador semiparamétrico da latencia de Peng & Dear (2000), implementado no paquete *smcure* de R, cuxa expresión está moi cerca de satisfacer o modelo de riscos proporcionais e foi truncada. Pódese ver que o estimador non paramétrico da latencia proposto é comparable ao estimador semiparamétrico en situacións onde se espera que este último obteña mellores resultados, como é o caso do primeiro modelo, e supera ao estimador semiparamétrico cando non se cumplen condicións paramétricas. Na terceira parte deste estudo de simulación, relativa ao comportamento do selector da ventá, móstrase que existe moi pouca diferenza en termos do erro cadrático medio integrado entre as estimacións da latencia utilizando a ventá bootstrap e a ventá óptima.

De forma similar que para o estimador da incidencia, na Sección 3.6 aplícanse os métodos propostos a unha base de datos de doentes de cancro colorrectal do CHUAC. Debido ao tamaño pequeno de mostras en cada estadio, os resultados están presentados en dous grupos: por unha parte, os estadios 1-2 e, por outra parte, os estadios 3-4. Pódese observar que nos estadios 1-2, a covariable idade non parece determinante para o estimador da latencia, xa que as diferentes funcións da latencia estimadas son moi similares para todo o rango de idades. Polo contrario, nos estadios 3-4, a estimación da latencia varía considerablemente dependendo da idade: a supervivencia a curto prazo é maior en doentes xoves, mentres que a supervivencia a longo prazo é mais alta en doentes con maior idade. Por exemplo, a probabilidade de que o tempo de seguemento desde a diagnose ata a morte sexa maior que 4.5 anos está ao redor do 0.2 para doentes con 35 e 50 anos, mentres que para individuos de 80 anos, dita probabilidade é maior que 0.4. O motivo é que cando o cancro colorrectal é diagnosticado nun doente xove, normalmente está nun estadio avanzado e con peor prognose, xa que as células de cancro son mais activas en individuos de corta idade. Os principais resultados deste capítulo están publicados en López-Cheda et al. (2017b).

Capítulo 4: Contrastes de significación de covariables

Os contrastes de significación de covariables teñen moita importancia en análise de regresión, debido a que o número de covariables potenciais que poden ser incluídas no modelo pode ser grande. En particular, en modelos de curación de tipo mixtura, a selección de variables ten especial importancia porque as covariables que teñen un efecto significativo na supervivencia dos individuos susceptibles non son necesariamente as mesmas que aquelas que inflúen na probabilidade de cura.

Neste capítulo preséntase un contraste de significación de covariables para a incidencia, baseado no método de Delgado & González-Manteiga (2001), quen introduciron un contraste para seleccionar covariables explicativas en contextos de regresión non paramétricos sen censura. A principal vantaxe sobre outros métodos de suavización é que este só require un parámetro de suavizado para o estimador non paramétrico da función de regresión que depende das covariables significativas baixo a hipótese nula. Esta característica resolve, en parte, o problema da "maldición da dimensionalidade"en contextos non paramétricos e da lugar a que o estatístico non dependa de ningún parámetro ventá se non se asume, baixo a hipótese nula, a dependencia da probabilidade de cura de ningunha covariable. Ademais, o método é estendido a casos de covariables non continuas: binarias, discretas e cualitativas. Na Sección 4.3 introdúcese o caso cunha soa covariable (isto é, cando se comproba se a probabilidade de cura, como unha función da covariable, pode ser considerada constante), denominado *caso 1*. A Sección 4.4 presenta o *caso 2* onde, baixo a hipótese nula, a probabilidade de cura depende dunha covariable unidimensional. Baixo a hipótese alternativa, a mesma probabilidade depende dunha covariable m-dimensional, con $m < 1$.

Na Sección 4.5 esta proposta esténdese a contextos cun grande número de covariables. Nestes casos, como en xenómica ou en outros campos relacionados coa bioloxía, a probabilidade de obter un resultado significativo simplemente debido á sorte é moi alta. Para tratar este problema, utilízase un algoritmo que controla a proporción esperada de rexeitar hipóteses falsas, isto é, o denominado *False Discovery Rate* (FDR). Esta taxa de erro é equivalente á *Family-wise Error Rate* (FWER) cando todas as hipóteses son verdadeiras, pero mais

pequena noutro caso. Cabe destacar que, cando usamos o FDR na práctica utilizamos dous métodos: un que controla o FDR e outro método mais conservador, que ademais do FDR controla o FWER.

Un extenso estudo de simulación inclúese na Sección 4.6. Para o caso 1, onde se estuda se a covariable Z ten un efecto estatísticamente significativo na incidencia (H_1) ou non (H_0), considéranse catro situacións dependendo do tipo da covariable Z: continua, discreta, binaria ou cualitativa. No caso 2, onde se contrasta se a incidencia depende únicamente da covariable X (H_0) ou tanto de X coma de Z (H_1), estúdanse os dezaseis escenarios resultantes da combinación de tipos de ambas covariables X e Z: ambas continuas, unha continua e outra discreta, etc. A simulación para o caso cun grande número de covariables levouse a cabo únicamente no caso 1, onde se consideran $Z_1, \dots, Z_m$ covariables e se estuda, para cada unha de forma independente, se a probabilidade de cura depende significativamente dela ou non.

Para concluír este capítulo, na Sección 4.7, aplícase a metodoloxía proposta a dúas bases de datos: a relacionada con doentes de cancro colorrectal do CHUAC, e outra relativa a doentes de sarcomas do CHUS. Esta última consiste en 261 observacións con 372452 covariables. Especificamente, inclúe 372420 covariables con información de metilacións de ADN e 32 covariables de datos clínicos. As covariables de metilacións son continuas, con valores entre 0 e 1. Un total de 195 observacións son censuradas, o que se corresponde cun 74.71 % dos datos. Para levar a cabo o algoritmo proposto dos contrastes de significación, utilízase o método de Benjamini & Hochberg (1995) e a súa alternativa conservadora, de Benjamini & Yekutieli (2001). Co método non conservador, para $B = 1000$ remostras bootstrap, obtense que co estatístico de Cramér-von Mises, 4179 covariables son significativas e 6924 non son concluíntes (polo que precisan volver a ser analizadas na seguinte iteración do proceso); mentres que co estatístico de Kolmogorov-Smirnov, 3457 covariables son significativas e 6263 son non concluíntes. No momento de depósito da tese, o programa está lanzado na seguinte fase do algoritmo con $B = 10000$ remostras boostrap. Respeto á alternativa conservadora, os resultados mostran que só unha covariable é significativa para a probabilidade de cura: *Year of initial pathologic diagnosis*. Utilizando o estimador non paramétrico da incidencia para esa covariable pódese ver que a probabilidade de cura está ao redor de 0.35

para datas anteriores ao ano 2010, mentres que entre ese ano e o ano 2013, dita probabilidade duplícase.

Capítulo 5: Traballo futuro

Considérase a posibilidade de aplicar a metodoloxía proposta a casos de alta dimensión que inclúan análise de imaxes, relacionadas con diagnose de cancro. Ademais, como o software utilizado en todos os estudos de simulación nesta tese é R, un entorno libre para gráficos e computación estatística, desenvolverase un paquete de R con todas as técnicas estudadas. As mesmo, introdúcense outros problemas de traballo futuro: estudar o estimador presuavizado da probabilidade de cura, estender os métodos a casos con truncamento ou datos censurados en intervalos, usar single-index models en análise de supervivencia para datos censurados e probar a consistencia dos métodos bootstrap, estudando os límites de converxencia.

Bibliography

Akritas, M. (1986). Bootstrapping the Kaplan-Meier estimator. *J. Am. Stat. Assoc.*, *81*, 1032–1038. doi: 10.1080/01621459.1986.10478369

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

Arcones, M. A. (1997). The law of the iterated logarithm for a triangular array of empirical processes. *Electron. J. Probab.*, *2*, 1–39. doi: 10.1214/EJP.v2-19

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B - Stat. Methodol.*, *57*, 289–300. doi: 10.2307/2346101

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, *29*, 1165–1188. doi: 10.1214/aos/1013699998

Bennani-Baiti, I. M. (2011). Epigenetic and epigenomic mechanisms shape sarcoma and other mesenchymal tumor pathogenesis. *Epigenomics*, *3*, 715–732. doi: 10.2217/epi.11.93

Beran, R. (1981). *Nonparametric regression with randomly censored survival data* (Tech. Rep.). Berkeley: University of California, Berkeley.

Betensky, R. A., & Schoenfeld, D. A. (2001). Nonparametric estimation in a cure model with random cure times. *Biometrics*, *57*, 282–286. doi: 10.1111/j.0006-341X.2001.00282.x

Billingsley, P. (1968). *Convergence of Probability Measures.* New York: Wiley.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc. Ser. B - Stat. Methodol.*, *11*, 15–53. doi: 10.2307/2983694

Breslow, N., & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Stat.*, *2*, 437–453. doi: 10.1214/aos/1176342705

Burningham, Z., Hashibe, M., Spector, L., & Schiffmann, J. D. (2012). The epidemiology of sarcoma. *Clin. Sarcoma Res.*, *421*, 14–29. doi: 10.1186/2045-3329-2-14

Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2012). *smcure: Fit Semiparametric Mixture Cure Models, R package version 2.0.* `http://CRAN.R-project.org/package=smcure`.

Cai, Z. (1998). Kernel density and hazard rate estimation for censored dependent data. *J. Multivar. Anal.*, *67*, 23–34. doi: 10.1006/jmva.1998.1752

Cantor, A. B., & Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, *11*, 931–937. doi: 10.1002/sim.4780110710

Cao, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Stat.*, *19*, 2226–2231. doi: 10.1214/aos/1176348394

Cao, R., & González-Manteiga, W. (1993). Bootstrap methods in regression smoothing. *J. Nonparametr. Stat.*, *2*, 379–388. doi: 10.1080/10485259308832566

Cao, R., Janssen, P., & Veraverbeke, N. (2001). Relative density estimation and local bandwidth selection for censored data. *Comput. Stat. Data Anal.*, *36*, 497–510. doi: 10.1016/S0167-9473(00)00055-4

Cao, R., & Van Keilegom, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *Can. J. Stat.*, *34*, 61–77. doi: 10.1002/cjs.5550340106

Chen, K., Jin, Z., & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, *89*, 659–668. doi: 10.1093/biomet/89.3.659

Chen, M. H., Ibrahim, J. G., & Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *J. Am. Stat. Assoc.*, *94*, 909–919. doi: 10.1080/01621459.1999.10474196

Chi, Y. Y., & Ibrahim, J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Stat. Sin.*, *17*, 445–462.

Corbière, F., Commenges, D., Taylor, J. M. G., & Joly, P. (2009). A penalized likelihood approach for mixture cure models. *Stat. Med.*, *28*, 510–524. doi: 10.1002/sim.3481

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Ann. Stat.*, *17*, 1157–1167. doi: 10.1214/aos/1176347261

Dabrowska, D. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scand. J. Stat.*, *19*, 351–361. doi: 10.2307/4616252

Delgado, M. A., & González-Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Ann. Stat.*, *29*, 1469–1507. doi: 10.1214/aos/1013203462

Denham, J. W., Denham, E., Dear, K. B., & Hudson, G. V. (1996). The follicular non-Hodgkin's lymphomas - I. The possibility of cure. *Eur. J. Cancer*, *32*, 470–479. doi: 10.1016/0959-8049(95)00607-9

Diehl, S., & Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivar. Anal.*, *25*, 299–310. doi: 10.1016/0047-259X(88)90053-X

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 831–853). University of California Press, Berkeley, California.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.*, *7*, 1–26. doi: 10.1214/aos/1176344552

Efron, B. (1981). Censored data and the bootstrap. *J. Am. Stat. Assoc.*, *76*, 312–319. doi: 10.1080/01621459.1981.10477650

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Vol. 57). New York: Monographs on Statistics and Applied Probability, Chapman and Hall.

Esteller, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.*, *358*, 1148–1159. doi: 10.1056/NEJMra072067

Esteller, M., García-Foncillas, J., Andion, E., Goodman, S. N., Hidalgo, O. F., Vanaclocha, V., ... Herman, J. G. (2000). Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.*, *343*, 1350–1354. doi: 10.1056/NEJM200011093431901

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, *38*, 1041–1046. doi: 10.2307/2529885

Farewell, V. T. (1986). Mixture models in survival analysis: are they worth the risk? *Can. J. Stat.*, *14*, 257–262. doi: 10.2307/3314804

Földes, A., & Rejtö, L. (1981). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. *Ann. Stat.*, *9*, 122–129. doi: 10.1214/aos/1176345337

Földes, A., Rejtö, L., & Winter, B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data, II: Estimation of density and failure rate. *Period. Math. Hung.*, *12*, 15–29. doi: 10.1007/BF01848168

Gao, J., & Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *J. Am. Stat. Assoc.*, *103*, 1584–1594. doi: 10.1198/016214508000000968

Gasser, T., & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.*, *11*, 171–185.

Ghitany, M. E., Maller, R. A., & Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *J. Multivar. Anal.*, *49*, 218–241. doi: 10.1006/jmva.1994.1023

Ghorai, J. K., & Pattanaik, L. M. (1990). L_1-Consistency of the kernel density estimators based on randomly right censored data. *Commun. Stat. - Theory Methods*, *19*, 2853–2870. doi: 10.1080/03610929008830353

Gill, R. D. (1980). *Censoring and stochastic integrals.* Amsterdam: Mathematisch Centrum, 124: Mathematical Centre Tracts.

González-Manteiga, W., & Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *J. Nonparametr. Stat.*, *4*, 65–78. doi: 10.1080/10485259408832601

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* New York: Springer Series in Statistics, Springer-Verlag.

Härdle, W., & Marron, J. S. (1985). Optimal bandwidth selection in non-parametric regression function estimation. *Ann. Stat.*, *13*, 1465–1481. doi: 10.1214/aos/1176349748

Haybittle, J. L. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br. J. Radiol.*, *32*, 725–733. doi: 10.1259/0007-1285-32-383-725

Haybittle, J. L. (1965). A two-parameter model for the survival curve of treated cancer patients. *J. Am. Stat. Assoc.*, *60*, 16–26. doi: 10.1080/01621459.1965.10480772

Iglesias-Pérez, M. C. (2007). Selección de la ventana en estimación de la distribución condicional. In *Libro de Actas del XXX Congreso Nacional de Estadística e Investigación Operativa.*

Iglesias-Pérez, M. C., & González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *J. Nonparametr. Stat.*, *10*, 213–244. doi: 10.1080/10485259908832761

Jácome, M. A., & Cao, R. (2004). Presmoothed kernel density estimator for censored data. *J. Nonparametr. Stat.*, *16*, 289–309. doi: 10.1080/10485250310001622622

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scand. J. Stat.*, *5*, 195–199. doi: 10.1080/03610928708829561

Kang, G. H. (2012). CpG island hypermethylation in gastric carcinoma and its premalignant lesions. *Korean J. Pathol.*, *46*, 1–9. doi: 10.4132/KoreanJPathol.2012.46.1.1

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, *53*, 457–481. doi: 10.2307/2281868

Kuk, A. Y. C., & Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, *79*, 531–541. doi: 10.1093/biomet/79.3.531

Kulasekera, K. B., & Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *J. Am. Stat. Assoc.*, *92*, 500–511. doi: 10.2307/2965699

Lai, X., & Yau, K. K. W. (2010). Extending the long-term survivor mixture model with random effects for clustered survival data. *Comput. Stat. Data Anal.*, *54*, 2103–2112. doi: 10.1016/j.csda.2010.03.017

Laska, E. M., & Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, *48*, 1223–1234. doi: 10.2307/2532714

Li, C., & Taylor, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Stat. Med.*, *21*, 3235–3247. doi: 10.1002/sim.1260

Li, C. S., Taylor, J. M. G., & Sy, J. P. (2001). Identifiability of cure models. *Stat. Probab. Lett.*, *54*, 389–395. doi: 10.1016/S0167-7152(01)00105-5

Li, G., & Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Ann. Inst. Stat. Math.*, *53*, 708–729. doi: 10.1023/A:1014644700806

Li, Q., & Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Stat. Sin.*, *14*, 485–512. doi: 10.1023/A:1014644700806

Liu, H., & Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. *J. Am. Stat. Assoc.*, *104*, 1168–1178. doi: 10.1198/jasa.2009.tm07494

Lo, S. H., Mack, Y. P., & Wang, J. L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Relat. Fields*, *80*, 461–473. doi: 10.1007/BF01794434

Lo, S. H., & Singh, K. (1986). The product-limit estimator and the bootstrap: some asymptotic representations. *Probab. Theory Relat. Fields*, *71*, 455–465. doi: 10.1007/BF01000216

López-Cheda, A., Cao, R., Jácome, M. A., & Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput. Stat. Data Anal.*, *105*, 144–165. doi: 10.1016/j.csda.2016.08.002

López-Cheda, A., Jácome, M. A., & Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *Test*, *26*, 353–376. doi: 10.1007/s11749-016-0515-1

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Stat. Sin.*, *20*, 661–674. doi: 10.1002/sim.1260

Lu, W., & Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, *91*, 331–343. doi: 10.1093/biomet/91.2.331

Maller, R. A., & Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, *79*, 731–739. doi: 10.1093/biomet/79.4.731

Maller, R. A., & Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. *J. Am. Stat. Assoc.*, *89*, 1499–1506. doi: 10.1080/01621459.1994.10476889

Maller, R. A., & Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*. Chichester, U. K.: Wiley. doi: 10.1002/cbm.318

Martínez-Camblor, P. (2010). Nonparametric k-sample test based on kernel density estimator for paired design. *Comput. Stat. Data Anal.*, *54*, 2035–2045. doi: 10.1016/j.csda.2010.03.009

Martínez-Camblor, P., & de Uña-Álvarez, J. (2013). Studying the bandwidth in k-sample smooth tests. *Comput. Stat.*, *28*, 875–892. doi: 10.1007/s00180-012-0333-1

Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *Ann. Stat.*, *14*, 766–773. doi: 10.1214/aos/1176349954

Miyamoto, Y., Hayashi, N., Sakamoto, Y., Ohuchi, M., Tokunagam, R., Kurashige, Y., Hiyoshi, Y., Baba, S., Iwagami, N., Yoshida, M., Yoshida, J., & Baba, H. (2015). Predictors of long-term survival in patients with stage IV colorectal cancer with multi-organ metastases: a single-center retrospective analysis. *Int. J. Clin. Oncol.*, *20*, 1140–1146. doi: 10.1007/s10147-015-0835-2

Muir, C. S., & Percy, C. (1991). Cancer registration: principles and methods. Classification and coding neoplasms. *IARC Sci. Publ.*, *95*, 64–81.

Müller, U. U., & Van Keilegom, I. (2018). Goodness-of-fit tests for the cure rate in a mixture cure model. *Submitted to Biometrika*.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.*, *9*, 141–142. doi: 10.1137/1109020

Othus, M., Li, Y., & Tiwari, R. C. (2009). A class of semiparametric mixture cure survival models with dependent censoring. *J. Am. Stat. Assoc.*, *104*, 1241–1250. doi: 10.1198/jasa.2009.tm08033

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, *33*, 1065–1076. doi: 10.1214/aoms/1177704472

Peng, Y. (2003). Fitting semiparametric cure models. *Comput. Stat. Data Anal.*, *41*, 481–490. doi: 10.1016/S0167-9473(02)00184-6

Peng, Y., & Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, *56*, 237–243. doi: 10.1111/j.0006-341X.2000.00237.x

Peng, Y., Dear, K. B., & Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Stat. Med.*, *17*, 813–830. doi: 10.1002/(SICI)1097-0258(19980430)17

Peng, Y., & Zhang, J. (2008). Identifiability of a mixture cure frailty model. *Stat. Probab. Lett.*, *78*, 2604–2608. doi: 10.1016/j.spl.2008.07.044

Phadia, E. G., & Shao, P. Y. (1999). Exact moments of the product limit estimator. *Stat. Probab. Lett.*, *41*, 277–286. doi: 10.1016/S0167-7152(98)00164-3

Reid, N. (1981). Estimating the median survival time. *Biometrika*, *68*, 601–608. doi: 10.1093/biomet/68.3.601

Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. *Ann. Math. Stat.*, *27*, 832–837. doi: 10.1214/aoms/1177728190

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer. doi: 10.1007/978-1-4612-4026-6

Strzalkowska-Kominiak, E., & Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *J. Multivar. Anal.*, *114*, 74–98. doi: 10.1016/j.jmva.2012.07.012

Suzuki, H., Watkins, D. N., Jair, K. W., Schuebel, K. E., Markowitz, S. D., Chen, W. D., ... Baylin, S. B. (2004). Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nature Genet.*, *36*, 417–422. doi: 10.1038/ng1330

Sy, J. P., & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, *56*, 227–236. doi: 10.1111/j.0006-341X.2000.00227.x

Sy, J. P., & Taylor, J. M. G. (2001). Standard errors for the Cox proportional hazards cure model. *Math. Comput. Model.*, *33*, 1237–1251. doi: 10.1016/S0895-7177(00)00312-5

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, *51*, 899–907. doi: 10.2307/2532991

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, *54*, 1508–1516. doi: 10.2307/2533675

Tsodikov, A. (2001). Estimation of survival based on proportional hazards when cure is a possibility. *Math. Comput. Model.*, *33*, 1227–1236. doi: 10.1016/S0895-7177(00)00311-3

Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *J. R. Stat. Soc. Ser. B - Stat. Methodol.*, *65*, 759–774. doi: 10.1111/1467-9868.00414

Tsuchiya, T., Sekine, K., Hinohara, S., Namiki, T., Nobori, T., & Kaneko, Y. (2000). Analysis of the p16INK4, p14ARF, p15, TP53, and MDM2 genes and their prognostic implications in osteosarcoma and Ewing sarcoma. *Cancer Genet. Cytogenet.*, *120*, 91–98. doi: 10.1016/S0165-4608(99)00255-1

Ushijima, T. (2005). Detection and interpretation of altered methylation patterns in cancer cells. *Nat. Rev. Cancer*, *5*, 223–231. doi: 10.1038/nrc1571

Van Keilegom, I. (2013). *Discussion on: 'An updated review of goodness-of-fit tests for regression models', by W. González-Manteiga and R.M. Crujeiras.* (Discussion paper). Louvain-la-Neuve, Belgium: Université catholique de Louvain.

Van Keilegom, I., & Veraverbeke, N. (1997a). Weak convergence of the bootstrapped conditional Kaplan-Meier process and its quantile process. *Commun. Stat. - Theory Methods*, *26*, 853–869. doi: 10.1080/03610929708831954

Van Keilegom, I., & Veraverbeke, N. (1997b). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Stat. Math.*, *49*, 467–491. doi: 10.1023/A:1003166728321

Wang, L., Du, P., & Lian, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, *68*, 726–735. doi: 10.1111/j.1541-0420.2011.01715.x

Wang, M. C. (1987). Product limit estimates: a generalized maximum likelihood study. *Commun. Stat. - Theory Methods*, *16*, 3117–3132. doi: 10.1080/03610928708829561

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, *26*, 359–372.

Xu, J., & Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Can. J. Stat.*, *42*, 1–17. doi: 10.1002/cjs.11197

Yakovlev, A. Y., Cantor, A. B., & Shuster, J. J. (1994). Parametric versus nonparametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, *13*, 983–986. doi: 10.1002/sim.4780130908

Yakovlev, A. Y., & Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*. Singapore: World Scientific Pub Co Inc.

Yamaguchi, K. (1992). Accelerated failure-time regression model with a regression model of surviving fraction: an analysis of permanent employment in Japan. *J. Am. Stat. Assoc.*, *87*, 284–292. doi: 10.1080/01621459.1992.10475207

Yin, G. S. (2005). Bayesian cure rate frailty models with application to a root canal therapy study. *Biometrics*, *61*, 552–558. doi: 10.1111/j.1541-0420.2005.040336.x

Yu, B., & Peng, Y. (2008). Mixture cure models for multivariate survival data. *Comput. Stat. Data Anal.*, *52*, 1524–1532. doi: 10.1016/j.csda.2007.04.018

Zeng, D., Yin, G., & Ibrahim, J. (2006). Semiparametric transformation models for survival data with a cure fraction. *J. Am. Stat. Assoc.*, *101*, 670–684. doi: 10.1198/016214505000001122

Zhang, J., & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Stat. Med.*, *26*, 3157–3171. doi: 10.1002/sim.2748

Zhang, J., Peng, Y., & Li, H. (2013). A new semiparametric estimation method for accelerated hazards mixture cure model. *Comput. Stat. Data Anal.*, *59*, 95–102. doi: 10.1111/j.1541-0420.2011.01592.x