



UNIVERSITE CATHOLIQUE DE LOUVAIN

INSTITUT DE STATISTIQUE, BIOSTATISTIQUE ET SCIENCES
ACTUARIELLES

Parametrically Guided Nonparametric Estimation and Inference with Censored Data

Membres du jury:

Prof. Anouar El Ghouh (*Promoteur*)

Prof. Ingrid Van Keilegom (*Promoteur*)

Prof. Irène Gijbels (KUL)

Prof. Ingrid Kristine Glad (University of Oslo)

Prof. Bernadette Govaerts

Prof. Catherine Legrand

Thèse présentée en vue de

l'obtention du grade de

Docteur en Sciences

(orientation statistique) par :

Majda Talamakrouni

Louvain-la-Neuve

2016

À mes chers parents et mon fils Haroun

وَمَ أَنَا دُونِ أُمِّي وَأَبِي.

Acknowledgements

Tout d'abord, je voudrais exprimer ma profonde gratitude à mes directeurs de thèse, Prof. Ingrid Van Keilegom et Prof. Anouar El Ghouch, pour le soutien continu et inconditionnel à ce travail. Je suis ravie d'avoir travaillé en leur compagnie car outre leur appui scientifique, ils ont toujours été là pour me soutenir et me conseiller durant les moments les plus difficiles de ces sept années. Je vous remercie pour les nombreuses discussions que nous avons eues, vos nombreuses relectures et corrections de mon travail. Merci pour votre oeil critique qui m'a été très précieux pour structurer et améliorer la qualité de ma recherche. Merci pour votre encouragement et patience et surtout merci d'avoir cru en moi.

"Pour tout cela, je vous dis MERCI"

I would like also to thank Prof. Ingrid Kristine Glad, first for being the first statistician who investigated the guided estimation and second for the very nice discussion and valuable comments.

Mes plus sincères remerciements vont aussi aux membres de mon comité, Prof. Catherine Legrand et Prof. Irène Gijbels, qui ont accepté avec enthousiasme de juger mon travail tout au long de ces années. Merci pour votre temps et remarques précieuses qui ont amélioré considérablement mon travail.

Je remercie également et très spécialement Prof. Bernadette Govaerts pour l'honneur qu'elle me fait d'être la présidente de mon jury de thèse et aussi d'avoir réduit ma charge d'assistante pour un de ses cours pour me permettre de me consacrer à la rédaction de ma thèse.

Je remercie toutes les personnes formidables que j'ai rencontrées par le biais de l'ISBA. Merci pour votre support et vos encouragements.

Enfin, je remercie mes chers parents et toute ma famille, surtout mon fils Haroun et ma sœur Mariam qui sans leurs sacrifices et soutien indéfectible ce travail n'aurait pu être accompli.

Abstract

Parametrically guided nonparametric estimation is an attractive method that allows to improve the bias of a nonparametric estimator by using a parametric pilot estimator. The aim of this dissertation is to generalize the parametrically guided nonparametric estimation to randomly right-censored data. The generalization is performed in three different contexts. First, based on the Kaplan-Meier (1958) estimator, we provide new parametrically guided kernel density and hazard rate estimators. Then, we investigate the parametrically guided local linear regression and the parametrically guided quasi-likelihood estimation using a synthetic data approach. The asymptotic properties of the new-guided estimators as well as their finite sample performance are investigated and compared with the corresponding unguided nonparametric estimators via numerical studies and applications to real data. The results confirm the bias reduction property and show that using an appropriate guide and the optimal bandwidth the guided estimators outperform the classical nonparametric estimators.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Models for survival distributions	4
1.2.1	Parametric approach	5
1.2.2	Nonparametric approach	7
1.2.3	Semi-parametric approach	11
1.3	Models with one covariate	12
1.3.1	Parametric and semi-parametric models	13
1.3.2	Nonparametric estimation	17
1.4	Parametrically guided nonparametric estimation	22
1.5	Brief review of the chapters	23
2	Guided kernel density and hazard estimation	25
2.1	Introduction	25
2.2	Methodology	26
2.3	Asymptotic results	28
2.3.1	The estimator with a fixed guide	29
2.3.2	The estimator with an estimated guide	30
2.3.3	Guided kernel hazard estimator	33
2.4	Simulation results	34
2.4.1	Normal mixture model	35
2.4.2	Weibull density with an exponential guide	39
2.5	Application	45
2.5.1	Density and hazard estimation for the UIS data	45

2.5.2	Pointwise asymptotic confidence intervals	47
2.5.3	Extending the PGK estimation to the Cox model	48
2.6	Appendix A	51
2.7	Appendix B	53
2.8	Appendix C	56
2.9	Appendix D	57
3	Guided local linear regression	61
3.1	Introduction	61
3.2	Model and estimation procedure	62
3.2.1	Definitions and notations	62
3.2.2	Parametrically guided nonparametric regression with censored data	63
3.3	Asymptotic results	65
3.3.1	The estimator with a fixed guide	66
3.3.2	The estimator with an estimated guide	67
3.3.3	Generalized parametrically guided local linear estimator	70
3.4	Simulations	71
3.4.1	Exponential proportional hazards model	71
3.4.2	Sinusoidal model	74
3.5	Real data analysis	77
3.6	Appendix	79
4	Guided local quasi-likelihood estimation	93
4.1	Introduction	93
4.2	Model and methodology	95
4.2.1	Guided local quasi-likelihood for complete data	96
4.2.2	Guided local quasi-likelihood and censoring	97
4.3	Theoretical properties	98
4.3.1	The model with a fixed guide	98
4.3.2	The model with an estimated guide	100
4.3.3	Extension to unified family of corrections	104

4.4	Simulation results	106
4.4.1	Poisson model	106
4.4.2	Exponential model	110
4.5	Application	113
4.6	Appendix A	115
4.7	Appendix B	118
4.8	Appendix C: Proofs	120
5	Conclusion and discussion	129

Chapter 1

Introduction

1.1 Motivation

Thanks to their widespread usage through many different disciplines, applications in the field of survival analysis have encountered significant developments over the last decades. However, statistical methods for survival data analysis are usually complicated by several censoring mechanisms. In fact, in many scientific fields such as engineering or biomedical research, it often happens that some observations are not completely observed and the problem of censoring is frequently encountered. There exists different type of censoring, the most common type that occurs in a broad variety of applications is the random right censoring. A variable of interest Y is right censored when it is only partially observed due to the earlier occurrence of a competing event C . As an example, we consider the UIS dataset from the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study. The goal of this study is to model time to return to drug use (in days) for 628 patients enrolled in two different residential treatment programs that differed in length. Actually, some patients might never return to drug use or they might return to drug use after the end of the study while other patients might decide to leave the study. For these patients (120 in total), the observations are right censored and only a lower bound of the time to return to drug use is observed. For a detailed description of the study we refer to the book of Hosmer *et al.* (2008).

In this statistical area, a first important objective is to visualize and explore the

distribution of data through the estimation of the probability density, the hazard or the survival function of Y . However, in many situations, the presence of an explanatory variable X may add an extra challenge to the censoring problem. For example, in addition to time to return to drug use, the age and the type of treatment of each patient were also recorded in the UIS dataset. One might be interested in the effect of patient age on the time to return to drug use for each type of treatment. The data are presented for each type of treatment in Figure 1.1.1.

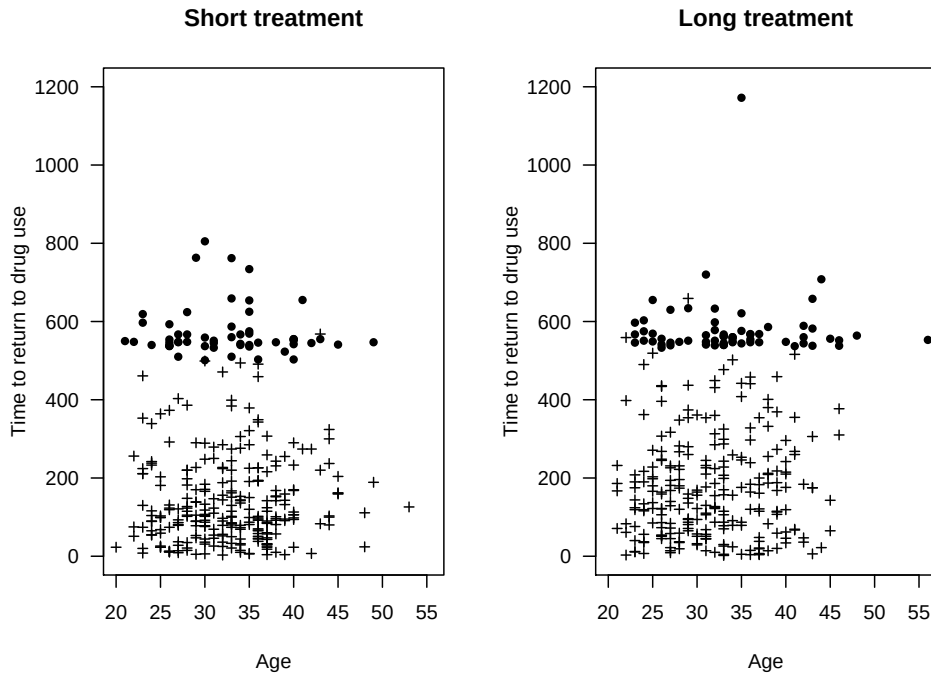


Figure 1.1.1: UIS dataset: censored data are indicated by $\bullet$ and uncensored data by $+$.

In such situation one can estimate the conditional probability density, conditional hazard, or conditional survival function as well as the mean regression function or a transformation of the mean regression function. The estimation under censoring mechanism can be carried out using different approaches. The parametric approach has the advantage of being powerful by its $\sqrt{n}$ -rate of convergence and also precise when the chosen family is correctly specified. However, a major complication that is

emphasized in parametric modeling is the risk of biased and inconsistent parameter estimation due to misspecification problem. In the fully nonparametric approach, the estimators suffer from the curse of dimensionality and have in general a slower rate of convergence. However, despite its drawbacks, nonparametric approach provides more flexibility since the estimation is not based on any parametrized family of functions and remains more robust and applicable in practice. There have been another interesting proposal in the literature which represents a mixture of the parametric and the nonparametric approaches, namely, the semi-parametric approach. Typically, a direct consequence of the semi-parametric analysis is the relaxation of the parametric assumptions as well as the reduction of the complexity of high dimensional problems.

Usually, even when the proposed model is misspecified, parametric estimation can provide a valuable information about the phenomenon under study. This motivates the consideration of an approach called parametrically guided nonparametric estimation, that contains both a parametric and a nonparametric component, and is therefore a good compromise between the two extreme approaches described above. The essential idea behind the guided estimation is to start with a crude parametric estimator which is not necessarily well specified, then to correct this parametric guide using a particular type of correction and a nonparametric estimator. Unlike the classical semi-parametric models (additive model, partially linear model, etc) a guided nonparametric estimator is completely nonparametric in the sense that it does not rely on any assumed global structure. On the other hand, a guided nonparametric estimator takes advantage of both parametric and nonparametric methods: It will always converge to the true model no matter if the parametric part is correct or not, and it will adapt automatically to the parametric model if the latter is locally or globally close to the true underlying curve.

In the complete data case, considerable attention has recently been paid to parametrically guided nonparametric estimation in the literature. The starting point for this method was Hjort and Glad (1995), who introduced the parametric guided kernel scheme and proved the bias reduction property of their guided estimator in the context of density estimation. Then, among others, Glad (1998a, 1998b), Fan and Ullah (1999), Gozalo and Linton (2000), Mays *et al.* (2001), Rahman and Ullah (2002) and Martins-Filho *et al.* (2008) investigated the guided nonparametric estimation in the

locally weighted least squares regression. Later, the method has been extended to local quantile regression by El Ghouh and Genton (2009) and to generalized linear models and local quasi-likelihood by Fan *et al.* (2009). Finally and very recently, Fan *et al.* (2014) applied the guided estimation to generalized additive models and Davenport *et al.* (2015) studied the parametrically guided nonparametric estimation in nonparametric varying coefficient models with quasi-likelihood. These papers noticed and showed the interesting property of bias reduction for their guided nonparametric estimators compared with the unguided nonparametric estimators without any increase in the variance, which is not at all trivial to achieve in a fully nonparametric context. As far as we are aware, parametrically guided nonparametric estimation has never been investigated in the presence of censoring mechanism.

The main objective of this thesis is to adapt and generalize the parametrically guided nonparametric estimation to the censored data case. The extension to censored data is performed according to whether the explanatory variable is taken into account or not. First, we study and extend the parametrically guided kernel density and hazard function estimators to accommodate the censored data. Then, we propose new parametrically guided nonparametric estimators for a particular transformation of the mean regression function in the local linear regression and local quasi-likelihood estimation.

1.2 Models for survival distributions

Let $Y_1, \dots, Y_n$ be independent and identically distributed (i.i.d) random variables each with unknown distribution function $F(t) = P(Y_1 \leq t)$. Throughout this thesis, we consider non-informative censoring, that is, the censoring distribution is not related to the time-to-event distribution. Under right censoring mechanism, there exist i.i.d random variables $C_1, \dots, C_n$ with unknown distribution function $G(t) = P(C_1 \leq t)$ and probability density $g(t)$ such that instead of observing $\{Y_i, i = 1, \dots, n\}$, one can only observe $\{(T_i, \delta_i), i = 1, \dots, n\}$, where $T_i = \min(Y_i, C_i)$ with unknown distribution function $H(t) = P(T_1 \leq t)$, $\delta_i = \mathbf{I}(Y_i \leq C_i)$ and $\mathbf{I}(\cdot)$ is the indicator function. Suppose for the moment that Y_i and C_i are independent. Let f be the probability density

function of Y and define the hazard function

$$\lambda(t) = f(t)/(1 - F(t)).$$

1.2.1 Parametric approach

In the parametric approach, the estimation is carried out within an assumed parametric family of densities $\{f(\cdot, \theta), \theta \in \Theta\}$, where θ is a vector of unknown parameters and Θ is a subset of $\mathbb{R}^p$. The Weibull, exponential and lognormal distribution are amongst the most commonly used distributions in the censored data framework. In practice, the parametric family of densities is chosen from previous experience and descriptive analysis of the underlying structure. For example, a constant hazard rate might refer to an exponential distribution. The most effective parametric methods that are extensively used in the presence of censoring mechanism are likelihood-based approaches. Under independence between Y and C , the likelihood for the sample $(T_i, \delta_i), i = 1, \dots, n$, has the form

$$L(\theta) = \prod_{i=1}^n \left\{ G(T_i)^{\delta_i} g(T_i)^{1-\delta_i} \right\} \prod_{i=1}^n \left\{ f(T_i, \theta)^{\delta_i} (1 - F(T_i, \theta))^{1-\delta_i} \right\}.$$

Since we consider non-informative censoring, the first product on the right side can be considered as a constant with respect to the parameter of interest θ . The contribution from uncensored observation to the likelihood is measured by $f(t, \theta)$ while the contribution from censored observation is evaluated by the parametric survival function defined as $1 - F(t, \theta)$, where $F(t, \theta)$ is the proposed parametric distribution function. Hence, the log-likelihood function is given by

$$l_n(\theta) = \sum_{i=1}^n \left\{ \delta_i \log f(T_i, \theta) + (1 - \delta_i) \log[1 - F(T_i, \theta)] \right\}.$$

Typically, the maximization of the log-likelihood function $l_n(\theta)$ requires iterative methods such as Newton-Raphson method, see for example Miller (1981) and Cox and Oakes (1984) for more details. The estimators of the density and hazard function are obtained by the plug-in approach, that is

$$\hat{f}(t) = f(t, \hat{\theta}), \quad \hat{\lambda}(t) = f(t, \hat{\theta})/(1 - F(t, \hat{\theta})),$$

where $\hat{\theta}$ is the so-called maximum-likelihood estimator (MLE) defined as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} l_n(\theta).$$

Figure 1.2.1, gives an application of the maximum likelihood estimation to the UIS data based on the Weibull distribution.

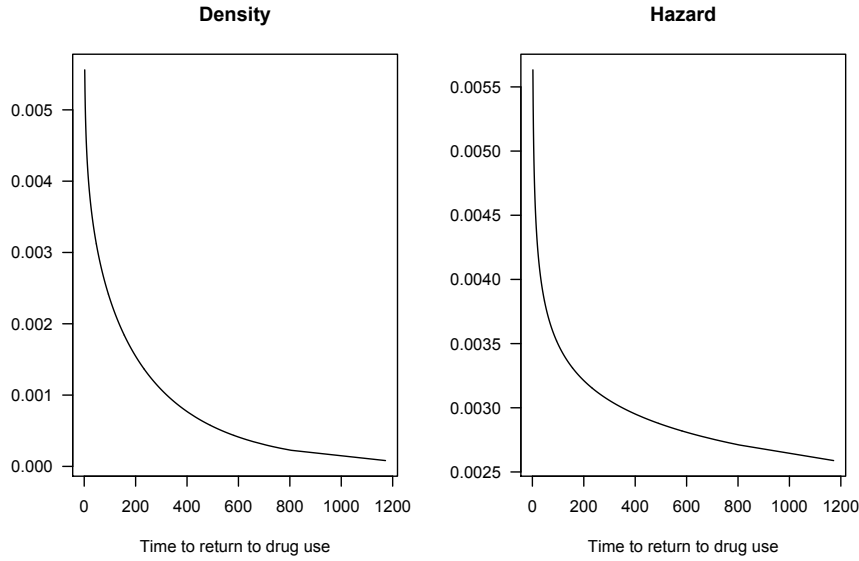


Figure 1.2.1: ML estimators of the density and hazard function for the UIS data based on the Weibull distribution.

Define

$$\theta^* = \arg \max_{\theta \in \Theta} \int \log(f(t, \theta)) dF(t).$$

We note that θ^* is the value of the parameter $\theta \in \Theta$ that minimizes the Kullback-Leibler distance measure between the true density $f(\cdot)$ and the parametric density $f(\cdot, \theta)$. If the parametric model is valid then there exists a unique parameter $\theta_0 \in \Theta$ such that $f(\cdot) = f(\cdot, \theta_0)$ and $\theta^* = \theta_0$. In such case, the MLE given above is $\sqrt{n}$ -consistent for θ^* and under some regularity assumptions, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0)),$$

where $I(\theta_0)$ is the Fisher information matrix. However, under right censoring, if the parametric model is misspecified, then the MLE is not consistent for θ^* and

converges to another value of θ different from θ^* . Some alternatives to the MLE under misspecification, have been investigated in the literature. We mention for example, Hjort (1992) who discussed the maximum weighted likelihood estimator and the modified maximum likelihood estimator that have better robustness properties under misspecification. Other interesting alternatives are the approximate maximum likelihood estimator (AMLE) investigated by Suzukawa *et al.* (2001) and the doubly robust estimator discussed by Yu and van der Laan (2006), among others. The next chapter of this thesis provides further discussions about the AMLE.

1.2.2 Nonparametric approach

The parametric approach described above is sometimes used in the estimation of the distribution of censored data. However, in many settings the exact distribution of the data is usually unknown and the consequences of misspecification of the chosen model are not trivial. Moreover, despite the importance of model specification and due to the considerable complexity added by the censoring mechanism, only few model misspecification and goodness-of-fit tests have been investigated in the literature. Consequently, using nonparametric methods is a common alternative in the current framework. The nonparametric techniques that are generally used in studies with right censored data are based on the well known nonparametric Kaplan-Meier (1958) estimator of the survival function $1 - F(\cdot)$. Also called the product limit estimator, the Kaplan-Meier estimator is defined by (in the absence of ties)

$$1 - \hat{F}(t) = \prod_{i: T_i \leq t} \left(1 - \frac{1}{\sum_{j=1}^n \mathbf{1}_{\{T_j \geq T_i\}}} \right)^{\delta_i}.$$

Figure 1.2.2, provides an illustrative example of the Kaplan-Meier estimator computed for the UIS data. $1 - \hat{F}(t)$ is a decreasing step function with jumps at the uncensored data. In the absence of censoring, the Kaplan-Meier estimator reduces to the empirical survival function $1 - F_n(t)$, where $F_n(t) = \sum_{i=1}^n I(Y_i \leq t)$ is the classical empirical distribution function. Asymptotic properties of the Kaplan-Meier estimator have been widely investigated in the literature, see for example Breslow and Crowley (1974), Peterson (1977), Burke *et al.* (1981), Gill (1983), Lo and Singh (1986), Lo *et al.* (1989) and Gijbels and Veraverbeke (1989), among others. In what follows, we give

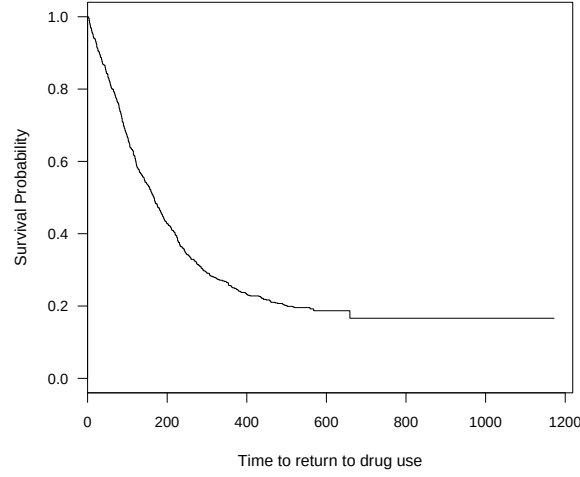


Figure 1.2.2: Estimation of the survival function for the UIS data, by means of Kaplan-Meier estimator.

some results that are useful for the next chapter of this dissertation. The first result provides an i.i.d. decomposition of $\hat{F}$. The proof of this Lemma is given in Lo *et al.* (1989).

Lemma 1.1. (Lo *et al.* (1989)). Let $\tau < \tau_H$, where $\tau_H = \sup\{t : H(t) < 1\}$ is the right endpoint of the distribution function $H(t) = P(T_1 \leq t)$. Also, let $H_1(t) = P(T_1 \leq t, \delta_1 = 1)$, then for $t \leq \tau$ we have

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

where $\xi_i(s) = -\int_{-\infty}^{T_i \wedge s} (1 - H(u))^{-2} dH_1(u) + (1 - H(T_i))^{-1} I\{T_i \leq s, \delta_i = 1\}$, $\sup_{s \leq \tau} |r_n(s)| = O_p(n^{-1} \log n)$ and $t \wedge s = \min(t, s)$.

The second result deals with the behavior of the modulus of continuity of the Kaplan-Meier estimator. The proof of this result can be found in Gijbels and Veraverbeke (1989).

Lemma 1.2. (Gijbels and Veraverbeke (1989)). Let $Y_{H(F^{-1})} = \sup\{t : H(F^{-1})(t) < 1\}$ and for $0 < p < Y_{H(F^{-1})}$ define ξ_p the p -th quantile of F . Assume f is bounded

in the neighborhood of ξ_p and $f(\xi_p) > 0$. Let a_n be a sequence of constants tending to zero with $a_n n^{1/2} (\log n)^{-1/2} > \Delta > 0$ for n sufficiently large. Then as $n \rightarrow \infty$, we have

$$\sup_{|\eta| \leq a_n} |[\widehat{F}(\xi_p + \eta) - \widehat{F}(\xi_p)] - [F(\xi_p + \eta) - F(\xi_p)]| = O(a_n^{1/2} n^{-1/2} (\log n)^{1/2}) \quad a.s.$$

Based on the Kaplan-Meier estimator, several nonparametric density estimators have been proposed in the literature. For example, Blum and Susarla (1980) proposed a kernel-type density estimator, Ramlau-Hansen (1980) investigated a nonparametric density estimator using the method of counting process and Tanner and Wong (1982) proposed another nonparametric density estimator based on Hajek's projection. Other kernel-type density and hazard rate estimators have been investigated by Singpurwalla and Wong (1983) and Padgett and McNichols (1984), among others. More recently, Berg and Politis (2007) investigated an infinite-order kernel density estimator and Bouezmarni *et al.* (2011) proposed new nonparametric estimators for the density and hazard rate based on the gamma kernels.

The nonparametric density estimator that we are currently investigating in this thesis is the kernel-type nonparametric estimator extended to censored data by Blum and Susarla (1980) and is defined as follows:

$$\widehat{f}(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) d\widehat{F}(s),$$

where K is a kernel function generally chosen to be a symmetric probability density function, $0 < h \equiv h_n$ is a bandwidth sequence and $\widehat{F}(\cdot)$ is the Kaplan-Meier estimator introduced above. A natural nonparametric estimator for the hazard function can be formed by dividing the kernel density estimator by the Kaplan-Meier estimator of the survival function $1 - F(\cdot)$:

$$\widehat{\lambda}(t) = \widehat{f}(t) / (1 - \widehat{F}(t)).$$

Figure 1.2.3, gives the kernel estimator for both density and hazard function for the UIS data, using the bandwidth $h = 663.63$ (selected by cross validation, see Chapter 2) and the Epanechnikov kernel.

The properties of the kernel density and hazard estimators defined above have been investigated by Blum and Susarla (1980), Földes *et al.* (1981), Tanner and

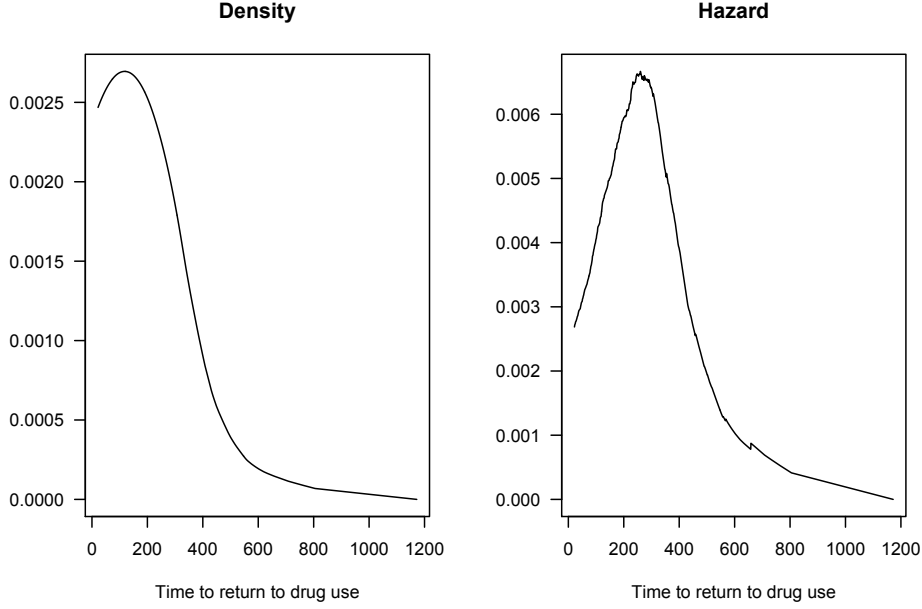


Figure 1.2.3: Estimation of the density and hazard function for the UIS data by means of the kernel estimator using Epanechnikov kernel function and $h = 663.63$.

Wong (1983), Padgett and McNichols (1984), Mielniczuk (1986), Lo *et al.* (1989), Xiang (1994) and Giné and Guillou (2001), among others. Under some regularity conditions, we have the following asymptotic normality result (see corollary 3.3 in Lo *et al.* (1989) for more details on the assumptions and the proof of this lemma):

Lemma 1.3. (*Lo et al. (1989)*). Let $\tau < \tau_H$, where $\tau_H = \sup\{t : H(t) < 1\}$ is the right endpoint of the distribution function $H(t) = P(T_1 \leq t)$ and suppose that the bandwidth sequence h satisfies $h \rightarrow 0$ and $(\log n)^2(nh)^{-1} \rightarrow 0$. Then, under some regularity assumptions, for $t \leq \tau$ we have

$$\sqrt{nh}(\hat{f}(t) - f(t) - b(t) + o(h^2)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(t)),$$

where $\sigma^2(t) = [f(t)/(1 - G(t))] \int K^2(u)du$ and $b(t) = \frac{f''(t)}{2}h^2 \int u^2 K(u)du$.

This result shows that the kernel estimator has a rate of convergence of $\sqrt{nh}$ which is slower compared with the $\sqrt{n}$ rate of convergence established in the parametric approach. It is worth noting that the asymptotic bias is the same as in the

uncensored data case, while the asymptotic variance depends on the censoring distribution $G(\cdot)$. In fact, increasing the censoring rate increases the asymptotic variance $\sigma^2(\cdot)$.

In practice, the choice of the bandwidth is a crucial issue in kernel-based density estimation; see Remark 2.3 in the next chapter for some guidelines on bandwidth selection under censorship.

1.2.3 Semi-parametric approach

Apart from the two extreme approaches discussed above, there have been other proposals for density estimators in the literature that combine the good properties of both the parametric and the nonparametric approach. These methods are quite different from the parametrically guided nonparametric method that we are currently investigating in this thesis, but can also achieve bias reduction compared to the fully nonparametric method. First of all, we find the projection pursuit density estimation developed by Friedman *et al.* (1984) for a multivariate density using a multiplicative correction. Hjort (1986) and Buckland (1992) introduced similar ideas using an estimated orthogonal expansion for the multiplicative correction factor. Hjort and Jones (1996) proposed a local parametric density estimator based on a local kernel smoothed likelihood function. This approach has a similar intention as the approach of Copas (1995) but is somehow more general. Another class of local likelihood methods has been discussed by Eguchi and Copas (1998). Efron and Tibshirani (1996) combined the maximum likelihood and the kernel estimator by putting an exponential family through a kernel estimator. Other semiparametric estimators involving an extra parameter have been proposed in the literature. For example, Olkin and Spiegelman (1987) and Faraway (1989) considered a convex combination of a parametric and a nonparametric estimate, and afterwards, Naito (2004) constructed a class of semi-parametric estimators using a local L_2 -fitting criterion to estimate the correction factor. Finally, more recently, Veraverbeke *et al.* (2014) discussed a parametrically pre-adjusted nonparametric density estimator. Unfortunately, except for the paper of Copas (1995) who adapted a local maximum likelihood estimator to censored data, none of the methods cited above has been considered so far in the context of censored

data. Another interesting method that was introduced in a semi-parametric fashion and extended to censored data by Kooperberg and Stone (1992), is the log-spline density estimator. The log-spline method models the logarithm of the density function using different types of splines and the maximum likelihood method to estimate the unknown coefficients. A comparison of the log-spline estimator with our proposed guided nonparametric estimator is performed in the next chapter based on simulation study as well as an application to the UIS data.

1.3 Models with one covariate

The methods discussed in the previous section focus on the statistical analysis of a single variable of interest. However, in engineering or biomedical applications the variable of interest Y may depend on some covariate that we denote by X . Modeling such dependency fits in a regression framework. Under right censoring, the response Y is only partially available. Instead of observing a sample (Y_i, X_i) , $i = 1, \dots, n$, from (Y, X) , one observes a random sample (T_i, δ_i, X_i) , $i = 1, \dots, n$, from (T, δ, X) , with $T = \min(Y, C)$, $\delta = \mathbf{I}(Y \leq C)$ and the censoring variable C is now supposed to be independent from the response Y given the covariate X . This condition allows censoring to depend on X , which is more realistic in practice. We define, respectively, by $F(t|x) = P(Y \leq t|X = x)$ and $G(t|x) = P(C \leq t|X = x)$ the conditional distribution function of Y and C given $X = x$. One way to analyze the relation between Y and X is via the estimation of the mean regression function

$$m(x) = E(\phi(Y)|X = x),$$

where ϕ is a known transformation introduced to include various functions of interest; see Chapter 3 for more details.

1.3.1 Parametric and semi-parametric models

Different regression models that can accommodate the problem of censoring have been investigated in the literature. In the linear regression, Buckley and James (1979) extended the least squares estimator to censored data. Their approach allows to

transform the censored data in an unbiased way. However, the computation of the transformed data requires an iterative scheme since the transformation depends on the unknown regression function m . The asymptotic properties of Buckley and James (1979) estimator have been investigated later by Lai and Ying (1991), among others. In the spirit of Buckley and James (1979), Koul *et al.* (1981) investigated another interesting unbiased transformation which does not involve the unknown regression function. Their procedure is known in the literature as the synthetic data approach. Several transformations have been proposed in the literature, see for instance Doksum and Yandell (1982), Zheng (1987), Leurgans (1987), Zhou (1992), Srinivasan and Zhou (1994) and Lai *et al.* (1995), among others. More recently, Ding and Nan (2011) investigated the sieve maximum likelihood estimation in the linear model for censored survival data. Finally, a considerable number of extensions to censored data of the nonlinear regression have been proposed in the literature. Some of these contributions are based on synthetic data, see for example Delecroix *et al.* (2008) and Heuchenne and Van Keilegom (2012), among others.

Basically, the need to develop such a procedure originates from the difference in terms of the conditional expectation between the observed variable T and the censored response Y . Koul *et al.* (1981) proposed to substitute the unknown response Y with a new variable Y^* satisfying

$$E(Y^*|X = x) = E(\phi(Y)|X = x). \quad (1.3.1)$$

Starting from the transformed data $(Y_i^*, X_i), i = 1, \dots, n$, one can now apply different statistical techniques to estimate $m(x)$. Under the independence between Y and C various transformations have been investigated. First, Koul *et al.* (1981) proposed a transformation that satisfies the condition (1.3.1) and depends only on the unconditional censoring distribution $G(t) = P(C \leq t)$ as follows:

$$Y_i^* = \frac{\delta_i \phi(T_i)}{1 - G(T_i^-)},$$

where $G(t^-) = P(C < t)$. In practice, the function $G(\cdot)$ is often unknown and needs to be estimated. Koul *et al.* (1981) proposed to substitute the unknown function $G(\cdot)$ with its Kaplan-Meier estimator $\hat{G}(\cdot)$. The estimated synthetic data are given by

$$\hat{Y}_i^* = \frac{\delta_i \phi(T_i)}{1 - \hat{G}(T_i^-)}.$$

Note that the Koul *et al.* (1981) transformation sets the censored observation to zero and multiplies the uncensored observation by a factor $1 - G(T_i^-)$, which can be very large if $G(T_i^-)$ is near to 1. Therefore, heavy censoring might affect the numerical stability of the transformation of Koul *et al.* (1981). Another appealing transformation satisfying also the condition (1.3.1) was introduced by Leurgans (1987) and is defined as

$$Y_i^* = \int \left(\frac{I(\phi(T_i) > t)}{1 - G(t)} - I(t < 0) \right) dt.$$

The presence of $I(\phi(T_i) > t)$ ensures that $1 - G(t) \neq 0$. Similarly, G needs to be estimated using Kaplan-Meier estimator $\hat{G}$, the resulting estimated synthetic data are now

$$\hat{Y}_i^* = \int \left(\frac{I(\phi(T_i) > t)}{1 - \hat{G}(t^-)} - I(t < 0) \right) dt.$$

Zheng (1987), suggested a more general class of transformations which includes the Leurgans transformation. His idea is based on the substitution of the censored response Y with the following variable

$$Y_i^* = \delta_i \phi_1(T_i) + (1 - \delta_i) \phi_2(T_i),$$

where the functions ϕ_1 and ϕ_2 satisfy the following expression

$$\int (1 - G(t^-)) \phi_1(t) dF(t|x) + \int (1 - F(t^-|x)) \phi_2(t) dG(t) = m(x). \quad (1.3.2)$$

Different propositions of functions ϕ_1 and ϕ_2 that satisfy the expression (1.3.2) have been developed in the literature, see for example Lai *et al.* (1995), among others. Later, the large sample behavior of the estimators proposed by Leurgans (1987) and Zheng (1987) has been studied by Zhou (1992), among many others. In the case of no censoring the synthetic response Y^* reduces to the response $\phi(Y)$. As stated before, the transformations introduced above have been proposed under the independence between Y and C , which happens for example when the censoring is due to the end of the study. However, this strong assumption is not always realistic in practice. Fortunately, most of these transformations can be easily extended to allow dependence between the censoring and the covariate X . The extension can be achieved by replacing the unconditional distribution function $G(\cdot)$ by the conditional distribution function $G(\cdot|x)$, and the Kaplan-Meier estimator by a particular nonparametric

estimator of the conditional distribution function $G(\cdot|x)$, see for example Fan and Gijbels (1994), El Ghouh and Van Keilegom (2008) and Lopez (2011), among others. However, when C and X are not independent, except for the Koul *et al.* (1981) transformation, the transformations cited above are more complicated. Due to its easy implementation, we will limit ourselves in this work to the transformation of Koul *et al.* (1981). Chapter 3 discusses in more detail the transformation of Koul *et al.* (1981) under independence of Y and C given X .

While linear and nonlinear models are customarily used in practice, their scope does not fulfill the expectations of several studies. Generalized linear models (Nelder and Wedderburn (1972)), expand the scope of such simple models to accommodate a wider class of regression problems. For example, in some applications, the conditional variance might depend on the target mean regression function, namely $Var(Y|X = x) = V(m(x))$ where V is a known positive function $V(\cdot)$, whereas in some other settings the response variable might be discrete. Furthermore, a crucial assumption under linear model is the normality of the errors which is not realistic in many applications. Generalized linear models allow to relax the normality assumption throughout the definition of a more general class of distributions, namely the exponential family defined as follows:

$$f(t|x) = \exp\left(\frac{\theta(x)t - b(\theta(x))}{a(\psi)} + c(t, \psi)\right),$$

where $f(t|x)$ is the conditional density of the response Y given $X = x$, $a(\cdot), b(\cdot)$ and $c(\cdot, \cdot)$ are known functions, θ is the canonical parameter and ψ is the dispersion parameter. The Binomial, Poisson, Normal, Gamma distributions are members of this family. In the parametric generalized linear models, the regression function $m(x)$ is modeled linearly via a link function $g(\cdot)$ as follows:

$$g(m(x)) = x'\beta.$$

Several applications of generalized linear models to survival data have been carried out; see for instance Hurley (1985), Hakulinen and Tenkanen (1987), Farrington (1996) and Wang *et al.* (2007), among others. The classical estimation procedure and statistical inferences about the generalized linear models parameters are based on the maximum likelihood estimation. However, in many statistical applications the

exact distribution of the response Y is unknown and the full likelihood is unavailable. In such situations, the quasi-likelihood estimation (Wedderburn (1974)) allows for more flexibility using only assumptions on the first two moments. Moreover, the corresponding quasi-likelihood function has similar statistical properties to the log-likelihood function. McCullagh and Nelder (1989) extended the quasi-likelihood method to parametric generalized linear models framework. The extension of quasi-likelihood to right censored data has been investigated by Yu *et al.* (2009), among others. In their work, Yu *et al.* (2009) adapted the parametric quasi-likelihood to censored data using a synthetic data approach. Recently, Yu (2011), Yu *et al.* (2012) and Yu and Peace (2012) proposed different semi-parametric quasi-likelihood estimators for right censored data.

Beside the class of generalized linear models and quasi-likelihood, two classes of models often discussed in the literature are the accelerated failure time model and the proportional hazards model. The accelerated failure time model (See by example, Kalbfleisch and Prentice (2002)) is widely used in survival analysis due to its easy interpretation. This class of models has different representations, the simplest one is the log-linear representation given by the following expression

$$\log(T) = \beta'X + \varepsilon,$$

where ε is an error term which follows a particular distribution.

An alternative approach that focuses directly on the hazard function has been proposed by Cox (1972), namely the proportional hazards model. In this class of models, one assumes that the covariate affects multiplicatively the hazard rate by a proportionate quantity at all observations. This condition is known in the literature as the proportional hazards assumption. Such models are expressed in terms of conditional hazard function as follows:

$$\lambda(t|x) = \lambda_0(t) \exp(\beta'X),$$

where $\lambda_0(t)$ is the baseline hazard which represent the hazard rate for a reference subject with zero covariate value. Assuming different parametric forms for the baseline hazard results on a different model in the parametric proportional hazards family. Parameters estimation for both accelerated failure time and parametric proportional

hazards models is performed using the maximum likelihood method. Using unspecified baseline hazard does not affect the proportionality assumption and leads to the famous semi-parametric Cox proportional hazards (PH) model. A partial likelihood approach has been developed by Cox (1972) to estimate the model parameters. Due to its flexibility, there have been a complete dominance of the Cox PH model in many statistical applications such as biostatistics and clinical trials. However, the proportional hazards assumption is often violated in scientific studies. Other semi-parametric models have been proposed in the literature, amongst them, the frailty model and the single index model. The frailty model generalizes the Cox model and allows to account for unmeasured covariate. For a wider discussion, see for example Vaupel *et al.* (1979), Nielsen *et al.* (1992), Legrand (2009) and Legrand *et al.* (2009), among others. The single index model allows to avoid the so-called curse of dimensionality in the multivariate nonparametric framework by assuming the existence of a vector of parameters θ_0 such that the response Y depends only on the projection $\theta_0'X$. The estimation of the single index model parameters and the choice of bandwidth are based on maximum likelihood method, see for example Strzalkowska-Kminak and Cao (2012) and González-Manteiga and Cadarso-Suarez (1993), among others.

1.3.2 Nonparametric estimation

The parametric modeling discussed above can provide a precise and accurate estimation when the chosen model is correct. However, in many situations, the functional form that describes the relation between a response and a covariate may be complex or unknown. Despite the fact that semi-parametric regression allows to relax the parametric restrictions, it still involves assumptions that can be restrictive in many settings. On the other hand, nonparametric regression offers more flexibility, by estimating the regression function directly from data without assuming any parametric form. For these reasons, using nonparametric regression is a common alternative in our framework.

In the fully observed data case, nonparametric smoothing techniques such as kernel, local linear, local polynomial and splines have been extensively investigated in the literature. For deeper understanding of the existing techniques see for example

Härdle (1990), Wahba (1990), Wand and Jones (1995), Fan and Gijbels (1996) and Simonoff (1996), among others. Moreover, the nonparametric regression techniques have been extended to the generalized linear models and quasi-likelihood by many authors. Among them, Green and Yandell (1985), O'Sullivan *et al.* (1986) investigated an extension of quasi-likelihood to smoothing splines. Staniswalis (1989), Fan *et al.* (1995) and Fan and Chen (1999) investigated a locally weighted maximum quasi-likelihood estimator by combining both quasi-likelihood estimation and local polynomial fitting. Chen *et al.* (2006) studied the local quasi-likelihood estimator when the response is missing at random. More recently, Wang *et al.* (2013) adapted the local quasi-likelihood to truncated and dependent data. Chapter 4 provides a more detailed and deeper study of the quasi-likelihood estimation.

For the censored data, the first study of regression problems in a fully nonparametric way was proposed by Beran (1981). In his paper, Beran (1981) extended the Kaplan-Meier estimator to the estimation of the conditional distribution of Y given the covariate X . The so-called conditional Kaplan-Meier estimator of Beran (1981) is defined by

$$\hat{F}(t|x) = 1 - \prod_{i=1}^n \left(1 - \frac{\delta_i \mathbf{1}_{\{T_i \leq t\}} W_{0i}(x, h_0)}{\sum_{j=1}^n \mathbf{1}_{\{T_i \leq T_j\}} W_{0j}(x, h_0)} \right), \quad (1.3.3)$$

for t less than $T_{(n)}$, the largest observation of the sample $(T_i)_{1 \leq i \leq n}$, $\{W_{0i}(x, h_0)\} \equiv \{W_{0i, K_0}(x, h_0)\}$ denotes a sequence of a suitable weights functions such as Gasser-Müller or Nadaraya-Watson weights, K_0 is a kernel function and $0 < h_0 \equiv h_{0,n}$ is a bandwidth parameter. Note that Beran's estimator reduces to the Kaplan-Meier estimator when all the weights $W_{0i}(x, h_0)$ are equal to $1/n$. Beran's estimator and related applications have been investigated by many authors, see for example Doksum and Yandell (1982), Dabrowska (1987), González-Manteiga and Cadarso-Suarez (1994), Van Keilegom and Veraverbeke (1997 b), Van Keilegom and Akritas (1999) and Du and Akritas (2002), among others. In the following lemma, we provide the rate of convergence as well as an i.i.d. representation of $\hat{F}(\cdot|x)$.

Lemma 1.4. *Let $\tau < \inf_x \tau_x$ with $\tau_x = \sup\{t : H(t|x) < 1\}$, i.e. the right endpoint of the support of $H(t|x) = P(T \leq t|X = x)$.*

1. We have

$$\sup_x \sup_{t \leq \tau_x} |\hat{F}(t^-|x) - F(t|x)| = O((nh_0)^{-1/2}(\log h_0^{-1}))^{1/2} \quad a.s.$$

2. For $t \leq \tau$, we have

$$\hat{F}(t^-|x) - F(t|x) = \sum_{j=1}^n W_{0j}(x, h_0) \eta_j(t^-, x) + O_p\left(\left(\frac{\log n}{nh_0}\right)^{3/4}\right),$$

where $\eta_j(t, x) = (1 - F(t|x)) \left[\frac{I(T_j \leq t, \delta_j = 1)}{1 - H(T_j|x)} - \int_0^t \frac{I(T_j \geq s)}{(1 - H(s|x))^2} dH_1(s|x) \right]$, and $H_1(t|x) = P(T \leq t, \delta = 1|X = x)$.

For more details about the assumptions and the proof of the first and the second point of Lemma 1.4, see Theorem 3.4 in Van Keilegom and Akritas (1999) and Theorem 3.2 in Du and Akritas (2002), respectively.

Considerable work has been done on nonparametric estimation of the conditional mean regression function $m(\cdot)$. First, Dabrowska (1987) and Zheng (1988) investigated nonparametric methods for univariate regression curve, among others. Later, Fan and Gijbels (1994) proposed a local linear estimator using a local version of the Buckley and James (1979) transformation. Amongst the investigations in multivariate local constant regression, Singh and Lu (1999) extended the Nadaraya-Watson estimator to censored data based on the Leurgans transformation. Several approaches have been proposed in the censored dependent data framework, including for example El Ghouch and Van Keilegom (2008) who investigated synthetic data in the local linear regression. In Chapter 3, we focus on the local linear regression technique discussed in Fan and Gijbels (1994). Under the independence of Y and C given X , the estimated transformed data can be defined by $\hat{Y}_i^* = \frac{\delta_i \phi(T_i)}{1 - \hat{G}(T_i^-|X_i)}$, where $\hat{G}(\cdot|\cdot)$ is Beran's estimator of the censoring conditional distribution function $G(\cdot|\cdot)$. Figure 1.3.1, provides the estimated transformed data for the UIS data, using the Epanechnikov kernel and a bandwidth of $h_0 = 5$.

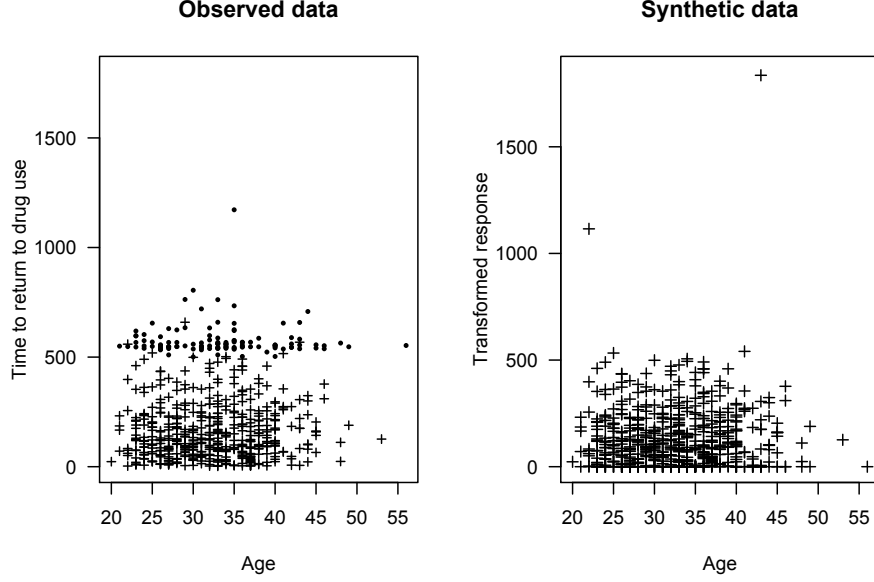


Figure 1.3.1: The observed data T and the estimated synthetic data $\hat{Y}^*$ for the UIS data, using the Koul *et al.* (1981) transformation, $h_0 = 5$ and Epanechnikov kernel.

Based on the synthetic sample $(\hat{Y}_i^*, X_i), i = 1, \dots, n$, the regression function $m(\cdot)$ can be estimated by minimizing the pseudo weighted mean squares criterion given by

$$\sum_{i=1}^n \{\hat{Y}_i^* - b_0 - b_1(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right),$$

where $0 < h \equiv h_n$ is a bandwidth and K is a kernel. The local linear estimator of m can be expressed as a weighted local average of the estimated transformed data as follows:

$$\hat{m}(x) = \hat{b}_0 = \sum_{i=1}^n W_i(x, h) \hat{Y}_i^*, \quad (1.3.4)$$

where

$$W_i(x, h) = \frac{s_{n,2}(x) - (X_i - x)s_{n,1}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)} K\left(\frac{X_i - x}{h}\right), \quad (1.3.5)$$

and

$$s_{n,l}(x) = \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) (X_i - x)^l, \quad l = 0, 1, 2. \quad (1.3.6)$$

The asymptotic properties of the estimator (1.3.4) are a special case of the asymptotic results given in Chapter 3, when one uses a linear guide. Fan and Gijbels (1994), proposed a data driven method to transform data and studied the asymptotic properties of the estimator (1.3.4) based on more general transformation. Note that the local linear estimator has many good statistical properties that are investigated in detail in the book of Fan and Gijbels (1996).

As discussed earlier in the context of kernel density estimation, the choice of the bandwidth parameters is also a crucial issue in nonparametric regression. Small bandwidths lead to an estimator with small bias and large variance. Large bandwidths lead to a small variance at the cost of increase in bias. Therefore, the bandwidths have to be chosen properly. A preview is given in Figure 1.3.2, in which the local linear estimation is applied to the UIS data, using two bandwidth values $h = 1$ and $h = 5$, $h_0 = 5$ and Epanechnikov kernel for both K_0 and K . It is obvious that the presence of censoring adds more complexity to the bandwidth selection problem. Remark 3.1 in Chapter 3 provides some guidelines on bandwidth selection in nonparametric regression with censored data.

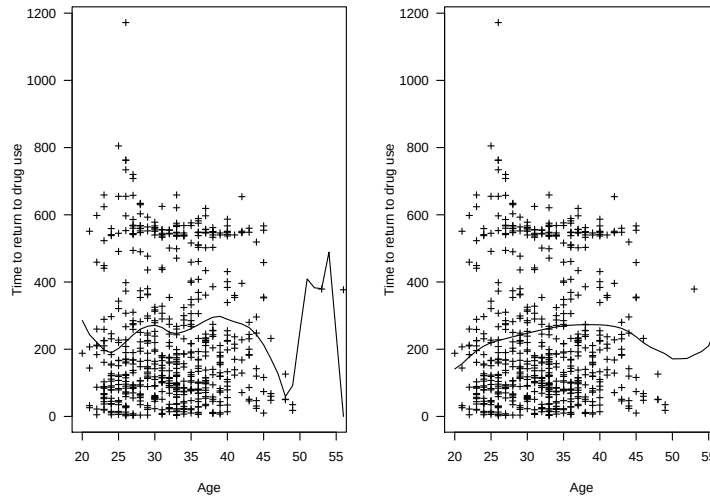


Figure 1.3.2: The estimation of regression function m for the UIS data, by means of local linear estimator, $h_0 = 5$ and two values for h : $h = 1$ on the left and $h = 5$ on the right.

1.4 Parametrically guided nonparametric estimation

As mentioned in the introduction, the parametrically guided nonparametric estimation under investigation in this work allows to decrease the bias of the nonparametric estimator without any increase in its variance, meaning that the well known bias-variance tradeoff can be bypassed. Let $\varphi(\cdot)$ be a target function that might be in our setting a density or a transformation of the mean regression function and $\varphi(x, \hat{\theta})$ is a proposed parametric estimator. As result, the estimator $\varphi(x, \hat{\theta})$ should capture some valuable information about the target function $\varphi(x)$. Consequently, the functions $\varphi(x) - \varphi(x, \hat{\theta})$ and $\varphi(x)/\varphi(x, \hat{\theta})$ are flatter and easier to estimate nonparametrically than the original function $\varphi(x)$.

There exist three different schemes for parametrically guided nonparametric estimation. The first scheme has been developed in the context of kernel density estimation by Hjort and Glad (1995) using the following multiplicative correction

$$\varphi(x) = \varphi(x, \hat{\theta}) r_{\hat{\theta}}(x),$$

where $r_{\hat{\theta}}(x) = \varphi(x)/\varphi(x, \hat{\theta})$. The parametrically guided nonparametric estimator is given by

$$\hat{\varphi}_{\hat{\theta}}(x) = \varphi(x, \hat{\theta}) \hat{r}_{\hat{\theta}}(x),$$

where $\hat{r}_{\hat{\theta}}(\cdot)$ is a nonparametric estimator of the correction factor $r_{\hat{\theta}}(\cdot)$. The multiplicatively corrected estimator achieves bias reduction relative to the unguided nonparametric estimator. However, the multiplicative scheme is not suitable in situations when the parametric part is not strictly bounded away from zero.

In the second scheme, the correction is carried out in an additive rather than a multiplicative scale. Such type of correction has been introduced in the kernel regression by Rahman and Ullah (2002) and used later in different frameworks due to its simplicity. The basic idea is based on the additive identity

$$\varphi(x) = \varphi(x, \hat{\theta}) + r_{\hat{\theta}}(x),$$

where $r_{\hat{\theta}}(x) = \varphi(x) - \varphi(x, \hat{\theta})$. Similarly, a parametrically guided nonparametric estimator can be defined by

$$\hat{\varphi}_{\hat{\theta}}(x) = \varphi(x, \hat{\theta}) + \hat{r}_{\hat{\theta}}(x).$$

Finally, the last guided scheme (see Martins-Filho *et al.* (2008)) combines both the additive and the multiplicative scheme in an unified family of corrections indexed by a new parameter α as follows:

$$\varphi(x) = \varphi(x, \hat{\theta}) + r_{\hat{\theta}, \alpha}(x) \varphi(x, \hat{\theta})^{\alpha},$$

where $r_{\hat{\theta}, \alpha}(x) = (\varphi(x) - \varphi(x, \hat{\theta})) / \varphi(x, \hat{\theta})^{\alpha}$ and $\alpha \in \mathbb{R}$. The generalized guided nonparametric estimator is now given by

$$\hat{\varphi}_{\hat{\theta}}(x) = \varphi(x, \hat{\theta}) + \hat{r}_{\hat{\theta}, \alpha}(x) \varphi(x, \hat{\theta})^{\alpha},$$

where $\hat{r}_{\hat{\theta}, \alpha}(x)$ is a nonparametric fit. Note that the unified family is more general, using $\alpha = 0, 1$ gives the additive and the multiplicative correction, respectively.

1.5 Brief review of the chapters

This dissertation is devoted to extending the parametrically guided nonparametric techniques to censored data framework. The dissertation is organized as follows.

In Chapter 2, we propose a parametrically guided kernel density and hazard estimators for censored data using a multiplicative correction. We investigate the asymptotic properties as well as the performance of our new estimators. We also perform a comparison with log-spline estimators using both simulation studies and an application to the UIS data. Moreover, we develop confidence intervals and provide an extension to the Cox model. The contents of Chapter 2 is mainly based on Talamakrouni *et al.* (2016a).

In Chapter 3, we adapt the parametrically guided local linear regression to censored data using synthetic data. First we use the additive correction, then we generalize our result to the unified family of corrections. We prove the asymptotic normality of the proposed estimator with $\sqrt{nh}$ rate of convergence and we confirm our theoretical results using simulation studies and an application to real data. The contents of Chapter 3 is based on Talamakrouni *et al.* (2015).

Finally, in Chapter 4, we investigate the parametrically guided local quasi-likelihood estimation with censored data. To deal with censoring, we also consider the synthetic data approach and the additive correction. Then, we generalize the asymptotic results to the guided scheme based on the unified family of corrections. Finally, we investigate simulation studies using both discrete and continuous responses and perform an application to real data. Chapter 4 is based on Talamakrouni *et al.* (2016b)

Chapter 2

Guided kernel density and hazard estimation ¹

2.1 Introduction

The estimation of the probability density and hazard function has received considerable attention in the literature, as it allows to visualize and explore the distribution of data. In this chapter we wish to estimate the density and hazard function when the data are subject to right censoring, by using a hybrid estimation method that has at the same time nonparametric and parametric ingredients.

As discussed in Chapter 1, when data are not subject to censoring, there is a large variety of approaches to estimate the density and the hazard function that are either semiparametric or that use aspects from both the nonparametric and the parametric school, and that are hence situated in between these two extreme approaches. One of these approaches is the parametrically guided nonparametric estimator proposed by Hjort and Glad (1995). Apart from reducing the bias compared to the classical kernel approach, the parametrically guided nonparametric approach allows for a theoretically unbiased estimator. In fact, when using a correct guide the asymptotic bias of the parametrically guided kernel estimator reduces to zero, which is impossible with

¹This chapter is based on Talamakrouni, M., Van Keilegom, I. and El Ghouch, A. (2016). Parametrically guided nonparametric density and hazard estimation with censored data. *Comput. Statist. Data Anal.* **93**, 308-323.

the classical kernel approach.

The basic idea of this approach is to start with any parametric density estimator and then to adjust this first stage parametric approximation using a nonparametric kernel-type estimator of a particular correction factor. Hjort and Glad (1995) defined the parametrically guided nonparametric estimator by

$$\hat{f}_{\hat{\theta}}(t) = f_{\hat{\theta}}(t)\hat{r}_{\hat{\theta}}(t), \quad (2.1.1)$$

where $\hat{r}_{\hat{\theta}}(\cdot)$ is a kernel-type nonparametric estimator of the correction factor

$$r_{\hat{\theta}}(\cdot) = \frac{f(t)}{f_{\hat{\theta}}(t)},$$

$f_{\hat{\theta}}(t)$, is a first stage parametric density approximation and $\hat{\theta}$ is an estimator of the least false value θ^* according to a certain distance measure between $f(\cdot)$ and $f_{\theta}(\cdot)$ (see Assumption 2.3, below).

Essentially, this multiplicative correction does not affect the variance but can reduce the bias. The intuitive idea behind this approach is that if the parametric estimator $f_{\hat{\theta}}(\cdot)$ is close to the true density $f(\cdot)$, the multiplicative correction function $r_{\hat{\theta}}(\cdot)$ will be smoother than the true density $f(\cdot)$ and therefore simpler to estimate using kernel smoothing, resulting in an improved $\hat{f}_{\hat{\theta}}(\cdot)$ compared to the traditional kernel estimator. If the true density is far from the parametric estimator, then there is not much loss in accuracy for the parametrically guided nonparametric estimator.

The aim of this chapter is to extend their method to the case of random right-censored data. In addition to studying the estimation of the density function, we also propose and study a parametrically guided nonparametric estimator of the hazard rate function in the presence of censoring. We establish the asymptotic properties of the proposed estimators and apply the proposed method to data on the time to return to drug use from a study of the AIDS research unit of the University of Massachusetts. The proofs are collected in the Appendix D.

2.2 Methodology

Let Y be a variable of interest with density f and distribution function F , and let C be a censoring variable with continuous distribution function G . We assume throughout

this chapter that Y is independent of C . Under random right censoring, the variable Y is not completely observed. One can only observe (T, δ) , where $T = \min(Y, C)$, $\delta = \mathbf{I}(Y \leq C)$. Our first objective is to estimate the probability density function f using the observed i.i.d sample $(T_i, \delta_i), i = 1, \dots, n$ of (T, δ) .

As discussed in Chapter 1, the kernel-based density and hazard function estimators that accommodate for censoring are given by

$$\hat{f}(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) d\hat{F}(s), \quad (2.2.1)$$

and

$$\hat{\lambda}(t) = \hat{f}(t)/(1 - \hat{F}(t)),$$

where K is a kernel function, $0 < h \equiv h_n$ is a bandwidth and $\hat{F}(t)$ is the Kaplan-Meier estimator defined by (in the absence of ties)

$$\hat{F}(t) = 1 - \prod_{i: T_i \leq t} \left(1 - \frac{1}{\sum_{j=1}^n \mathbf{1}_{\{T_j \geq T_i\}}}\right)^{\delta_i}. \quad (2.2.2)$$

Note that, these estimators defined above are by construction completely nonparametric. In the uncensored data context, Hjort and Glad (1995) proposed a parametrically guided kernel density estimator (PGK) as an alternative to the traditional kernel density estimator (TK). As argued in the introduction, the PGK estimator combines the advantages of both parametric and nonparametric approaches and includes a prior information that allows the bias reduction of the PGK estimator compared to the TK estimator.

For censored data, we propose to multiply the first stage parametric estimator $f_{\hat{\theta}}(t)$ in expression (2.1.1) with the following nonparametric kernel-type estimator of the correction function $r_{\hat{\theta}}(t)$ adapted to censored data:

$$\hat{r}_{\hat{\theta}}(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) \frac{1}{f_{\hat{\theta}}(s)} d\hat{F}(s).$$

The ensuing PGK density estimator is

$$\hat{f}_{\hat{\theta}}(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) \frac{f_{\hat{\theta}}(t)}{f_{\hat{\theta}}(s)} d\hat{F}(s)$$

$$= \frac{1}{h} \sum_{i=1}^n K\left(\frac{t - T_i}{h}\right) \frac{f_{\hat{\theta}}(t)}{f_{\hat{\theta}}(T_i)} W_i, \quad (2.2.3)$$

where W_i is the size of the jump of $\hat{F}$ at T_i .

Note that when the data are completely observed, the weights $W_{(i)}$ are all equal to $1/n$ and the PGK estimator given above reduces to the estimator defined by Hjort and Glad (1995). Naturally the PGK estimator that we propose for the hazard function $\lambda(\cdot)$ is

$$\hat{\lambda}_{\hat{\theta}}(t) = \hat{f}_{\hat{\theta}}(t) / (1 - \hat{F}(t)).$$

As we will see in the following section, the multiplicative correction used in the PGK density and hazard function estimators does not affect the variance but can reduce the bias compared to the traditional kernel estimators defined above.

2.3 Asymptotic results

This section is devoted to the development of the asymptotic normality of the PGK estimators $\hat{f}_{\hat{\theta}}(\cdot)$ and $\hat{\lambda}_{\hat{\theta}}(\cdot)$. For the PGK density estimator, we split the problem into two parts. First, we establish in Theorem 2.1 the asymptotic normality of $\hat{f}_*(\cdot)$, an estimator of $f(\cdot)$ based on a given non-random guide $f_*(\cdot)$. Then, in Theorem 2.2, we extend this result to the case of a data-driven guide. Finally, in Theorem 2.3 we prove the asymptotic normality of the PGK estimator of the hazard function $\hat{\lambda}_{\hat{\theta}}(\cdot)$.

As stated in the previous section, under random right censoring the PGK estimator depends on the Kaplan-Meier estimator $\hat{F}$, which is defined as a product (see expression (2.2.2)). This adds some extra complexity to the PGK estimation approach compared to the uncensored case when developing asymptotic results. To circumvent these technical difficulties we mainly use the asymptotic i.i.d. representation of the Kaplan-Meier estimator investigated in Lo *et al.* (1989).

Let $\tau < \tau_H$, where $\tau_H = \sup\{t : H(t) < 1\}$ is the right endpoint of the distribution function $H(t) = P(T \leq t)$. Also, let $H_1(t) = P(T \leq t, \delta = 1)$, and define $\mu_K^2 = \int u^2 K(u) du$. The kernel function $K : \mathfrak{R} \rightarrow \mathfrak{R}$, the bandwidth h and the density $f(\cdot)$ are assumed to satisfy the following conditions for a fixed value $t \leq \tau$.

Assumption 2.1.

- (A.1) *The kernel K is a symmetric, continuously differentiable probability density function with compact support $[-1, 1]$.*
- (A.2) *The bandwidth sequence h satisfies $h \rightarrow 0$ and $nh^2(\log n)^{-2} \rightarrow \infty$.*
- (A.3) *f is twice continuously differentiable in a neighborhood of t and $f(t) > 0$.*

2.3.1 The estimator with a fixed guide

Let $f_*(t)$ be a non-random density function that approximates $f(t)$, and let $\hat{f}_*(t)$ be the corresponding PGK estimator defined as

$$\hat{f}_*(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) \frac{f_*(t)}{f_*(s)} d\hat{F}(s). \quad (2.3.1)$$

In the next section we will replace $f_*(\cdot)$ by the best approximation of $f(\cdot)$ within a certain parametric class, but for the time being $f_*(\cdot)$ can be any deterministic density.

Note that if $f_*(t)$ would be a uniform density, then $\hat{f}_*(t)$ reduces to the TK estimator, which means that the PGK estimator is a generalization of the TK estimator. The following additional conditions are required for a fixed point $t \leq \tau$ at which we want to estimate the density.

Assumption 2.2.

- (B.1) *The density $f_*(\cdot)$ is twice continuously differentiable in a neighborhood of t .*
- (B.2) *The density $f_*(\cdot)$ satisfies $f_*(t) > 0$.*

The following theorem provides the asymptotic distribution of the PGK estimator $\hat{f}_*(\cdot)$ using a non-random guide.

Theorem 2.1. *Suppose Assumptions 2.1 and 2.2 hold and let $t \leq \tau$.*

1. *Then,*

$$\begin{aligned} \hat{f}_*(t) - f(t) &= \frac{1}{nh} \sum_{i=1}^n U_{in}(t) + \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) \frac{f_*(t)}{f_*(s)} dF(s) - f(t) \\ &\quad + O_p(n^{-1/2}), \end{aligned}$$

where

$$U_{in}(t) = \int_{-1}^1 \xi_i(t - uh) K'(u) du,$$

$$\xi_i(t) = \int_{-\infty}^{T_i \wedge t} \frac{dH_1(s)}{(1 - H(s))^2} + \frac{I\{T_i \leq t, \delta_i = 1\}}{1 - H(T_i)}.$$

2. Moreover,

$$\sqrt{nh} \left(\hat{f}_*(t) - f(t) - B_*(t) + o(h^2) \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(t)), \quad (2.3.2)$$

where

$$B_*(t) = \frac{1}{2} h^2 \mu_K^2 r_*''(t) f_*(t),$$

$$r_*(t) = f(t)/f_*(t) \text{ and } \sigma^2(t) = [f(t)/(1 - G(t))] \int_{-1}^1 K^2(u) du.$$

Note that the choice of the guide has an obvious impact on the expression of the asymptotic bias $B_*(t)$, whilst the variance $\sigma^2(t)$ is invariant under this choice and is the same as for the TK estimator.

Remark 2.1. We note that when the parametric guide and the true density are equal in a small neighborhood around t , then $B_*(t) = 0$; see expression (2.9.4) in the Appendix D.

2.3.2 The estimator with an estimated guide

In this section, we investigate the situation where the guide is derived from the data by a first stage estimation procedure. We consider a possibly misspecified parametric model $\{f_\theta(\cdot) : \theta \in \Theta\}$ and assume that there exists an estimator $\hat{\theta}$ that converges in probability to a finite limit θ^* . We need the following additional conditions for a fixed point $t \leq \tau$.

Assumption 2.3.

(C.1) The parametric density function f_θ belongs to a parametrically indexed class defined by the following characteristics:

1. $\theta \in \Theta$, where Θ is a compact subset of $\mathbb{R}^p$.

2. The function $(t, \theta) \mapsto f_\theta(t)$ is twice continuously differentiable with respect to t and the components of θ in a neighborhood of t and θ^* .

(C.2) The parameter $\theta^* \in \Theta$ satisfies the following conditions:

1. $\hat{\theta} - \theta^* = O_p(n^{-1/2})$.
2. The density $f_{\theta^*}(\cdot)$ satisfies $f_{\theta^*}(t) > 0$.

In order to be as general as possible, in this chapter, we don't restrict ourselves to a particular parametric estimation procedure. However, to illustrate the idea and give an example of an estimator that satisfies the conditions above, especially assumption (C.2.1), we discuss now the case of the maximum likelihood estimator (MLE). Define

$$\theta^* = \arg \max_{\theta \in \Theta} \int_{-\infty}^{+\infty} \log(f_\theta(t)) dF(t).$$

This is the minimizer of $KL(f, f_\theta) = \int \log \frac{f(t)}{f_\theta(t)} dF(t)$, the Kullback-Leibler distance measure between the true density f and the parametric density model f_θ . If the parametric model is correct, i.e. if there exists a $\theta_0 \in \Theta$ such that $f(\cdot) = f_{\theta_0}(\cdot)$, then $\theta^* = \theta_0$. In the uncensored case, it is well known that the usual MLE given by $\hat{\theta} = \arg \max_{\theta \in \Theta} n^{-1} \sum_i \log(f_\theta(T_i))$ is consistent for θ^* , even under misspecification. In the censored case, the analogue of $\hat{\theta}$ is the approximate maximum likelihood estimator (AMLE) proposed by Oakes (1986) and defined as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \int_{-\infty}^{+\infty} \log(f_\theta(t)) d\hat{F}(t),$$

where $\hat{F}(\cdot)$ is the Kaplan-Meier estimator. Note that, in the uncensored case, the Kaplan-Meier estimator coincides with the empirical distribution function and therefore the AMLE reduces to the MLE. The properties of the AMLE estimator have been investigated by Suzukawa *et al.* (2001). Assuming that $\tau_F := \sup\{t : F(t) < 1\} \leq \tau_G := \sup\{t : G(t) < 1\}$ and under certain regularity assumptions, the authors prove that $\hat{\theta}$ is $\sqrt{n}$ -consistent.

Remark 2.2. *Even if the data are censored, the usual MLE estimator $\hat{\theta}$ can still be used. However, under misspecification, this estimator does not converge to θ^* but to another quantity; see Suzukawa *et al.* (2001) for more details.*

The following theorem is the most important result of this chapter. It establishes that the PGK estimator with an estimated guide $f_{\hat{\theta}}(t)$ is asymptotically equivalent to the PGK estimator with the fixed guide $f_{\theta^*}(t)$.

Theorem 2.2. *Suppose Assumptions 2.1 and 2.3 hold and let $t \leq \tau$. Then,*

$$\sqrt{nh} \left(\hat{f}_{\hat{\theta}}(t) - f(t) - B_{\theta^*}(t) + o(h^2) \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(t)\right),$$

where $B_{\theta^*}(t) = \frac{1}{2}h^2\mu_K^2 r_{\theta^*}''(t)f_{\theta^*}(t)$ and $r_{\theta^*}(t) = f(t)/f_{\theta^*}(t)$.

First, notice that the expression of the asymptotic variance is independent of the parametric estimating procedure and is equal to that of the TK estimator. As revealed in the previous section by Theorem 2.1, the main difference between the behavior of the PGK estimator and the TK estimator appears in the term of the asymptotic bias $B_{\theta^*}(t)$, which depends on the parametric guide. Remind that, if the parametric guide is the uniform density, then the PGK estimator becomes the traditional kernel density estimator and $B_{\theta^*}(t)$ coincides with $B(t) = \frac{1}{2}h^2\mu_K^2 f''(t)$, the asymptotic bias of the TK estimator; see, for example, Lo *et al.* (1989). So, with an appropriate choice of the guide, i.e. when $|r_{\theta^*}''(t)f_{\theta^*}(t)| \leq |f''(t)|$, the bias of the PGK estimator will be reduced in absolute value compared to that of the TK estimator, whilst the variance remains unchanged. If the parametric density is a good guess, then the correction function $r_{\theta^*}(\cdot)$ will be nearly constant and its second derivative $r_{\theta^*}''(\cdot)$ should be very small. In this case the bias reduction will be attained. Finally, in the ideal case when the parametric guide coincides with the true density we have that $B_{\theta^*}(t) = 0$. In such a case, the PGK estimator is unbiased and one can choose an arbitrarily large bandwidth to reduce the variance to its minimal possible value, which can never be achieved by the TK estimator even with an optimal bandwidth.

Remark 2.3. *In practice, the choice of the bandwidth is a crucial issue in kernel-based density estimation. The theoretical optimal bandwidth that minimizes the asymptotic mean integrated squared error (MISE) criterion is given by*

$$h_{opt} = \left(\frac{\int \sigma^2(t)dt}{\mu_K^4 \int (r_{\theta^*}''(t)f_{\theta^*}(t))^2 dt} \right)^{1/5} n^{-1/5}.$$

This expression can hardly be used in practice, since it depends on many unknown components. To select the bandwidth h in our case, one can use for example the least

squares cross validation method or the bootstrap method discussed in Sánchez-Sellero *et al.* (1999). In our data analysis, see Section 2.5, we adopt the cross validation method as discussed in Wang and Wang (2007). We also point out that, the PGK method will work even if the parametric guide is not optimally chosen. However, an optimal choice of the parametric guide will improve the quality of the PGK estimator. One can for example use goodness-of-fit tests to choose the parametric guide; see for example Castro-Kuriss (2011).

Remark 2.4. As noticed by Hjort and Glad (1994, 1995), our new PGK estimator does not integrate to one. In the uncensored data case, a number of proposals have been investigated in the literature to overcome this problem. Amongst them, Hjort and Glad (1994) discussed a possible correction by dividing the PGK estimator by its integral. A corrected PGK estimator can be defined in our case by

$$\widehat{f}_{\widehat{\theta}}^c(t) = \frac{\widehat{f}_{\widehat{\theta}}(t)}{\int_0^\tau \widehat{f}_{\widehat{\theta}}(s)ds}.$$

However, as noticed by Hjort and Glad (1994), this correction does not lead to better performance in terms of MISE. Glad *et al.* (2003) proposed a promising correction that improves the MISE by subtracting a constant such that the integral of the corrected estimator equals 1. Following their proposal, one may correct the PGK estimator in our case as follows:

$$\widehat{f}_{\widehat{\theta}}^c(t) = \max\{0, \widehat{f}_{\widehat{\theta}}(t) - \xi\},$$

where ξ is chosen such that $\int_0^\tau \widehat{f}_{\widehat{\theta}}^c(t)dt = 1$.

2.3.3 Guided kernel hazard estimator

The hazard function $\lambda(\cdot)$ has been extensively studied in the literature. The estimation by means of kernel methods has been investigated by Gefeller and Dette (1992), Gefeller and Michels (1992), Patil (1993), Müller and Wang (1994) and González-Manteiga *et al.* (1996), among others. The PGK estimator for the hazard function $\lambda(\cdot)$ that we proposed in the previous section is

$$\widehat{\lambda}_{\widehat{\theta}}(t) = \widehat{f}_{\widehat{\theta}}(t)/[1 - \widehat{F}(t)],$$

where $\hat{f}_\theta(t)$ is the PGK density estimator given in (2.2.3) and $\hat{F}(t)$ is the Kaplan-Meier estimator. Note that one can also replace the Kaplan-Meier estimator by a parametrically guided nonparametric version of the distribution function F or by any other estimator that has parametric and nonparametric ingredients (see e.g. Veraverbeke *et al.* (2014), Section 7.1, for an overview of possible estimators). However, given that the rate of convergence of the estimator of $F(t)$ will always be faster than the rate of convergence of the density estimator $\hat{f}_\theta(t)$, the choice of the estimator of $F(t)$ has no impact on the asymptotic distribution of the estimator of $\lambda(t)$. For simplicity we therefore estimate $F(t)$ by the Kaplan-Meier estimator $\hat{F}(t)$.

The following theorem deals with the asymptotic normality of the PGK hazard rate estimator.

Theorem 2.3. *Suppose Assumptions 2.1 and 2.3 hold and let $t \leq \tau$. Then,*

$$\sqrt{nh} \left(\hat{\lambda}_\theta(t) - \lambda(t) - \beta_{\theta^*}(t) + o(h^2) \right) \xrightarrow{d} \mathcal{N}(0, \tau^2(t)),$$

where

$$\beta_{\theta^*}(t) = \frac{1}{2} h^2 \mu_K^2 r_{\theta^*}''(t) f_{\theta^*}(t) / [1 - F(t)],$$

$$\text{and } \tau^2(t) = [\lambda(t)/(1 - H(t))] \int K^2(u) du.$$

As for the density estimator, the asymptotic bias of the PGK hazard rate estimator depends on the parametric guide, while the asymptotic variance remains unchanged compared to the TK hazard estimator.

2.4 Simulation results

In this section we evaluate the finite sample performance of the PGK estimator by means of Monte Carlo simulations. To check the theoretical results and compare the PGK estimator with the TK estimator we investigate two examples. In the first example we study the classical class of normal mixture densities of Marron and Wand (1992), and in the second example we investigate the Weibull density and perform a comparison with the logspline approach of Kooperberg and Stone (1991, 1992). The logspline method estimates $\log(f)$ by a function from a space of cubic splines that

are linear in the tails and have a finite number of knots. Along the simulations we consider the Epanechnikov kernel function K , and, for every estimator, we only show the results corresponding to the optimal tuning parameters, i.e. those which minimize the empirical mean squared error (MSE).

2.4.1 Normal mixture model

The class of normal mixture densities of Marron and Wand (1992) includes fifteen densities that cover a broad variety of shapes. In the context of uncensored guided density estimation, this class was investigated by Hjort and Glad (1994) and Naito (2004), among others. We studied all the fifteen densities but for the sake of brevity we only show the results of the following ones: the normal density, #1, the bimodal density, #6, the separated bimodal density, #7, the asymmetric bimodal density, #8, the trimodal density, #9, the claw density, #10, the double claw density, #11, the asymmetric double claw density, #13, and the smooth comb density, #14. See Figure 2.4.1 for a plot of all these densities. In each case, independent and identically distributed variables Y_i , $i = 1, \dots, 400$, are drawn. Independently, we drawn the censoring variables C_i from the same distribution. This leads to 50% rate of censoring. As for the parametric guide, we consider a standard normal density whose parameters are estimated by maximum likelihood. So the first case (#1) is the only situation where the guide is correctly specified. We compute the PGK and the TK density estimators at $t = 0$ taking 100 equally spaced bandwidths over the interval $[0.01, 4]$. The squared bias ($\text{Bias}^2 \times 10^4$), the variance ($\text{Var} \times 10^4$) and the empirical mean squared error ($\text{MSE} \times 10^4$) of each estimator were computed as well as the ratio of the empirical MSEs ($\text{MSE}_{PGK}/\text{MSE}_{TK}$). Table 2.4.1 provides the results with 1000 replications for cases #1, #6, #7, #8, #9, #10, #11, #13, and #14. As already mentioned, #1 corresponds to the case where the guide is a good guess of the true density. As expected, in this case, the bias of the PGK estimator is almost zero and is substantially reduced compared to that of the TK estimator. For cases #8, #9 and #11 the PGK estimator is significantly better than the TK estimator. For case #10 the MSE becomes very large and reveals unstable behavior for both estimators. In addition, although the MSE is not greatly enhanced, the bias and the MSE of the

PGK estimator remain the smallest. Cases #7 and #14 show a quite similar behavior of the PGK and the TK estimators. However, even if the bias is still reduced, the TK estimator beats the PGK estimator in terms of the MSE, for cases #6 and #13. Finally, we point out that the selected bandwidths for both competitors were the same for most of the non-normal situations.

Figure 2.4.2 shows how the MSE changes with h for three cases: #1, #10 and #13. As expected, in case #1, where the guide is correct, the bias is almost zero. In this situation increasing the bandwidth reduces the MSE of the PGK estimator compared to that of the TK estimator. For case #10 the parametric guide is well specified and the MSE of the PGK estimator is significantly reduced especially for large bandwidths. For case #13 the parametric guide is completely misspecified and the MSE of the PGK estimator is now larger than that of the TK estimator for most choices of h . Finally, in all cases, similar behavior of both competitors is observed for small values of h . This is not surprising since with a small bandwidth only the T_i 's quite close to t are used and the ratio $f_{\hat{\theta}}(t)/f_{\hat{\theta}}(T_i)$ defined in expression (2.2.3) is close to 1 in that case.

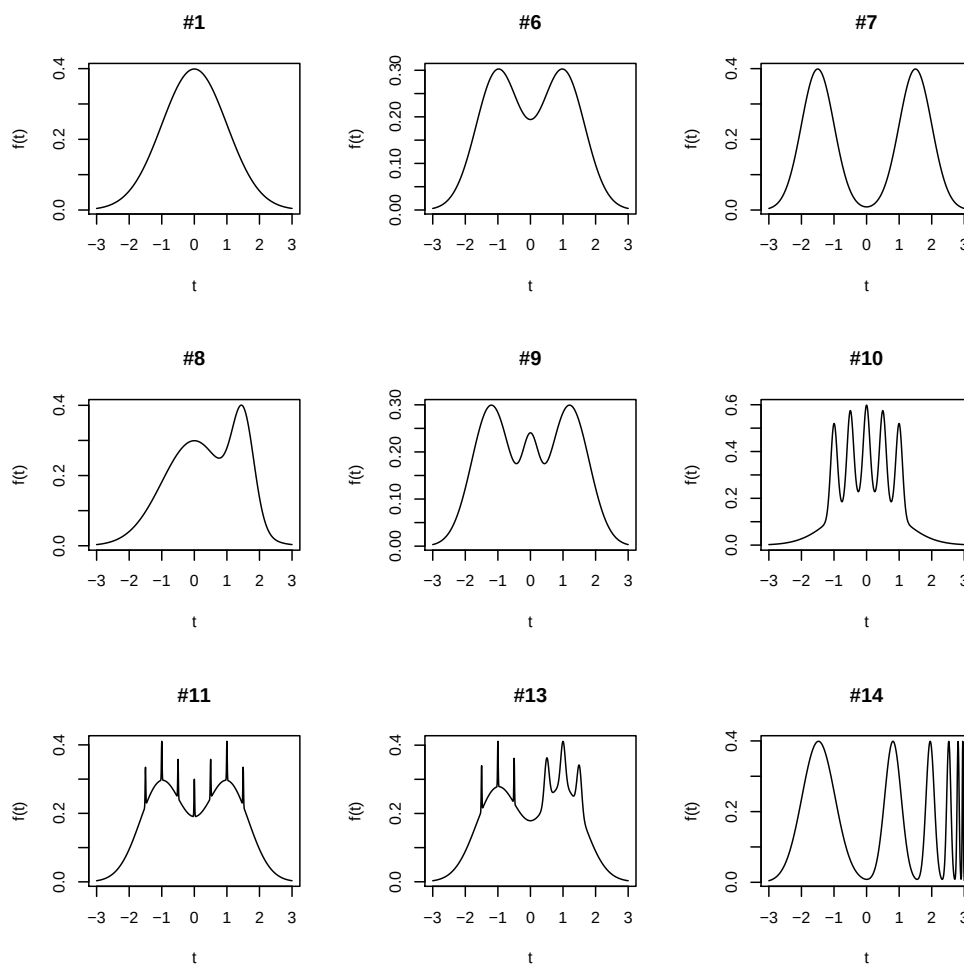


Figure 2.4.1: Standard normal and normal mixture target densities.

	Method	Bias ²	Var	MSE	h	MSE-ratio
#1	PGK	0.015	4.376	4.391	0.332	0.356
	TK	3.764	8.556	12.32	0.736	
#6	PGK	0.003	2.352	2.355	3.234	6.578
	TK	0.008	0.350	0.358	1.582	
#7	PGK	0.207	0.866	1.073	0.171	1.016
	TK	0.196	0.860	1.056	0.171	
#8	PGK	0.008	3.067	3.075	3.435	0.392
	TK	2.467	5.372	7.839	0.777	
#9	PGK	0.001	2.663	2.664	2.307	0.574
	TK	3.479	1.161	4.640	0.856	
#10	PGK	18.72	135.3	154.0	0.050	0.996
	TK	19.40	135.1	154.5	0.050	
#11	PGK	0.006	2.977	2.983	0.775	0.075
	TK	37.99	1.590	39.58	0.735	
#13	PGK	0.001	1.563	1.564	3.476	4.933
	TK	0.012	0.305	0.317	1.783	
#14	PGK	0.381	1.210	1.591	0.131	1.012
	TK	0.368	1.204	1.572	0.131	

Table 2.4.1: Squared bias ($\times 10^4$), Variance ($\times 10^4$), MSE ($\times 10^4$), the optimal bandwidth h and the MSE-ratio (MSE_{PGK}/MSE_{TK}) of the estimators of several normal mixture densities, for samples of size $n = 400$, with a censoring rate of 50% and $N = 1000$ replications.

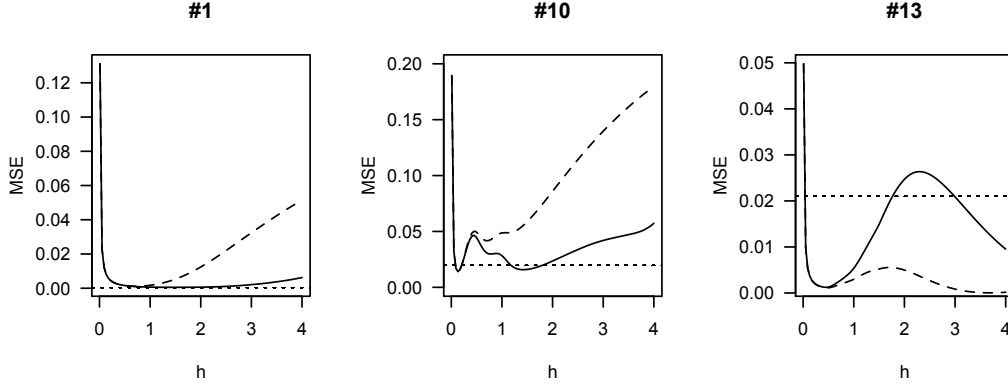


Figure 2.4.2: The MSE as a function of h for cases #1, #10 and #13 for the PGK estimator (solid curve), the TK estimator (dashed curve), and the parametric guide (dotted curve).

2.4.2 Weibull density with an exponential guide

In this model, the variable of interest Y is generated from a Weibull distribution with a scale parameter $b = 4$ and a shape parameter taking three values $a = 1, 2, 4$. The graphs of the resulting densities are plotted in Figure 2.4.3. The censoring variable is also drawn from a Weibull distribution with shape parameter a and scale parameter given by $b((1 - p)/p)^{1/a}$, ensuring a degree of censoring equal to p . We consider two censoring rates $p = 10\%$ and $p = 40\%$, and two sample sizes $n = 150$ and $n = 400$. As a parametric guide we use the exponential density $f_\theta(t) = \theta \exp(-\theta t)$, where θ is estimated using the approximated maximum likelihood estimator given by $\hat{\theta} = 1 / \sum_{i=1}^n W_{(i)} T_{(i)}$, where $T_{(i)}$ are the ordered values of the observed variables $T_i = \min(Y_i, C_i)$ and $W_{(i)}$ is the size of the jump of the Kaplan-Meier estimator at $T_{(i)}$ (see Suzukawa *et al.* (2001)). Note that the case $a = 1$ is the only situation where the guide is correctly specified. If $a \neq 1$ then the parametric guide is incorrect and deviates gradually from the true density. Our goal is to compare the performance of the PGK estimator with that of the TK estimator and the log-spline estimator (LSP) (Kooperberg and Stone (1992)) for both the density and the hazard function.

To this end, we run 1000 simulations and for every generated data set we calculate the estimators at the point $t = 3$. We look for the optimal bandwidths via a grid on $[0.1, 5]$. For the LSP estimator we select the optimal number of knots which minimizes the MSE over a set on $\{3, \dots, 10\}$. The results are summarized in Tables 2.4.2 and 2.4.3 for the density and the hazard function, respectively.

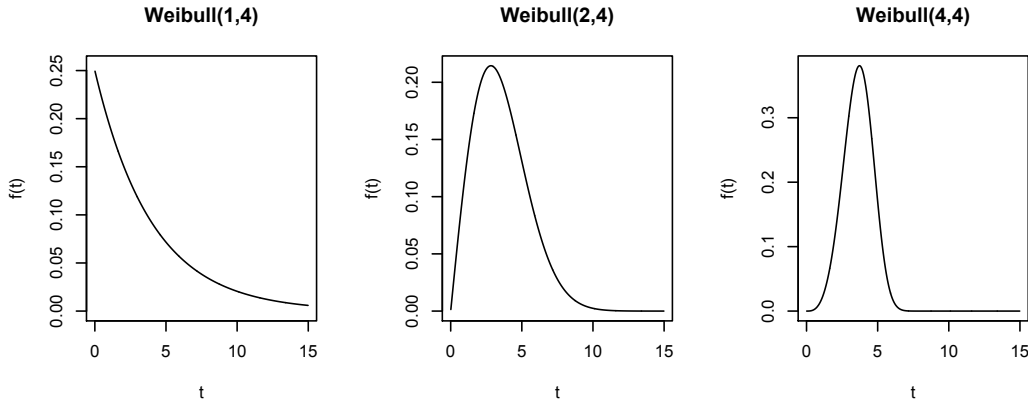


Figure 2.4.3: Weibull density with shape parameters $a = 1, 2, 4$ and scale parameter $b = 4$.

We start with the simulation results for the density estimators. As expected, with a correct parametric guide ($a = 1$) we get the best results for the PGK estimator. The bias of the PGK estimator is significantly reduced compared to that of the TK and the LSP estimator. Regarding the MSE, it is also reduced for the PGK estimator compared to the MSE of the LSP and the TK estimator, except for the case of sample size 150 and censoring rate 40%, where we observe a slightly larger variance and MSE for the PGK estimator compared to the TK estimator, but this is corrected with a larger sample size $n = 400$. For $a = 2$ and $a = 4$, even if the parametric guide is incorrect, the PGK estimator remains significantly better than the TK estimator, while the LSP estimator has a significantly smaller bias than both the PGK and the TK estimator. Regarding the variance, as expected, the PGK and the TK estimator have similar behavior except for the case $a = 4$ where the TK estimator has a larger variance. The LSP estimator has in general substantially larger variance and MSE, compared to the two kernel-based estimators, except for the case $a = 2$ with sample

size 400, where the LSP estimator outperforms both kernel-based estimators. In general, the MSE of the PGK estimator is not substantially reduced compared to that of the TK estimator, because in this example the variance dominates the bias.

For the hazard function, we computed the PGK, the TK and the LSP estimator using the same data generating procedure as for density function. The results are summarized in Table 2.4.3 and show that the PGK estimator generally outperforms the TK and the LSP estimator even if the parametric guide is not correctly specified. Note that for both the density and the hazard function, increasing the sample size enhances the performance of the PGK estimator. Another point to remark is that for the density and the hazard function, the selected optimal bandwidths for the PGK and the TK estimator are often close. In addition, we compared the PGK estimator based on the MLE and the PGK estimator based on the approximate MLE (AMLE). Simulation results given in Tables 2.6.1 and 2.6.2 in Appendix A show that when the guide is correct the PGK estimator based on the MLE outperforms the PGK estimator based on the AMLE (APGK). This is to be expected, since in this case the MLE is consistent, while the PGK estimator based on the AMLE behaves better when the guide is misspecified (see Remark 2.2 and Suzukawa *et al.* (2001)).

Finally, we compared the performance of the PGK and the TK density estimator at different time points. We computed both competitors at 100 different equally spaced time points from 0 to 7. We compared the performance for all cases given in Table 2.4.2, but, for sake of brevity, we only discuss three cases. The other cases are similar and are given in Appendix B. Figure 2.4.4 gives the squared bias, the variance and the MSE as a function of t for the cases $a = 1, 2$ and 4, respectively. When the guide is correct ($a = 1$) the bias of the PGK estimator is reduced to zero for most values of t . This is not the case for the MSE because as said before, the variance dominates the bias. For the case $a = 2$, the bias and the MSE are reduced considerably for many values of t . In the last case $a = 4$, the PGK and the TK estimator behave similarly for most values of t .

p		10%						40%					
n	a	Method	Bias ²	Var	MSE	$h/nknot$	MSE-ratio	Bias ²	Var	MSE	$h/nknot$	MSE-ratio	
150													
	1	PGK	0.009	114.6	114.6	5	0.929	0.579	194.2	194.8	5	1.080	
		TK	23.60	99.70	123.3	5		26.60	153.7	180.3	5		
		LSP	22.95	162.2	185.2	6		20.97	251.6	272.6	6		
	2	PGK	99.60	251.1	350.7	3.367	0.902	123.7	334.2	457.9	3.713	0.918	
		TK	114.5	274.1	388.6	3.020		152.9	345.7	498.6	3.268		
		LSP	81.90	332.7	414.6	3		10.90	627.9	638.8	6		
	4	PGK	98.91	258.1	357.0	3.169	0.469	94.40	377.0	471.4	3.268	0.505	
		TK	246.6	514.4	761.0	2.228		303.7	629.2	932.9	2.327		
		LSP	21.10	835.0	856.1	5		410.2	900.8	1311	4		
n	a	Method	Bias ²	Var	MSE	$h/nknot$	MSE-ratio	Bias ²	Var	MSE	$h/nknot$	MSE-ratio	
400													
	1	PGK	0.020	40.58	40.60	5	0.671	0.01	68.66	68.67	5	0.848	
		TK	20.63	39.89	60.52	4.802		22.35	58.59	80.94	5		
		LSP	3.36	55.81	59.17	6		8.340	85.10	93.44	6		
	2	PGK	50.70	124.8	175.5	2.872	0.876	62.80	177.7	240.5	3.020	0.909	
		TK	54.20	146.0	200.2	2.525		75.00	189.3	264.3	2.723		
		LSP	0.450	150.4	150.9	5		0.778	207.5	208.3	6		
	4	PGK	44.20	136.3	180.5	2.822	0.439	51.80	187.6	239.4	2.921	0.507	
		TK	117.5	293.8	411.3	1.882		141.8	330.5	472.3	1.981		
		LSP	20.90	319.1	340.0	5		0.523	520.2	520.7	6		

Table 2.4.2: Squared bias ($\times 10^6$), Variance ($\times 10^6$), MSE ($\times 10^6$), the optimal bandwidth h , the optimal number of knots $nknot$ and the ratio of MSE (MSE_{PGK}/MSE_{TK}) for the estimators of several Weibull densities for $a = (1, 2, 4)$, two censoring rates $p = (10\%, 40\%)$, two sample sizes $n = (150, 400)$ and $N = 1000$ replications.

p			10%					40%				
n	a	Method	Bias ²	Var	MSE	$h/nknot$	MSE-ratio	Bias ²	Var	MSE	$h/nknot$	MSE-ratio
150												
	1	PGK	0.360	90.77	91.13	5	0.714	1.300	145.0	146.3	5	0.782
		TK	14.69	113.0	127.7	4.901		18.60	168.6	187.2	5	
		LSP	20.91	73.22	94.13	5		5.430	189.5	194.9	6	
	2	PGK	37.10	128.0	165.1	3.664	0.818	43.90	167.1	211.0	4.010	0.814
		TK	35.40	166.4	201.8	3.168		52.40	206.9	259.3	3.515	
		LSP	0.140	179.2	179.3	5		1.105	325.3	326.4	6	
	4	PGK	24.20	91.4	115.6	3.366	0.529	24.70	119.6	144.2	3.465	0.551
		TK	48.00	169.2	218.2	2.327		71.90	189.4	261.3	2.525	
		LSP	6.343	258.4	264.7	5		4.548	411.3	415.8	5	
n	a	Method	Bias ²	Var	MSE	$h/nknot$	MSE-ratio	Bias ²	Var	MSE	$h/nknot$	MSE-ratio
400												
	1	PGK	0.190	34.91	35.10	5	0.631	0.220	57.13	57.35	5	0.701
		TK	8.200	47.40	55.60	4.307		13.30	68.50	81.80	4.901	
		LSP	0.810	41.42	42.23	6		1.860	75.25	77.11	6	
	2	PGK	17.20	63.50	80.70	3.020	0.835	21.80	85.10	106.9	3.218	0.855
		TK	15.80	80.90	96.70	2.574		22.90	102.3	125.1	2.822	
		LSP	0.340	66.92	67.26	5		0.870	120.4	121.3	6	
	4	PGK	10.36	46.39	56.75	2.971	0.508	12.06	56.66	68.72	3.070	0.549
		TK	24.50	87.10	111.6	1.981		26.40	98.70	125.1	2.030	
		LSP	5.300	101.8	107.1	5		0.800	153.2	154.0	5	

Table 2.4.3: Squared bias ($\times 10^5$), Variance ($\times 10^5$), MSE ($\times 10^5$), the optimal bandwidth h , the optimal number of knots $nknot$ and the ratio of MSE (MSE_{PGK}/MSE_{TK}) for the estimators of several Weibull hazards for $a = (1, 2, 4)$, two censoring rates $p = (10\%, 40\%)$, two sample sizes $n = (150, 400)$ and $N = 1000$ replications.

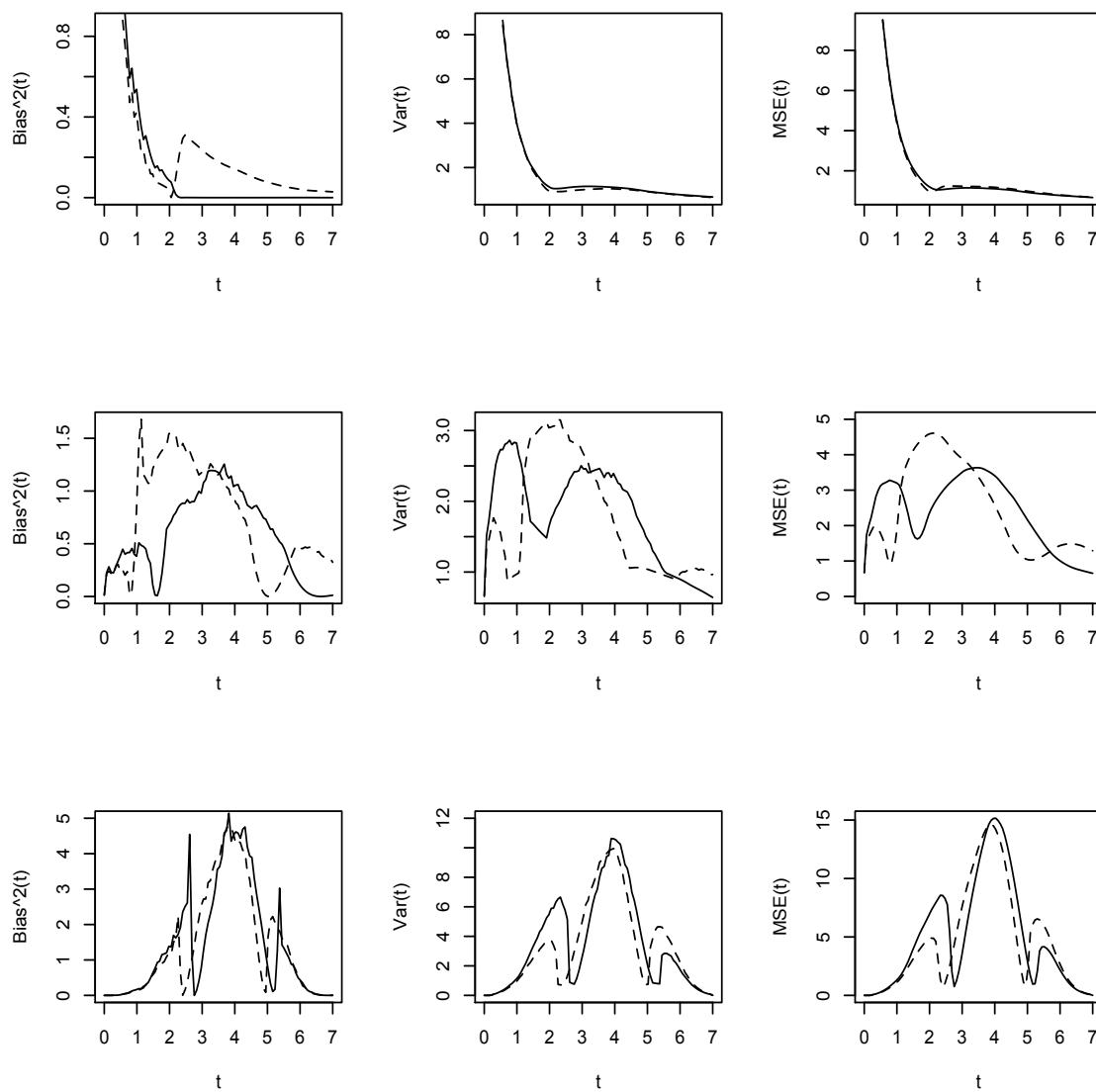


Figure 2.4.4: Squared bias, variance and $\text{MSE} (\times 10^4)$ as a function of t , for the PGK estimator (solid curve) and the TK estimator (dotted curve). The figures correspond to $a = 1$ (top), $a = 2$ (middle) and $a = 4$ (bottom). The sample size is $n = 150$ and the proportion of censoring is $p = 0.10$.

2.5 Application

To illustrate our method with a real dataset, we consider the UIS dataset from the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study. As discussed in Chapter 1, the goal of this study is to model time until return to drug use for patients enrolled in two different residential treatment programs that differed in length (treat=0 is the short program and treat=1 is the long program). Among a total of 628 observations, there are 120 censored observations, which corresponds to a censoring rate of 19.12%.

As a first step, before proceeding to more complicated analyses, it is always useful to have an idea about the distribution of the variable of interest without considering explanatory variables. This basic univariate analysis is given in Section 2.5.1 and is in general informative but not sufficient in practice. A more complete analysis including explanatory variables is performed in a second step as an extension of the PGK method to the Cox model, see Section 2.5.3 for more details.

We also explain in Section 2.5.2 how to obtain pointwise confidence intervals for the density and hazard function.

2.5.1 Density and hazard estimation for the UIS data

In this section, we use the PGK, the TK and the logspline (LSP) estimator (Kooperberg and Stone (1992)) for the analysis of the time to return to drug use in the UIS dataset. We consider the PGK estimator using two different guides, an exponential density and a Weibull density for which the parameters are estimated using maximum likelihood. For the kernel-based estimators, we use the Epanechnikov kernel and the choice of the bandwidths is achieved by a data driven bandwidth selection based on the least squares cross-validation method adapted to each estimator (see for example Wang and Wang (2007)). The selected bandwidths are: $h = 671.21$ for the PGK estimator with an exponential guide, $h = 716.66$ for the PGK estimator with a Weibull guide and $h = 663.63$ for the TK estimator. For the LSP estimator, we specify that the density equals zero to the left of 0, we use a BIC penalty and a stepwise knot deletion procedure (see Kooperberg and Stone (1991)), which selected four knots of the original 10 knots.

The plots of the different estimators of the density and the hazard function are given in Figure 2.5.1. From the plot of the hazard function it can be seen that the risk to return to drug use increases slowly during the first 200-300 days, after which it decreases to zero. While the overall shape is the same for all four estimators, there are some absolute differences especially during the first 400 days. We do not know which estimator is closer to the true curve, but given the results of our simulation study, we suspect that the TK method has tendency to either overestimate or underestimate the hazard function and so the real risks to return to drug use. Finally, concerning the density function we see that the PGK and LSP estimators are quite close. Moreover, it seems that our PGK estimators remove a considerable part of the boundary effect on the left endpoint of the distribution compared to the TK estimator.

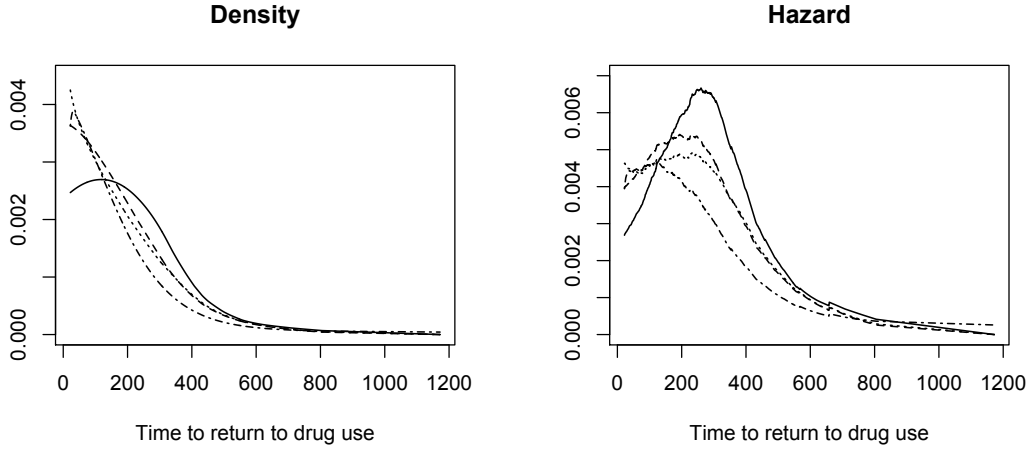


Figure 2.5.1: Estimation of the density and hazard function for the UIS data by means of the TK estimator (solid curve), the LSP estimator (dot-dashed curve), the PGK estimator with exponential guide (dashed curve), and the PGK estimator with Weibull guide (dotted curve).

Remark 2.5. *As stated earlier, one can also define a parametrically guided kernel estimator for the cumulative distribution function (CDF) $F(\cdot)$ as follows:*

$$\hat{F}_{\hat{\theta}}(t) = \int_0^t \hat{f}_{\hat{\theta}}(s) ds, t \leq \tau$$

Figure 2.5.2 gives the plot of the guided estimator of $F(\cdot)$ with exponential guide as well as its corresponding Kaplan-Meier estimator.

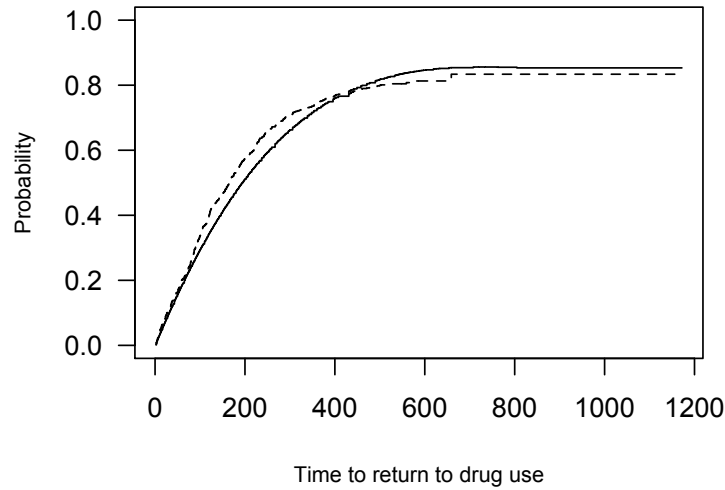


Figure 2.5.2: Estimation of the CDF for the UIS data by means of the PGK estimator with exponential guide (solid curve) and the Kaplan-Meier estimator (dashed curve), $h = 671.21$.

2.5.2 Pointwise asymptotic confidence intervals

Based on the asymptotic normality of our PGK estimators, we construct pointwise confidence intervals for the density and the hazard function of the time to return to drug use for the UIS data. In order to remove the asymptotic bias, we use the undersmoothing technique, that is, we use a smaller bandwidth than the one chosen to compute the density estimator in order to make the bias converge to zero more rapidly; see for example Horowitz (2001) and Fiorio (2004). The pointwise $100(1 - \alpha)\%$ confidence interval for the density function $f(t)$ and the hazard function $\lambda(t)$ are given respectively by

$$\hat{f}_{\hat{\theta}}(t) \pm z_{1-\alpha/2} \hat{\sigma}(t) / \sqrt{nh}, \quad \hat{\lambda}_{\hat{\theta}}(t) \pm z_{1-\alpha/2} \hat{\tau}(t) / \sqrt{nh},$$

where $\hat{\sigma}^2(t) = [\hat{f}_{\hat{\theta}}(t)/(1 - \hat{G}(t))] \int K^2(u)du$, $\hat{\tau}^2(t) = [\hat{\lambda}_{\hat{\theta}}(t)/(1 - \hat{H}(t))] \int K^2(u)du$, and $h = h_{opt}n^{1/5}/n^{\tau} = 486.36$. Here, $\tau = 1/4$ is the undersmoothing parameter, h_{opt} is the optimal bandwidth selected via the cross-validation method used in the previous section, $\hat{G}(\cdot)$ is the Kaplan Meier estimator of $G(\cdot)$ and $\hat{H}(\cdot)$ is the empirical distribution function of $H(\cdot)$. The undersmoothed PGK density and hazard estimator with exponential guide and their respective confidence intervals are plotted in Figure 2.5.3. After the 500th day, the intervals become larger because most censoring occurs after this date.

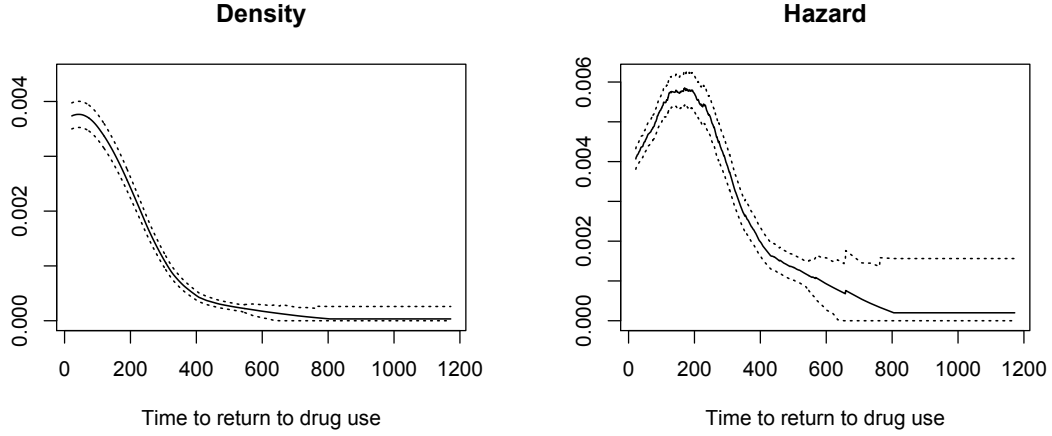


Figure 2.5.3: Estimation of the density and hazard function for the UIS data by means of the undersmoothed PGK estimator with exponential guide (solid curve), together with 95% pointwise asymptotic confidence intervals (dotted curve).

2.5.3 Extending the PGK estimation to the Cox model

In this section, we extend the PGK method to the estimation of the baseline and the conditional density and hazard functions in the Cox model. First, the PGK estimator that we propose for the baseline hazard function is $\hat{\lambda}_{\hat{\theta},B}(t) = \hat{f}_{\hat{\theta},B}(t)/(1 - \hat{F}_B(t))$, where $\hat{F}_B(\cdot)$ is the baseline distribution function estimated after fitting a semiparametric Cox

model and $\hat{f}_{\hat{\theta},B}(t)$ is the PGK estimator of the baseline density defined as

$$\hat{f}_{\hat{\theta},B}(t) = \frac{1}{h} \sum_{i=1}^n K\left(\frac{t - T_i}{h}\right) \frac{f_{\hat{\theta},B}(t)}{f_{\hat{\theta},B}(T_i)} W_{Bi},$$

where $f_{\hat{\theta},B}(\cdot)$ is a parametric baseline guide estimated via a parametric proportional hazards model (PPH) and W_{Bi} is the jump size of $\hat{F}_B(\cdot)$ at T_i .

Now, to extend the PGK estimator to the conditional survival and hazard function, we propose the following estimators:

$$\hat{\lambda}_{\hat{\theta}}(t | X) = \hat{\lambda}_{\hat{\theta},B}(t) \exp(X^t \hat{\beta}), \quad \hat{S}_{\hat{\theta}}(t | X) = \exp\left(-\int_0^t \hat{\lambda}_{\hat{\theta}}(s | X) ds\right),$$

where X is a vector of covariates and $\hat{\beta}$ are the estimated parameters of the Cox model using the partial likelihood maximization. For more details about the steps of this procedure, we refer to Appendix C.

We apply the proposed estimators to the UIS data. In addition to the time to return to drug use, we now consider $X = (X_1, X_2, X_3, X_4)$ a vector of four covariates: X_1 is the age in years, X_2 is the drug use history (1=never, 2=previous, 3=recent), X_3 is the number of prior drug treatments at admission (0-40), and X_4 is the treatment randomization assignment (0=short, 1=long). As parametric guide, we use a Weibull baseline density estimated after fitting a PPH model. The bandwidth is selected as before using the cross-validation method. The selected bandwidth is $h_{cv} = 590$. Figure 2.5.4 shows the plots of the PGK baseline density and hazard estimators. We see that, except for the beginning of the study, the estimators are monotonically decreasing, meaning that at the start of the study the participants are at high risk to restart using drugs. The estimated parameters for the fitted Cox model are $\hat{\beta} = (-0.030, 0.208, 0.029, -0.235)$. Based on this model, we conclude that increasing the participant age or the treatment duration reduces the risk of returning to drug use, while increasing the drug use history or the number of previous treatments increases the risk of returning to drug use. Figure 2.5.5 shows the plots of the PGK conditional survival and hazard function estimators for $X = (36, 2, 1, 1)$. It appears from the plot of the conditional survival function that for a participant aged 36 years, receiving a long treatment, with a previous history of drug use and one prior treatment, the median time to return to drug use is about 200 days.

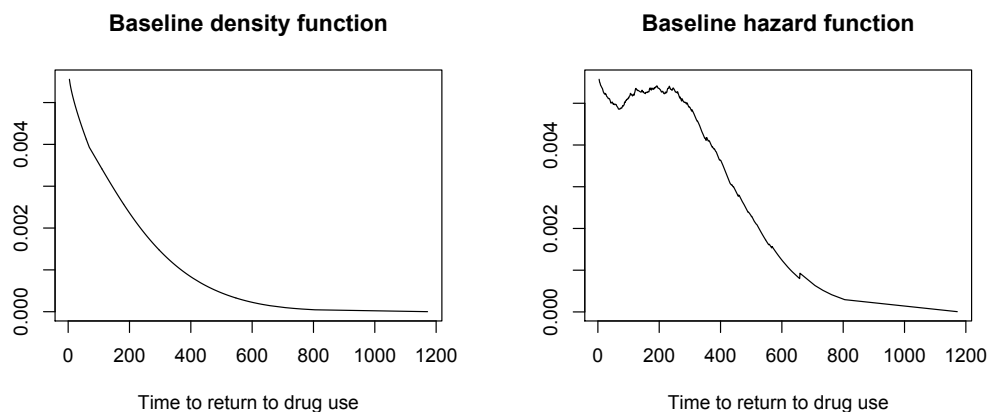


Figure 2.5.4: Estimation of the baseline density and hazard function in the Cox model for the UIS data, by means of the PGK estimator with Weibull guide.

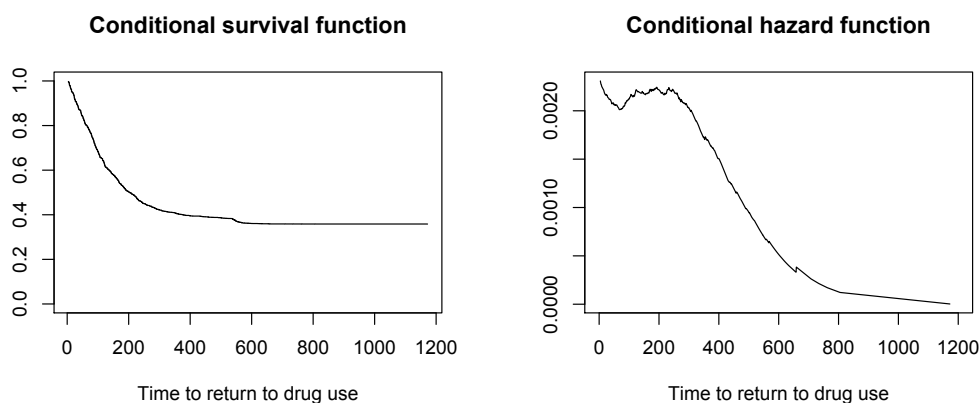


Figure 2.5.5: Estimation of the conditional survival and hazard function in the Cox model for the UIS data, by means of the PGK estimator with Weibull guide. The covariate equals $X = (36, 2, 1, 1)$.

2.6 Appendix A

p			10%			40%		
n	a	Method	$Bias^2$	Var	MSE	$Bias^2$	Var	MSE
150	1	PGK	0.165	107.7	107.9	0.548	179.6	180.1
		APGK	0.180	108.1	108.3	0.910	189.6	190.5
	2	PGK	81.24	248.3	329.5	104.4	352.5	456.9
		APGK	80.50	243.7	324.2	108.4	331.1	439.5
	4	PGK	96.60	300.7	397.3	170.2	479.8	650.0
		APGK	89.10	267.3	356.4	91.64	383.9	475.5
n	a	Method	$Bias^2$	Var	MSE	$Bias^2$	Var	MSE
400	1	PGK	0.037	41.12	41.16	0.009	67.71	67.72
		APGK	0.039	41.18	41.22	0.025	69.43	69.46
	2	PGK	45.94	129.8	175.8	61.22	186.9	248.1
		APGK	46.02	126.32	172.3	61.13	170.9	232.0
	4	PGK	48.42	149.1	197.5	83.37	248.2	331.6
		APGK	41.33	134.4	175.7	38.37	196.0	234.4

Table 2.6.1: Squared bias $Bias^2$ ($\times 10^6$), Variance Var ($\times 10^6$) and MSE ($\times 10^6$) for density estimators for $a = (1, 2, 4)$, two censoring rate $p = (10\%, 40\%)$, two sample sizes $n = (150, 400)$ and $N = 500$ replications.

p			10%			40%		
n	a	Method	$Bias^2$	Var	MSE	$Bias^2$	Var	MSE
150	1	PGK	0.950	91.25	92.20	1.60	140.5	142.1
		APGK	0.970	91.30	92.27	1.94	142.6	144.5
	2	PGK	29.89	139.6	169.5	40.12	189.5	229.6
		APGK	31.23	134.8	166.0	35.11	178.6	213.7
	4	PGK	25.60	109.7	135.3	44.33	158.7	203.0
		APGK	23.90	100.4	124.3	29.36	125.0	154.4
n	a	Method	$Bias^2$	Var	MSE	$Bias^2$	Var	MSE
400	1	PGK	0.260	33.43	33.69	0.580	58.18	58.76
		APGK	0.270	33.43	33.70	0.640	58.45	59.09
	2	PGK	16.18	62.83	79.01	23.29	89.47	112.8
		APGK	16.05	61.27	77.32	20.29	83.93	104.2
	4	PGK	11.60	50.26	61.86	19.24	73.86	93.10
		APGK	10.71	45.56	56.27	11.35	57.47	68.82

Table 2.6.2: Squared bias $Bias^2$ ($\times 10^5$), Variance Var ($\times 10^5$) and MSE ($\times 10^5$) for hazard estimators for $a = (1, 2, 4)$, two censoring rate $p = (10\%, 40\%)$, two sample sizes $n = (150, 400)$ and $N = 500$ replications.

2.7 Appendix B

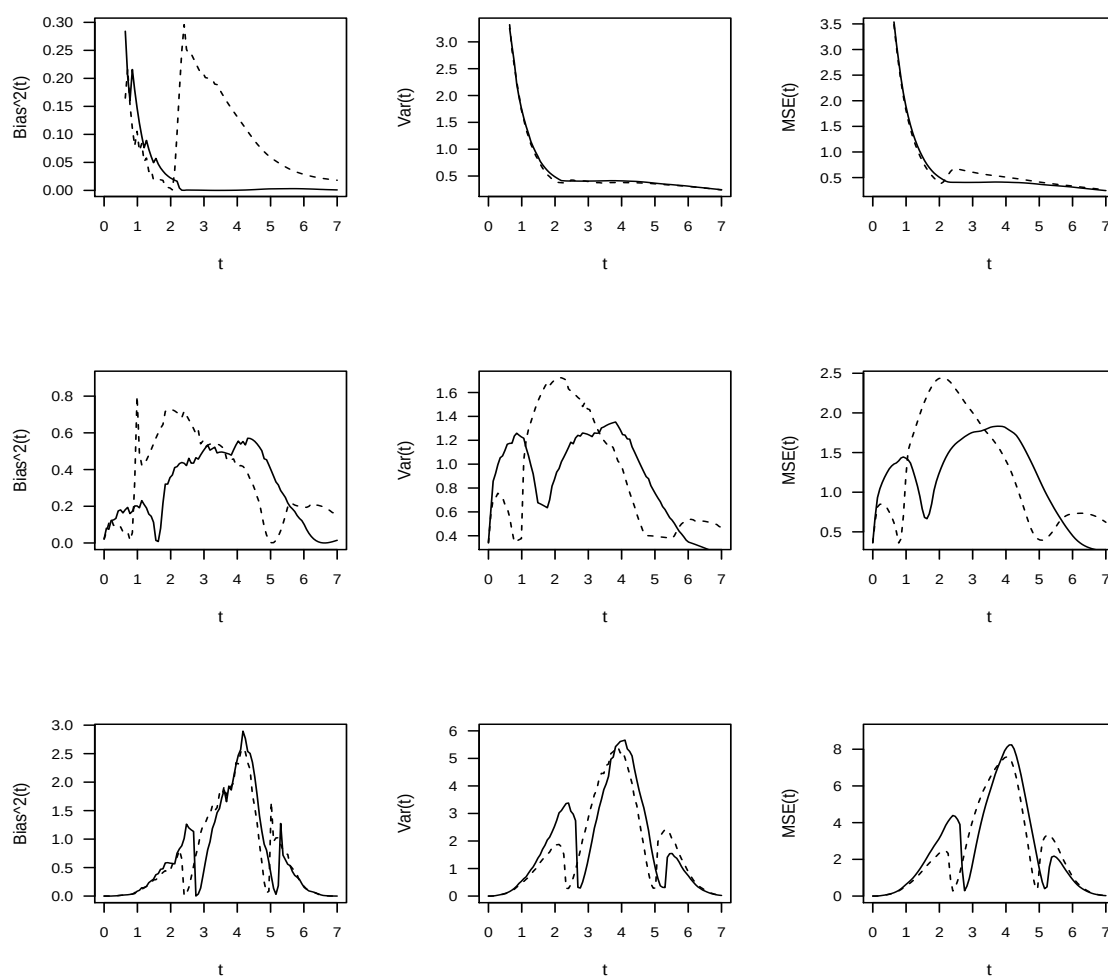


Figure 2.7.1: Squared bias, variance and MSE ($\times 10^4$) as a function of t , for the PGK estimator (solid curve) and the TK estimator (dotted curve). The figures correspond to $a = 1$ (top), $a = 2$ (middle) and $a = 4$ (bottom). The sample size is $n = 400$ and the proportion of censoring is $p = 0.10$.

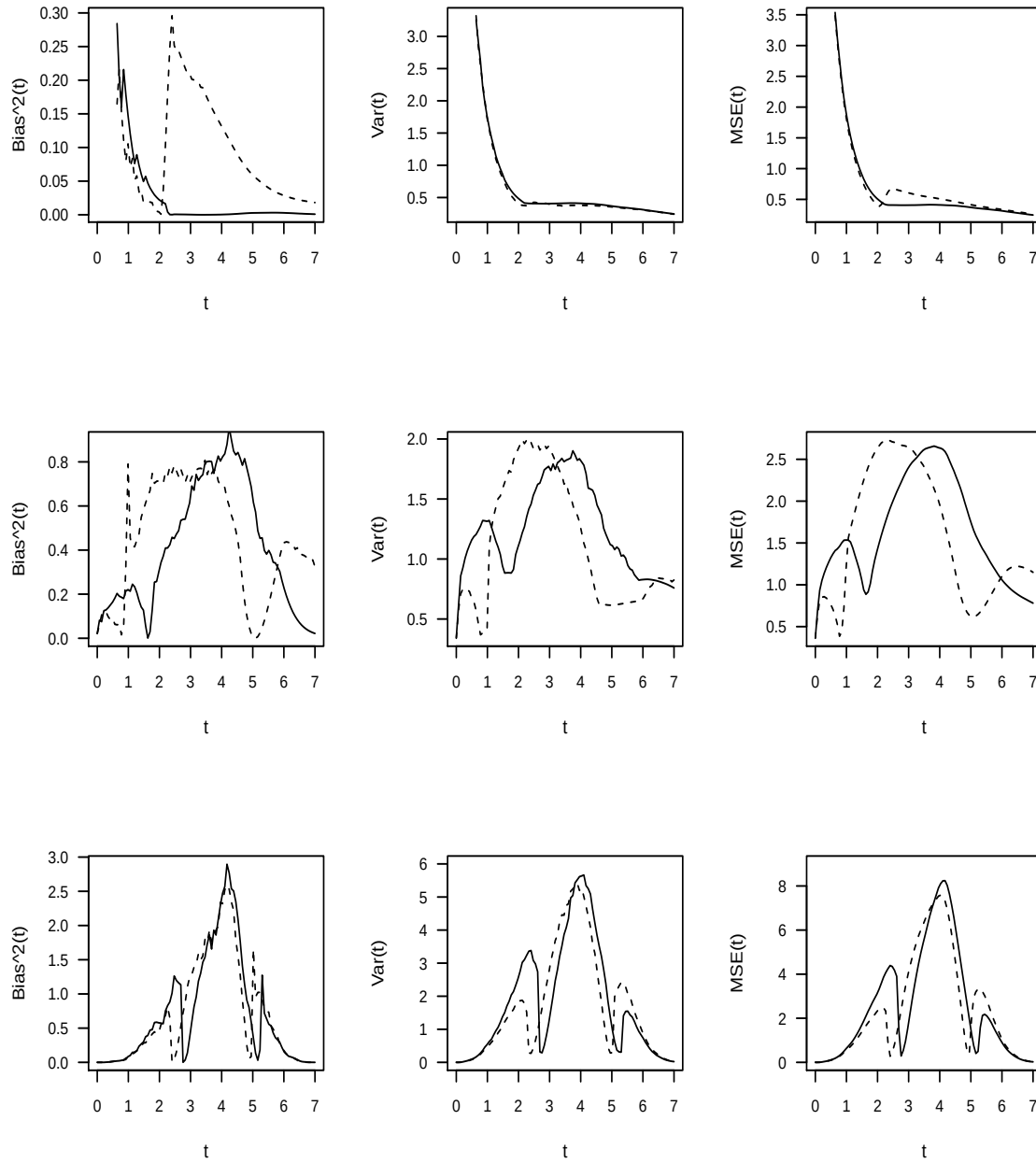


Figure 2.7.2: Squared bias, variance and MSE ($\times 10^4$) as a function of t , for the PGK estimator (solid curve) and the TK estimator (dotted curve). The figures correspond to $a = 1$ (top), $a = 2$ (middle) and $a = 4$ (bottom). The sample size is $n = 400$ and the proportion of censoring is $p = 0.40$.

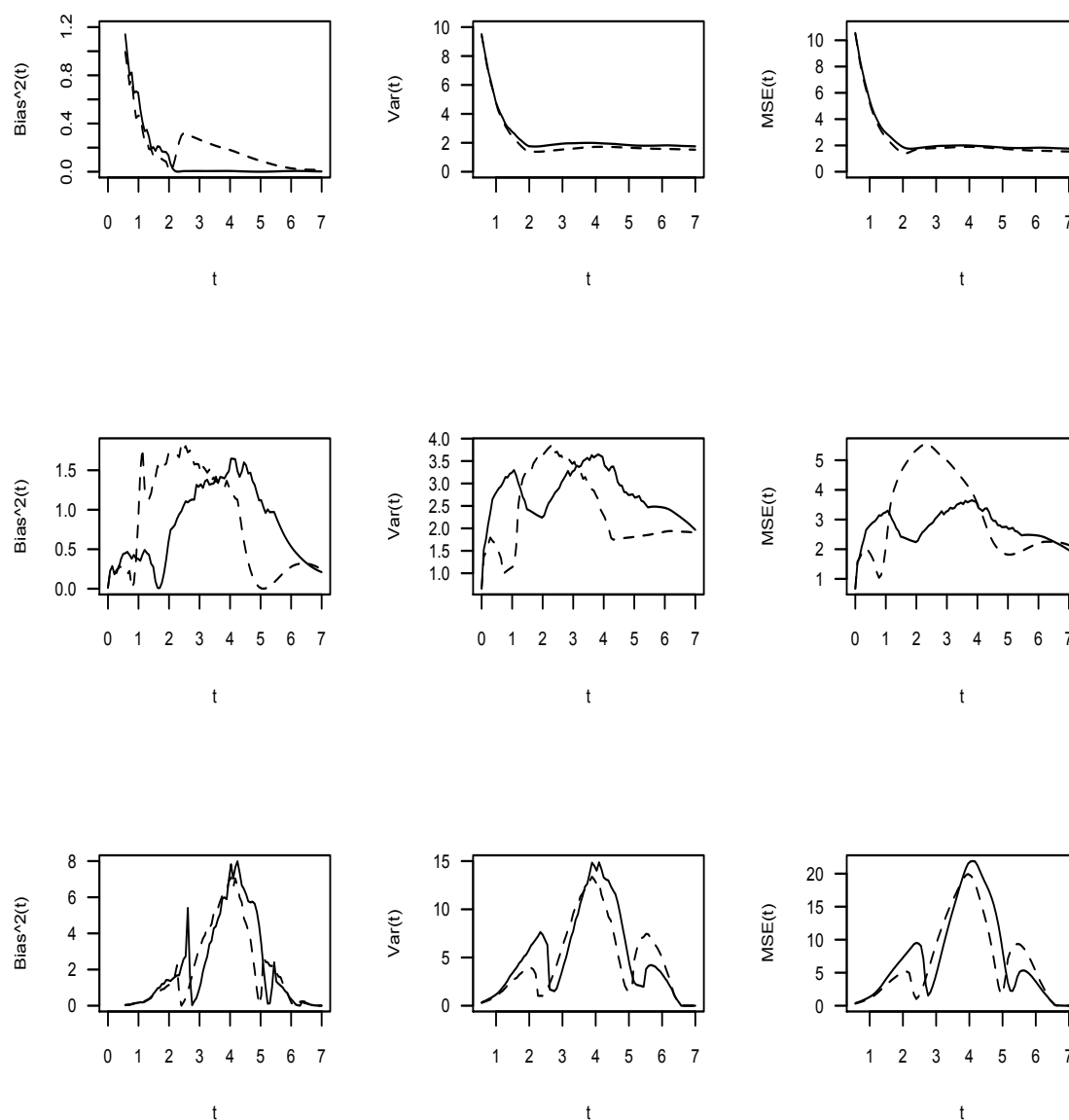


Figure 2.7.3: Squared bias, variance and MSE ($\times 10^4$) as a function of t , for the PGK estimator (solid curve) and the TK estimator (dotted curve). The figures correspond to $a = 1$ (top), $a = 2$ (middle) and $a = 4$ (bottom). The sample size is $n = 150$ and the proportion of censoring is $p = 0.40$.

2.8 Appendix C

The extension of the PGK estimation to the Cox model is carried out according to the following steps:

1. Compute $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ the estimated parameters of the Cox model using partial likelihood as well as $\hat{\Lambda}_B(\cdot)$ the estimated baseline cumulative hazard function.
2. Compute $\hat{F}_B(\cdot) = 1 - \exp(-\hat{\Lambda}_B(\cdot))$ the estimated baseline CDF and the W_{Bi} the size of the jump of $\hat{F}_B$ at T_i .
3. Estimate the parametric guide via the following Weibull-PPH:

$$\lambda(t) = \lambda_{\theta,B}(t) \exp(X^t \hat{\beta}).$$

- (a) Compute $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ the estimated scale and shape of a Weibull baseline distribution using MLE.
 - (b) Get the Weibull baseline density $f_{\hat{\theta},B}(t) = \hat{\theta}_1 \hat{\theta}_2 t^{(\hat{\theta}_2-1)} \exp(-\hat{\theta}_1 t^{\hat{\theta}_2})$ (the parametric guide).
4. Compute the PGK estimator for the baseline density as follows:

$$\hat{f}_{\hat{\theta},B}(t) = \frac{1}{h} \sum_{i=1}^n K\left(\frac{t - T_i}{h}\right) \frac{f_{\hat{\theta},B}(t)}{f_{\hat{\theta},B}(T_i)} W_{Bi}.$$

5. Compute the PGK estimator of the conditional survival and hazard function using the following estimators:

$$\hat{\lambda}_{\hat{\theta}}(t | X) = \hat{\lambda}_{\hat{\theta},B}(t) \exp(X^t \hat{\beta}), \quad \hat{S}_{\hat{\theta}}(t | X) = \exp\left(-\int_0^t \hat{\lambda}_{\hat{\theta}}(s | X) ds\right).$$

2.9 Appendix D

Proof of Theorem 2.1. The PGK density estimator based on the non random parametric guide $f_*(\cdot)$ can be decomposed as follows:

$$\widehat{f}_*(t) - f(t) = (\widehat{f}_*(t) - \widetilde{f}(t)) + (\widetilde{f}(t) - f(t)), \quad (2.9.1)$$

where

$$\widetilde{f}(t) = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{t-s}{h}\right) \frac{f_*(t)}{f_*(s)} dF(s).$$

1. For $t \leq \tau$, we have

$$\begin{aligned} \widehat{f}_*(t) - \widetilde{f}(t) &= \frac{f_*(t)}{h} \int_{t-h}^{t+h} K\left(\frac{t-s}{h}\right) \frac{1}{f_*(s)} d(\widehat{F}(s) - F(s)) \\ &= \frac{f_*(t)}{h} \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) d(K(u)/f_*(t-uh)) \\ &= \frac{f_*(t)}{h} \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) \frac{K'(u)}{f_*(t-uh)} du \\ &\quad + f_*(t) \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) \frac{K(u)f'_*(t-uh)}{f_*^2(t-uh)} du \\ &= A_{1,n} + A_{2,n}. \end{aligned}$$

First, write

$$\begin{aligned} A_{1,n} &= \frac{1}{h} \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) K'(u) du \\ &\quad + \frac{f_*(t)}{h} \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) \left(\frac{K'(u)}{f_*(t-uh)} - \frac{K'(u)}{f_*(t)} \right) du \\ &= A_{11,n} + A_{12,n}. \end{aligned}$$

We start with $A_{12,n}$. We have

$$\begin{aligned} A_{12,n} &= \frac{1}{h} \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) \frac{f_*(t) - f_*(t-uh)}{f_*(t-uh)} K'(u) du \\ &= \frac{1}{h} \int_{-1}^1 (\widehat{F}(t-uh) - F(t-uh)) \frac{f'_*(t+\rho)uh}{f_*(t-uh)} K'(u) du, \end{aligned}$$

for some ρ between 0 and $-uh$. Therefore,

$$|A_{12,n}| \leq \sup_{s \in \mathbb{N}_t} |\widehat{F}(s) - F(s)| \sup_{s \in \mathbb{N}_t} |f'_*(s)| (\inf_{s \in \mathbb{N}_t} f_*(s))^{-1} \int_{-1}^1 |K'(u)| |u| du,$$

where $\aleph_t$ is a small neighborhood around t . Hence, under assumptions 2.1 and 2.2, and using the uniform rate of the Kaplan-Meier estimator (see e.g. Theorem 1 in Lo and Singh (1986)) we have that $A_{12,n} = O_p(n^{-1/2})$. Now, we treat the term $A_{11,n}$. We consider the i.i.d. decomposition of $\hat{F}$ given in Lemma 2.1 in Lo *et al.* (1989):

$$\hat{F}(s) - F(s) = n^{-1} \sum_{i=1}^n \xi_i(s) + r_n(s),$$

where $\xi_i(s) = -\int_{-\infty}^{T_i \wedge s} (1 - H(x))^{-2} dH_1(x) + (1 - H(T_i))^{-1} I\{T_i \leq s, \delta_i = 1\}$ and $\sup_{s \in \aleph_t} |r_n(s)| = O_p(n^{-1} \log n)$. Note that Lo *et al.* (1989) assume that the lifetimes are non-negative, whereas we work with random variables defined on the whole real line. However, it can be easily seen that their results remain valid in our setup. Then,

$$A_{11,n} = \frac{1}{nh} \sum_{i=1}^n U_{in}(t) + O_p((nh)^{-1} \log n),$$

where $U_{in}(t) = \int_{-1}^1 \xi_i(t - uh) K'(u) du$. Therefore,

$$A_{1,n} = \frac{1}{nh} \sum_{i=1}^n U_{in}(t) + O_p(n^{-1/2}), \quad (2.9.2)$$

thanks to assumption (A.2).

Finally, we consider the term $A_{2,n}$. Under assumptions 2.1 and 2.2, we have

$$\begin{aligned} |A_{2,n}| &\leq f_*(t) \sup_{s \in \aleph_t} |\hat{F}(s) - F(s)| \sup_{s \in \aleph_t} |f'_*(s)| \left(\inf_{s \in \aleph_t} f_*(s) \right)^{-2} \\ &= O_p(n^{-1/2}). \end{aligned}$$

Therefore,

$$\hat{f}_*(t) - \tilde{f}(t) = \frac{1}{nh} \sum_i U_{in}(t) + O_p(n^{-1/2}). \quad (2.9.3)$$

The result now follows from expressions (2.9.1) and (2.9.3).

2. We have

$$\begin{aligned}
\tilde{f}(t) - f(t) &= \frac{f_*(t)}{h} \int_{t-h}^{t+h} K\left(\frac{t-s}{h}\right) \frac{f(s)}{f_*(s)} ds - f(t) \\
&= f_*(t) \int_{-1}^1 K(u) r_*(t - uh) du - f(t) \\
&= f_*(t) \int_{-1}^1 K(u) (r_*(t) - r'_*(t)uh + \frac{1}{2}h^2 r''_*(t)u^2 + o(h^2)) du - f(t) \\
&= \frac{1}{2}h^2 r''_*(t) f_*(t) \mu_K^2 + o(h^2).
\end{aligned} \tag{2.9.4}$$

Now, the result is an immediate consequence of the first point, and of Theorem 3.2 and Corollary 3.3 in Lo *et al.* (1989). $\square$

Proof of Theorem 2.2. Write

$$\sqrt{nh}(\hat{f}_{\hat{\theta}}(t) - f(t)) = \sqrt{nh}(\hat{f}_{\hat{\theta}}(t) - \hat{f}_{\theta^*}(t)) + \sqrt{nh}(\hat{f}_{\theta^*}(t) - f(t)),$$

where $\hat{f}_{\theta^*}(t)$ is the PGK density estimator based on the parametric guide $f_{\theta^*}(\cdot)$. From the second point in Theorem 2.1 it follows that

$$\sqrt{nh}\left(\hat{f}_{\theta^*}(t) - f(t) - \frac{1}{2}h^2\mu_K^2 r''_{\theta^*}(t)f_{\theta^*}(t) + o(h^2)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(t)\right). \tag{2.9.5}$$

On the other hand, we have

$$\begin{aligned}
\hat{f}_{\hat{\theta}}(t) - \hat{f}_{\theta^*}(t) &= \frac{1}{h}(f_{\hat{\theta}}(t) - f_{\theta^*}(t)) \int_{t-h}^{t+h} K\left(\frac{t-s}{h}\right) \frac{1}{f_{\hat{\theta}}(s)} d\hat{F}(s) \\
&\quad - \frac{f_{\theta^*}(t)}{h} \int_{t-h}^{t+h} K\left(\frac{t-s}{h}\right) \frac{f_{\hat{\theta}}(s) - f_{\theta^*}(s)}{f_{\hat{\theta}}(s)f_{\theta^*}(s)} d\hat{F}(s) \\
&= I_{1n} - I_{2n}.
\end{aligned}$$

By a Taylor expression we have, for an intermediate point θ_m between $\hat{\theta}$ and θ^* and a constant $C < \infty$,

$$\begin{aligned}
|I_{1n}| &\leq \frac{C}{h} \|\nabla_{\theta} f_{\theta_m}(t)\| |\hat{\theta} - \theta^*| |\hat{F}(t+h) - \hat{F}(t-h)| \\
&\leq \frac{C}{h} \|\nabla_{\theta} f_{\theta_m}(t)\| |\hat{\theta} - \theta^*| |\hat{F}(t+h) - \hat{F}(t-h) - F(t+h) + F(t-h)| \\
&\quad + \frac{C}{h} \|\nabla_{\theta} f_{\theta_m}(t)\| |\hat{\theta} - \theta^*| |F(t+h) - F(t-h)|,
\end{aligned}$$

where $\nabla_{\theta} f_{\theta}(t) = (\partial f_{\theta}(t) / \partial \theta_j)_{j=1}^p$. From Lemma 3 in Gijbels and Veraverbeke (1989), we have

$$\widehat{F}(t+h) - \widehat{F}(t-h) - F(t+h) + F(t-h) = O_p(n^{-1/2}h^{1/2}(\log n)^{1/2}).$$

Hence, $I_{1n} = o_p((nh)^{-1/2})$. In similar way it can be shown that $I_{2n} = o_p((nh)^{-1/2})$, and so

$$\widehat{f}_{\widehat{\theta}}(t) - \widehat{f}_{\theta^*}(t) = o_p((nh)^{-1/2}). \quad (2.9.6)$$

The result of Theorem 2.2 now follows from equations (2.9.5) and (2.9.6). $\square$

Proof of Theorem 2.3. We have

$$\sqrt{nh}(\widehat{\lambda}_{\widehat{\theta}}(t) - \lambda(t)) = \sqrt{nh}\widehat{f}_{\widehat{\theta}}(t) \left[\frac{\widehat{F}(t) - F(t)}{(1 - \widehat{F}(t))(1 - F(t))} \right] + \sqrt{nh} \left[\frac{\widehat{f}_{\widehat{\theta}}(t) - f(t)}{1 - F(t)} \right].$$

Since the first term on the right hand side converges to zero in probability, the result of Theorem 2.3 is a direct consequence of Theorem 2.2. $\square$

Chapter 3

Guided local linear regression¹

3.1 Introduction

In this chapter we consider the estimation of the mean regression function using the additive correction as discussed in Martins-Filho *et al.* (2008). Suppose for the moment that we have completely observed data $(Y_i, X_i), i = 1, \dots, n$, let $m(x) = E(Y|X = x)$ be the true regression function and let $m(x, \hat{\theta})$ be any parametric estimator of m , where $\hat{\theta}$ is an estimator of the least false value θ^* according to a certain distance measure between m and the parametric regression model $m(., \theta)$ (see Assumption 3.2, below). As discussed in Chapter 1, the additive correction is based on the identity

$$m(x) = m(x, \hat{\theta}) + r_{\hat{\theta}}(x),$$

where $r_{\theta}(x) = m(x) - m(x, \theta) = E(Y - m(X, \theta)|X = x)$. Based on the observed data, any classical parametric approach can be used in the first stage while a nonparametric estimator, say $\hat{r}_{\hat{\theta}}(x)$, based on the (pseudo) estimated data $(Y_i - m(X_i, \hat{\theta}), X_i)$ will be used for the correction term $r_{\hat{\theta}}(x)$. Hence, a parametrically guided nonparametric estimator can be defined by

$$\hat{m}(x) = m(x, \hat{\theta}) + \hat{r}_{\hat{\theta}}(x). \quad (3.1.1)$$

¹This chapter is based on Talamakrouni, M., El Ghouch, A. and Van Keilegom, I. (2015). Guided censored regression. *Scand. J. Stat.* **42**, 214-233.

Intuitively, if the parametric model is "close" to the true curve m , the additive correction $r_{\hat{\theta}}$ will be easier to estimate than m and this should improve the quality of the resulting nonparametric estimator $\hat{m}(x)$ compared to the traditional direct approach. On the other hand, if the parametric model is not adequate, the nonparametric correction should compensate.

The aim of this chapter is to propose a new estimator of the mean regression function, which is an extension of the guided nonparametric estimator to the situation where the response is randomly right censored. To deal with censoring we use the Koul *et al.* (1981) synthetic data approach. Thus, we first transform the observed data in an unbiased way before making any inference. We establish the asymptotic normality of the proposed estimator and illustrate its performance via finite sample simulations and an application to the UIS dataset investigated in the previous chapters.

3.2 Model and estimation procedure

3.2.1 Definitions and notations

We are interested in the relationship between a variable of interest $Y \in \mathfrak{R}^+$ and some covariate $X \in \mathfrak{R}$. In the presence of random right censoring, the response Y is not always available. Instead of observing a sample (Y_i, X_i) , $i = 1, \dots, n$, from (Y, X) , one observes a random sample (T_i, δ_i, X_i) , $i = 1, \dots, n$, from (T, δ, X) , with

$$T = \min(Y, C), \quad \delta = \mathbf{I}(Y \leq C),$$

where C is the censoring variable, supposed to be independent of the response Y given the covariate X .

Denote, respectively, by $F(y|x)$ and $G(y|x)$ the conditional distribution function of Y and C given $X = x$. Our goal is to estimate the function

$$m(x) = E(\phi(Y)|X = x),$$

where ϕ is a known transformation introduced to include various functions of interest. For example, when using the transformation $\phi(y) = y \cdot \mathbf{I}(y \leq \tau)$ for some known τ , the function m will be the truncated conditional mean $m(x) = \int_0^\tau y dF(y|x)$.

3.2.2 Parametrically guided nonparametric regression with censored data

Let us first consider the case where the response is not censored. In principle, the nonparametric method used in the estimation of the correction term $r_{\hat{\theta}}(x)$ might be any of the kernel type regression estimators available in the literature. In the following, we will focus on the local linear regression estimators. Since for any given parametric model $m(x, \theta)$, $m(x) = E[m(x, \theta) + (Y - m(X, \theta)) | X = x]$, the regression function m can be estimated using local linear regression techniques by $\hat{b}_0$, where $(\hat{b}_0, \hat{b}_1)$ minimize

$$\sum_{i=1}^n \{Z_i(x, \hat{\theta}) - b_0 - b_1(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right)$$

with respect to b_0 and b_1 , $Z_i(x, \theta) = m(x, \theta) + (Y_i - m(X_i, \theta))$, $i = 1, \dots, n$, $\hat{\theta}$ is an estimator of θ , K is a kernel function and $0 < h \equiv h_n$ is a bandwidth. This estimator can be explicitly expressed as

$$\hat{m}(x) = \hat{b}_0 = \sum_{i=1}^n W_i(x, h) Z_i(x, \hat{\theta}), \quad (3.2.1)$$

where W_i 's are the local linear weights given in Chapter 1 by expression (1.3.5).

Note that since $\sum_{i=1}^n W_i(x, h) = 1$, the estimator defined in (3.2.1) has the form given in (3.1.1) with

$$r_{\theta}(x) = \sum_{i=1}^n W_i(x, h) (Y_i - m(X_i, \theta)).$$

Since, in the case of censored data, the observed variable T does not have the same conditional expectation as the variable of interest Y , we will use some unbiased transformation based on (T, δ, X) . This procedure is known in the literature as the synthetic data approach. Several investigations exist in the literature, for an overview of existing transformations, we refer to Chapter 1. In this chapter, we will use the transformation proposed by Koul *et al.* (1981), which is given by

$$Y^* = \frac{\delta \phi(T)}{1 - G(T|X)}. \quad (3.2.2)$$

Since Y and C are independent given X we have

$$E(Y^*|X = x) = E(\phi(Y)|X = x). \quad (3.2.3)$$

Unfortunately, in real data analysis, $G(y|x)$ is typically unknown and so the transformation (3.2.2) can not be computed without estimating G . In the uncensored case, the estimation of the conditional distribution function has been considered by Stone (1977) among many others. In the censored data framework, $G(y|x)$ can be estimated using Beran's estimator, see Beran (1981), defined by

$$1 - \hat{G}(y|x) = \prod_{i=1}^n \left(1 - \frac{(1 - \delta_i) \mathbf{1}_{\{T_i \leq y\}} W_{0i}(x, h_0)}{\sum_{j=1}^n \mathbf{1}_{\{T_i \leq T_j\}} W_{0j}(x, h_0)} \right), \quad (3.2.4)$$

for y less than $T_{(n)}$, the largest observation of the sample $(T_i)_{1 \leq i \leq n}$. Here, W_0 's are Nadaraya-Watson weights defined by

$$W_{0i}(x, h_0) = \frac{K_0((X_i - x)/h_0)}{\sum_{j=1}^n K_0((X_j - x)/h_0)},$$

where K_0 is a kernel function and $0 < h_0 \equiv h_{0,n}$ is a bandwidth parameter. Plugging-in Beran's estimator $\hat{G}(y|x)$ in the transformation (3.2.2), provides an estimator of the unknown transformed response Y^* , that will be denoted by $\hat{Y}^*$. Starting from the transformed data $(\hat{Y}_i^*, X_i)$, $i = 1, \dots, n$, we now apply the guided local linear method (GLL) to estimate the function $m(x) = E(\phi(Y)|X = x)$. Define

$$(\hat{b}_0, \hat{b}_1) = \arg \min_{b_0, b_1} \left[\sum_{i=1}^n \{Z_{\hat{G},i}(x, \hat{\theta}) - b_0 - b_1(X_i - x)\}^2 K\left((X_i - x)/h\right) \right],$$

where $Z_{\hat{G},i}(x, \theta) = m(x, \theta) + (\hat{Y}_i^* - m(X_i, \theta))$, $i = 1, \dots, n$. The parametrically guided local linear estimator (GLL) is given by

$$\hat{m}_{\hat{G},\hat{\theta}}(x) = \hat{b}_0 = \sum_{i=1}^n W_i(x, h) Z_{\hat{G},i}(x, \hat{\theta}), \quad (3.2.5)$$

where $W_i(x, h)$, $i = 1, \dots, n$, are defined by expression (1.3.5).

Note that if $m(x, \hat{\theta})$ is linear in x , then $\hat{m}_{\hat{G},\hat{\theta}}(x)$ reduces to the classical local linear estimator, i.e. $\sum_{i=1}^n W_i(x, h) \hat{Y}_i^*$, which means that our new estimator $\hat{m}_{\hat{G},\hat{\theta}}(x)$ is a generalization of the classical one. In principle, the parametric guide can be

obtained using any parametric technique adapted to censored data. Often, even a simple parametric guide can improve the regression estimator compared to the purely non parametric version. However, using a completely non correct parametric pilot will not improve or may even harm the quality of the nonparametric estimator. Without any prior knowledge about the structure under study, one can use any of the well known model selection criterions such as AIC adapted to censored data; see Collett (1994) and more recently Liang and Zou (2008). However, the problem of parametric model selection is not in the scope of this chapter.

In the next section we study the asymptotic properties of the proposed estimator.

3.3 Asymptotic results

The presence of censorship adds considerable complexity to the method described in the previous sections. In fact, the quantities $Z_{\hat{G},i}(x, \hat{\theta})$ in expression (3.2.5) depend on Beran's estimator $\hat{G}(y|x)$ defined by (3.2.4), which is computed from the whole sample and this makes the estimators (3.2.5) more difficult to study than (3.2.1).

To establish the asymptotic normality of $\hat{m}_{\hat{G},\hat{\theta}}(x)$ we follow the traditional approach of breaking the problem into two parts. First, we establish in Theorem 3.1 the asymptotic normality of $\tilde{m}_{\hat{G}}(x)$, an estimator of $m(x)$ constructed using a given non random approximation $\tilde{m}(x)$. Then, in Theorem 3.2 we provide sufficient conditions for the asymptotic equivalence of $\hat{m}_{\hat{G},\theta^*}(x)$ and $\hat{m}_{\hat{G},\hat{\theta}}(x)$, where θ^* is a fixed parameter that will be defined later. Throughout, c will be a positive constant that may take different values. We now provide a list of regularity conditions that are required in these theorems.

Assumption 3.1.

- (A.1) X belongs to a compact subset $S \subset \mathfrak{R}$ with marginal density $f_X(\cdot)$, which is continuously differentiable and $\inf_{x \in S} f_X(x) > 0$.
- (A.2) 1. ϕ is a bounded function that vanishes outside the interval $[0, \tau]$ for some $\tau < \inf_{x \in S} \tau_x$ with $\tau_x = \sup\{y : H(y|x) < 1\}$, i.e. the right endpoint of the support of $H(y|x) = P(T \leq y|X = x) = 1 - (1 - F(y|x))(1 - G(y|x))$.

2. The functions $H_j(y|x) = P(T \leq y, \delta = j|X = x)$, $j = 0, 1$, have four derivatives with respect to x . Furthermore, the derivatives are bounded uniformly for all $y \leq \tau$ and $x \in S$.
- (A.3) 1. $f_{Y|X}(y|x)$, the conditional density of Y given $X = x$, is differentiable in y for all $x \in S$.
2. $f_{\varepsilon^*|X}(y|x)$, the density of the synthetic errors $\varepsilon^* = Y^* - m(X)$ given $X = x$, is uniformly bounded and continuous in y for all $x \in S$.
3. For all $x \in S$, $\sigma_*^2(x) = \text{Var}(\varepsilon^*|X = x) \neq 0$ and is continuous in x .
4. $E(\phi(Y)^2) < \infty$.
- (A.4) The kernel K is a symmetric, continuously differentiable probability density function with compact support S_K , and $\int x^2 K(x) dx = \mu_K^2 < \infty$. The kernel K_0 is a symmetric, continuously differentiable function of order four with compact support S_{K_0} , $\int K_0(x) dx = 1$ and $\int x^4 K_0(x) dx = \mu_{K_0}^4 < \infty$.
- (A.5) $(\log n)^3/nh_0^3 = O(1)$, $nh_0^8 = O(1)$ and $nh^3 \rightarrow \infty$ as $n \rightarrow \infty$.
- (A.6) $m(x)$ is twice continuously differentiable.

3.3.1 The estimator with a fixed guide

Let $\tilde{m}(x)$ be any non-random parametric function that approximates $m(x)$. Define the corresponding GLL estimator as

$$\tilde{m}_{\hat{G}}(x) = \sum_{i=1}^n W_i(x, h) \tilde{Z}_{\hat{G}, i}(x),$$

where $\tilde{Z}_{\hat{G}, i}(x) = \tilde{m}(x) + (\hat{Y}_i^* - \tilde{m}(X_i))$, $i = 1, \dots, n$, and $W_i(x, h)$ are the local linear weight functions defined in (1.3.5). The following Theorem establishes the asymptotic normality of $\tilde{m}_{\hat{G}}(\cdot)$.

Theorem 3.1. *Suppose that Assumption 3.1 holds. Then,*

$$\begin{aligned} (nh)^{1/2} \left(\tilde{m}_{\hat{G}}(x) - m(x) - \tilde{B}(x) + O_p \left(\left(\frac{\log n}{nh_0} \right)^{1/2} \right) + o_p(h^2) \right) \\ \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du \right), \end{aligned}$$

with

$$\tilde{B}(x) = \frac{1}{2}h^2\mu_K^2(m^{(2)}(x) - \tilde{m}^{(2)}(x)).$$

The effect of the parametric guide appears in the expression of the bias $\tilde{B}(x)$, while the variance remains unchanged compared to the local linear estimator. The term $O_p((\frac{\log n}{nh_0})^{1/2})$ comes from the estimation of the function G by Beran's estimator and the term $o_p(h^2)$ comes from the local linear method as in the classical nonparametric approach.

3.3.2 The estimator with an estimated guide

Now, we consider the case where the parametric pilot $m(x, \hat{\theta})$ is obtained from a first stage estimation procedure. We assume that the vector of parameters $\theta \in \mathbb{R}^p$ is estimated through a least squares procedure (LS) applied to the transformed data, i.e. $\hat{\theta} = \arg \min_{\theta \in \Theta} \hat{M}_n(\theta)$ with $\hat{M}_n(\theta) = n^{-1} \sum_{i=1}^n (\hat{Y}_i^* - m(X_i, \theta))^2$. This choice is motivated by the simplicity of the (LS) method and its nice properties. In the case of independence between (Y, X) and C , even if the model $m(x, \theta)$ is incorrectly specified, under weak assumptions, Lopez and Patilea (2009) proved that $\hat{\theta}$ converges in probability to $\theta^* = \arg \min_{\theta \in \Theta} E(\phi(Y) - m(X, \theta))^2$. If the parametric model is correct then θ^* is the true parameter usually denoted by θ_0 otherwise θ^* is just a pseudo-true parameter such that $\hat{\theta}$ converges in probability to θ^* under standard conditions. In the following proposition, we extend the result of Lopez and Patilea to a more general setting where the response Y and the censoring variable C are independent given the covariate X . The proof of this proposition is postponed to the Appendix. The following additional conditions are required.

Assumption 3.2.

(B.1) *The parametric regression function $m(x, \theta)$ belongs to a parametrically indexed class defined by the following characteristics:*

1. $\theta \in \Theta$, Θ a compact subset of $\mathbb{R}^p$.
2. *The function $(x, \theta) \mapsto m(x, \theta)$ is twice continuously differentiable with respect to x and θ .*

(B.2) *There exists a unique $\theta^* = \arg \min_{\theta \in \Theta} E(\phi(Y) - m(X, \theta))^2$.*

(B.3) The matrix of second derivatives $\nabla_{\theta}^2 E(Y^* - m(X, \theta^*))^2$ is nonsingular.

Note that from equality (3.2.3) one can prove that $\theta^* = \arg \min_{\theta \in \Theta} M(\theta)$ with $M(\theta) = E(Y^* - m(X, \theta))^2$.

Proposition 3.1. *Under Assumptions 3.1 and 3.2,*

$$\hat{\theta} - \theta^* = O_p(n^{-1/2}).$$

The following Theorem is the main result of the chapter. It establishes that the guided estimator with an estimated guide $m(x, \hat{\theta})$ is asymptotically equivalent to the guided estimator with the fixed guide $m(x, \theta^*)$.

Theorem 3.2. *Suppose that Assumptions 3.1 and 3.2 hold. Then,*

$$\begin{aligned} (nh)^{1/2}(\hat{m}_{\hat{G}, \hat{\theta}}(x) - m(x) - B(x, \theta^*) + O_p\left(\left(\frac{\log n}{nh_0}\right)^{1/2}\right) + o_p(h^2)) \\ \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du\right), \end{aligned}$$

with

$$B(x, \theta^*) = \frac{1}{2} h^2 \mu_K^2 (m^{(2)}(x) - m^{(2)}(x, \theta^*)). \quad (3.3.1)$$

If $h_0(h \log n)^{-1} \rightarrow \infty$ and $nh^5 = O(1)$, then,

$$(nh)^{1/2}(\hat{m}_{\hat{G}, \hat{\theta}}(x) - m(x) - B(x, \theta^*)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du\right).$$

It is clear from expression (3.3.1) that the choice of the parametric guide $m(x, \theta)$ will have an impact on the bias of the GLL estimator. If $m(x, \theta) = \theta_1 + \theta_2 x$ is a linear guide, then $B(x, \theta^*)$ is the asymptotic bias of the local linear estimator, which means as said before that the GLL estimator is a generalization of the classical local linear estimator. On the other hand, if the parametric function $m(x, \theta)$ is twice differentiable with respect to x and $m^{(2)}(x) = m^{(2)}(x, \theta^*)$, then the asymptotic bias $B(x, \theta^*) = 0$, which means that one can increase the bandwidth to reduce the variance without increasing bias. Finally, if the second derivatives of $m(x, \theta)$ and $m(x)$ are close to each other, then the bias will be reduced in absolute value compared to that of the local linear estimator when $|m^{(2)}(x) - m^{(2)}(x, \theta^*)| < |m^{(2)}(x)|$. Note that the

expression of the asymptotic variance is equal to that of the local linear estimator, which means that with a suitable parametric guide one can reduce the bias of the local linear estimator without any increase in the variance, and thus the bias-variance problem can be bypassed. The condition $h_0(h \log n)^{-1} \rightarrow \infty$ means that the selected bandwidth for Beran's estimator must be asymptotically larger than the bandwidth of the GLL estimator. In this case ($h \ll h_0$) and under the condition $nh_0^8 = O(1)$, the asymptotic bias from the estimation of $G(y|x)$ will have no asymptotic effect on the GLL estimator. Finally, the condition $nh^5 = O(1)$ is exactly as in the classical nonparametric approach.

Remark 3.1. *In practice, choosing the bandwidth parameters is crucial. To select h_0 , the bandwidth used in the estimation of the conditional cumulative distribution $G(.|x)$, one can use, for example, the plug-in method proposed by Dabrowska (1992) or the bootstrap method developed by Van Keilegom and Veraverbeke (1997 a). As for the regression's bandwidth h , which is more important in our setting, the theoretical optimal value that minimizes the asymptotic mean integrated squared error (see Theorem 3.2) is given by*

$$h_{opt} = \left(\frac{\int \sigma_*^2(x) f_X^{-1}(x) dx \int K^2(u) du}{\mu_K^4 \int (m^{(2)}(x) - m^{(2)}(x, \theta^*))^2 dx} \right)^{1/5} n^{-1/5}.$$

Unfortunately, this expression is not directly applicable in practice, since it depends on many unknown components. In addition, the presence of censoring adds more complexity to the bandwidth selection problem. To the best of our knowledge, no consistent method has been proposed so far in the literature. In the context of nonparametric regression with censored data, limited investigation about bandwidth selection can be found in the literature. For instance, Fan and Gijbels (1994) applied the cross validation criterion to the estimated transformed data $(X_i, \hat{Y}_i^*)$, $i = 1, \dots, n$. Finally, in the spirit of the proposed method by El Ghouch and Van Keilegom (2008), one can adapt the cross validation criterion to our setting by simultaneously minimizing the function $CV(h_0, h) = n^{-1} \sum_i (\hat{m}_{\hat{G}, \hat{\theta}}(X_i) - \hat{Y}_i^*)^2$ with respect to h_0 and h .

3.3.3 Generalized parametrically guided local linear estimator

With completely observed data Martins-Filho *et al.* (2008) proposed a class of guided local linear estimators that generalizes the estimator (3.1.1). Their approach can be easily extended to censored data. The starting point is the following identity:

$$m(x) = m(x, \hat{\theta}) + r_{\hat{\theta}, \alpha}(x)m(x, \hat{\theta})^\alpha,$$

where $m(x, \hat{\theta})$ is a parametric estimator, $r_{\hat{\theta}, \alpha}(x) = (m(x) - m(x, \hat{\theta}))/m(x, \hat{\theta})^\alpha$ and $\alpha \in \mathbb{R}$. In the same way as in Martins-Filho *et al.* (2008), we propose the estimator

$$\hat{m}(x, \hat{\theta}) = m(x, \hat{\theta}) + \hat{r}_{\hat{\theta}, \alpha}(x)m(x, \hat{\theta})^\alpha,$$

where $\hat{r}_{\hat{\theta}, \alpha}(x)$ is now a nonparametric fit based on the pseudo-estimated data given by

$$(X_i, (\hat{Y}_i^* - m(X_i, \hat{\theta}))/m(X_i, \hat{\theta})^\alpha).$$

It is easy to verify that the case $\alpha = 0$ leads back to the estimator (3.1.1). The generalized parametrically guided local linear estimator is now given by

$$\hat{m}_{\hat{G}, \hat{\theta}, \alpha}(x) = \sum_{i=1}^n W_i(x, h) Z_{\hat{G}, i}(x, \hat{\theta}, \alpha),$$

where $Z_{\hat{G}, i}(x, \theta, \alpha) = m(x, \theta) + (\hat{Y}_i^* - m(X_i, \theta))\rho_{x, \theta}(X_i)$, $i = 1, \dots, n$, and $\rho_{x, \theta}(v) = (m(x, \theta)/m(v, \theta))^\alpha$.

All the results proved before can be extended to the generalized guided local linear estimator. The generalization of the proof is postponed to the Appendix. The most important result is the following theorem, which is a generalization of Theorem 3.2:

Theorem 3.3. *Suppose that Assumptions 3.1 and 3.2 hold. In addition, suppose that $m(x, \theta^*) \neq 0$. Then,*

$$\begin{aligned} (nh)^{1/2}(\hat{m}_{\hat{G}, \hat{\theta}, \alpha}(x) - m(x) - B(x, \theta^*, \alpha) + O_p\left(\left(\frac{\log n}{nh_0}\right)^{1/2}\right) + o_p(h^2)) \\ \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du\right), \end{aligned}$$

with

$$B(x, \theta^*, \alpha) = \frac{1}{2} h^2 \mu_K^2 (\Gamma_{\theta^*, \alpha}^{(2)}(x) - \Lambda_{\theta^*, \alpha}^{(2)}(x)), \quad (3.3.2)$$

and

$$\Gamma_{\theta, \alpha}(v) = m(v) \left(\frac{m(x, \theta)}{m(v, \theta)} \right)^\alpha, \quad \Lambda_{\theta, \alpha}(v) = m(v, \theta) \left(\frac{m(x, \theta)}{m(v, \theta)} \right)^\alpha.$$

Theorem 3.3 shows that, in addition to $m(x, \theta^*)$ the asymptotic bias $B(x, \theta^*, \alpha)$ is now also depending on the parameter α . In practice, the parameter α needs also to be selected in order to minimize the bias, Martins-Filho *et al.* (2008) proposed some guidelines for choosing α . Their method can be extended to censored data, however, adding another tuning parameter may not be very practical in the present context. For simplicity, in the following simulations studies we will limit ourselves to the case where $\alpha = 0$.

3.4 Simulations

In this section, we conduct some simulations in order to compare the proposed guided local linear (GLL) estimator with the fully nonparametric competitor. To achieve this goal, we consider two examples. Along the simulations we consider Epanechnikov kernel functions for K_0 and K , and select the value of the bandwidths h_0 and h by minimizing the average mean squared error (MSE). We estimate the truncated conditional mean function $m(x) = \int_0^\tau y dF(y|x)$, where $\tau = \inf_x \{\tau_x\}$ and τ_x is the 0.99 upper quantile of the distribution function $H(\cdot|x)$.

3.4.1 Exponential proportional hazards model

Along this section the following model is considered; see Bender *et al.* (2005):

$$Y_i = -\log(U_i) r(X_i), \quad i = 1 \dots, n,$$

where $r(x) = \exp(-\theta x) + \lambda(\sin(2\pi x))^2$, X_i and U_i are independent uniform variables on $[0, 1]$, and $\theta = 2$. The bandwidths are selected over a grid on $[0.1, 0.8]$ with steps of length 0.1. The censoring variable C_i is independent of the response Y_i given X_i , and is defined by $C_i = -\log(V_i) \exp(\tilde{r}(X_i))$, where $\tilde{r}(x) = \log(r(x)) + \log(b^2)$ and V_i is a

uniform variable on $[0, 1]$ independent of X_i and U_i . The parameter b allows to control the rate of censoring (RC) which is given by $RC(x) = P(Y_i > C_i | X_i = x) = (b^2 + 1)^{-1}$. The value $b = 2$ corresponds to a fixed censoring rate of 20% and $b = 1.22$ corresponds to a fixed censoring rate of 40%. Note that the model defined above is equivalent to the case where conditional on $X_i = x$, the variables Y_i and C_i are independent and exponentially distributed with rates $1/r(x)$ and $\exp(-\tilde{r}(x))$. We aim to compare the GLL estimator with the fully nonparametric estimator. For the parametric guide we consider an exponential proportional hazards model (PPH) $m(x, \theta) = \exp(-\theta x)$ and for the nonparametric one we use the local linear (LL) estimator. The parameter $\lambda \geq 0$ in the expression of $r(x)$ allows to control the deviation of the parametric guide from the true regression function. The case where $\lambda = 0$ corresponds to the situation where the parametric guide is correct, which is the ideal situation for the GLL estimator. Note that this case corresponds to the situation where the response $Y_i \sim \exp(1)$ is the survival time of an exponential proportional hazards model; see Bender *et al.* (2005). If $\lambda \neq 0$ then the parametric pilot is incorrect and deviates increasingly from the true model when the value of λ increases. We consider five values of λ : 0, 0.04, 0.2, 1, 2, the graphs of the truncated conditional mean are plotted in Figure 3.4.1 for each value of λ . We compare the performance of the estimators in the situations of: correct, approximately correct and incorrect parametric pilots. We conducted $N = 200$ simulations for samples of size $n = 100$. For each model and for each set of data, we evaluate the different estimators at 30 equally spaced points, x_i , taken from 0 to 1. Let $\hat{m}_{\hat{G}, \hat{\theta}}^k(x_i)$ be the GLL estimator for the k^{th} replication. At every data point $x_i, i = 1, \dots, 30$, we approximate the bias and the variance by

$$B(x_i) = N^{-1} \sum_{k=1}^N [\hat{m}_{\hat{G}, \hat{\theta}}^k(x_i) - m(x_i)],$$

$$S^2(x_i) = N^{-1} \sum_{k=1}^N [\hat{m}_{\hat{G}, \hat{\theta}}^k(x_i) - N^{-1} \sum_{k=1}^N \hat{m}_{\hat{G}, \hat{\theta}}^k(x_i)]^2.$$

Let $Bias^2 = 30^{-1} \sum_{i=1}^{30} B^2(x_i)$, $Var = 30^{-1} \sum_{i=1}^{30} S^2(x_i)$ and $MSE = Bias^2 + Var$ be the average squared bias, the average variance and the average mean squared error of the estimates, respectively. The results are summarized in Table 3.4.1 and show that with a correct parametric (PPH) guide ($\lambda = 0$) we get the best results for the

GLL estimator, the observed bias of GLL estimator is approximately 10^{-1} times that of the LL estimator. With roughly correct parametric guide ($\lambda = 0.04$), we observe that the GLL estimator remains as good as before and the observed bias is reduced (by a factor 10) compared to the LL estimator. For $\lambda = 0.2$ the bias is still reduced by a factor 2 compared to the LL estimator. In the two last cases the parametric guide is incorrect, the bias of the GLL estimator is very close to that of the LL estimator. Regarding the variance, as expected, the LL and GLL estimator behave similarly. The MSE is not obviously reduced, this is because in this case the variance dominates the bias. Finally, it seems that increasing the censoring rate does not harm the performance of the GLL estimator, since the estimated transformed data depend on Beran's estimator $\hat{G}$ of which the performance increases with the rate of censorship.

Censoring rate		20%				40%			
λ	Method	$Bias^2$	Var	MSE	MSE-Ratio	$Bias^2$	Var	MSE	MSE-Ratio
0	GLL	0.082	3.489	3.572	0.864	0.431	4.678	5.109	0.867
	LL	0.556	3.578	4.134		1.179	4.709	5.888	
0.04	GLL	0.476	7.607	8.083	0.809	0.403	4.944	5.347	0.891
	LL	2.242	7.740	9.982		1.034	4.969	6.004	
0.2	GLL	2.061	13.259	15.320	0.874	1.934	11.727	13.661	0.905
	LL	4.049	13.480	17.530		3.358	11.728	15.086	
1	GLL	10.879	28.916	39.795	0.978	9.384	16.714	26.098	0.981
	LL	11.772	28.902	40.674		10.034	16.562	26.596	
2	GLL	25.541	26.865	52.406	0.993	27.379	23.691	51.070	0.997
	LL	25.950	26.838	52.789		27.815	23.395	51.210	

Table 3.4.1: Average squared bias ($Bias^2 \times 10^4$), average variance ($Var \times 10^4$), average MSE ($\times 10^4$) and the ratio of the average MSE ($AMSE_{GLL}/AMSE_{LL}$) for $\lambda = 0, 0.04, 0.2, 1, 2$, two censoring rates (20% and 40%), sample size $n = 100$, and $N = 200$ replications.

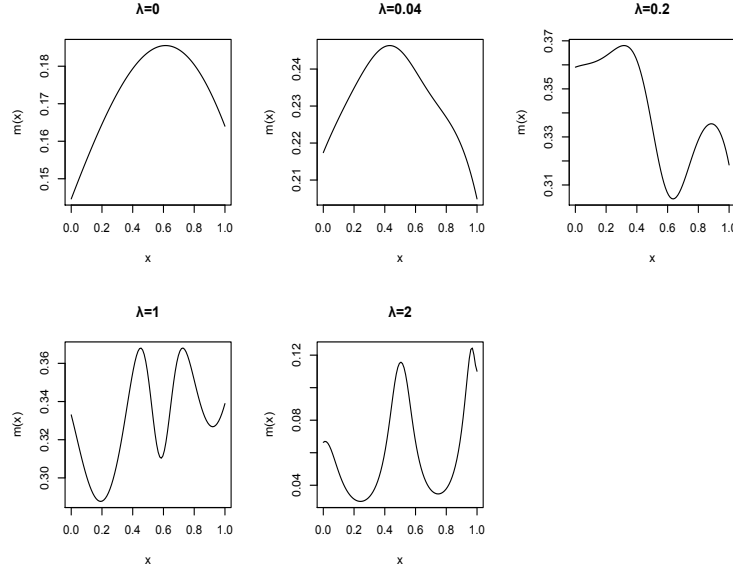


Figure 3.4.1: Truncated conditional mean for different values of λ .

3.4.2 Sinusoidal model

The data are generated from the following model:

$$Y_i = \sin(2\pi X_i) + \varepsilon_i,$$

where X_i is drawn from a uniform density on $[0, 1]$ and ε_i is normally distributed with mean 0 and variance 1. The bandwidths are selected over grid on $[0.2, 0.8]$ with steps of length 0.1. The censoring variable C_i is defined as $C_i|_{X_i=x} \sim N(\mu(x), 1)$ with $\mu(x) = \sin(2\pi x) + (a_1 + a_2 x)^2$. The variables X_i and ε_i are independent. As before, the parameters a_1 and a_2 control the rate of censoring (RC), which is given by $RC(x) = 1 - \Phi((a_1 + a_2 x)^2 / \sqrt{2})$, where Φ is the cumulative distribution function of the standard normal variable. We selected $a_1 = 0.84$, $a_2 = 0$. This leads to a constant rate of censoring $p = 30\%$. We present results from a sinusoidal regression curve with three various forms for the parametric guide. The first guide is a linear model $m(x, \theta) = \theta_0 + \theta_1 x$, the second guide is a cubic model $m(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$ and the last guide is guessed correctly and belongs to the true parametric family $m(x, \theta) = \theta \sin(2\pi x)$. The parameters $\hat{\theta}$ are obtained using least squares estimation. We carried out the simulations for various sample sizes (25, 50, 100, 200) and for 30

equidistant data points in $[0, 1]$. The results are obtained by using 200 simulations and are presented in Table 3.4.2. As expected in Section 3.2, with a linear guide, the GLL estimator (denoted L-GLL) behaves exactly as the LL estimator, and is therefore only calculated for $n = 25$. Regarding the bias, the GLL estimator with cubic (C-GLL) or sinus guide (S-GLL) outperforms the LL estimator even for small sample sizes. It is clear from the results that the bias is substantially reduced for all sample sizes. The variance decreases when the sample size increases, which confirms the asymptotic results of the previous section. The MSE is also reduced compared to the local linear estimator, except for the GLL estimator with cubic guide, which has a slightly larger variance especially with few data. Finally, we note that throughout the simulations the bias is generally smaller when the bandwidth of Beran's estimator is larger than the bandwidth used for the LL estimator or the GLL estimator ($h < h_0$), which is in concordance with the results of Theorem 3.2. Figure 3.4.2 illustrates this point. We also notice that the bandwidths h_0 , used in Beran's estimator, and h , used in the local linear approximation, do not affect the results in the same manner. In fact, h seems to have a more significant effect on the MSE than h_0 . Figure 3.4.3 illustrates this remark.

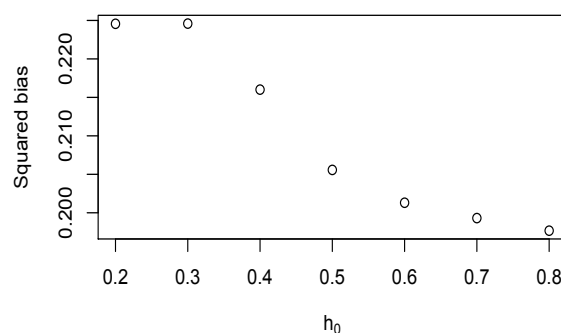


Figure 3.4.2: Squared bias ($\times 10^2$) of the C-GLL estimator, for $n = 200$, $B = 200$, $h = 0.1$ and h_0 varies from 0.2 to 0.8.

n	Method	$Bias^2$	Var	MSE	MSE-Ratio
25	L-GLL	2.751	4.057	6.808	1.000
	C-GLL	0.621	5.589	6.210	0.912
	S-GLL	0.466	3.756	4.223	0.620
	LL	2.751	4.057	6.808	
50	C-GLL	0.503	2.617	3.121	0.670
	S-GLL	0.444	1.979	2.423	0.520
	LL	2.778	1.880	4.658	
100	C-GLL	0.586	1.342	1.929	0.457
	S-GLL	0.498	1.065	1.564	0.370
	LL	3.252	0.972	4.224	
200	C-GLL	0.540	0.728	1.269	0.354
	S-GLL	0.449	0.545	0.994	0.277
	LL	3.118	0.470	3.588	

Table 3.4.2: Average squared bias ($Bias^2 \times 10^2$), average variance ($Var \times 10^2$), average MSE ($\times 10^2$) and the ratio of the average MSE ($AMSE_{GLL}/AMSE_{LL}$) for samples of size $n = 25, 50, 100, 200$, with a censoring rate $p = 30\%$, and $N = 200$ replications.

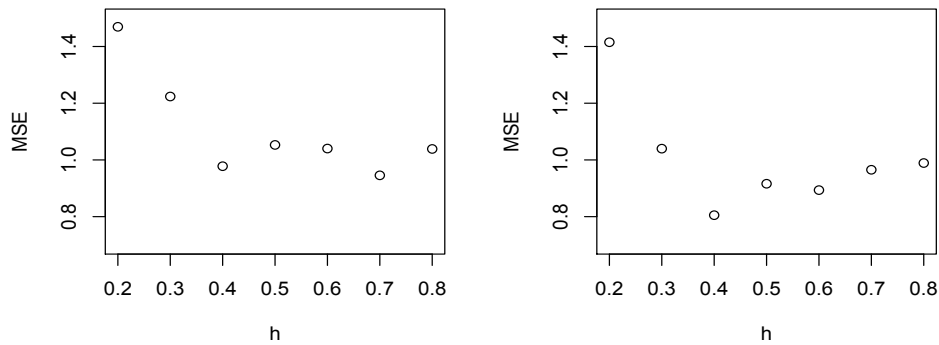


Figure 3.4.3: MSE ($\times 10^2$) of the S-GLL estimator, for $n = 200$, $B = 200$, h varies from 0.2 to 0.8 and two values of h_0 : $h_0 = 0.3$ on the left and $h_0 = 0.7$ on the right.

3.5 Real data analysis

In this section, we apply the proposed methodology to the UIS dataset from the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study. As stated in the previous chapters, one of the main objectives of this study is to model (Y) time to return to drug use (in days) for 628 patients enrolled in two different residential treatment programs that differed in length. The censoring rate is about 19.12%. Among the three available continuous covariables, we chose (X) the length of treatment measured from admission (in days). In the following, we investigate the effect of the length of treatment on the time to return to drug use. To select h_0 , the bandwidth used in Beran's estimator, we compared different values of h_0 and chose $h_0 = 120$. This heuristic choice is motivated by the results of our simulations which state that the bandwidth h_0 used in Beran's estimator seems to have no significant effect on the performance of our estimator. Regarding the bandwidths used in local linear approximation, which is more crucial in this setting, we use the least squares cross validation adapted to censored data. Thus, we minimize with respect to h the functions $CV_T(h) = n^{-1} \sum_i (\hat{m}_{h,-i}(X_i) - \hat{Y}_i^*)^2$ for the classical LL estimator and $CV_G(h) = n^{-1} \sum_i (\hat{m}_{\hat{G}, \hat{\theta}, h, -i}(X_i) - \hat{Y}_i^*)^2$ for the GLL estimator, where $\hat{m}_{h,-i}(\cdot)$ and $\hat{m}_{\hat{G}, \hat{\theta}, h, -i}(\cdot)$ are the leave-one-out LL and GLL estimators, respectively. The selected bandwidths are $h = 55.6$ for the classical LL estimator and $h = 110$ for the GLL estimator.

We apply both the classical LL and the GLL regression to estimate the truncated conditional mean function $E(\phi(Y)|X = x)$ with $\phi(t) = t \cdot \mathbf{I}(t \leq \tau)$ and $\tau = 805$ is the right endpoint of the support of the empirical conditional distribution function $H_n(t|x)$. As a guide, we chose a cubic polynomial for which the parameters are estimated using least squares estimation based on the synthetic data. The resulting LL and GLL estimators are given in Figure 3.5.1. Based on the fitted model, it seems that longer duration of treatment increases the time to return to drug use. While the LL and the GLL competitors seem to be quite close, there is a non negligible correction of the boundary effect on the right tail by the GLL estimator compared with the LL estimator.

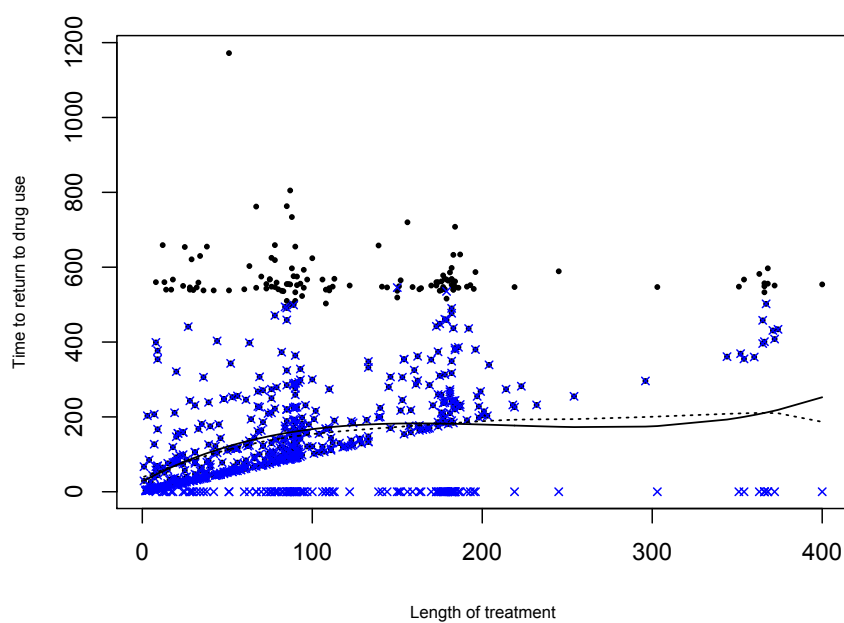


Figure 3.5.1: Estimation of the truncated conditional mean, by means of the LL estimator (dotted curve) and the GLL estimator (solid curve). Synthetic data are indicated by $\times$.

3.6 Appendix

Proof of Theorem 3.1. We start with the case when G is known. The guided local linear estimator based on the non random parametric model $\tilde{m}(x)$ and transformation (3.2.2) is $\tilde{m}_G(x) = \sum_{i=1}^n W_i(x, h) \tilde{Z}_{G,i}(x)$, where $\tilde{Z}_{G,i}(x) = \tilde{m}(x) + (Y_i^* - \tilde{m}(X_i))$, $i = 1, \dots, n$. We have the following result.

Lemma 3.1. *Suppose that the assumptions of Theorem 3.1 hold. Then,*

$$\sqrt{nh} \left(\tilde{m}_G(x) - m(x) - \tilde{B}(x) + o_p(h^2) \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du \right), \quad (3.6.1)$$

where $\tilde{B}(x) = \frac{1}{2} h^2 \mu_K^2 (m^{(2)}(x) - \tilde{m}^{(2)}(x))$.

Proof. Write

$$\begin{aligned} \tilde{m}_G(x) - m(x) &= \left[\sum_{i=1}^n W_i(x, h) Y_i^* - m(x) \right] - \left[\sum_{i=1}^n W_i(x, h) (\tilde{m}(X_i) - \tilde{m}(x)) \right] \\ &= I_{1,n}(x) - I_{2,n}(x). \end{aligned}$$

From Theorem 5.2 in Fan and Gijbels (1996) it follows that

$$\sqrt{nh} \left(I_{1,n}(x) - \frac{h^2}{2} m^{(2)}(x) \mu_K^2 + o(h^2) \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du \right). \quad (3.6.2)$$

Since $\sum_{i=1}^n W_i(x, h)(X_i - x) = 0$, write $I_{2,n}(x) = \sum_{i=1}^n W_i(x, h) R(X_i)$ where

$$R(X_i) = \tilde{m}(X_i) - \tilde{m}(x) - \tilde{m}^{(1)}(x)(X_i - x).$$

We have

$$\begin{aligned} I_{2,n}(x) &= \frac{s_{n,2}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}(x)^2} \left[\sum_{i=1}^n K \left(\frac{X_i - x}{h} \right) R(X_i) \right. \\ &\quad \left. - \sum_{i=1}^n K \left(\frac{X_i - x}{h} \right) R(X_i)(X_i - x) \right]. \end{aligned}$$

Using Lemma 4 in Fan and Gijbels (1992), we get

$$\begin{aligned} \sum_{i=1}^n K \left(\frac{X_i - x}{h} \right) R(X_i) &= \frac{n}{2} h^3 \tilde{m}^{(2)}(x) f_X(x) \mu_K^2 (1 + o_p(1)), \text{ and} \\ \sum_{i=1}^n K \left(\frac{X_i - x}{h} \right) R(X_i)(X_i - x) &= o_p(1). \end{aligned}$$

Hence,

$$I_{2,n}(x) = \frac{h^2}{2} \tilde{m}^{(2)}(x) \mu_K^2 + o_p(h^2).$$

The result of Lemma 3.1 now follows from this last expression and (3.6.2). $\square$

Now we turn to the case when G is unknown.

Lemma 3.2. *Suppose that the assumptions of Theorem 3.1 hold. Then,*

$$\sup_{x \in S} |\tilde{m}_{\hat{G}}(x) - \tilde{m}_G(x)| = O_p \left(\sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)| \right).$$

Proof. Write $|\tilde{m}_{\hat{G}}(x) - \tilde{m}_G(x)| \leq \max_i |\hat{Y}_i^* - Y_i^*| \sum_{i=1}^n |W_i(x, h)|$. Lemma 4 in and Gijbels (1992) yields that $\sum_{i=1}^n |W_i(x, h)| = O_p(1)$ and we have

$$\max_i |\hat{Y}_i^* - Y_i^*| \leq O_p(1) \times \sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)|. \quad (3.6.3)$$

The result of Lemma 3.2 now follows from (3.6.3). $\square$

Now return to the proof of Theorem 3.1. From Proposition 4.3 in Van Keilegom and Akritas (1999), it follows that if $nh_0^9(\log n)^{-1} = O(1)$, then,

$$\sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)| = O_p((nh_0)^{-1/2}(\log n)^{1/2}), \quad (3.6.4)$$

(note that their condition $nh_0^5(\log n)^{-1} = O(1)$ is replaced by $nh_0^9(\log n)^{-1} = O(1)$ since we work with a kernel of order 4), and from Lemma 3.2 we have,

$$\sqrt{nh}(\tilde{m}_{\hat{G}}(x) - \tilde{m}_G(x)) = O_p \left(\left(\frac{h \log n}{h_0} \right)^{1/2} \right). \quad (3.6.5)$$

The result of Theorem 3.1 is now a direct consequence of Lemma 6.1 and equation (3.6.5). $\square$

For the proof of Proposition 3.1, the following Lemmas are needed.

Lemma 3.3. *Under the assumptions of Proposition 3.1, we have,*

$$\hat{\theta} - \theta^* = o_p(1).$$

Proof. Define the following functions:

$M(\theta) = E(Y^* - m(X, \theta))^2$, $M_n(\theta) = n^{-1} \sum_{i=1}^n (Y_i^* - m(X_i, \theta))^2$ and $\hat{M}_n(\theta) = n^{-1} \sum_{i=1}^n (\hat{Y}_i^* - m(X_i, \theta))^2$. To prove the result of Lemma 3.3 we have to show that $\sup_{\theta \in \Theta} |\hat{M}_n(\theta) - M(\theta)| = o_p(1)$. Note that $|\hat{M}_n(\theta) - M(\theta)| \leq |\hat{M}_n(\theta) - M_n(\theta)| + |M_n(\theta) - M(\theta)|$. First consider

$$\begin{aligned} |\hat{M}_n(\theta) - M_n(\theta)| &\leq \left| n^{-1} \sum_{i=1}^n (\hat{Y}_i^{*2} - Y_i^{*2}) \right| + 2 \left| n^{-1} \sum_{i=1}^n (\hat{Y}_i^* - Y_i^*) m(X_i, \theta) \right| \\ &:= |A_{1,n}| + 2|A_{2,n}(\theta)|. \end{aligned}$$

Then,

$$\begin{aligned} |A_{1,n}| &\leq O_p(1) \times \sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)| \times \sup_{t \leq \tau, x \in S} |(1 - \hat{G}(t^-|x)) + (1 - G(t|x))| \\ &\quad \times \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \phi(T_i)^2}{(1 - G(T_i|X_i))^4}. \end{aligned}$$

The empirical sum converges almost surely to $E(\phi(Y)^2[1 - G(Y|X)]^{-3})$ (which is finite by assumption (A.3)) by the strong law of large numbers, the first supremum tends to zero in probability by Proposition 4.3 in Van Keilegom and Akritas (1999) and the second supremum is bounded by 2. Hence, $|A_{1,n}| = o_p(1)$. Now consider

$$\begin{aligned} \sup_{\theta \in \Theta} |A_{2,n}(\theta)| &\leq O_p(1) \times \sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)| \times \sup_{x \in S, \theta \in \Theta} |m(x, \theta)| \\ &\quad \times \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |\phi(T_i)|}{(1 - G(T_i|X_i))^2}. \end{aligned}$$

The empirical sum converges almost surely to $E(|\phi(Y)|[1 - G(Y|X)]^{-1})$ (which is finite by assumption (A.2)), the first supremum tends to zero in probability and the second supremum is bounded by a constant. Hence, $\sup_{\theta \in \Theta} |A_{2,n}(\theta)| = o_p(1)$, and $\sup_{\theta \in \Theta} |\hat{M}_n(\theta) - M_n(\theta)| = o_p(1)$.

Now, under assumptions (A.3) and (B.1) and by Lemma 2.4 in Newey and McFadden (1994), we have $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = o_p(1)$. Then the result of Lemma 3.3 is a consequence of Theorem 2.1 in Newey and McFadden (1994). $\square$

Define the following functions:

$$\begin{aligned} D(\theta) &= \nabla_{\theta} M(\theta), & H(\theta) &= \nabla_{\theta}^2 M(\theta), \\ D_n(\theta) &= \nabla_{\theta} M_n(\theta), & H_n(\theta) &= \nabla_{\theta}^2 M_n(\theta), \\ \hat{D}_n(\theta) &= \nabla_{\theta} \hat{M}_n(\theta), & \hat{H}_n(\theta) &= \nabla_{\theta}^2 \hat{M}_n(\theta), \end{aligned}$$

where $\nabla_{\theta} f(x, \theta) = \partial f(x, \theta) / \partial \theta$ and $\nabla_{\theta}^2 f(x, \theta) = \partial^2 f(x, \theta) / \partial \theta \partial \theta'$ for a twice differentiable function $\theta \rightarrow f(x, \theta)$. For any matrix L let $\|L\|_2$ denote the 2-norm of L , that is $\|L\|_2 = \sup_{u \neq 0} \|Lu\| / \|u\|$, where $\|u\|$ is the Euclidean norm of the vector u .

Lemma 3.4. *Under the assumptions of Proposition 3.1, we have,*

1. $\sup_{\theta \in \Theta} \|\hat{H}_n(\theta) - H_n(\theta)\|_2 = o_p(1)$.
2. $\sup_{\theta \in \Theta} \|H_n(\theta) - H(\theta)\|_2 = o_p(1)$.
3. $\hat{D}_n(\theta^*) - D_n(\theta^*) = O_p(n^{-1/2})$.
4. $D_n(\theta^*) = O_p(n^{-1/2})$.

Proof.

1. We have

$$\begin{aligned} \sup_{\theta \in \Theta} \|\hat{H}_n(\theta) - H_n(\theta)\|_2 &\leq O_p(1) \times \sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)| \\ &\quad \times \sup_{x \in S, \theta \in \Theta} \|\nabla_{\theta}^2 m(x, \theta)\|_2 \times \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |\phi(T_i)|}{(1 - G(T_i|X_i))^2}. \end{aligned}$$

The first supremum converges in probability to zero, the second supremum is bounded by assumption (B.1) and the empirical sum converges almost surely. Hence, $\sup_{\theta \in \Theta} \|\hat{H}_n(\theta) - H_n(\theta)\|_2 = o_p(1)$.

2. The proof of the second point is a consequence of Lemma 2.4 in Newey and McFadden (1994), under assumption 3.2.

3. Using the uniform rate of convergence of Beran's estimator $\hat{G}$, observe that

$$\begin{aligned} \hat{D}_n(\theta^*) - D_n(\theta^*) &= -2n^{-1} \sum_{i=1}^n \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) (\hat{G}(T_i^-|X_i) - G(T_i|X_i))}{(1 - G(T_i|X_i))^2} \\ &\quad + O_p\left(\frac{\log n}{nh_0}\right). \end{aligned}$$

Now, we consider the i.i.d. representation of $\hat{G}(t|x)$ given in Du and Akritas (2006):

$$\hat{G}(t^-|x) - G(t|x) = \sum_{j=1}^n W_{0j}(x, h_0) \eta_j(t^-, x) + O_p\left(\left(\frac{\log n}{nh_0}\right)^{3/4}\right), \quad (3.6.6)$$

where

$$\eta_j(t, x) = (1 - G(t|x)) \left[\frac{I(T_j \leq t, \delta_j = 0)}{1 - H(T_j|x)} - \int_0^t \frac{I(T_j \geq s)}{(1 - H(s|x))^2} dH_0(s|x) \right],$$

and where as before the usual condition $nh_0^5(\log n)^{-1} = O(1)$ is replaced by $nh_0^9(\log n)^{-1} = O(1)$, since we work with a kernel of order of 4. Then, with $\hat{f}_X(x) = \frac{1}{nh_0} \sum_{k=1}^n K_0\left(\frac{X_k - x}{h_0}\right)$,

$$\begin{aligned} \hat{D}_n(\theta^*) - D_n(\theta^*) &= -2n^{-1} \sum_{i,j} \frac{W_{0j}(X_i, h_0) \delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_j(T_i^-, X_i)}{(1 - G(T_i|X_i))^2} \\ &\quad + O_p((\log n / nh_0)^{3/4}), \\ &= \frac{-2}{n^2 h_0} \sum_{i,j} K_0\left(\frac{X_j - X_i}{h_0}\right) \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_j(T_i^-, X_i)}{(1 - G(T_i|X_i))^2 \hat{f}_X(X_i)} \\ &\quad + O_p(n^{-1/2}), \end{aligned}$$

provided that $(\log n)^3 / nh_0^3 = O(1)$ and $nh_0^8 = O(1)$. The main term can be decomposed into $U_{1n} + U_{2n} + U_{3n}$ with

$$\begin{aligned} U_{1n} &= \frac{-2}{n^2 h_0} \sum_{i,j} K_0\left(\frac{X_j - X_i}{h_0}\right) \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_j(T_i^-, X_i)}{(1 - G(T_i|X_i))^2 f_X(X_i)} \\ U_{2n} &= \frac{2}{n^2 h_0} \sum_{i,j} \left\{ K_0\left(\frac{X_j - X_i}{h_0}\right) \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_j(T_i^-, X_i)}{(1 - G(T_i|X_i))^2} \right. \\ &\quad \left. \times \frac{(\hat{f}_X(X_i) - f_X(X_i))}{f_X^2(X_i)} \right\} \\ U_{3n} &= \frac{-2}{n^2 h_0} \sum_{i,j} \left\{ K_0\left(\frac{X_j - X_i}{h_0}\right) \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_j(T_i^-, X_i)}{(1 - G(T_i|X_i))^2} \right. \\ &\quad \left. \times \frac{(\hat{f}_X(X_i) - f_X(X_i))^2}{\hat{f}_X(X_i) f_X^2(X_i)} \right\} \end{aligned}$$

We treat each term separately. Observe that U_{1n} can be written as

$$U_{1n} = -\frac{2}{n^2} \sum_{i=1}^n h_{ii} - \frac{2}{n^2} \sum_{i<j} \psi_n(Z_i, Z_j) = U_{11n} + U_{12n},$$

where $Z_i = (X_i, T_i, \delta_i)$, $\psi_n(Z_i, Z_j) = h_{ij} + h_{ji}$ is a symmetric kernel function and

$$h_{ij} = \frac{1}{h_0} K_0 \left(\frac{X_j - X_i}{h_0} \right) \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_j(T_i^-, X_i)}{(1 - G(T_i|X_i))^2 f_X(X_i)}.$$

We start with U_{11n} . We have

$$\begin{aligned} U_{11n} &= \frac{-2}{n^2 h_0} K_0(0) \sum_{i=1}^n \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*) \eta_i(T_i^-, X_i)}{(1 - G(T_i|X_i))^2 f_X(X_i)} \\ &= O_p((nh_0)^{-1}) = O_p(n^{-1/2}), \end{aligned}$$

since $nh_0^2 \rightarrow \infty$. Now, we treat the second term U_{12n} :

$$U_{12n} = \frac{-2}{n^2} \sum_{i<j} \psi_n(Z_i, Z_j) = -\frac{(n-1)}{n} \tilde{U}_{12n},$$

where $\tilde{U}_{12n} = \frac{2}{n(n-1)} \sum_{i<j} \psi_n(Z_i, Z_j)$ is a second order U-statistic. By Theorem 1 in Yao and Martins-Filho (2015),

$$\tilde{U}_{12n} = \theta_n + O_p((n^{-1} \sigma_{1n}^2)^{1/2}) + O_p((n^{-2} \sigma_{2n}^2)^{1/2}), \quad (3.6.7)$$

where $\theta_n = E(\psi_n(Z_i, Z_j))$, $\sigma_{1n}^2 = \text{Var} \left[E(\psi_n(Z_i, Z_j) | Z_i) \right]$ and $\sigma_{2n}^2 = \text{Var} [\psi_n(Z_i, Z_j)]$.

Next, write

$$\theta_n = 2E \left[\frac{1}{h_0} \frac{\delta_i \phi(T_i) \nabla_{\theta} m(X_i, \theta^*)}{(1 - G(T_i|X_i))^2 f_X(X_i)} E \left\{ K_0 \left(\frac{X_j - X_i}{h_0} \right) \eta_j(T_i^-, X_i) \middle| Z_i \right\} \right].$$

Now consider

$$E \left[K_0 \left(\frac{X_j - x}{h_0} \right) \eta_j(t^-, x) \right] := E \left[K_0 \left(\frac{X_j - x}{h_0} \right) r_{t,x}(X_j) \right],$$

where $r_{t,x}(X_j) = E(\eta_j(t^-, x)|X_j)$. Since $r_{t,x}(x) = E(\eta_j(t^-, x)|X_j = x) = 0$ and using Taylor's Theorem we have, for all $x \in S$ and $t \leq \tau$,

$$\begin{aligned} E\left[K_0\left(\frac{X_j - x}{h_0}\right)\eta_j(t^-, x)\right] &= E\left[K_0\left(\frac{X_j - x}{h_0}\right)\left[(X_j - x)r'_{t,x}(x) \right. \right. \\ &\quad \left. \left. + \frac{1}{2}(X_j - x)^2 r''_{t,x}(x) + \frac{1}{3!}(X_j - x)^3 r^{(3)}_{t,x}(x) \right. \right. \\ &\quad \left. \left. + \frac{1}{4!}(X_j - x)^4 r^{(4)}_{t,x}(\xi_j)\right]\right] \\ &\leq \frac{\mu_{K_0}^4}{4!} h_0^5 \sup_{x \in S, t \leq \tau, u \in V_x} |r^{(4)}_{t,x}(u)| \\ &= O(h_0^5), \end{aligned}$$

uniformly in x and t , where V_x is a neighborhood of x and ξ_j is an intermediate point between x and X_j . Hence, $\theta_n = O(h_0^4) = O(n^{-1/2})$, since $nh_0^8 = O(1)$. We now consider the term σ_{1n}^2 . We have the following

$$\sigma_{1n}^2 \leq 2 \left\{ E[E(h_{ij}|Z_i)^2] + E[E(h_{ji}|Z_i)^2] \right\},$$

and

$$\begin{aligned} E[E(h_{ij}|Z_i)^2] &= E\left[\frac{1}{h_0^2} \frac{\delta_i \phi^2(T_i) \nabla_{\theta} m(X_i, \theta^*)^2}{(1 - G(T_i|X_i))^4 f_X^2(X_i)} \right. \\ &\quad \left. \times \left\{ E\left(K_0\left(\frac{X_j - X_i}{h_0}\right)\eta_j(T_i^-, X_i) \middle| Z_i\right) \right\}^2 \right] \\ &= O(h_0^8). \end{aligned}$$

Hence, $E[E(h_{ij}|Z_i)^2] = O(1)$. Moreover,

$$E[E(h_{ji}|Z_i)^2] \leq cE\left[E\left(\frac{1}{h_0} K_0\left(\frac{X_i - X_j}{h_0}\right) \middle| Z_i\right)^2\right] = O(1).$$

From these two equations, it follows that

$$\sigma_{1n}^2 = O(1). \tag{3.6.8}$$

Now consider the second term σ_{2n}^2 . By Lyapunov's inequality,

$$\sigma_{2n}^2 \leq E(\psi(Z_i, Z_j)^2) \leq 4E(h_{ij}^2),$$

and

$$\begin{aligned}
h_0 E(h_{ij}^2) &= E\left(\frac{1}{h_0} K_0^2\left(\frac{X_j - X_i}{h_0}\right) \frac{\delta_i \phi^2(T_i) \nabla_{\theta} m(X_i, \theta^*)^2 \eta_j^2(T_i^-, X_i)}{(1 - G(T_i|X_i))^4 f_X^2(X_i)}\right) \\
&\leq c E\left(\frac{1}{h_0} K_0^2\left(\frac{X_j - X_i}{h_0}\right) \frac{1}{f_X^2(X_i)}\right) \\
&\longrightarrow c \int K_0^2(u) du \text{ as } n \rightarrow \infty.
\end{aligned}$$

Hence,

$$\sigma_{2n}^2 = O(h_0^{-1}). \quad (3.6.9)$$

From equations (3.6.7), (3.6.8) and (3.6.9) we finally get $U_{1n} = O_p(n^{-1/2})$ under the conditions $nh_0^2 \rightarrow \infty$ and $nh_0^8 = O(1)$. Similarly, we can prove that the second term $U_{2n} = O_p(n^{-1/2})$ using a third order U-statistic and the condition $(\log n)^3/nh_0^3 = O(1)$; see Martins-Filho and Yao (2006) and Yao and Martins-Filho (2015).

Finally, we consider the term U_{3n} . Using the uniform rate of convergence of $\hat{f}$; see Einmahl and Mason (2005), for a constant $c > 0$,

$$\begin{aligned}
|U_{3n}| &\leq \frac{c}{n^2 h_0} \sup_{x \in S} \left| \frac{(\hat{f}_X(x) - f_X(x))^2}{\hat{f}_X(x)} \right| \left| \sum_{i,j} K_0\left(\frac{X_j - X_i}{h_0}\right) \phi(T_i) \right| \\
&= O_p((nh_0)^{-1} \log n) = O_p(n^{-1/2}).
\end{aligned}$$

4. This is obvious, since

$$D_n(\theta^*) = n^{-1} \sum_{i=1}^n \zeta(\delta_i, T_i, X_i, \theta^*) = O_p(n^{-1/2}),$$

where $\zeta(\delta_i, T_i, X_i, \theta^*) = 2(m(X_i, \theta^*) - Y_i^*) \nabla_{\theta} m(X_i, \theta^*)$, $i = 1, \dots, n$, are i.i.d. and have zero mean. $\square$

Proof of Proposition 3.1. This proposition is an immediate consequence of Lemmas 3.3 and 3.4, and of Theorem 3.1 in Newey and McFadden (1999). $\square$

Proof of Theorem 3.2. Consider the following decomposition:

$$(nh)^{1/2}(\hat{m}_{\hat{G}, \hat{\theta}}(x) - m(x)) = (nh)^{1/2}(\hat{m}_{\hat{G}, \hat{\theta}}(x) - \tilde{m}_{\hat{G}, \theta^*}(x)) + (nh)^{1/2}(\tilde{m}_{\hat{G}, \theta^*}(x) - m(x)),$$

where $\tilde{m}_{\hat{G},\theta^*}(x)$ is the guided local linear estimator of $m(x)$ using the nonrandom function $\tilde{m}(x, \theta^*)$. The second term in the sum is asymptotically normal by Theorem 3.1. To prove the result of Theorem 3.2 it is sufficient to establish the following:

$$B_n(x) = (nh)^{1/2}(\hat{m}_{\hat{G},\hat{\theta}}(x) - \tilde{m}_{\hat{G},\theta^*}(x)) = o_p(1).$$

Since $\sum_{i=1}^n W_i(x, h) = 1$, we have

$$\begin{aligned} B_n(x) &= \frac{(nh)^{1/2}s_{n,2}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) q(X_i) \\ &\quad - \frac{(nh)^{1/2}s_{n,1}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) q(X_i)(X_i - x) \\ &\quad - (nh)^{1/2}q(x) \\ &= q_{n,1}(x) - q_{n,2}(x) - q_{n,3}(x), \end{aligned}$$

where $q(x) = m(x, \theta^*) - m(x, \hat{\theta})$. Now we treat each term $q_{n,i}(x)$, $i = 1, \dots, 3$, separately. By Taylor's Theorem we have $|m(x, \hat{\theta}) - m(x, \theta^*)| = \|\nabla_{\theta} m(x, \theta_m)\| \|\hat{\theta} - \theta^*\|$, $\theta_m = \lambda\hat{\theta} - (1 - \lambda)\theta^*$, where $\lambda \in [0, 1]$. Since $\nabla_{\theta} m(x, \theta_m)$ is uniformly bounded by condition (B.1), there exists a constant $c > 0$ such that for all $\theta \in \Theta$ and $x \in S$ we have $|m(x, \hat{\theta}) - m(x, \theta^*)| \leq c|\hat{\theta} - \theta^*| = O_p(n^{-1/2})$ by Lemma 3.4. Therefore, $q(x) = O_p(n^{-1/2})$ and $q_{n,3}(x) = h^{1/2}O_p(1) = o_p(1)$. From Lemma 4 in Fan and Gijbels (1992), we have for $k = 1, 2$, $q_{n,k}(x) = h^{1/2}O_p(1) = o_p(1)$. This concludes the proof of the first part of Theorem 3.2. The second part is a direct consequence of the first part under the conditions $h_0(h \log n)^{-1} \rightarrow \infty$ and $nh^5 = O(1)$. $\square$

Proof of Theorem 3.3. To establish the asymptotic normality of $\hat{m}_{\hat{G},\hat{\theta},\alpha}(x)$ we also break the problem into two parts. First, we establish in the Proposition 3.2 the asymptotic normality of $\tilde{m}_{\hat{G},\alpha}(x)$, an estimator of $m(x)$ constructed using a given non random approximation $\tilde{m}(x)$. Then, Theorem 3.3 provides sufficient conditions for the asymptotic equivalence of $\hat{m}_{\hat{G},\theta^*,\alpha}(x)$ and $\hat{m}_{\hat{G},\hat{\theta},\alpha}(x)$.

Let $\tilde{m}(x)$ be any non-random parametric function that approximates $m(x)$. Define the corresponding generalized guided local linear estimator (GGLL) as

$$\tilde{m}_{\hat{G},\alpha}(x) = \sum_{i=1}^n W_i(x, h) \tilde{Z}_{\hat{G},i}(x, \alpha),$$

where $\tilde{Z}_{\hat{G},i}(x, \alpha) = \tilde{m}(x) + (\hat{Y}_i^* - \tilde{m}(X_i))\rho_x(X_i)$, $i = 1, \dots, n$, and $\rho_x(v) = \left(\frac{\tilde{m}(x)}{\tilde{m}(v)}\right)^\alpha$. The following Proposition is an extension of Theorem 3.1:

Proposition 3.2. *Suppose that Assumption 3.1 holds and suppose that $\tilde{m}(x) \neq 0$. Then,*

$$(nh)^{1/2} \left(\tilde{m}_{\hat{G},\alpha}(x) - m(x) - \tilde{B}(x, \alpha) + O_p \left(\left(\frac{\log n}{nh_0} \right)^{1/2} \right) + o_p(h^2) \right) \\ \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du \right),$$

with

$$\tilde{B}(x, \alpha) = \frac{1}{2} h^2 \mu_K^2 (\Gamma_\alpha^{(2)}(x) - \Lambda_\alpha^{(2)}(x)),$$

and

$$\Gamma_\alpha(v) = m(v)\rho_x(v), \quad \Lambda_\alpha(v) = \tilde{m}(v)\rho_x(v).$$

Proof. We start with the case when G is known. The GLL estimator based on the non random parametric model $\tilde{m}(x)$ and transformation (3.2.2) is

$$\tilde{m}_{G,\alpha}(x) = \sum_{i=1}^n W_i(x, h) \tilde{Z}_{G,i}(x, \alpha),$$

where $\tilde{Z}_{G,i}(x, \alpha) = \tilde{m}(x) + (Y_i^* - \tilde{m}(X_i))\rho_x(X_i)$, $i = 1, \dots, n$. We have the following result.

Lemma 3.5. *Suppose that Assumptions of Proposition 3.2 hold. Then,*

$$(nh)^{1/2} (\tilde{m}_{G,\alpha}(x) - m(x) - \tilde{B}(x, \alpha) + o_p(h^2)) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du \right).$$

Proof. Note that

$$E(Y_i^* \rho_x(X_i) | X_i = v) = \rho_x(v) m(v) = \Gamma_\alpha(v),$$

and $\Gamma_\alpha(x) = m(x)$ since $\rho_x(x) = 1$. Write

$$\tilde{m}_{G,\alpha}(x) - m(x) = J_{1,n}(x) - J_{2,n}(x),$$

where

$$J_{1,n}(x) = \sum_{i=1}^n W_i(x, h) Y_i^* \rho_x(X_i) - \Gamma_\alpha(x)$$

and

$$J_{2,n}(x) = \sum_{i=1}^n W_i(x, h) (\Lambda_\alpha(X_i) - \Lambda_\alpha(x)).$$

From Theorem 5.2 in Fan and Gijbels (1996) it follows that

$$\sqrt{nh} \left(J_{1,n}(x) - \frac{h^2}{2} \Gamma_\alpha^{(2)}(x) \mu_K^2 + o(h^2) \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_*^2(x)}{f_X(x)} \int K^2(u) du \right).$$

Since $\sum_{i=1}^n W_i(x, h)(X_i - x) = 0$, we can redefine $J_{2,n}(x)$ as follows:

$$J_{2,n}(x) = \sum_{i=1}^n W_i(x, h) \bar{R}(X_i),$$

where

$$\bar{R}(X_i) = \Lambda_\alpha(X_i) - \Lambda_\alpha(x) - \Lambda_\alpha^{(1)}(x)(X_i - x).$$

We have

$$J_{2,n}(x) = \frac{s_{n,2}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}(x)^2} \left[\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \bar{R}(X_i) - \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \bar{R}(X_i)(X_i - x) \right].$$

From the Lemma 4 in Fan and Gijbels (1992), we have,

$$J_{2,n}(x) = \frac{h^2}{2} \Lambda_\alpha^{(2)}(x) \mu_K^2 + o_p(h^2).$$

and the result of Lemma 3.5 is now a direct consequence. $\square$

Now we turn to the case when G is unknown. We have the following Lemma.

Lemma 3.6. *Suppose that Assumptions of Proposition 3.2 hold. Then,*

$$\sup_{x \in S} |\tilde{m}_{\hat{G}, \alpha}(x) - \tilde{m}_{G, \alpha}(x)| = O_p \left(\sup_{t \leq \tau, x \in S} |\hat{G}(t^-|x) - G(t|x)| \right).$$

The proof of this result is very similar to the proof of Lemma 3.2 and so it is omitted.

The result of Proposition 3.2 is now a consequence of Lemmas 3.5, 3.6 and expression (3.6.4). $\square$

Now we go back to the proof of Theorem 3.3. We consider $\widehat{m}_{\widehat{G},\widehat{\theta},\alpha}(x)$ the GLL estimator with an estimated guide $m(x, \widehat{\theta})$ and we will prove that $\widehat{m}_{\widehat{G},\widehat{\theta},\alpha}(x)$ is asymptotically equivalent to $\widetilde{m}_{\widehat{G},\theta^*,\alpha}(x)$ the GLL estimator with the fixed guide $m(x, \theta^*)$. Consider the following decomposition:

$$\sqrt{nh}(\widehat{m}_{\widehat{G},\widehat{\theta},\alpha}(x) - m(x)) = \sqrt{nh}(\widehat{m}_{\widehat{G},\widehat{\theta},\alpha}(x) - \widetilde{m}_{\widehat{G},\theta^*,\alpha}(x)) + \sqrt{nh}(\widetilde{m}_{\widehat{G},\theta^*,\alpha}(x) - m(x)).$$

The second term in the sum is asymptotically normal by Proposition 3.2. To prove the result of Theorem 3.3 it is sufficient to establish the following:

$$A_n(x) = \sqrt{nh}(\widehat{m}_{\widehat{G},\widehat{\theta},\alpha}(x) - \widetilde{m}_{\widehat{G},\theta^*,\alpha}(x)) = o_p(1).$$

Write

$$A_n(x) = \sqrt{nh}[Q_{n,1}(x) + Q_{n,2}(x) - Q_{n,3}(x)],$$

where

$$\begin{aligned} Q_{n,1}(x) &= \sum_i^n W_i(x, h)(m(x, \widehat{\theta}) - m(x, \theta^*)), \\ Q_{n,2}(x) &= \sum_i^n W_i(x, h)\widehat{Y}_i^*(\rho_x(X_i, \widehat{\theta}) - \rho_x(X_i, \theta^*)), \\ Q_{n,3}(x) &= \sum_i^n W_i(x, h)(\Lambda_{\widehat{\theta},\alpha}(X_i) - \Lambda_{\theta^*,\alpha}(X_i)). \end{aligned}$$

Under Assumptions 3.1 and 3.2 and using the same arguments as in the proof of Theorem 3.2, we get $\sqrt{nh}Q_{n,1}(x) = o_p(1)$ and $\sqrt{nh}Q_{n,3}(x) = o_p(1)$. Now for $Q_{n,2}(x)$ we have

$$|Q_{n,2}(x)| \leq O_p(1) \max_i |\rho_x(X_i, \widehat{\theta}) - \rho_x(X_i, \theta^*)|.$$

We have, for $i = 1, \dots, n$,

$$\begin{aligned} |\rho_x(X_i, \hat{\theta}) - \rho_x(X_i, \theta^*)| &\leq \frac{|m(X_i, \hat{\theta})^\alpha - m(X_i, \theta^*)^\alpha| |m(x, \theta^*)|^\alpha}{|m(X_i, \hat{\theta})|^\alpha |m(X_i, \theta^*)|^\alpha} \\ &\quad + \frac{||m(x, \theta^*)^\alpha - m(x, \hat{\theta})^\alpha| |m(x, \theta^*)|^\alpha}{|m(X_i, \hat{\theta})|^\alpha |m(X_i, \theta^*)|^\alpha}. \end{aligned}$$

By Taylor's Theorem, there exist $\theta_{m,i}$ such that $\theta_{m,i} = \lambda_i \hat{\theta} + (1 - \lambda_i) \theta^*$, $\lambda_i \in [0, 1]$ for $i = 1, \dots, n$ and

$$|m(X_i, \hat{\theta})^\alpha - m(X_i, \theta^*)^\alpha| \leq |\alpha| |m(X_i, \theta_{m,i})^{\alpha-1}| |\nabla_\theta m(X_i, \theta_{m,i})| |\hat{\theta} - \theta^*|.$$

Therefore,

$$\max_i |m(X_i, \hat{\theta})^\alpha - m(X_i, \theta^*)^\alpha| = O_p(n^{-1/2}).$$

Similar arguments give $|m(x, \hat{\theta})^\alpha - m(x, \theta^*)^\alpha| = O_p(n^{-1/2})$. Then, from Assumption 3.2 we have

$$\max_i |\rho_x(X_i, \hat{\theta}) - \rho_x(X_i, \theta^*)| = O_p(n^{-1/2}).$$

Thus,

$$\sqrt{nh} Q_{n,2}(x) = h^{1/2} O_p(1) = o_p(1).$$

□

Chapter 4

Guided local quasi-likelihood estimation¹

4.1 Introduction

The concept of quasi-likelihood estimation was proposed by Wedderburn (1974) as a flexible extension of the maximum likelihood estimation method for generalized linear models (GLMs). The latter, as introduced by Nelder and Wedderburn (1972), relies on strong parametric assumption about the distribution of the data that can be hard to verify in practice. In such situations, the quasi-likelihood estimation may be a suitable alternative, since it relies only on assumptions about the first two moments. Moreover, the quasi-likelihood function has similar properties as the classical full log-likelihood function; see McCullagh and Nelder (1989) for more details.

Likelihood and quasi-likelihood provide consistent and powerful estimators if the required assumptions imposed on the data are met. However, a misspecified model can create an important bias in the estimation of the underlying target function. For this reason, nonparametric techniques, that are more robust, have been investigated by many studies. This include Green and Yandell (1985), O'Sullivan *et al.* (1986), Staniswalis (1989), Hunsberger (1994) and Severini and Staniswalis (1994), to cite just a few examples. More recent contributions include the work of Fan *et al.* (1995)

¹This chapter is based on Talamakrouni, M., El Ghouch, A. and Van Keilegom, I. (2016). Parametrically guided nonparametric local quasi-likelihood with censored data. (submitted).

and Fan and Chen (1999), who investigated local polynomial fitting for likelihood and quasi-likelihood in the context of GLMs, Chen *et al.* (2006), who studied local quasi-likelihood for missing data.

As argued in Chapter 1, even when the proposed model is misspecified, parametric quasi-likelihood estimation can provide a useful information about the target function. Using a parametrically guided nonparametric estimation, this information can be injected into the local quasi-likelihood estimator in order to improve its performance in terms of bias and mean-squared error (MSE). In the following, we give a brief description of the guided additive scheme in the context of local quasi-likelihood, more details are provided in Section 4.2.

We are interested in the relationship between a variable of interest $Y \in \mathfrak{R}$ and some covariate $X \in \mathfrak{R}$. Suppose for the moment that we have completely observed and i.i.d. data $(Y_i, X_i), i = 1, \dots, n$, and let $m(x) = E(Y|X = x)$ be the true mean regression function. In classical parametric GLMs, $m(x)$ is modeled linearly using a known link function $g(\cdot)$, that is $g(m(x)) = \eta(x)$, with $\eta(x) = \theta_0 + \theta_1 x$. The parameter of interest $\theta = (\theta_0, \theta_1)^T$ can be estimated via the likelihood or the quasi-likelihood. However, in practice, the linearity assumption is not met in many situations. In such cases, local quasi-likelihood is more appropriate since it allows the estimation of $\eta(x)$ without any explicit specification of its form. In between these two “extreme” approaches, Fan *et al.* (2009) proposed a guided local quasi-likelihood estimator with the objective of combining the advantages of both parametric and local quasi-likelihood estimators. As stated before, we focus on the additively guided local quasi-likelihood estimator. The additive scheme starts with a parametric quasi-likelihood estimator which is not necessarily correctly specified. Then, in a second step, this crude parametric approximation is adjusted using a local quasi-likelihood estimator. More formally, let $\eta(x, \hat{\theta})$ be a “naive” quasi-likelihood estimator of $\eta(x, \theta)$, a given, possibly misspecified, parametric model for $\eta(x)$. Fan *et al.* (2009) proposed to estimate the error term $r_{\hat{\theta}}(x) := \eta(x) - \eta(x, \hat{\theta})$ using a nonparametric weighted local quasi-likelihood (LQL) estimator that we denote by $\hat{r}_{\hat{\theta}}(x)$. The additive parametrically guided local quasi-likelihood (GLQL) estimator is defined by $\hat{\eta}(x) = \eta(x, \hat{\theta}) + \hat{r}_{\hat{\theta}}(x)$. When the parametric model is properly chosen, $r_{\hat{\theta}}$ may be flatter and easier to adjust non-parametrically than the original function η . In this case, the guided local

quasi-likelihood estimator should be of smaller MSE than the classical LQL estimator. Otherwise, the nonparametric correction is expected to correct for the misspecification and there should not be much loss in accuracy for the resulting GLQL estimator compared to the classical LQL estimator.

The first extension of quasi-likelihood to the right censored data case has been established in the framework of partially linear single-index models by Lu and Cheng (2007). In the generalized linear model, Yu *et al.* (2009) adapted the parametric quasi-likelihood to censored data. Recently, Yu (2011), Yu *et al.* (2012) and Yu and Peace (2012) proposed different semi-parametric quasi-likelihood estimators in the framework of accelerated failure time models. Note that, none of the papers mentioned above has considered a fully nonparametric quasi-likelihood. Thus, one of the main objectives of this chapter is to extend the local quasi-likelihood of Fan *et al.* (1995) to the censored data case.

4.2 Model and methodology

Regression techniques are commonly used to describe a relationship between a variable of interest $Y \in \mathfrak{R}$ and a covariate $X \in \mathfrak{R}$. In the presence of a censoring variable C one can only observe an i.i.d. random sample $(X_i, T_i, \delta_i), i = 1, \dots, n$, from (X, T, δ) , where $T = \min(Y, C)$ and $\delta = \mathbf{I}(Y \leq C)$. Similarly as in Chapter 3, we suppose that given the covariate X , the censoring variable C is independent of the variable of interest Y . Set $F(y|x) = P(Y \leq y|X = x)$ and $G(y|x) = P(C \leq y|X = x)$ the conditional distribution function of Y and C given $X = x$, respectively. Suppose that there exists a known positive function $V(\cdot)$ that relates the conditional mean and the conditional variance of $\phi(Y)$ given X as follows:

$$m(x) = E(\phi(Y)|X = x) \text{ and } Var(\phi(Y)|X = x) = V(m(x)),$$

where ϕ is a known function used to cover various parameters of interest. For example, when $\phi(y) = y\mathbf{1}_{\{y \leq \tau\}}$, for some known τ , we get the truncated mean $m(x) = \int_{-\infty}^{\tau} y dF(y|x)$. Our main objective is to estimate $\eta(x) = g(m(x))$, where $g(\cdot)$ is a known link function. Since only the relationship between the conditional mean and the conditional variance is known, the likelihood estimation method can

not be used. In the following, we first introduce the guided local quasi-likelihood for complete data, and then we adapt the method to handle censoring.

4.2.1 Guided local quasi-likelihood for complete data

Wedderburn (1974) defined the quasi-log-likelihood function as any function $Q(\mu, y)$ satisfying

$$\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y - \mu}{V(\mu)}.$$

Assuming that $\eta(x) = \theta_1 + \theta_2 x$, the parameters θ_1 and θ_2 can be estimated via maximizing the parametric quasi-likelihood $\sum_{i=1}^n Q(g^{-1}(\theta_1 + \theta_2 X_i), Y_i)$, that plays the role of the log-likelihood in the classical GLM model. Because the assumption of linearity does not hold in many situations, Fan *et al.* (1995) proposed a local polynomial quasi-likelihood method to estimate $\eta(\cdot)$ without assuming any specific form for it. The maximum local quasi-likelihood estimator of order p ($p \geq 1$) of $\eta(x)$ is $\hat{\eta}(x) = \hat{\beta}_0$, where $(\hat{\beta}_0, \dots, \hat{\beta}_p)$ is the maximizer of

$$\sum_{i=1}^n Q(g^{-1}(\beta_0 + \dots + \beta_p(X_i - x)^p), Y_i) K_h(X_i - x),$$

where $K(\cdot)$ is a kernel density, $h \equiv h_n$ is a smoothing bandwidth, and $K_h(\cdot) = h^{-1}K(\cdot/h)$.

Let $\eta(x, \theta)$ be a parametric model which belongs to some family of parameterized functions $\{\eta(x, \theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ and define the parametric maximum quasi-likelihood estimator of θ as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \sum_{i=1}^n Q(g^{-1}(\eta(X_i, \theta)), Y_i). \quad (4.2.1)$$

As discussed in the introduction, the parametric estimator given in (4.2.1) provides some useful and interesting information about the target function $\eta(\cdot)$ that may help us to improve the local quasi-likelihood estimator. To simplify the presentation, we focus on the local linear case ($p = 1$). Under the additive scheme, Fan *et al.* (2009) proposed to estimate the error term $r_{\hat{\theta}}(x) = \eta(x) - \eta(x, \hat{\theta})$ by $\hat{r}_{\hat{\theta}}(x) = \hat{\beta}_0$, where $(\hat{\beta}_0, \hat{\beta}_1)$ is the maximizer, with respect to (β_0, β_1) , of the local quasi-likelihood function $\sum_{i=1}^n Q(g^{-1}(\beta_0 + \beta_1(X_i - x) + \eta(X_i, \hat{\theta})), Y_i) K_h(X_i - x)$. $\eta(x)$ can then be

estimated by $\hat{\eta}(x) = \eta(x, \hat{\theta}) + \hat{r}_{\hat{\theta}}(x)$. The latter can, equivalently and directly, be derived by maximizing

$$\sum_{i=1}^n Q(g^{-1}(\beta_0 + \beta_1(X_i - x) + \eta(X_i, \hat{\theta}) - \eta(x, \hat{\theta})), Y_i) K_h(X_i - x),$$

and taking $\hat{\eta}(x) = \hat{\beta}_0$, where $(\hat{\beta}_0, \hat{\beta}_1)$ maximizes the above function with respect to (β_0, β_1) .

4.2.2 Guided local quasi-likelihood and censoring

In the presence of censoring, as $E(\phi(T)|X = x) \neq m(x)$, one cannot directly use the observed data to estimate $\eta(x) = g(m(x))$. In order to overcome this problem, we will use the synthetic data approach. In this approach, the observed response T is substituted by a synthetic response Y^* , such that, under the conditional independence of Y and C given X , $E(Y^*|X = x) = m(x)$. Different transformations satisfying this equality exist in the literature, see for instance Leurgans (1987) and Zheng (1987), among others. Similarly as in Chapter 3, we limit ourselves to the transformation of Koul *et al.* (1981) given by

$$Y^* = \frac{\delta\phi(T)}{1 - G(T^-|X)}. \quad (4.2.2)$$

This transformation is not directly applicable in practice, since it depends on $G(y|x)$, the conditional distribution function of C given $X = x$, which is unknown. In the censored data framework, $G(y|x)$ can be estimated using Beran's estimator defined in expression (3.2.4). We obtain the synthetic response $\hat{Y}^*$ by plugging-in Beran's estimator $\hat{G}(\cdot|\cdot)$ into the transformation (4.2.2).

Following Fan *et al.* (2009), we define our parametrically guided local quasi-likelihood estimator of η , based on the synthetic sample $(\hat{Y}_i^*, X_i)$, $i = 1, \dots, n$, as $\hat{\eta}_{\hat{G}, \hat{\theta}}(x) = \hat{\beta}_0$, where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$ is the maximizer of

$$\sum_{i=1}^n Q(g^{-1}(\beta_0 + \beta_1(X_i - x) + \eta(X_i, \hat{\theta}) - \eta(x, \hat{\theta})), \hat{Y}_i^*) K_h(X_i - x), \quad (4.2.3)$$

with respect to $\beta = (\beta_0, \beta_1)^T$, and $\hat{\theta}$ is a pseudo parametric quasi-likelihood estimator of θ adapted to censored data. The estimation approach that we adopt will be discussed in detail in Section 4.3.2.

Note that the parametrically guided local quasi-likelihood given in (4.2.3) raises new challenges when compared to the equivalent estimator with completely observed data since the synthetic observations $Y_i^*, i = 1, \dots, n$ defined by expression (4.2.2) must be estimated using the whole sample.

Remark 4.1. *We didn't find any results in the literature concerning the estimation of a general misspecified parametric model using quasi-likelihood under censoring. We also note that using a linear guide reduces the estimator to the classical local quasi-likelihood estimator of η , which means that our GLQL estimator $\hat{\eta}_{\hat{G}, \hat{\theta}}(x)$ is a generalization of the classical LQL estimator that can be obtained by maximizing (4.2.3) with $\eta = 0$.*

4.3 Theoretical properties

In order to show the bias reduction property of our new estimator, we investigate in this section the asymptotic distribution of $\hat{\eta}_{\hat{G}, \hat{\theta}}(\cdot)$. First of all, we derive in Theorem 4.1 the asymptotic properties of $\tilde{\eta}_{\hat{G}}(\cdot)$ an estimator of $\eta(\cdot)$ guided by a given non stochastic approximation $\tilde{\eta}(\cdot)$. Then, in Theorem 4.2 we generalize the results to cover the case of a data-driven parametric guide.

4.3.1 The model with a fixed guide

Let $\tilde{\eta}(x)$ be a non stochastic guide that approximates the true function $\eta(x)$ and let $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)$ maximize the following function:

$$\sum_{i=1}^n Q(g^{-1}(\beta_0 + \beta_1(X_i - x) + \tilde{\eta}(X_i) - \tilde{\eta}(x)), \hat{Y}_i^*) K_h(X_i - x). \quad (4.3.1)$$

Define the corresponding GLQL estimator as $\tilde{\eta}_{\hat{G}}(x) = \tilde{\beta}_0$. In the following, we provide the assumptions required for the main results.

Assumption 4.1.

- A1. *i. X has a compact support $S_X \subset \mathfrak{R}$.*
- ii. $f_X(\cdot)$, the marginal density of X , is twice continuously differentiable and $\inf_{x \in S_X} f_X(x) > 0$.*

- A2. The function ϕ is bounded and vanishes outside the interval $[0, \tau]$ for some $\tau < \inf_{x \in S_X} \tau_x$ with $\tau_x = \sup\{y : H(y|x) < 1\}$, that is, the right endpoint of the support of $H(y|x) = P(T \leq y|X = x)$.
- A3. The functions $H_j(y|x) = P(T \leq y, \delta = j|X = x)$, $j = 0, 1$, have four derivatives with respect to x . Furthermore, the derivatives are bounded uniformly for all $y \leq \tau$ and $x \in S_X$.
- A4. $E(\phi(Y)^2) < \infty$.
- A5. i. K is a symmetric probability density function with compact support, say $S_K = [-1, 1]$.
 ii. K_0 is a symmetric, twice continuously differentiable probability density function with compact support S_{K_0} .
 iii. $\int x^2 K(x) dx = \mu_2^K < \infty$, $\int x^2 K_0(x) dx = \mu_2^{K_0} < \infty$ and $\int x^j K^2(x) dx = \nu_j^K < \infty$ for $j = 0, 1, 2$.
- A6. $nh^5 = O(1)$, $nh_0^5/\log n = O(1)$, $nh \rightarrow \infty$ and $nh_0 \rightarrow \infty$ as $n \rightarrow \infty$.
- A7. $\eta(\cdot)$, $V(\cdot)$, $\tilde{\eta}(\cdot)$ and $g'(\cdot)$ are twice continuously differentiable.

Let $q_l(x, y) = \frac{\partial^l}{\partial x^l} Q(g^{-1}(x), y)$ and $\rho_l(x) = (g'(g^{-1}(x))^l V(g^{-1}(x)))^{-1}$, $l = 1, 2$. Note that q_l is linear in y for a fixed x , $q_1(\eta(x), m(x)) = 0$ and $q_2(\eta(x), m(x)) = -\rho_2(\eta(x))$. The following additional assumptions are also required.

Assumption 4.2.

- B1. The function $q_2(x, y) < 0$ for all $x \in S_X$ and $y \leq \tau$.
- B2. The function $\sigma_*^2(x) = \text{Var}(Y^*|X = x)$ is continuous on S_X .
- B3. For all $x \in S_X$, $\rho_2(x) \neq 0$, $\sigma_*^2(x) \neq 0$ and $g'(m(x)) \neq 0$.

Condition A2 is related to the inconsistency of Beran's estimator in the right tail of the distribution of Y . Conditions A1, A3 and A4-A6 are usually required in kernel-based estimation with censored data. Finally, assumption A7 and B1-B3 are similar to the assumptions in Fan *et al.* (1995) and Fan *et al.* (2009) for uncensored case. The following Theorem provides the asymptotic distribution of $\tilde{\eta}_{\hat{G}}(\cdot)$.

Theorem 4.1. *Suppose that Assumptions 4.1 and 4.2 hold. Then,*

$$(nh)^{1/2} \left\{ \tilde{\eta}_{\tilde{G}}(x) - \eta(x) - \tilde{B}(x) + O_p \left(\left(\frac{\log n}{nh_0} \right)^{1/2} \right) \right\} \\ \xrightarrow{d} \mathcal{N} \left(0, \sigma_*^2(x) g'(m(x))^2 f_X^{-1}(x) \nu_0^K \right),$$

with

$$\tilde{B}(x) = \frac{1}{2} h^2 \mu_2^K (\eta''(x) - \tilde{\eta}''(x)) \{1 + o(1)\}.$$

Remark 4.2. *The bias produced by Beran's estimator is bounded by $\left(\frac{\log n}{nh_0}\right)^{1/2}$. This extra term vanishes when the bandwidths are chosen such that $\frac{h_0}{h \log n} \rightarrow \infty$. Therefore, there is no loss of accuracy when one replaces the response by synthetic data, provided that the bandwidth for Beran's estimator is asymptotically larger than the bandwidth used in the local linear fit. This fact has also been pointed out by Talamakrouni et al. (2015) in the context of guided nonparametric regression with censored data. The bias term $\tilde{B}(x)$ is similar to the fully observed data case and reveals the effect of the parametric guide. If the guide is chosen such that $|\eta''(x) - \tilde{\eta}''(x)| \leq |\eta''(x)|$, then the bias of the GLQL estimator will be smaller compared with that of the classical LQL estimator. If the second derivatives of the parametric guide and the true function are equal, then the bias term $\tilde{B}(x)$ vanishes. Regarding the variance, there is no difference compared with the classical LQL under censorship. The only difference appears when one compares the variance term of the GLQL estimator in the presence and the absence of censoring. In fact, the term $\sigma_*^2(x) = \text{Var}(Y^*|X = x) = \sigma^2(x) + E[\phi(Y)^2 G(Y|X)/(1 - G(Y|X))|X = x]$ replaces $\sigma^2(x) = \text{Var}(\phi(Y)|X = x)$ and this is due to the synthetic data. Note that, if the parametric guide is chosen to be constant, then the GLQL estimator reduces to the classical LQL estimator. Therefore, the result of our Theorem 4.1 is a generalization of Theorem 1.a (for $p = 1, r = 0$) in Fan et al. (1995) to right censored data.*

4.3.2 The model with an estimated guide

In the previous section, Theorem 4.1 investigated the simple case of a fixed guide. However, in practice, the guide needs to be estimated. In the following, we consider

the case where the parametric guide $\eta(x, \hat{\theta})$ is obtained from a first stage estimation procedure. Following Fan *et al.* (2009), we denote by

$$f(x, y) = f_X(x) \exp(Q(g^{-1}(\eta(x)), y)),$$

the true unknown joint density of (X, Y) and by

$$f(x, y; \theta) = f_X(x) \exp(Q(g^{-1}(\eta(x, \theta)), y)),$$

the proposed parametric joint density. Define $\theta_* \in \Theta \subset \mathbb{R}^d$, the value of θ which maximizes the following function:

$$\iint_{\Delta} Q(g^{-1}(\eta(x, \theta)), y) dF(x, y), \quad (4.3.2)$$

where $F(x, y)$ is the joint distribution function of (X, Y) and $\Delta = S_X \times (-\infty, \tau]$ is needed because $F(x, y)$ cannot be estimated consistently in the presence of censoring. θ_* is the parameter value that minimizes the Kullback-Leibler distance between the true joint density $f(x, y)$ and the parametric joint density $f(x, y; \theta)$, that is, $\theta_* = \arg \min_{\theta \in \Theta} E_{\Delta}[\log(f(X, Y)/f(X, Y; \theta))]$ with $E_{\Delta}(A) = E(A \cdot \mathbf{I}\{(X, Y) \in \Delta\})$. If the parametric model is correct, i.e. there exists $\theta_0 \in \Theta$ such that $f(x, y) = f(x, y; \theta_0)$, then $\theta_0 = \theta_*$.

In the spirit of Suzukawa *et al.* (2001), we estimate θ by $\hat{\theta}$, the maximizer of a suitable analogue of (4.2.1) that we define as

$$\iint_{\Delta} Q(g^{-1}(\eta(x, \theta)), y) d\hat{F}(x, y), \quad (4.3.3)$$

where $\hat{F}$ is an estimator of F satisfying the following assumptions:

Assumption 4.3.

1.

$$\sup_{\theta \in \Theta} \left| \iint_{\Delta} \nabla_{\theta}^r \log f(x, y; \theta) d(\hat{F} - F)(x, y) \right| = o_p(1), \quad \text{for } r = 0, 2.$$

2.

$$\sqrt{n} \iint_{\Delta} \nabla_{\theta}^1 \log f(x, y; \theta_*) d(\hat{F} - F)(x, y) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where Σ is a nonnegative-definite matrix and $\nabla_{\theta}^r \Phi(x, y; \theta) = \partial^r \Phi(x, y; \theta) / \partial \theta^r$ for a twice differentiable function $\theta \rightarrow \Phi(x, y; \theta)$ and $r = 0, 1, 2$.

When the data are completely observed, the estimator $\hat{F}$ may be replaced by the usual bivariate empirical distribution function $F_n(x, y) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq x, Y_i \leq y}$. In this case, the pseudo quasi-likelihood defined by (4.3.3) reduces to (4.2.1), meaning that our approach is more general. In the censored data framework, there have been few proposals for estimating $F(x, y)$ in the literature. For example, Lopez (2011) has developed an estimator of $F(x, y)$ satisfying Assumption 4.3 (see Theorem 3.1 and Theorem 3.6 in Lopez (2011)) and given by the following expression

$$\hat{F}_L(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i 1_{X_i \leq x, T_i \leq y}}{1 - \hat{G}(T_i^- | X_i)}. \quad (4.3.4)$$

Another different and interesting approach was introduced by Van Keilegom and Akritas (1999). Their estimator is constructed through an integrated version of Beran's estimator as follows

$$\hat{F}_{VA}(x, y) = \int_{-\infty}^x \hat{F}(y|u) dF_n(u), \quad (4.3.5)$$

where $F_n(x)$ is the empirical distribution function of X and $\hat{F}(y|x)$ is Beran's estimator of $F(y|x) = P(Y \leq y | X = x)$. We note that both estimators can be used in practice. However, Assumption 4.3 has not yet been investigated for $\hat{F}_{VA}$. Therefore, for sake of consistency, we only investigate the estimator of Lopez (2011) in our simulation studies. Next, we give additional conditions that are also needed.

Assumption 4.4.

D1. $\eta(x, \theta)$ belongs to a parametrically indexed class of functions defined by the following characteristics:

1. $\theta \in \Theta$, Θ is a compact subset of $\mathbb{R}^d$.
2. The function $(x, \theta) \mapsto \eta(x, \theta)$ is twice continuously differentiable with respect to x and θ .

D2. The function $\log f(x, y; \theta)$ is twice continuously differentiable with respect to θ .

D3. $E_{\Delta}(\log f(X, Y))$ exists and there exists a function $\ell(x, y)$ such that:

$$|\log f(x, y; \theta)| \leq \ell(x, y) \text{ for any } \theta \in \Theta \text{ and } E_{\Delta} \ell(X, Y) < \infty.$$

D4. $|\partial^2 \log f(x, y; \theta) / \partial \theta_i \partial \theta_j|$ and $|\partial \log f(x, y; \theta) / \partial \theta_i \times \partial \log f(x, y; \theta) / \partial \theta_j|$, for $i, j = 1, \dots, d$, are dominated by integrable functions with respect to $F(., .)$ for all (x, y) in Δ and all θ in Θ .

D5. $\theta_* = \arg \min_{\theta \in \Theta} E_{\Delta}(\log(f(X, Y) / f(X, Y; \theta)))$ is unique.

D6. The matrix of second derivatives $\nabla_{\theta}^2 \log f(x, y; \theta_*)$ is nonsingular.

Conditions D2, D3 and D4 are classical conditions in the uncensored case that allow to take derivatives under the integrals. The following proposition provides the weak consistency and the asymptotic normality of the estimator $\hat{\theta}$.

Proposition 4.1. *Under Assumptions 4.3 and 4.4, we have*

1. $\hat{\theta}$ converges to θ_* in probability as $n \rightarrow \infty$.
2. $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{N}(0, \Omega^{-1} \Sigma \Omega^{-1})$, with $\Omega \equiv \Omega(\theta_*)$ and $\Omega(\theta) = E_{\Delta}[\nabla_{\theta}^2 \log f(X, Y; \theta)]$.

Note that, the results of Proposition 4.1 reveal the $\sqrt{n}$ -consistency of the estimator $\hat{\theta}$, that is $\sqrt{n}(\hat{\theta} - \theta_*) = O_p(1)$. Now, given this result and some additional conditions, the next Theorem states that there is no loss in accuracy when the parametric guide is estimated.

Theorem 4.2. *Suppose that Assumptions 4.3 and 4.4 hold. Then, under assumptions of Theorem 4.1, we have*

$$(nh)^{1/2} \left\{ \hat{\eta}_{\hat{G}, \hat{\theta}}(x) - \eta(x) - B(x, \theta_*) + O_p\left(\left(\frac{\log n}{nh_0}\right)^{1/2}\right) \right\} \\ \xrightarrow{d} \mathcal{N}\left(0, \sigma_*^2(x) g'(m(x))^2 f_X^{-1}(x) \nu_0^K\right),$$

with

$$B(x, \theta_*) = \frac{1}{2} h^2 \mu_2^K (\eta''(x) - \eta''(x, \theta_*)) \{1 + o(1)\}.$$

Comparing this last result with the result of Theorem 4.1, we notice that the estimation of the parameter θ_* does not affect the asymptotic bias and the asymptotic variance. A crucial issue that arises in any nonparametric method is the choice of the

bandwidth parameters. From Theorem 4.2, the asymptotic mean integrated squared error is given by

$$\frac{1}{4}h^4(\mu_2^K)^2 \int_{S_X} (\eta''(x) - \eta''(x, \theta_*))^2 dx + \frac{\nu_0^K}{nh} \int_{S_X} \sigma_*^2(x) g'(m(x))^2 f_X^{-1}(x) dx. \quad (4.3.6)$$

If $\eta''(x) - \eta''(x, \theta_*) = 0$, then $B(x, \theta_*) = 0$. In such a case, one can choose an arbitrary large bandwidth so that the variance is reduced to its minimum possible value, which is impossible in a fully nonparametric framework (except for a linear η). If $\eta''(x) - \eta''(x, \theta_*) \neq 0$, then minimizing (4.3.6) with respect to h gives the following theoretical optimal bandwidth:

$$h_{opt} = \left(\frac{\nu_0^K \int_{S_X} \sigma_*^2(x) g'(m(x))^2 f_X^{-1}(x) dx}{(\mu_2^K)^2 \int_{S_X} (\eta''(x) - \eta''(x, \theta_*))^2 dx} \right)^{1/5} n^{-1/5}. \quad (4.3.7)$$

This last expression indicates that, if the parametric guide is chosen so that its second derivatives $\eta''(x, \theta_*)$ is close to the second derivative of the true function $\eta''(x)$, then the optimal bandwidth for the GLQL estimator will be larger than the optimal bandwidth of the classical LQL estimator. This allows to reduce also the variance compared with the classical LQL estimator. This fact is widely noticed in our simulation studies. In practice, expression (4.3.7) cannot be used directly since it depends on a number of unknown quantities. Fan and Gijbels (1996) (see Section 4.9) and Fan *et al.* (1995) proposed some guidelines for the selection of the bandwidth h based on the bias-variance tradeoff. Their procedures can be easily extended to censored data framework by simply substituting the censored response Y_i by the synthetic data $\hat{Y}_i^*$. Finally, the bandwidth for Beran's estimator can be chosen using for example the plug-in method (see Dabrowska (1992)) or the bootstrap method investigated by Van Keilegom and Veraverbeke (1997).

4.3.3 Extension to unified family of corrections

As mentioned in the introduction, we investigate the additive correction in order to simplify our presentation. However, in addition to the additive scheme, Martins-Filho, *et al.* (2008) proposed a unified family of corrections in the uncensored data case. In the following we give some guidelines allowing to generalize their proposal

to our framework. Starting from a parametric model $\eta(x, \hat{\theta})$, the basic idea of the guided estimation can be generalized using the following more general identity:

$$\eta(x) = \eta(x, \hat{\theta}) + r_{\hat{\theta}, \alpha}(x) \eta(x, \hat{\theta})^\alpha,$$

where $r_{\hat{\theta}, \alpha}(x) = [\eta(x) - \eta(x, \hat{\theta})] / \eta(x, \hat{\theta})^\alpha$ and $\alpha \geq 0$. We propose to estimate the correction factor $r_{\hat{\theta}, \alpha}(x)$ by $\hat{r}_{\hat{\theta}, \alpha}(x) = \hat{\beta}_0$, where $(\hat{\beta}_0, \hat{\beta}_1)$ is the maximizer of

$$\sum_{i=1}^n Q(g^{-1}(\eta(X_i, \hat{\theta}) + (\beta_0 + (X_i - x)\beta_1)\eta(X_i, \hat{\theta})^\alpha), \hat{Y}_i^*) K_h(X_i - x).$$

Therefore, the extended guided local quasi-likelihood estimator is given by $\hat{\eta}_{\hat{G}, \hat{\theta}, \alpha}(x) = \eta(x, \hat{\theta}) + \hat{r}_{\hat{\theta}, \alpha}(x) \eta(x, \hat{\theta})^\alpha$. Similarly as in Section 4.2.1, the extended guided estimator $\hat{\eta}_{\hat{G}, \hat{\theta}, \alpha}(x)$ can be defined directly as the first component of the maximizer of

$$\sum_{i=1}^n Q(g^{-1}(\eta(X_i, \hat{\theta}) + (\beta_0 + (X_i - x)\beta_1 - \eta(x, \hat{\theta}))\eta(X_i, \hat{\theta})^\alpha / \eta(x, \hat{\theta})^\alpha), \hat{Y}_i^*) K_h(X_i - x)$$

with respect to $\beta = (\beta_0, \beta_1)$. All the results established before can be generalized to the guided estimator based on the unified family of corrections, the generalization of the proof is straightforward and is omitted here. Theorem 4.3 generalizes the result of Theorem 4.2.

Theorem 4.3. *Suppose that Assumption 4.3 and 4.4 hold and $\eta(x, \theta_*) \neq 0$. Then, under the Assumptions of Theorem 4.1, we have*

$$(nh)^{1/2} \left\{ \hat{\eta}_{\hat{G}, \hat{\theta}, \alpha}(x) - \eta(x) - B(x, \theta_*, \alpha) + O_p \left(\left(\frac{\log n}{nh_0} \right)^{1/2} \right) \right\} \\ \xrightarrow{d} \mathcal{N} \left(0, \sigma_*^2(x) (g'(m(x)))^2 f_X^{-1}(x) \nu_0^K \right),$$

with

$$B(x, \theta_*, \alpha) = \frac{1}{2} h^2 \mu_2^K \eta(x, \theta_*)^\alpha r_{\theta_*, \alpha}''(x) \{1 + o(1)\}, \\ r_{\theta_*, \alpha}(x) = [\eta(x) - \eta(x, \theta_*)] / \eta(x, \theta_*)^\alpha.$$

Note that, the additive correction is a special case of the unified family for $\alpha = 0$. The choice of the parameter α was investigated by Fan *et al.* (2009). However, according to their simulation results, using the best α does not enhance the performance considerably compared with the additive correction. Therefore, to simplify our simulation studies we investigate the additive correction.

4.4 Simulation results

This section is concerned with the evaluation of the finite sample performance of the GLQL estimator. To this end, we conduct two Monte Carlo simulation studies. In the first study, a Poisson model is investigated under right censoring. Such model is widely used in studies dealing with quasi-likelihood and discrete responses, see for example Fan *et al.* (2009) and Davenport *et al.* (2015). Then, in a second time, an exponential model is considered to cover the case of continuous responses. Our target function is $\eta(x) = g\left(\int_0^\tau y dF(y|x)\right)$ where g is the canonical link, $\tau = \inf_x\{\tau_x\}$ and τ_x is the 99.99% upper quantile of $H(y|x) = P(T \leq y|X = x)$. The parametric guides are estimated via maximizing the pseudo QL given in (4.3.3) combined with the estimator (4.3.4) proposed by Lopez (2011). Along the simulations we use local linear fitting and the Epanechnikov kernel for both K_0 and K . To reduce our calculation time, we first selected the value of the bandwidths h_0 and h by minimizing the average mean squared error (MSE) using a small number of simulations. Then, we applied both guided and traditional LQL to 1000 other simulated data sets using the selected “optimal” bandwidths for each method.

4.4.1 Poisson model

In this model, the covariate X is generated from a uniform distribution over the interval $[-2, 2]$. On the other hand, given $X = x$, the response Y is generated from a Poisson distribution with mean $\exp(\eta(x))$ and $\eta(x) = 6 + 3\sin(\frac{\pi}{4}x - \frac{\pi}{2})$. Given $X = x$ and independently from Y , the censoring variable C is also drawn from a Poisson distribution with mean $\exp(\eta(x) + \lambda)$. The parameter λ allows us to control the rate of censoring. The values $\lambda = 0.22, 0.135$, and 0.078 correspond to a fixed censoring rate of 10%, 20% and 30%, respectively. Following Fan *et al.* (2009), three different parametric guides are investigated. The first two guides are misspecified and are given by $\eta_1(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2$ and $\eta_2(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3$, respectively. The third parametric guide is correctly specified and is given by the following sinusoidal function $\eta_3(x, \theta) = \theta_1 + \theta_2 \sin(\frac{\pi}{4}x - \frac{\pi}{2})$. As a quasi-likelihood function we used $Q(\mu, y) = y \log |\mu| - \mu$. We investigate the performance of both the GLQL estimator and the LQL estimator at the median point $x = 0$ for three sample

sizes $n = 100, 250$ and 500 .

As stated before, to select the bandwidths we repeat the simulation 200 times. Figure 4.4.1 shows how the squared bias, the variance and the MSE change with the bandwidth h , for sample size $n = 250$ and a censoring rate of 20%. The other cases are similar and are presented in Appendix A. As established in the asymptotic results, the bias is substantially reduced for the three guided estimators compared with the unguided estimator, while the variance remains unchanged or is slightly reduced especially when a large bandwidth is used. We also note that when the appropriate guide (sinusoidal) is used, the bias of the GLQL estimator is almost zero. This allows us to choose a larger bandwidth and so to reduce the variance substantially.

Now, using the selected bandwidths, we compute the different estimators 1000 times. The squared bias ($Bias^2 \times 10^3$), the variance ($Var \times 10^3$), the empirical mean squared error ($MSE \times 10^3$), the selected bandwidths as well as the ratio of the empirical MSEs (MSE_{GLQL}/MSE_{LQL}) are given in Table 4.4.1 for each setting. Generally speaking, the results show that the GLQL estimators have lower MSE compared to the classical LQL even if the parametric guide is not completely correct. As expected, the best results are obtained when the guide is correctly specified, namely with the sinusoidal guide. The selected bandwidths for the GLQL estimators are larger than or equal to those for the classical LQL. Overall, we can say that the GLQL estimator considerably outperforms the classical LQL estimator. As expected, increasing the sample size improves the quality of all the estimators, in terms of MSE, but increasing the censoring rate affects negatively the results.

Finally, we evaluate the precision of our simulations as well as the computation time. Table 4.7 in Appendix B gives the empirical mean $\hat{\eta}(x) = \frac{1}{N} \sum_{k=1}^N \hat{\eta}_k(x)$, the standard deviation, the variance ($\times 10$) ($\frac{1}{N-1} \sum_{k=1}^N (\hat{\eta}_k(x) - \hat{\eta}(x))^2$) of our simulation results and the corresponding 95% confidence intervals under normality assumption, using three sample sizes (100, 250, 500), censoring rate of 20% and $x = 0$. Figure 4.7.1 in Appendix B shows the boxplot of the same simulation results. The results reveal a smaller standard deviation and more precise confidence intervals for the GLQL estimator especially with the correct guide (GLQL3).

Regarding the computation time, we investigate ten replications with samples of size

100 and censoring rate of 20%. The computation time for the GLQL estimator is about 0.351 seconds while that of the LQL is about 0.339 seconds, meaning that the computation time for the GLQL does not increase too much (almost 3.8%) compared with the classical LQL.

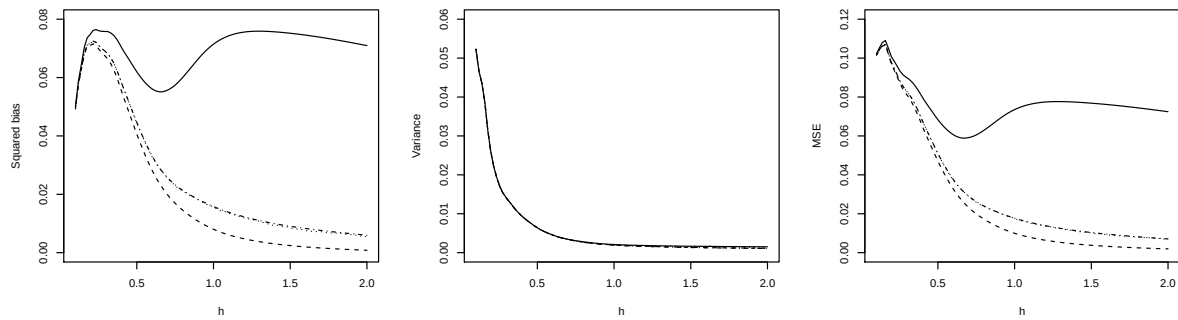


Figure 4.4.1: Squared bias, variance and MSE as a function of h , for the LQL estimator (solid curve), GLQL1 estimator (dot-dashed curve), GLQL2 (dotted curve), GLQL3 (dashed curve). The sample size is $n = 250$, the proportion of censoring is $p = 20\%$.

p	10%					20%					30%					
n	Method	Bias ²	Var	MSE	h	Ratio-MSE	Bias ²	Var	MSE	h	Ratio-MSE	Bias ²	Var	MSE	h	Ratio-MSE
100	GLQL1	2.369	2.177	4.546	2	0.133	5.056	3.079	8.135	2	0.136	7.290	4.050	11.34	2	0.149
	GLQL2	2.249	2.222	4.471	2	0.131	4.617	3.312	7.929	2	0.133	6.810	4.240	11.05	2	0.145
	GLQL3	0.003	2.207	2.210	2	0.065	0.568	3.069	3.637	2	0.061	1.495	3.981	5.476	2	0.072
	LQL	25.12	9.080	34.20	0.6		48.31	11.40	59.71	0.7		70.03	6.230	76.26	2	
250	GLQL1	2.683	0.957	3.640	2	0.122	5.825	1.260	7.085	2	0.122	8.720	1.570	10.29	2	0.129
	GLQL2	2.551	1.147	3.698	2	0.124	5.447	1.583	7.030	2	0.121	8.153	1.765	9.918	2	0.124
	GLQL3	0.013	0.981	0.994	2	0.033	0.752	1.264	2.016	2	0.035	1.962	1.558	3.520	2	0.044
	LQL	26.41	3.470	29.88	0.6		53.51	4.380	57.89	0.7		77.41	2.320	79.73	2	
500	GLQL1	2.827	0.410	3.237	2	0.109	5.817	0.595	6.412	2	0.113	8.975	0.730	9.705	2	0.122
	GLQL2	2.683	0.567	3.249	2	0.110	5.553	0.751	6.305	2	0.112	8.438	0.983	9.421	2	0.120
	GLQL3	0.018	0.426	0.444	2	0.150	0.712	0.591	1.303	2	0.023	2.031	0.715	2.746	2	0.034
	LQL	27.81	1.860	29.67	0.6		54.64	1.920	56.56	0.7		78.73	1.100	79.83	2	

Table 4.4.1: Squared bias ($\times 10^3$), Variance ($\times 10^3$), MSE ($\times 10^3$), the optimal bandwidth h , the ratio of MSE (MSE_{GLQL}/MSE_{LQL}) three censoring rates $p = (10\%, 20\%, 30\%)$, three sample sizes $n = (100, 250, 500)$ and $N = 1000$ replications.

4.4.2 Exponential model

This section addresses the case of a continuous response. Given $X = x$, the response Y is generated from an exponential distribution with parameter $\eta(x) = (0.5x^2 + 1) + a(\sin(2\pi x))^2$, where $a = 0, 0.1, 0.3, 0.5, 0.8, 1$, while the covariate X is uniformly distributed on $[0, 4]$. The censoring variable C is independent of Y given $X = x$ and is also generated from an exponential distribution with parameter $\eta(x)/2$ which leads to almost 33.4% rate of censoring. Regarding the parametric guide, we consider a second order polynomial guide $\eta(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2$. The parameter a allows to control the difference between the true function and the parametric guide. Figure 4.4.2 gives the shapes of different target functions. The case $a = 0$ is the only situation where the guide is correct. Similarly to the Poisson model, the bandwidths are selected using 200 simulations. Note that the six settings are not comparable to each other since the target function changes for each value of a . Therefore, one can only compare the GLQL and the LQL estimators within each setting. As a quasi-likelihood function we choose $Q(\mu, y) = -\frac{y}{\mu} - \log |\mu|$. Using $N = 1000$ replications and samples of size $n = 400$, we calculate the estimators at ten equidistant data points in the interval $[0, 4]$. For a data point $x_i, i = 1, \dots, 10$, we calculate the empirical bias by $b_i = N^{-1} \sum_{j=1}^N [\hat{\eta}_{\hat{G}, \hat{\theta}}^j(x_i) - \eta(x_i)]$ and the empirical variance by $v_i^2 = N^{-1} \sum_{j=1}^N [\hat{\eta}_{\hat{G}, \hat{\theta}}^j(x_i) - N^{-1} \sum_{j=1}^N \hat{\eta}_{\hat{G}, \hat{\theta}}^j(x_i)]^2$, where $\hat{\eta}_{\hat{G}, \hat{\theta}}^j(x_i)$ is the GLQL estimator for the j^{th} replication. Then we calculate $B^2 = 10^{-1} \sum_{i=1}^{10} b_i^2$, the average squared bias, $V = 10^{-1} \sum_{i=1}^{10} v_i^2$, the average variance and $MSE = B^2 + V$, the average mean squared error. The obtained results are summarized in table 4.4.2. When the guide is correct ($a = 0$) the GLQL estimator clearly outperforms the LQL estimator. In fact, in this case, the average squared bias is approximately reduced by half. For $a = 0.1, 0.3, 0.5, 0.8$, even if the parametric guide is not correctly specified, the GLQL estimator behaves better than the classical LQL estimator. Regarding the variance, the guided estimator has generally smaller variance, except for the cases $a = 0.5, 0.8$ where we observe a larger variance for the GLQL estimator. For $a = 1$, the GLQL and the LQL estimators behave similarly in terms of bias while the variance and MSE of the GLQL are now larger than those of the LQL. Finally, as noticed in the first example, the bandwidth selected for the GLQL method is generally larger than the one selected for the classical

LQL method when the parametric guide is not too far from the true function.

a	Method	$Bias^2$	Var	AMSE	h	AMSE-ratio
0	GLQL	0.269	1.298	1.567	1.612	0.989
	LQL	0.489	1.369	1.858	1.263	
0.1	GLQL	0.268	1.319	1.587	1.574	0.842
	LQL	0.493	1.391	1.884	1.264	
0.3	GLQL	0.287	1.320	1.608	1.573	0.842
	LQL	0.549	1.369	1.918	1.263	
0.5	GLQL	0.204	1.472	1.676	1.186	0.856
	LQL	0.518	1.440	1.958	1.147	
0.8	GLQL	0.210	2.199	2.409	0.720	1.112
	LQL	0.438	1.729	2.167	0.953	
1	GLQL	0.442	3.213	3.655	0.527	1.237
	LQL	0.441	2.513	2.954	0.643	

Table 4.4.2: Average squared bias ($\times 10$), average variance ($\times 10$), average MSE ($\times 10$), the optimal bandwidth h and the ratio of the average MSE ($AMSE(GLQL)/AMSE(LQL)$) of the estimators for different conditional mean functions ($a = 0, 0.1, 0.3, 0.5, 0.8, 1$) computed for ten equidistant data points. The samples are of size $n = 400$ with a censoring rate of 33.4%, and $N = 1000$ replications.

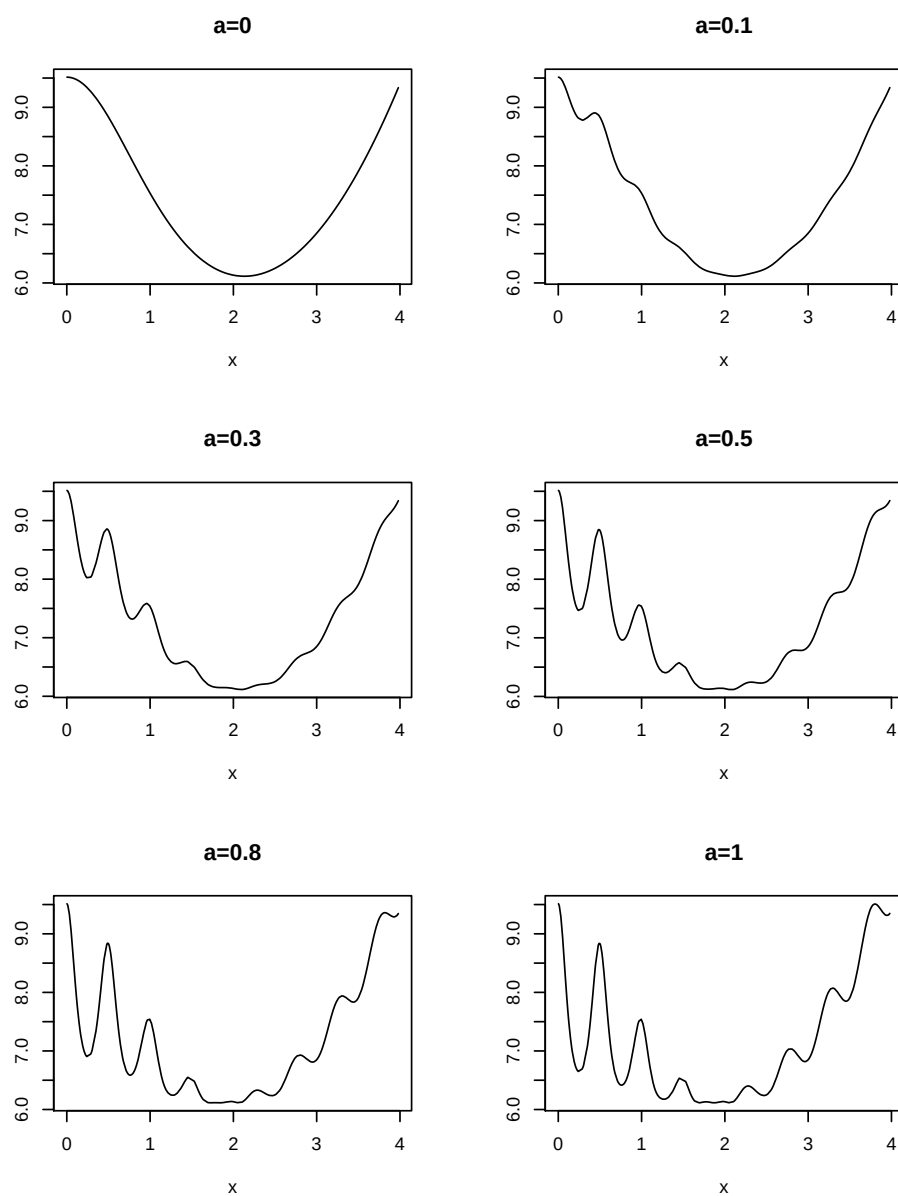


Figure 4.4.2: Target function $\eta(x)$ for different values of a .

4.5 Application

In this section, we apply the GLQL and the LQL estimation to the UIS dataset. We consider an outcome usually of interest in the public health studies, which is whether or not a participant remained drug free for at least one year from randomization to treatment (see for example the book of Hosmer and Lemeshow (2013)). As for the explanatory variable we consider the age of a participant at the enrollment. Our objective is to analyze the effect of the age of a participant on the probability of remaining drug free at least one year. To this end, we consider the target function $\eta(x) = g(E[\phi(Y)|X = x])$, where g is the Logit link, $\phi(Y) = \mathbf{I}(365 \leq Y \leq \tau)$, Y is the corresponding time to return to drug use and $\tau = 807$ is the right endpoint of the support of $H_n(t|x)$ the empirical distribution function of the observed time T given $X = x$. The bandwidth for Beran's estimator is selected using similar heuristic method as in Chapter 3 and is $h_0 = 12$. For the bandwidths used in the LQL estimator and the GLQL estimator we minimize with respect to h , respectively, the functions $CV_T(h) = n^{-1} \sum_i (g^{-1}(\hat{\eta}_{h,-i}(X_i)) - \hat{Y}_i^*)^2$ and $CV_G(h) = n^{-1} \sum_i (g^{-1}(\hat{\eta}_{\hat{G}, \hat{\theta}, h, -i}(X_i)) - \hat{Y}_i^*)^2$, where $\hat{\eta}_{h,-i}(\cdot)$ and $\hat{\eta}_{\hat{G}, \hat{\theta}, h, -i}(\cdot)$ are the leave-one-out LQL and GLQL estimators. The selected bandwidths are $h = 9.7$ for the classical LQL estimator and $h = 11$ for the GLL estimator. As a guide we chose a cubic polynomial for which the parameters are estimated via maximizing the pseudo QL given in (4.3.3) combined with the estimator (4.3.4) proposed by Lopez (2011). The LQL and the GLQL estimators of the log odds defined as $\log \left(\frac{P(365 \leq Y \leq \tau | X=x)}{1 - P(365 \leq Y \leq \tau | X=x)} \right)$ are given in Figure 4.5.1. The log odds estimates show a slow decline of the probability of remaining drug free for at least one year from randomization to treatment with increasing age.

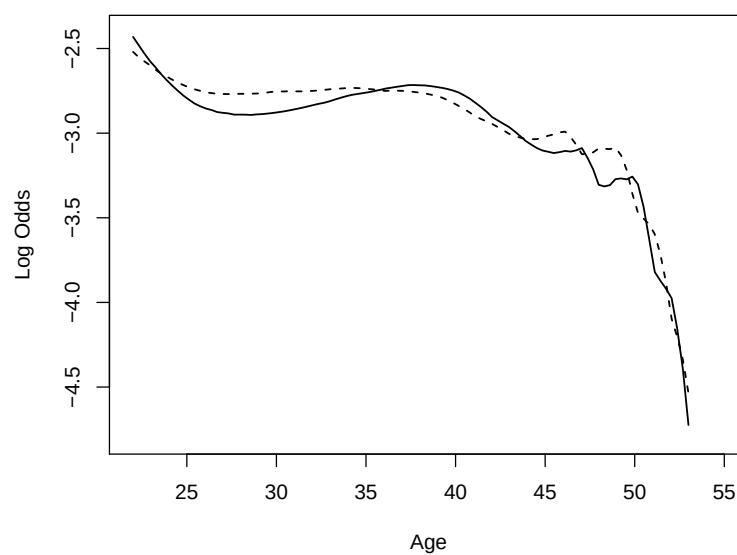


Figure 4.5.1: Estimation of the log odds function, by means of the LQL estimator (dotted curve) and the GLQL estimator (solid curve).

4.6 Appendix A

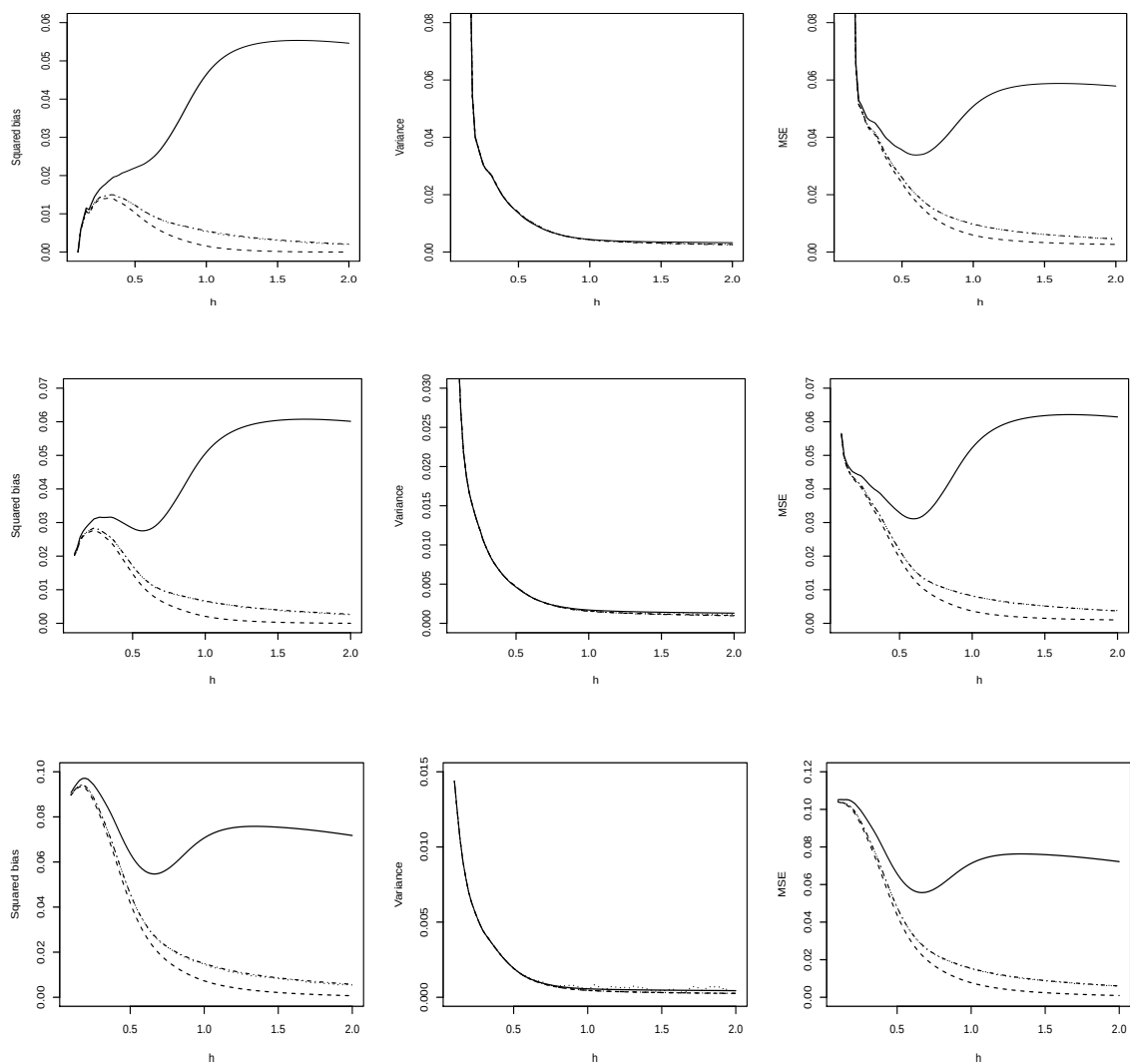


Figure 4.6.1: Squared bias, variance and MSE as a function of h , for the LQL estimator (solid curve), GLQL1 estimator (dot-dashed curve), GLQL2 (dotted curve), GLQL3 (dashed curve). The sample size is $n = 100$ (top), $n = 250$ (middle), $n = 500$ (bottom), the proportion of censoring is $p = 10\%$.

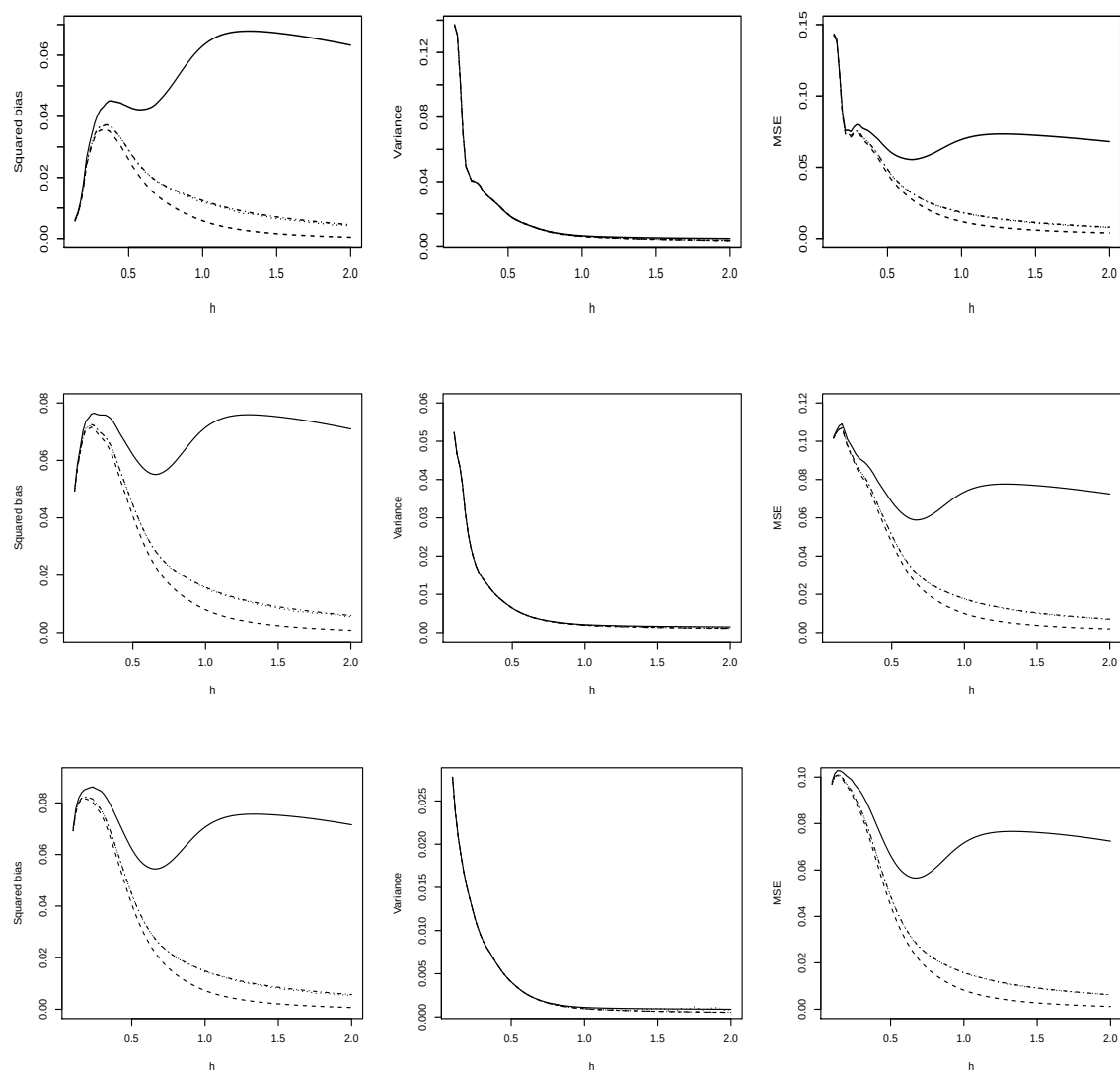


Figure 4.6.2: Squared bias, variance and MSE as a function of h , for the LQL estimator (solid curve), GLQL1 estimator (dot-dashed curve), GLQL2 (dotted curve), GLQL3 (dashed curve). The sample size is $n = 100$ (top), $n = 250$ (middle), $n = 500$ (bottom), the proportion of censoring is $p = 20\%$.

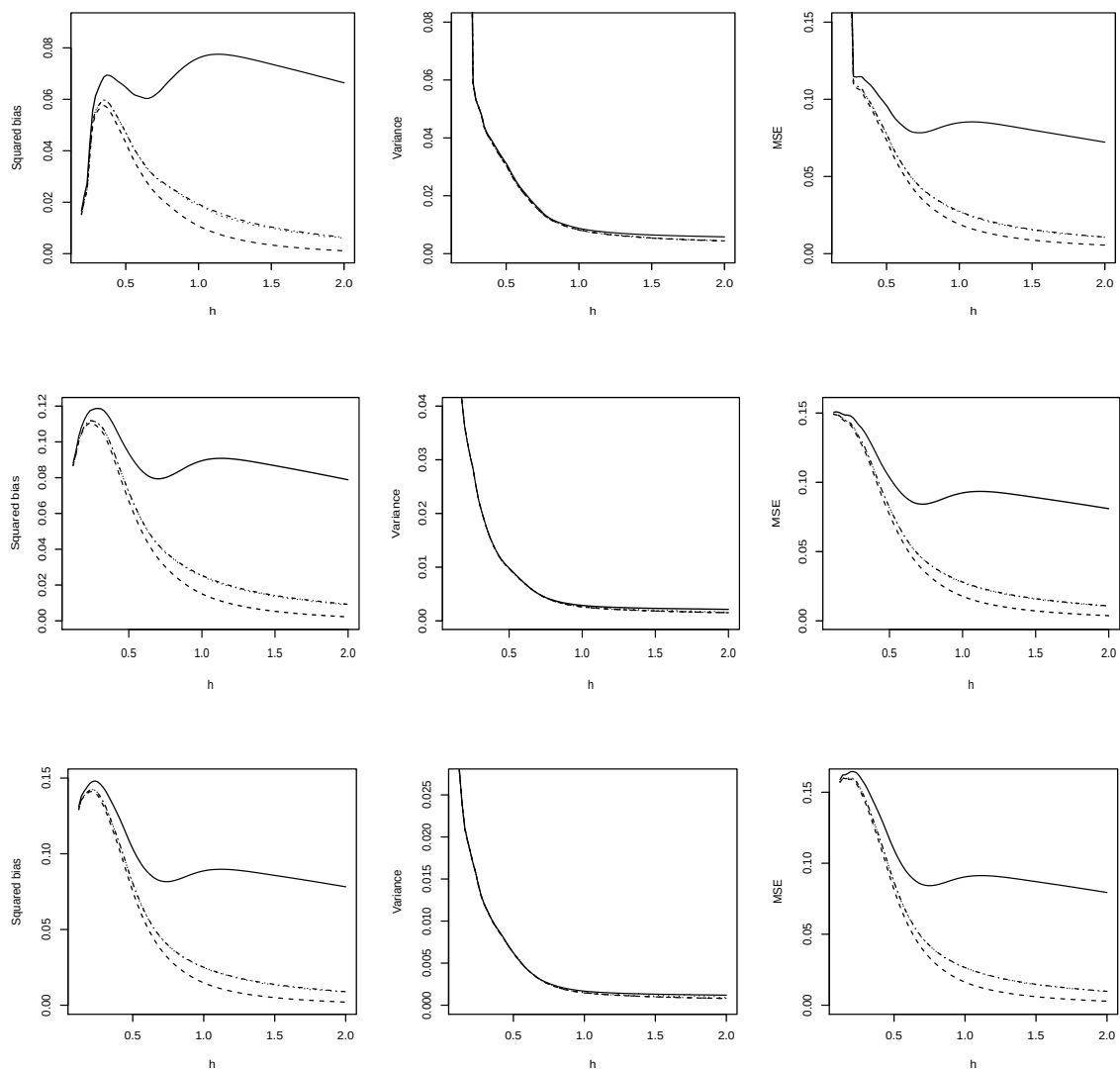


Figure 4.6.3: Squared bias, variance and MSE as a function of h , for the LQL estimator (solid curve), GLQL1 estimator (dot-dashed curve), GLQL2 (dotted curve), GLQL3 (dashed curve). The sample size is $n = 100$ (top), $n = 250$ (middle), $n = 500$ (bottom), the proportion of censoring is $p = 30\%$.

4.7 Appendix B

n	Method	Mean	$CI_{95\%}$	SD	Var ($\times 10$)	$CI_{95\%}(\times 10)$
100	GLQL1	3.077	[3.073; 3.080]	0.056	0.031	[0.028; 0.034]
	GLQL2	3.073	[3.070; 3.077]	0.058	0.034	[0.031; 0.037]
	GLQL3	3.028	[3.024; 3.031]	0.055	0.030	[0.028; 0.034]
	LQL	3.098	[3.094; 3.102]	0.066	0.044	[0.041; 0.049]
250	GLQL1	3.082	[3.080; 3.084]	0.036	0.013	[0.012; 0.014]
	GLQL2	3.079	[3.077; 3.082]	0.040	0.016	[0.015; 0.018]
	GLQL3	3.031	[3.029; 3.033]	0.035	0.012	[0.011; 0.013]
	LQL	3.105	[3.103; 3.108]	0.041	0.017	[0.016; 0.019]
500	GLQL1	3.081	[3.080; 3.083]	0.027	0.007	[0.006; 0.008]
	GLQL2	3.079	[3.077; 3.081]	0.030	0.009	[0.008; 0.010]
	GLQL3	3.030	[3.029; 3.032]	0.026	0.006	[0.005; 0.007]
	LQL	3.105	[3.102; 3.106]	0.031	0.010	[0.009; 0.011]

Table 4.7.1: The mean and its 95% pointwise confidence interval, the standard deviation (SD), the variance ($Var \times 10$) and its 95% pointwise confidence interval ($\times 10$), computed for the 1000 estimators of the function $\eta(x)$ at $x = 0$. The censoring rate is $p = 20\%$ and three sample sizes (100, 250, 500).

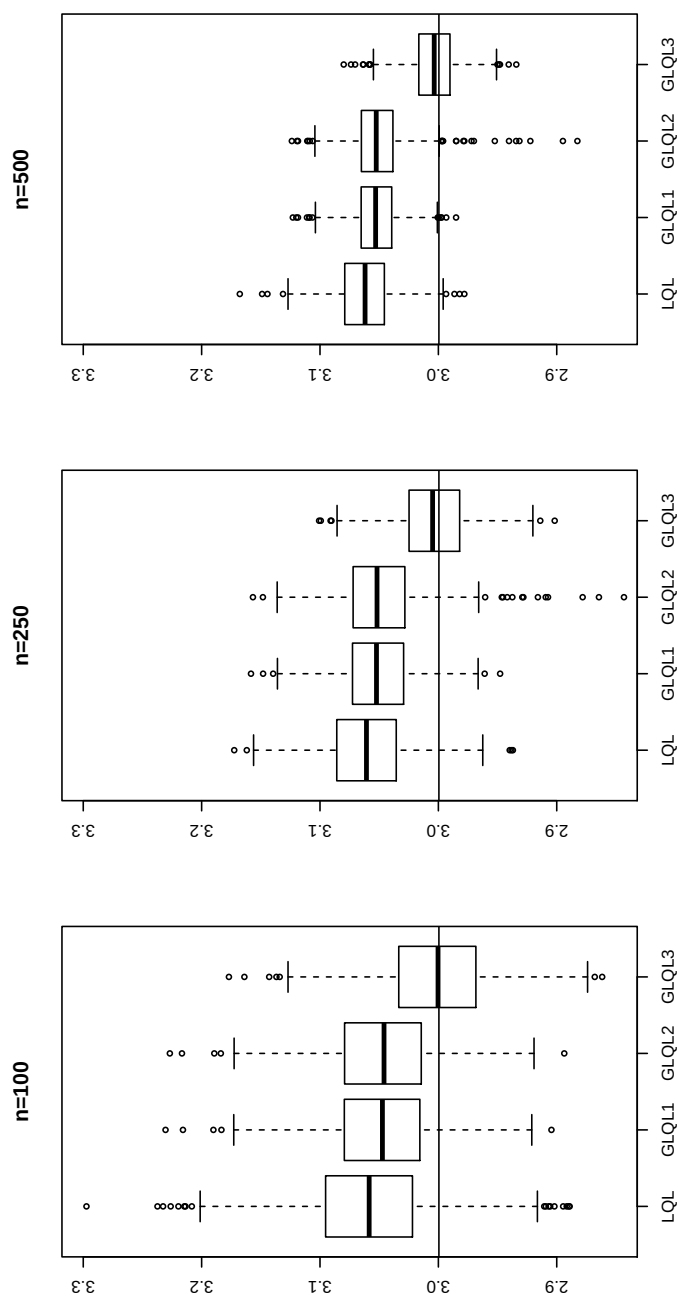


Figure 4.7.1: Boxplot of 1000 estimators of the function η at $x = 0$, three sample sizes (100, 250, 500) and a rate of censoring of 20%.

4.8 Appendix C: Proofs

We start this section with some notations. Set $\psi(x) = \eta(x) - \tilde{\eta}(x)$, $\bar{\eta}(x, u) = \psi(x) + \psi'(x)(u - x)$, $\mathbf{X}_k = (1, (X_k - x)/h)^T$ for $k = 1, \dots, n$, $S = (\mu_{i+j-2}^K)_{1 \leq i, j \leq 2}$, $S^* = (\nu_{i+j-2}^K)_{1 \leq i, j \leq 2}$, $U^K = (\mu_2^K, \mu_3^K)^T$ and $\mu_l = \int x^l K(x) dx$ for $l = 0, \dots, 3$. Let $\tilde{\beta}^*$ be a normalized estimator defined as follows:

$$\tilde{\beta}^* = (nh)^{1/2} (\tilde{\beta}_0 - \eta(x), h\{\tilde{\beta}_1 - \psi'(x)\})^T.$$

If $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)$ maximizes (4.3.1), then $\tilde{\beta}^*$ maximizes

$$\sum_{k=1}^n Q(g^{-1}(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k) + (nh)^{-1/2} \tilde{\beta}^{*T} \mathbf{X}_k), \hat{Y}_k^*) K\left(\frac{X_k - x}{h}\right),$$

with respect to $\tilde{\beta}^*$. Also define the following quantities:

$$\begin{aligned} \mathbf{V}_{n, \hat{G}} &= (nh)^{-1/2} \sum_{k=1}^n q_1(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k), \hat{Y}_k^*) \mathbf{X}_k K\left(\frac{X_k - x}{h}\right), \\ \mathbf{B}_{n, \hat{G}} &= (nh)^{-1} \sum_{k=1}^n q_2(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k), \hat{Y}_k^*) \mathbf{X}_k \mathbf{X}_k^T K\left(\frac{X_k - x}{h}\right), \end{aligned}$$

where $q_1(x, y) = (y - g^{-1}(x))\rho_1(x)$ and $q_2(x, y) = (y - g^{-1}(x))\rho_1'(x) - \rho_2(x)$. In order to prove Theorem 4.1, the following lemmas are needed.

Lemma 4.1. *Under the assumptions of Theorem 4.1 we have,*

$$\mathbf{B}_{n, \hat{G}} = -\rho_2(\eta(x))f_X(x)S + o_p(1) \equiv \mathbf{B} + o_p(1).$$

Proof. Set $\mathbf{B}_{n, G} = (nh)^{-1} \sum_{k=1}^n q_2(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k), Y_k^*) \mathbf{X}_k \mathbf{X}_k^T K\left(\frac{X_k - x}{h}\right)$. In view of conditions A1 and A7, for $1 \leq i, j \leq 2$, we have

$$\begin{aligned} |\mathbf{B}_{n, \hat{G}} - \mathbf{B}_{n, G}|_{ij} &= (nh)^{-1} \left| \sum_{k=1}^n \frac{\delta_k \phi(T_k) [\hat{G}(T_k^- | X_k) - G(T_k^- | X_k)]}{(1 - \hat{G}(T_k^- | X_k))(1 - G(T_k^- | X_k))} \rho_1'(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k)) \right. \\ &\quad \left. \times K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i+j-2} \right| \\ &\leq O_p(1) \times \sup_{t \leq \tau, x \in S_X} |\hat{G}(t^- | x) - G(t^- | x)| \\ &\quad \times (nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left| \frac{X_k - x}{h} \right|^{i+j-2}. \end{aligned}$$

The above supremum tends to zero in probability by the results of Du and Akritas (2002) and the empirical sum is bounded in probability by assumptions A1 and A5. Hence,

$$\mathbf{B}_{n,\hat{G}} - \mathbf{B}_{n,G} = o_p(1). \quad (4.8.1)$$

Now, note that $(\mathbf{B}_{n,G})_{ij} = (E\mathbf{B}_{n,G})_{ij} + O_p(\text{Var}\{(\mathbf{B}_{n,G})_{ij}\}^{1/2})$. Since q_2 is linear in y and using A2, A5 and A7, we obtain that

$$\begin{aligned} (E\mathbf{B}_{n,G})_{ij} &= h^{-1} E \left[q_2(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), m(X_1)) K\left(\frac{X_1 - x}{h}\right) \left(\frac{X_1 - x}{h}\right)^{i+j-2} \right] \\ &= h^{-1} \int q_2(\tilde{\eta}(u) + \bar{\eta}(x, u), m(u)) f_X(u) K\left(\frac{u - x}{h}\right) \left(\frac{u - x}{h}\right)^{i+j-2} du \\ &= \int q_2(\tilde{\eta}(x + vh) + \bar{\eta}(x, x + vh), m(x + vh)) f_X(x + vh) K(v) v^{i+j-2} dv \\ &\longrightarrow -\rho_2(\eta(x)) f_X(x) \mu_{i+j-2}^K \text{ as } n \rightarrow \infty. \end{aligned}$$

In view of Assumption 4.1, we have

$$\begin{aligned} (nh)\text{Var}(\mathbf{B}_{n,G})_{ij} &\leq h^{-1} E \left[q_2(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), Y_1^*)^2 \right. \\ &\quad \left. \times K^2\left(\frac{X_1 - x}{h}\right) \left(\frac{X_1 - x}{h}\right)^{2(i+j-2)} \right] \\ &\leq \int E[q_2(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), Y_1^*)^2 | X_1 = x + vh] f_X(x + vh) \\ &\quad \times K^2(v) v^{2(i+j-2)} dv \\ &= O(1). \end{aligned}$$

Therefore,

$$\mathbf{B}_{n,G} = -\rho_2(x) f_X(x) S + o_p(1). \quad (4.8.2)$$

The result of Lemma 4.1 is now a direct consequence of (4.8.1) and (4.8.2). $\square$

Lemma 4.2. *Suppose that the assumptions of Theorem 4.1 hold. Then,*

$$V_{n,G} - EV_{n,G} \xrightarrow{d} \mathcal{N}\left(0, \sigma_*^2(x) f_X(x) \rho_1^2(\eta(x)) S^*\right),$$

where $V_{n,G} = (nh)^{-1/2} \sum_{k=1}^n q_1(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k), Y_k^*) \mathbf{X}_k K\left(\frac{X_k - x}{h}\right)$.

Proof. Using Taylor's expansion of $\psi(\cdot)$, we have

$$\tilde{\eta}(x + vh) + \bar{\eta}(x, x + vh) = \eta(x + vh) - \frac{\psi''(x)}{2}(vh)^2 + o(h^2). \quad (4.8.3)$$

Since $q_1(\eta(\cdot), m(\cdot)) = 0$ and using Taylor expansion and (4.8.3), we obtain

$$q_1(\tilde{\eta}(x + vh) + \bar{\eta}(x, x + vh), m(x + vh)) = \frac{\psi''(x)}{2}(vh)^2 \rho_2(\eta(x + vh)) + o(h^2). \quad (4.8.4)$$

Note that $q_1(\cdot, y)$ is linear in y . For $1 \leq i \leq 2$, we have

$$\begin{aligned} E(V_{n,G})_i &= (n/h)^{1/2} E \left[q_1(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), m(X_1)) K \left(\frac{X_1 - x}{h} \right) \left(\frac{X_1 - x}{h} \right)^{i-1} \right] \\ &= (nh)^{1/2} \int q_1(\tilde{\eta}(x + vh) + \bar{\eta}(x, x + vh), m(x + vh)) f_X(x + vh) \\ &\quad \times K(v) v^{i-1} dv \\ &= (nh)^{1/2} \rho_2(\eta(x)) f_X(x) h^2 \frac{\psi''(x)}{2} \mu_{i+1}^K \{1 + o(1)\}. \end{aligned} \quad (4.8.5)$$

Thus,

$$E(V_{n,G}) = (nh)^{1/2} \rho_2(\eta(x)) f_X(x) h^2 \frac{\psi''(x)}{2} U^K \{1 + o(1)\}.$$

Now, from Assumption 4.1, we obtain

$$\begin{aligned} Var(V_{n,G}) &= h^{-1} Var \left[q_1(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), Y_1^*) \mathbf{X}_1 K \left(\frac{X_1 - x}{h} \right) \right] \\ &= h^{-1} E \left[Var \left\{ q_1(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), Y_1^*) \mathbf{X}_1 K \left(\frac{X_1 - x}{h} \right) \middle| X_1 \right\} \right] \\ &\quad + h^{-1} Var \left[E \left\{ q_1(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), Y_1^*) \mathbf{X}_1 K \left(\frac{X_1 - x}{h} \right) \middle| X_1 \right\} \right] \\ &= W_{1n,G} + W_{2n,G}. \end{aligned}$$

For $1 \leq i \leq 2$, we have

$$\begin{aligned} (W_{1n,G})_{ij} &= h^{-1} E \left\{ Var(Y_1^* | X_1) \rho_1^2(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1)) \left(\frac{X_1 - x}{h} \right)^{i+j-2} K^2 \left(\frac{X_1 - x}{h} \right) \right\} \\ &= \int \sigma_*^2(x + vh) f_X(x + vh) \rho_1^2(\tilde{\eta}(x + vh) + \bar{\eta}(x, x + vh)) K^2(v) v^{i+j-2} dv \\ &= \sigma_*^2(x) f_X(x) \rho_1^2(\eta(x)) \nu_{i+j-2}^K + O(h). \end{aligned}$$

The second term $W_{2n,G}$ can be bounded as follows:

$$W_{2n,G} \leq h^{-1} E \left[q_1^2(\tilde{\eta}(X_1) + \bar{\eta}(x, X_1), m(X_1)) \mathbf{X}_1 \mathbf{X}_1^T K^2 \left(\frac{X_1 - x}{h} \right) \right].$$

In view of expression (4.8.4) and conditions A1 and A7, we get

$$\begin{aligned} (W_{2n,G})_{ij} &\leq \int q_1^2(\tilde{\eta}(x + vh) + \bar{\eta}(x, x + vh), m(x + vh)) v^{i+j-2} K^2(v) dv \\ &= O(h^4), \quad \text{for } 1 \leq i, j \leq 2. \end{aligned}$$

Thus,

$$\text{Var}(V_{n,G}) = \sigma_*^2(x) f_X(x) \rho_1^2(\eta(x)) S^* + o(1).$$

Finally, it suffices to check the Lyapunov condition. Let $c \in \mathfrak{R}^2$, based on similar arguments to those used to develop (4.8.5), we can easily show that

$$\{c^T \text{Var}(V_{n,G}) c\}^{-3/2} \sum_{k=1}^n |c^T v_{G,k} - E c^T v_{G,k}| = O_p((nh)^{-1/2}),$$

where $v_{G,k} = q_1(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k), Y_k^*) \mathbf{X}_k K\left(\frac{X_k - x}{h}\right)$. The result of Lemma 4.2 is now a direct consequence of the Cramér-Wold device. $\square$

Proof of Theorem 4.1. Consider $\ell_{n,\hat{G}}(\beta^*)$ the normalized function defined as follows

$$\begin{aligned} \ell_{n,\hat{G}}(\beta^*) &= \sum_{k=1}^n \left\{ Q(g^{-1}(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k) + (nh)^{-1/2} \beta^{*T} \mathbf{X}_k), \hat{Y}_k^*) \right. \\ &\quad \left. - Q(g^{-1}(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k)), \hat{Y}_k^*) \right\} K\left(\frac{X_k - x}{h}\right). \end{aligned}$$

Then, $\tilde{\beta}^*$ maximizes $\ell_{n,\hat{G}}(\beta^*)$. Using a Taylor expansion of $Q\{g^{-1}(\cdot), \hat{Y}_i^*)\}$, we have

$$\ell_{n,\hat{G}}(\beta^*) = \mathbf{V}_{n,\hat{G}}^T \beta^* + \frac{1}{2} \beta^{*T} \mathbf{B}_{n,\hat{G}} \beta^* \{1 + o_p(1)\}.$$

Now by Lemma 4.1 and the quadratic approximation lemma of Fan and Gijbels (1996), we obtain

$$\tilde{\beta}^* = \mathbf{B}^{-1} \mathbf{V}_{n,\hat{G}} + o_p(1). \quad (4.8.6)$$

Next, write $\mathbf{V}_{n,\hat{G}} = (\mathbf{V}_{n,\hat{G}} - \mathbf{V}_{n,G}) + (\mathbf{V}_{n,G} - E\mathbf{V}_{n,G}) + E\mathbf{V}_{n,G}$. For $i = 1, 2$, we have

$$\begin{aligned} |\mathbf{V}_{n,\hat{G}} - \mathbf{V}_{n,G}|_i &= (nh)^{-1/2} \left| \sum_{k=1}^n \frac{\delta_k \phi(T_k) [\hat{G}(T_k^- | X_k) - G(T_k^- | X_k)]}{(1 - \hat{G}(T_k^- | X_k))(1 - G(T_k^- | X_k))} \right. \\ &\quad \left. \times \rho_1(\tilde{\eta}(X_k) + \bar{\eta}(x, X_k)) K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i-1} \right| \\ &\leq (nh)^{1/2} O_p(1) \times \sup_{t \leq \tau, x \in S_X} |\hat{G}(t^- | x) - G(t^- | x)| \\ &\quad \times (nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left| \frac{X_k - x}{h} \right|^{i-1}. \end{aligned}$$

From the results of Du and Akritas (2002), it follows that if $\frac{nh_0^5}{\log n} = O(1)$, then,

$$\sup_{t \leq \tau, x \in S_X} |\hat{G}(t^- | x) - G(t^- | x)| = O_p((nh_0)^{-1/2} (\log n)^{1/2}).$$

Since $(nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left| \frac{X_k - x}{h} \right|^{i-1} = O_p(1)$, we get

$$\mathbf{V}_{n,\hat{G}} - \mathbf{V}_{n,G} = O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right). \quad (4.8.7)$$

Finally, from Lemma 4.2 and equation (4.8.7), we obtain

$$\begin{aligned} \tilde{\beta}^* - (nh)^{1/2} \frac{1}{2} h^2 (\eta''(x) - \tilde{\eta}''(x)) S^{-1} U^K \{1 + o(1)\} + O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right) \\ \xrightarrow{d} \mathcal{N}(0, \sigma_*^2(x) f_X(x) \rho_1^2(\eta(x)) S^{-1} S^* S^{-1}). \end{aligned} \quad (4.8.8)$$

The result of Theorem 4.1 is a special case of (4.8.8). $\square$

Proof of Proposition 4.1.

1. Define

$$\begin{aligned} \mathbb{L}(\theta) &= E_\Delta \log f(X, Y; \theta), \\ \hat{\mathbb{L}}(\theta) &= \iint_\Delta \log f(x, y; \theta) d\hat{F}(x, y). \end{aligned}$$

Then, it is obvious that $\theta_* = \arg \max_{\theta \in \Theta} \mathbb{L}(\theta)$ and $\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{\mathbb{L}}(\theta)$. The proof of the first point is a direct consequence of Theorem 2.1 in Newey and McFadden (1994), under conditions D1.1, D2, D3, D5 and the first condition in Assumption 4.3.

2. In view of Corollary 5.8 in Bartle (1966), conditions D2, D3 and D4 ensure an interchange of differentiation and integration. Since $\Omega = \Omega(\theta^*)$ is non-singular by condition D6 and using Assumption 4.3, we get

$$\sup_{\theta \in \Theta} \|\nabla_{\theta}^2 \widehat{\mathbb{L}}(\theta) - \Omega(\theta)\| = o_p(1),$$

and

$$\sqrt{n} \nabla_{\theta}^1 \widehat{\mathbb{L}}(\theta_*) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Therefore, the second point results directly from the first point together with Theorem 3.1 in Newey and McFadden (1994). $\square$

Proof of Theorem 4.2. Similarly to the proof of Theorem 4.1, we now define $\widehat{\beta}^*$ and $\widetilde{\beta}^*$ the normalized guided estimators of β^* based on the estimated guide $\eta(\cdot, \widehat{\theta})$ and the fixed guide $\eta(\cdot, \theta_*)$, respectively. Define $\psi(x, \theta) = \eta(x) - \eta(x, \theta)$, $\bar{\eta}(x, u; \theta) = \psi(x, \theta) + \psi'(x, \theta)(u - x)$ and

$$\begin{aligned} B_{n, \hat{G}}(\theta) &= (nh)^{-1} \sum_{k=1}^n q_2(\eta(X_k, \theta) + \bar{\eta}(x, X_k; \theta), \hat{Y}_k^*) \mathbf{X}_k \mathbf{X}_k^T K\left(\frac{X_k - x}{h}\right), \\ V_{n, \hat{G}}(\theta) &= (nh)^{-1/2} \sum_{k=1}^n q_1(\eta(X_k, \theta) + \bar{\eta}(x, X_k; \theta), \hat{Y}_k^*) \mathbf{X}_k K\left(\frac{X_k - x}{h}\right). \end{aligned}$$

Write $B_{n, \hat{G}}(\hat{\theta}) = [B_{n, \hat{G}}(\hat{\theta}) - B_{n, G}(\hat{\theta})] + [B_{n, G}(\hat{\theta}) - B_{n, G}(\theta_*)] + B_{n, G}(\theta_*)$, where $B_{n, G}(\theta) = (nh)^{-1} \sum_{k=1}^n q_2(\eta(X_k, \theta) + \bar{\eta}(x, X_k; \theta), Y_k^*) \mathbf{X}_k \mathbf{X}_k^T K\left(\frac{X_k - x}{h}\right)$. Using a Taylor expansion, for $i, j = 1, 2$, we have

$$\begin{aligned} (B_{n, G}(\hat{\theta}) - B_{n, G}(\theta_*))_{i,j} &= (nh)^{-1} \sum_{k=1}^n \nabla_{\theta}^1 q_2(\eta(X_k, \tilde{\theta}) + \bar{\eta}(x, X_k; \tilde{\theta}), Y_k^*) (\hat{\theta} - \theta_*) \\ &\quad \times K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i+j-2}, \end{aligned}$$

for $\tilde{\theta}$ between $\hat{\theta}$ and θ_* . By assumptions A1, A7 and D1, there exists a constant $c > 0$ such that

$$|B_{n, G}(\hat{\theta}) - B_{n, G}(\theta_*)|_{i,j} \leq c \|\hat{\theta} - \theta_*\| \times (nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left|\frac{X_k - x}{h}\right|^{i+j-2}.$$

Note that $\hat{\theta} - \theta_*$ converges to zero in probability by Proposition 4.1 and the empirical sum is bounded in probability by assumptions A1 and A5. Thus, $B_{n,G}(\hat{\theta}) - B_{n,G}(\theta_*) = o_p(1)$. Now, we have

$$|B_{n,\hat{G}}(\hat{\theta}) - B_{n,G}(\hat{\theta})| \leq |D_{1n}| + |D_{2n}|,$$

where

$$\begin{aligned} (D_{1n})_{i,j} &= (nh)^{-1} \sum_{k=1}^n \frac{\delta_k \phi(T_k) [\hat{G}(T_k^- | X_k) - G(T_k^- | X_k)]}{(1 - \hat{G}(T_k^- | X_k))(1 - G(T_k^- | X_k))} K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i+j-2} \\ &\quad \times [\rho_1(\eta(X_k, \hat{\theta}) + \bar{\eta}(x, X_k; \hat{\theta})) - \rho_1(\eta(X_k, \theta_*) + \bar{\eta}(x, X_k; \theta_*))], \end{aligned}$$

and

$$\begin{aligned} (D_{2n})_{i,j} &= (nh)^{-1} \sum_{k=1}^n \frac{\delta_k \phi(T_k) [\hat{G}(T_k^- | X_k) - G(T_k^- | X_k)]}{(1 - \hat{G}(T_k^- | X_k))(1 - G(T_k^- | X_k))} \rho_1(\eta(X_k, \theta_*) + \bar{\eta}(x, X_k; \theta_*)) \\ &\quad \times K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i+j-2}. \end{aligned}$$

By a Taylor expression and assumptions A1, A7 and D1, we have

$$\begin{aligned} |(D_{1n})_{i,j}| &\leq O_p(1) \times \sup_{t \leq \tau, x \in S_X} |\hat{G}(t^- | x) - G(t^- | x)| \|\hat{\theta} - \theta_*\| \\ &\quad \times (nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left|\frac{X_k - x}{h}\right|^{i+j-2}. \end{aligned}$$

From Assumptions 4.1, 4.4, the results of Du and Akritas (2002) and Proposition 4.1, we get $D_{1n} = o_p(1)$. Similar arguments give $D_{2n} = o_p(1)$. Thus, $B_{n,\hat{G}}(\hat{\theta}) = B_{n,G}(\theta_*) + o_p(1)$. Therefore, by Lemma 4.1, we get

$$B_{n,\hat{G}}(\hat{\theta}) = \mathbf{B} + o_p(1). \quad (4.8.9)$$

Using (4.8.9) and similar arguments to those used to get equation (4.8.6), we have

$$\hat{\beta}^* - \tilde{\beta}^* = \mathbf{B}^{-1} [V_{n,\hat{G}}(\hat{\theta}) - V_{n,\hat{G}}(\theta_*)] + o_p(1).$$

Now, write $V_{n,\hat{G}}(\hat{\theta}) - V_{n,\hat{G}}(\theta_*) = [V_{n,G}(\hat{\theta}) - V_{n,G}(\theta_*)] + [V_{n,\hat{G}}(\hat{\theta}) - V_{n,G}(\hat{\theta})] + [V_{n,G}(\theta_*) - V_{n,\hat{G}}(\theta_*)]$, where $V_{n,G}(\theta) = (nh)^{-1/2} \sum_{k=1}^n q_1(\eta(X_k, \theta) + \bar{\eta}(x, X_k; \theta), Y_k^*) \mathbf{X}_k K\left(\frac{X_k - x}{h}\right)$.

Using a Taylor expansion, for $\dot{\theta}$ between $\hat{\theta}$ and θ_* , we have

$$\begin{aligned} V_{n,G}(\hat{\theta}) - V_{n,G}(\theta_*) &= (nh)^{-1/2} \sum_{k=1}^n \nabla_{\dot{\theta}}^1 q_1(\eta(X_k, \dot{\theta}) + \bar{\eta}(x, X_k; \dot{\theta}), Y_k^*)(\hat{\theta} - \theta_*) \\ &\quad \times \mathbf{X}_k K\left(\frac{X_k - x}{h}\right). \end{aligned}$$

By assumptions A1, A7 and D1, there exists a constant $c > 0$ such that for $i = 1, 2$, we get

$$|V_{n,G}(\hat{\theta}) - V_{n,G}(\theta_*)|_i \leq (nh)^{1/2} c \|\hat{\theta} - \theta_*\| \times (nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left|\frac{X_k - x}{h}\right|^{i-1}.$$

Since $\|\hat{\theta} - \theta_*\| = O_p(n^{-1/2})$, we have $V_{n,G}(\hat{\theta}) - V_{n,G}(\theta_*) = O_p(h^{1/2})$. Using expression (4.8.7), we get $V_{n,G}(\theta_*) - V_{n,\hat{G}}(\theta_*) = O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right)$. Finally, write

$$V_{n,\hat{G}}(\hat{\theta}) - V_{n,G}(\hat{\theta}) = I_{1n} + I_{2n},$$

where (for $i = 1, 2$)

$$\begin{aligned} (I_{1n})_i &= (nh)^{-1/2} \sum_{k=1}^n \frac{\delta_k \phi(T_k) [\hat{G}(T_k^- | X_k) - G(T_k^- | X_k)]}{(1 - \hat{G}(T_k^- | X_k))(1 - G(T_k^- | X_k))} K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i-1} \\ &\quad \times [\rho_1(\eta(X_k, \hat{\theta}) + \bar{\eta}(x, X_k; \hat{\theta})) - \rho_1(\eta(X_k, \theta_*) + \bar{\eta}(x, X_k; \theta_*))], \\ (I_{2n})_i &= (nh)^{-1/2} \sum_{k=1}^n \frac{\delta_k \phi(T_k) [\hat{G}(T_k^- | X_k) - G(T_k^- | X_k)]}{(1 - \hat{G}(T_k^- | X_k))(1 - G(T_k^- | X_k))} K\left(\frac{X_k - x}{h}\right) \left(\frac{X_k - x}{h}\right)^{i-1} \\ &\quad \times \rho_1(\eta(X_k, \theta_*) + \bar{\eta}(x, X_k; \theta_*)). \end{aligned}$$

By a Taylor expression and conditions A1, A7 and D1, we have

$$\begin{aligned} |I_{1n}|_i &\leq O_p(1) \times (nh)^{1/2} \times \sup_{t \leq \tau, x \in S_X} |\hat{G}(t^- | x) - G(t^- | x)| \\ &\quad \times \|\hat{\theta} - \theta_*\| \times (nh)^{-1} \sum_{k=1}^n K\left(\frac{X_k - x}{h}\right) \left|\frac{X_k - x}{h}\right|^{i-1}. \end{aligned}$$

From the results of Du and Akritas (2002) and Proposition 4.1, we get $I_{1n} = O_p\left(\left(\frac{h \log n}{nh_0}\right)^{1/2}\right)$. Similar arguments give $I_{2n} = O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right)$. Hence,

$$V_{n,\hat{G}}(\hat{\theta}) - V_{n,\hat{G}}(\theta_*) = O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right) + o_p(1).$$

Therefore,

$$\widehat{\beta}^* - \widetilde{\beta}^* = O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right) + o_p(1). \quad (4.8.10)$$

Finally, write $\widehat{\beta}^* = [\widehat{\beta}^* - \widetilde{\beta}^*] + \widetilde{\beta}^*$. From expressions (4.8.8) and (4.8.10), we get

$$\begin{aligned} \widehat{\beta}^* - (nh)^{1/2} \frac{1}{2} h^2 (\eta''(x) - \eta''(x, \theta_*)) S^{-1} U \{1 + o(1)\} + O_p\left(\left(\frac{h \log n}{h_0}\right)^{1/2}\right) \\ \xrightarrow{d} \mathcal{N}(0, \sigma_*^2(x) f_X(x) \rho_1^2(\eta(x)) S^{-1} S^* S^{-1}). \end{aligned}$$

This concludes the proof of Theorem 4.2 which is a special case of this last result. $\square$

Chapter 5

Conclusion and discussion

Thanks to its bias reduction property, parametrically guided nonparametric estimation is more and more investigated in different statistical areas. The application of the guided nonparametric method to density estimation, nonparametric regression, local quasi-likelihood, additive models and very recently varying coefficient models has revealed an improved performance for the guided estimator compared with the classical nonparametric estimator. However, these investigations are based on completely observed data. This dissertation was devoted to the generalization of the parametrically guided nonparametric estimation to the censored data case.

First, we investigated in Chapter 2 new parametrically guided kernel density and hazard function estimators for censored data. The proposed estimators are obtained by multiplying an initial parametric estimator by a nonparametric kernel type estimator of a suitable correction function. We established the asymptotic normality of the proposed estimators and obtained asymptotic expressions of the bias and variance. Under certain regularity conditions, we proved that the bias of the proposed estimator can be reduced compared to that of the traditional kernel estimator, while the variance does not change. Simulations confirmed the theoretical results and provide the following remarks for the density and the hazard functions: the guided kernel estimator with censored data outperforms the traditional kernel estimator if the parametric guide is equal or close to the true target function and performs as the traditional kernel estimator if the parametric guide is misspecified. The comparison to the log-spline estimator shows that the guided kernel estimator is generally better

in terms of the mean squares error. Moreover, confidence intervals and an extension to the Cox model are developed and applied to the UIS data.

As pointed out by Hjort and Glad (1995) in the uncensored case, the advantages of the multiplicative PGK method come with some drawback caused by the correction factor $f_{\hat{\theta}}(t)/f_{\hat{\theta}}(T(i))$, see equation (2.2.3). Small values of $f_{\hat{\theta}}(T(i))$ may affect the numerical stability of the resulting estimator especially with a “large” bandwidth and this affects also the MSE. One may correct for this by adding a small ϵ to both the numerator and the denominator or by adopting an additive parametric correction (instead of a multiplicative one). This is a very interesting subject for future research. Note also that the kernel approach investigated in this thesis is based on a symmetric kernel density with a bounded support. However, in many situations, using the classical kernel leads to the well known boundary effect problem, that is, the estimator has a large bias near the endpoints. In our real data analysis, we found out that our guided kernel estimator seems to correct in an automatic way for possible boundary effects. Bouezmarni *et al.* (2011) proposed a gamma kernel estimator that corrects for the boundary effects. It would be interesting to perform a comparison between the guided kernel estimator and the gamma kernel estimator. It would be also of interest to develop a guided kernel estimator based on a gamma kernel, which we expect may reduce more significantly the boundary effects.

In Chapter 3, using a synthetic data approach, we adapted and proved the asymptotic normality of the parametrically guided local linear estimator. We investigated both the additive correction and the unified family of corrections. Under certain conditions, we found that the bias of the guided local linear estimator can be reduced compared to that of the local linear estimator, while the variance remains unchanged. Simulations confirm the theoretical results and provide the following conclusions: as in the uncensored framework, the guided local linear estimator with censored data outperforms the local linear estimator if the parametric guide is equal or close to the true regression curve and performs as the local linear estimator if the guide is linear or misspecified. In this thesis, we focused on the kernel smoothing and the local linear fitting. A generalization to other smoothing techniques such as local polynomial and splines will be studied in the future.

In Chapter 4, we focused on the parametrically guided local quasi-likelihood esti-

mation with censored data. To deal with censoring, we also considered the synthetic data approach and investigated the additive correction as well as the unified family of corrections. Our results provide a generalization to the censored data case of both the results of Fan *et al.* (1995) and Fan *et al.* (2009). The asymptotic results confirm the bias reduction property of the guided local quasi-likelihood estimator in the presence of censoring mechanism. The results also show that when an optimal bandwidth and an appropriate parametric guide are used the variance can also be reduced. Our finite sample simulation investigated both the case of discrete and continuous responses. The simulation results corresponded quite closely to the theoretical results and proved that the guided local quasi-likelihood estimator outperforms the unguided local-quasi-likelihood estimator in terms of bias and mean squared error. We emphasize that in the current study we considered the simple case where the variance function V is known. An interesting extension would be to generalize the guided local quasi-likelihood estimator to the more realistic case of unknown variance function. We also note that, in our simulation studies, we observed that the guided local quasi-likelihood is not much more computational demanding compared with the unguided local quasi-likelihood. An issue with great importance for any method based on local quasi-likelihood estimation is the computational time. Despite the advantages of the guided local quasi-likelihood estimation it still shares same drawback as the classical local quasi-likelihood, which is related to its intensive computation. In fact, the local quasi-likelihood equations must be solved at every data point. In the uncensored data case, Fan and Chen (1999) proposed a one-step local quasi-likelihood estimation that overcomes the computational drawbacks. A similar approach may be investigated and generalized to censored data in order to improve the computation time of the guided local quasi-likelihood estimator. Apart from Koul *et al.* (1981) transformation, other transformations can also be investigated. We mention for example, Buckley and James (1979) and Zheng (1987) who proposed iterative procedures for the implementation of the synthetic data.

It is also worth mentioning that in our different simulation studies, we have seen that the guided approach improves the estimation when the parametric guide is correct by reducing the bias or even reducing the variance in many situations because it allows to choose a larger bandwidth. However, even if it still enhances the

estimation when the parametric guide is misspecified, the use of a completely wrong guide is not helping here or may even affect negatively the quality of the estimation in some cases. Consequently, it is important to choose an appropriate guide. One can for example choose the parametric guide based on the purely nonparametric fit or uses the goodness-of-fit tests to check the validity of the parametric guide; see for example, amongst others, Pardo-Fernández *et al.* (2007) who proposed a goodness-of-fit test based on a comparison between the estimator of the error distribution relying on parametric residuals and the one based on nonparametric residuals. Nevertheless, further investigation is needed to address the goodness-of-fit issue in the framework of guided estimation.

It is also worth noting that the estimators proposed in this dissertation focused on the case with a univariate explanatory variable. The extension to the multivariate covariate case is straightforward. However, increasing the dimensionality leads to the well known "curse of dimensionality" for the nonparametric part, while selecting an appropriate multivariate parametric guide is more challenging than in the univariate covariate case. As alternative, one can extend the parametrically guided generalized additive models of Fan *et al.* (2014) to censored data using a synthetic data approach, see for example Debbarh and Viallon (2007, 2008) for the adaptation of additive models to censored data.

Another point that may be considered in future research is quantile regression. In fact, the regression techniques that we discussed in this thesis summarize the average relationship between a response Y and a covariate X based on the conditional mean function. However, with a skewed conditional distribution $F_{Y|X}$, the conditional median or other quantiles of the response are more robust and appropriate measures of central tendency. In such a case, one could use a quantile regression rather than a mean regression. First introduced by Koenker and Bassett (1978) in the context of linear regression, quantile regression allows to estimate the effect of a covariate on various quantiles in the conditional distribution of Y . In the censored data framework, quantile regression becomes attractive as an alternative to popular regression techniques like Cox proportional hazards model or the accelerated failure time model; see Koenker and Biliias (2001) and Koenker and Geling (2001), among many others. Considerable work has been devoted to the nonparametric quantile regression with

censored data, see for example Dabrowska (1992), Van Keilegom and Veraverbeke (1998) and El Ghouch and Van Keilegom (2009), among others. With completely observed data, El Ghouch and Genton (2009) proposed a guided local polynomial quantile estimator using an additive correction and a local polynomial fit. Their approach may be generalized to censored data using a synthetic data approach, see El Ghouch and Van Keilegom (2009).

Another important future research task is to extend the guided estimation to other types of censoring such as left censoring and interval censoring. An observation is left censored when the event of interest has already occurred before the start of the study, while the problem of interval censoring arises when a variable of interest is known only to lie in an interval. Moreover, another more complex issue in survival analysis, is the presence of truncation which is due to sampling bias, since only individuals whose lifetimes lie within a certain interval can be observed. Note also that in addition to censoring and truncation, it is more and more observed that some patients are cured permanently. In the literature, this is referred to as the "cure models". An adaptation of the guided estimation to accommodate the cured subjects, other types of censoring, and truncation may broaden the applicability of the guided estimation. Finally, it would be useful to develop an R package that efficiently implements the parametrically guided nonparametric estimator for every type of corrections and different contexts.

Publications

- [1] Talamakrouni, M., El Ghouh, A. and Van Keilegom, I. (2015). Guided censored regression. *Scand. J. Stat.* **42**, 214-233.

- [2] Talamakrouni, M., Van Keilegom, I. and El Ghouh, A. (2016a). Parametrically guided nonparametric density and hazard estimation with censored data. *Comput. Statist. Data Anal.* **93**, 308-323.

- [3] Talamakrouni, M., El Ghouh, A. and Van Keilegom, I. (2016b). Parametrically guided local quasi-likelihood with censored data. (Submitted).

Bibliography

- [1] Bartle, R. (1966). *The Elements of Integration*. John Wiley and Sons, New York.
- [2] Bender, R., Augustin T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **24**, 1713-1723.
- [3] Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report, Univ. California, Berkeley*.
- [4] Berg, A. and Politis, D. N. (2007). Density estimation of censored data with infinite-order kernels. unpublished manuscript.
- [5] Blum, J. R. and Susarla, V. (1980). Maximal deviation theory of density and failure rate estimates based on censored data. *J. Multiv. Anal.* **5**, 213-222.
- [6] Bouezmarni, T., El Ghouh, A. and Mesfioui, M. (2011). Gamma kernel estimators for density and hazard rate of right-censored data. *J. Probab. Stat.* **2011**, 16 pages.
- [7] Breslow, N. E. and Crowley, J. J. (1974). A large sample study of the life table and the product limit estimates under random censorship. *Ann. Statist.* **2**, 437-453.
- [8] Buckland, S.T. (1992). Maximum likelihood fitting of Hermite and simple polynomial densities. *J. Roy. Statist. Soc. Ser. C*, **41**, 241-266.
- [9] Buckley, J. and James, I. (1979). Linear-regression with censored data. *Biometrika* **66(3)**, 429-436.
- [10] Burke, M. D., Csörgö, S. and Horváth, L. (1981). Strong approximations of some biometric estimates under random censorship. *Z. Wahrsch. Verw. Gebiete*

- 56**, 87-112.
- [11] Castro-Kuriss, C. (2011). On a goodness-of-fit test for censored data from a location-scale distribution with applications. *Chil. J. Stat.* **2**, 115-136.
 - [12] Chen, J., Fan, J., Li, K. and Zhou, H. (2006). Local quasi-likelihood estimation with data missing at random. *Statist. Sinica* **16**, 1071-1100.
 - [13] Collett, D. (1998). *Modeling survival data in medical research*. Chapman & Hall, New York.
 - [14] Copas, J. B. (1995). Local likelihood based on kernel censoring. *J. Roy. Statist. Soc. Ser. B*, **57**, 221-235.
 - [15] Cox, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B*, **34**, 187-220.
 - [16] Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Stat.* **14**, 181-197.
 - [17] Dabrowska, D. M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scand. J. Stat.* **19**, 351-361.
 - [18] Davenport, C., Maity, A. and Wu, Y. (2015). Parametrically guided estimation in nonparametric varying coefficient models with quasi-likelihood. *J. Nonparametr. Statist.* **27**, 195-213.
 - [19] Debbbarh, M. and Viallon, V. (2007). Mean square convergence for an estimator of the additive regression function under random censorship. *C. R. Acad. Sci. Paris Ser. I*, **344(3)**, 205-210.
 - [20] Debbbarh, M. and Viallon, V. (2008). Uniform limit laws of the logarithm for estimators of the additive regression function in the presence of right censored data. *Electron. J. Stat.* **2**, 516-541.
 - [21] Delecroix, M., Lopez, O. and Patilea, V. (2008). Nonlinear censored regression using synthetic data. *Scand. J. Stat.* **35**, 248-265.
 - [22] Ding, Y. and Nan, B. (2011). A sieve M-theorem for bundled parameters in semi-parametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.* **39**, 3032- 3061.

-
- [23] Doksum, K. A. and Yandell, B. S. (1982). Properties of regression estimates based on censored survival. In Festschrift for Erich L. Lehmann (eds P. J. Bickel, K. A. Doksum, & J. L. Hodges), 140-156.
- [24] Du, Y. and Akritas, M. G. (2002). I.i.d. representations of the conditional Kaplan-Meier process for arbitrary distributions. *Math. Methods Statist.* **11**, 152-182.
- [25] Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel type function estimators. *Ann. Statist.* **33**, 1380-1403.
- [26] Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *J. Amer. Statist. Assoc.* **24**, 2431-2461.
- [27] Eguchi, S. and Copas, J. B. (1998). A class of local likelihood methods and near-parametric asymptotics. *J. Roy. Statist. Soc. Ser. B*, **60**, 709-724.
- [28] El Ghouch, A. and Genton, M. G. (2009). Local polynomial quantile regression with parametric features. *JASA.* **104**, 1416-1429.
- [29] El Ghouch, A. and Van Keilegom, I. (2008). Nonparametric regression with dependent censored data. *Scand. J. Stat.* **35**, 228-247.
- [30] El Ghouch and Van Keilegom. (2009). Local linear quantile regression with dependent censored data. *Stat. Sinica.* **19**, 1621-1640.
- [31] Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *J. Roy. Statist. Soc. Ser. B*, Vol. **64**, 927-943.
- [32] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008-2036.
- [33] Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89**, 560-570.
- [34] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall, London.
- [35] Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

-
- [36] Fan, J., Maity, A., Wang, Y. and Wu, Y. (2014). Parametrically guided generalized additive models with application to mergers and acquisitions data. *J. Nonparametr. Statist.* **25**, 109-128.
- [37] Fan, Y. and Ullah, A. (1999). Asymptotic normality of a combined regression estimator. *J. Multivariate Anal.* **71**, 191-240.
- [38] Fan, J., Wu, Y. and Feng, Y. (2009). Local quasi-likelihood with a parametric guide. *Ann. Statist.* **37**, 4153-4183.
- [39] Faraway, J. (1989). Implementing semiparametric density estimation. *Statist. Probab. Lett.* **10**, 141-143.
- [40] Farrington, C. P. (1996). Interval censored survival data: a generalized linear modelling approach. *Stat. Probab. Med.* **15**, 283-292.
- [41] Fiorio, C. V. (2004). Confidence intervals for kernel density estimation. *Stata J.* **4**(2), 168-179.
- [42] Földes, A., Rejtő, L. and Winter, B. 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.
- [43] Friedman, J. H., Stuetzle, W. and Schroeder, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79**, 599-608.
- [44] Gefeller, O. and Dette, H. (1992). Nearest neighbor kernel estimation of the hazard function from censored data. *J. Statist. Comput. Simul.* **43**, 93-101.
- [45] Gefeller, O. and Michels, P. (1992). A review on smoothing methods for the estimation of the hazard rate based on kernel functions, in Dodge, Y. and Whittaker, J. (eds). *Computational Statistics*, Physica-Verlag, Switzerland, 459-464.
- [46] Gijbels, I. and Veraverbeke, N. (1989). Sequential fixed width confidence intervals for quantiles in the presence of censoring. *J. Statist. Plan. Infer.* **22**, 213-222.
- [47] Gill, R.D. (1983). Large sample behaviour of the product limit estimator on the whole line. *Ann. Statist.* **11**, 49-58.

- [48] Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.* **37**, 503-522.
- [49] Glad, I. K. (1998 a). Parametrically guided nonparametric regression. *Scand. J. Stat.* **25**, 649-668.
- [50] Glad, I. K. (1998 b). A note on unconditional properties of a parametrically guided Nadaraya-Watson estimator. *Statistics and Probability Letters.* **37**, 101-108.
- [51] Glad, I. K., Hjort, N. L. and Ushakov, N.G. (2003). Correction of density estimators that are not densities. *Scand. J. Stat.* **30(2)**, 415-427.
- [52] González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of generalized Kaplan-Meier estimator with some applications. *J. Nonparametr. Statist.* **4**, 65-78.
- [53] González-Manteiga, W., Cao, R. and Marron, J. S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *J. Amer. Statist. Assoc.* **91**, 1130-1140.
- [54] Gozalo, P. and Linton, O. (2000). Local nonlinear least squares: using parametric information in nonparametric regression. *J. Econometrics* **99**, 63-106.
- [55] Green, P. J. and Yandell, B. S. (1985). Semi-parametric generalized linear models. In *Lecture Notes in Statistics* **32**, 44-55.
- [56] Hakulinen, T. and Tenkanen L. (1987). Regression analysis of relative survival rates. *Appl Statist.* **36**, 309-317.
- [57] Härdle, W. (1990), *Applied Nonparametric Regression*. Cambridge University Press, New York.
- [58] Heuchenne, C. and Van Keilegom, I. (2012). Estimation of a general parametric location in censored regression. In : Van Keilegom, I. and Wilson, P.W. (eds.), *Exploring research frontiers in contemporary statistics and econometrics- Festschrift in honor of L. Simar*, Springer, 177-187

-
- [59] Hjort, N.L. (1986). *Theory of Statistical Symbol Recognition*. Norwegian Computing Centre, Oslo.
- [60] Hjort, N.L. (1992). On inference in parametric survival data models. *Int. Stat. Rev.* **60**, 355-387.
- [61] Hjort, N. L. and Glad, I. K. (1994). *Nonparametric density estimation with a parametric start*. Statistical research report, Dept. Mathematics, Univ. Oslo.
- [62] Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23**, 882-904.
- [63] Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619-1647.
- [64] Horowitz, J. L. (2001). *The bootstrap*. In Handbook of Econometrics, ed. J. J. Heckeman and E. Leamer. **5**, 3159-3228. Amsterdam: North-Holland.
- [65] Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition*, John Wiley and Sons Inc., New York.
- [66] Hosmer, D., Lemeshow, S., Sturdivant, R.X. (2013). *Applied Logistic Regression*, 3rd Ed. A Wiley-Interscience Publication, John Wiley and Sons Inc., New York, NY, 500 pages.
- [67] Hunsberger, S. (1994). Semiparametric regression in likelihood-based models. *J. Amer. Statist. Assoc.* **89**, 1354-1365.
- [68] Hurley, M. A. (1985). An application of generalized linear models to survival analysis with two types of failures. *J. Roy. Statist. Soc. Ser. C*, **34**, 273-281.
- [69] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical Analysis of Failure Time Data*. 2nd ed. Wiley, New York.
- [70] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457-481.
- [71] Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis. Techniques for Censored and Truncated Data*. Springer, New York.

-
- [72] Koenker, R. (2005). *Quantile Regression*, Volume 38 of Econometric Society Monographs. Cambridge: Cambridge University Press.
- [73] Koenker, R. and G. Bassett, Jr. (1978). Regression quantiles. *Econometrica* **46**, 33-50.
- [74] Koenker, R. and Y. Biliass (2001). Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments. *Empirical Economics* **26**, 199-220.
- [75] Koenker, R. and O. Geling (2001). Reappraising medfly longevity: a quantile regression survival analysis. *J. Amer. Statist. Assoc.* **96**, 458-468.
- [76] Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation. *Comput. Stat. Data Anal.* **12**, 327-347.
- [77] Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *J. Comp. Graph. Stat.* **1**, 301-328.
- [78] Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *Ann. Statist.* **9**, 1276-1288.
- [79] Lai, T. L. and Ying, Z. (1991). Large-sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist.* **19**, 1370-1402.
- [80] Lai, T. L., Ying, Z. and Zheng, Z. K. (1995). Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. *J. Multivariate Anal.* **52**, 259-279.
- [81] Legrand, C. (2009). *Assessing heterogeneity in clinical trials using the frailty model: Quantifying, interpreting and explaining heterogeneity over centers in multicenter cancer clinical trials*, VDM Verlag Dr. Müller e.K., 978-3-639-21334-8. 188 p.
- [82] Legrand, C., Duchateau, L., Janssen, P., Ducrocq, V. and Sylvester, R. (2009). Validation of prognostic indices using the frailty model. *Lifetime Data Analysis* **15**, 59-78.

-
- [83] Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* **74**, 301-309.
- [84] Liang, H. and Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Comput. Statist. Data Anal.* **52**, 2538-2548.
- [85] Lo, S. H., Mack, Y.P. and 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.
- [86] Lo, S. H. and Singh, K. (1986). The product-limit estimator and the bootstrap: Some asymptotic representation. *Probab. Theory Relat. Fields* **71**, 455-465.
- [87] Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Comm. Statist. Theory Methods* **40**, 2639-2660.
- [88] Lopez, O. and Patilea, V. (2009). Nonparametric lack-of-fit tests for parametric mean-regression models with censored data. *J. Multivariate. Anal.* **100**, 210-230.
- [89] Lu, X. and Cheng, T. L. (2007). Randomly censored partially linear single-index models. *J. Multivariate Anal.* **98(10)**, 1895-1922.
- [90] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712-736.
- [91] Martins-Filho, C., Mishra, S. and Ullah, A. (2008). A class of improved parametrically guided nonparametric regression estimators. *Econometric Reviews* **27**, 542-573.
- [92] Martins-Filho, C. and Yao, F. (2006). A note on the use of V and U statistics in nonparametric models of regression. *Ann. Inst. Statist. Math.* **58**, 389-406.
- [93] Mays, J. E., Birch, J. B. and Starnes, B. A. (2001). Model robust regression: combining parametric, nonparametric and semiparametric methods. *J. Nonparametr. Stat.* **13**, 245-277.
- [94] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall.

- [95] Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in the case of censored data. *Ann. Statist.* **14**, 766-773.
- [96] Miller, R. G. (1981). *Survival Analysis*. John Wiley and Sons.
- [97] Müller, H. G. and Wang, J. L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61-76.
- [98] Naito, K. (2004). Density estimation by local L2-fitting. *Ann. Statist.* **32**, 1162-1191.
- [99] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A*, **135**, 370-384.
- [100] Newey, W. K. and McFadden, D. (1999). Large sample estimation and hypothesis testing. *In Handbook of Econometrics*. Vol. 4. Edited by D. McFadden & R. Engle, Amsterdam, The Netherlands.
- [101] Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Stat.* **19**, 25-43.
- [102] Oakes, D. (1986). An approximate likelihood procedure for censored data. *Biometrics* **42**, 177-182.
- [103] Olkin, I. and Spiegelman, C.H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **82**, 858-865.
- [104] O'Sullivan, F., Yandell, B. S. and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81**, 96-103.
- [105] Padgett, W. J. and McNichols, D. T. (1984). Nonparametric density estimation from censored data. *Commun. Statist.-Theory Meth.* **13**, 1581-1611.
- [106] Pardo-Fernández, J. C., Van Keilegom, I. and González-Manteiga, W. (2007). Goodness-of-fit tests for parametric models in censored regression. *Can. J. Stat.* **35(2)**, 249-264.
- [107] Patil, P. N. (1993). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inf.* **35**, 15-30.

-
- [108] Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.* **72**, 854-858.
- [109] Rahman, M. and Ullah, A. (2002). Improved combined parametric and non-parametric regression: estimation and hypothesis testing. In *Handbook of Applied Econometrics and Statistical Inference*, Edited by Ullah, A., Wan, A., Chaturvedi, A. Marcel Dekker, New York.
- [110] Sánchez-Sellero, C., González-Manteiga, W. and Cao, R. (1999). Bandwidth selection in density estimation with truncated and censored data. *Ann. Inst. Statist. Math.* **51**, 51-70.
- [111] Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semi-parametric models. *J. Amer. Statist. Assoc.* **89**, 501-511.
- [112] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Verlag, New York.
- [113] Singh, R. S. and Lu, X. (1999). Nonparametric synthetic data regression estimation for censored survival data. *J. Nonparametr. Stat.* **11**, 13-31.
- [114] Singpurwalla, N. D. and Wong, M. Y. (1983), Kernel estimators of the failure-rate function and density estimation: an analogy. *J. Amer. Statist. Assoc.* **78**, 478-481.
- [115] Srinivasan, C. and Zhou, M. (1994). Linear regression with censoring. *J. Multivariate Anal.* **49**, 179-201.
- [116] Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.* **84**, 276-283.
- [117] Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595-645.
- [118] Strzalkowska-Kominiak, E., Cao, R., (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *J. Multiv. Anal.* **114**, 74-98.
- [119] Suzukawa, A., Imai, H. and Sato, Y. (2001). Kullback-Leibler information consistent estimation for censored data. *Ann. Inst. Statist. Math.* **53**, 262-276.

-
- [120] Talamakrouni, M., El Ghouh, A. and Van Keilegom, I. (2015). Guided censored regression. *Scand. J. Stat.* **42**, 214-233.
 - [121] Talamakrouni, M., Van Keilegom, I. and El Ghouh, A. (2016a). Parametrically guided nonparametric density and hazard estimation with censored data. *Comput. Statist. Data Anal.* **93**, 308-323.
 - [122] Talamakrouni, M., El Ghouh, A. and Van Keilegom, I. (2016b). Parametrically guided local quasi-likelihood with censored data. (Submitted).
 - [123] Tanner, M. and Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989-993.
 - [124] Van Keilegom, I. and Akritas, M. G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.* **27**, 1745-1784.
 - [125] Van Keilegom, I. and Veraverbeke, N. (1997 a). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Statist. Math.* **49**, 467-491.
 - [126] Van Keilegom, I. and Veraverbeke, N. (1997 b). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles. *Comm. Statist. Theory Methods* **25**, 2251-2265.
 - [127] Van Keilegom, I. and Veraverbeke N. (1998). Bootstrapping quantiles in a fixed design regression model with censored data. *J. Statist. Plann. Inference* **69**, 115-131.
 - [128] Veraverbeke, N., Gijbels, I. and Omelka, M. (2014). Preadjusted non-parametric estimation of a conditional distribution function. *J. Roy. Statist. Soc. Ser. B* **76**, 399-438.
 - [129] Vaupel, J. W., Manton, K. G., Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-454.
 - [130] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
 - [131] Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Chapman and Hall.

-
- [132] Wang, Y., He, S., Zhu, L. and Yuen, K.C. (2007). Asymptotics for censored generalized linear model with unknown link function. *Probab. Theory Relat. Fields*, **138**, 235-267.
- [133] Wang, B. and Wang, X-F. (2007). *Bandwidth selection for weighted kernel density estimation* (unpublished manuscript).
- [134] Wang, J. F., Liang, H. Y. and Fan, G. H. (2013). Local polynomial quasi-likelihood regression with truncated and dependent data. *J. Theoret. Appl. Statist.* **47** (4), 744-761.
- [135] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.
- [136] Xiang, X. J. (1994). Law of the logarithm for density and hazard rate estimation for censored data, *J. Multiv. Anal.* **49**, 278-286.
- [137] Yao, F. and Martins-Filho, C. (2015). An asymptotic characterization of finite order U-statistics with sample size dependent kernels: applications to non-parametric estimators and test statistics. *Comm. Statist. Theory Methods* **44**, 3251-3265.
- [138] Yu, L., Yu, R. and Liu, L. (2009). Quasi-likelihood for right-censored data in the generalized linear model. *Comm. Statist. Theory Methods* **38**, 2187-2200.
- [139] Yu, L. (2011). Nonparametric quasi-likelihood for right censored data. *Lifetime Data Anal.* **17**, 594-607.
- [140] Yu, L. and Peace, K. E. (2012). Spline nonparametric quasi-likelihood regression within the frame of the accelerated failure time model. *Comput. Statist. Data Anal.* **56**, 2657-2687.
- [141] Yu, Z. and van der Laan, M. (2006). Double robust estimation in longitudinal marginal structural models. *J. Statist. Plann. Inference* **136**, 1061-1089.
- [142] Yu, L., Yu, R., Liu, L. and Chen, D. G. (2012). Extended quasi-likelihood with fractional polynomials in the frame of the accelerated failure time model. *Statistic in Medicine* **31** (13), 1369-1379.

-
- [143] Zheng, Z. (1987). A class of estimators of the parameters in linear regression with censored data. *Acta Mathem. Applic. Sinica* **3**, 231-241.
- [144] Zheng, Z. (1988). Strong consistency of nonparametric regression estimates with censored data. *I. Math. Res. Exposit.* **8**, 307-313.
- [145] Zhou, M. (1992). Asymptotic normality of the synthetic data regression estimator for censored survival data. *Ann. Statist.* **20**, 1002-1021.