Bernoulli **16**(2), 2010, 514–542
DOI: 10.3150/09-BEJ221

Conditional density estimation in a censored single-index regression model

OLIVIER BOUAZIZ and OLIVIER LOPEZ*

*Laboratoire de Statistique Théorique et Appliquée, Université Paris VI, 175 rue du Chevaleret, 75013 Paris, France. E-mail: *olivier.lopez@ens-cachan.org*

Under a single-index regression assumption, we introduce a new semiparametric procedure to estimate a conditional density of a censored response. The regression model can be seen as a generalization of the Cox regression model and also as a profitable tool for performing dimension reduction under censoring. This technique extends the results of Delecroix *et al.* [*J. Multivariate Anal.* **86** (2003) 213–226]. We derive consistency and asymptotic normality of our estimator of the index parameter by proving its asymptotic equivalence with the (uncomputable) maximum likelihood estimator, using martingales results for counting processes and arguments from empirical processes theory. Furthermore, we provide a new adaptive procedure which allows us both to choose the smoothing parameter involved in our approach and to circumvent the weak performances of the Kaplan–Meier estimator [*Amer. Statist. Assoc.* **53** (1958) 457–481] in the right-tail of the distribution. By means of a simulation study, we study the behavior of our estimator for small samples.

Keywords: asymptotic normality; censoring; empirical processes; martingales for counting processes; pseudo-maximum likelihood; single-index model

1. Introduction

A major concern in recent papers dealing with censored regression is to propose alternatives to the popular Cox regression model. This model, also known as the multiplicative hazard regression model (see Cox (1972)), states some semi-parametric assumptions on the conditional hazard function. Estimation in this model is traditionally performed using pseudolikelihood techniques and the theoretical properties of these procedures are covered in a large number of papers (see, for example, Fleming and Harrington (1991)). However, in some situations, the assumptions of the Cox regression model are obviously not satisfied by the data set. In this paper, our aim is to perform estimation in a semi-parametric regression model which allows more flexibility than the Cox regression model. This new technique can be seen as a particularly interesting alternative since it is valid in a larger number of situations than the multiplicative hazard model.

Alternatives to the Cox regression model mostly focus on the estimation of a conditional expectation or of a quantile regression model. Koul *et al.* (1981), Stute (1999) and Delecroix *et al.* consider mean regression models where the regression function belongs to a parametric family, but with an unknown distribution of the residuals. Parametric quantile regression was studied by Gannoun *et al.* (2005). On the other hand, Lu and Burke (2005) and Lopez (2009) considered a semi-parametric single-index regression model. Single-index regression models were initially introduced to circumvent the so-called "curse of dimensionality" in nonparametric regression

1350-7265

(see, for example, Ichimura (1993)), by assuming that the conditional expectation depends only on an unknown linear combination of the covariates. Another appealing aspect of such models is that they include the Cox regression model as a particular case. The main assumption of this model is that the conditional density depends only on an unknown linear combination of the covariates, while the multiplicative hazard model makes a similar assumption on the conditional hazard rate. In this paper, we focus on estimation of the parameter in a regression model in which the conditional density of the response satisfies a single-index assumption. We provide asymptotic results for a new M-estimation procedure for the index parameter. This procedure can be seen as a generalization of the method of Delecroix *et al.* (2003) to the case of censored regression.

As in the uncensored case, we show that, as regards the estimation of the parametric part of our model, there is an asymptotic equivalence between our semi-parametric approach and a parametric one relying on some prior knowledge of the family of regression functions. For the nonparametric part, we use kernel estimators of conditional densities as in Delecroix *et al.* (2003). Since the performance of kernel estimators strongly relies on the choice of the smoothing parameter, we also provide a method to choose this parameter adaptively. Another technical issue in our approach concerns a truncation parameter involved in our procedure. This problem of truncation comes directly from the censored framework, where estimators of the underlying distribution functions sometimes fail to correctly estimate the tail of the distribution. This problem is traditionally circumvented by, for example, introducing integrability assumptions on the response and censoring distribution; see, for example, Stute (1999). On the other hand, the truncation procedure consists of removing the observations which are too large in the estimation of the regression function; see, for example, Heuchenne and Van Keilegom (2007) or condition (2.2) in Brunel and Comte (2006), which can be interpreted as such a type of truncation. Until now, the truncation bounds which have been used are arbitrarily fixed and, usually, no method is proposed to discuss a method for choosing this truncation bound in practical situations. Therefore, in the new method we propose, we also provide a data-driven procedure to choose the truncation parameter. In our practical implementations, we use a criterion based on an asymptotic discussion which focuses on the mean square error associated with the estimation of the single-index parameter. We also suggest some possible adaptations to other types of criteria which are covered by our theoretical results.

In Section 2, we introduce our censored regression model and present our estimation procedure. It relies on the Kaplan–Meier estimator (1958) of the distribution function and on semi-parametric estimators of the conditional density. Following the procedure of Delecroix *et al.* (2003), we considered kernel-based estimators. Our theoretical results are presented in Section 3. In Section 4, we report simulation results and analysis on real data. Section 5 contains the detailed proof of our main lemma, which states the asymptotic equivalence of estimating the parameter in the semi-parametric and parametric models. All of the technicalities are postponed to the Appendix.

2. Censored regression model and estimation procedure

2.1. Notation and general setting

Let $Y_1, \ldots, Y_n$ be i.i.d. copies of a random response variable $Y \in \mathbb{R}$ and let $X_1, \ldots, X_n$ be i.i.d. copies of a random vector of covariates $X \in \mathcal{X}$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Introducing $C_1, \ldots, C_n$, i.i.d. replications of the censoring variable $C \in \mathbb{R}$, we consider the following censored regression framework, where the observations are

$$\begin{cases} Z_i = Y_i \wedge C_i, & 1 \le i \le n, \\ \delta_i = \mathbb{1}_{\{Y_i \le C_i\}}, & 1 \le i \le n, \\ X_i \in \mathcal{X} \subset \mathbb{R}^d, & 1 \le i \le n. \end{cases}$$

Let us introduce some notation for the distribution functions of the random variables appearing in this model, that is, $H(t) = \mathbb{P}(Z \le t)$, $F_X(t) = \mathbb{P}(X \le t)$, $F_Y(t) = \mathbb{P}(Y \le t)$, $F_{X,Y}(x, y) = \mathbb{P}(X \le x, Y \le y)$ and $G(t) = \mathbb{P}(C \le t)$. A major difficulty arising in censored regression models is the unavailability of the empirical distribution function to estimate functions F_Y, $F_{X,Y}$ and G, which must be replaced by Kaplan–Meier estimators.

We are interested in estimating $f(y|x)$, where $f(y|x)$ denotes the conditional density of Y given $X = x$ evaluated at point y. If one has no insight into the function f, it becomes necessary to perform nonparametric estimation of the conditional density. In the absence of censoring, a classical way to proceed is to use kernel smoothing; see, for example, Bashtannyk and Hyndman (2001). However, the so-called "curse of dimensionality" prevents this approach from being of practical interest when the number of covariates is important ($d > 3$ in practice). Therefore, it becomes relevant to consider semi-parametric models which appear to be a good compromise between the parametric (which relies on strong assumptions on the function f which may not hold in practice) and the nonparametric approach (which relies on fewer assumptions). In the following, we will consider the following semi-parametric single-index regression model,

$$\exists \theta_0 \in \Theta \subset \mathbb{R}^d \text{ such as} \qquad f(y|x) = f_{\theta_0}(y, x'\theta_0), \tag{1}$$

where $f_\theta(y, u)$ denotes the conditional density of Y given $X'\theta = u$ evaluated at y. For identifiability reasons, we will impose the condition that the first component of θ_0 is 1. In comparison to the Cox regression model for absolute continuous variables, our model (1) is more general since it only assumes that the law of Y given X depends on an unknown linear combination of the covariates, without imposing additional conditions on the conditional hazard rate.

Model (1) was considered by Delecroix *et al.* (2003) in the uncensored case. However, their procedure cannot be directly applied in the censored framework since the response variables are not directly observed. As a consequence, the empirical distribution function is unavailable and most of the tools used in this context are not at our disposal. A solution consists of using procedures relying on Kaplan–Meier estimators for the distribution function. An important difficulty arising in techniques of this type is the poor behavior of Kaplan–Meier estimators in the tail of the distribution. A practical way to avoid this kind of problem is to consider the truncated version of the variable Y. In the following, we will consider A_τ, a sequence of compacts included in the set $\{t : \tau_1 \le t \le \tau\}$ for $\tau \le \tau_0$, where $\tau_0 < \inf\{t : H(t) = 1\}$. Using only the observations

in A_τ allows us to avoid the bad behavior of the usual Kaplan–Meier estimators in the tail of the distribution. Moreover, this technique of truncation is particularly well adapted to our problem of estimating θ_0. In our framework, this truncation does not lead to any asymptotic bias since, denoting by $f^\tau(\cdot|x)$ the conditional density of Y given $X = x$ and $Y \in A_\tau$, for any $\tau < \infty$, we have, under (1),

$$f^\tau(y|x) = f^\tau_{\theta_0}(y, x'\theta_0), \tag{2}$$

where $f^\tau_\theta(y, u)$ denotes the conditional density of Y given $X'\theta = u$ and $Y \in A_\tau$ evaluated at y, and where the parameter is the same in (1) as in (2). In Section 2.6, we will discuss a new method allowing us to choose τ from the data in order to improve the performance in estimating θ_0.

2.2. Estimation procedure

We will extend the idea behind the procedure developed by Delecroix *et al.* (2003), adapting it to our censored framework. First, assume that we know the family of functions f^τ_θ. This approach is a modification of the maximum likelihood estimation procedure. Define, for any function $J \geq 0$,

$$L^\tau(\theta, J) = E[\log f^\tau_\theta(Y, \theta' X) J(X) \mathbb{1}_{Y \in A_\tau}] = \int \log f^\tau_\theta(y, \theta' x) J(x) \mathbb{1}_{y \in A_\tau} \, \mathrm{d}F_{X,Y}(x, y).$$

Here, J is a positive trimming function which will be defined later in order to avoid denominator problems in the nonparametric part of the model; see Section 2.4. From (2), θ_0 maximizes $L^\tau(\theta, J)$ for any $\tau < \infty$, this maximum being unique under some additional conditions on the regression model and J. Since, in our framework, $F_{X,Y}$ and f^τ_θ are unknown, it is natural to estimate them in order to produce an empirical version of $L^\tau(\theta, J)$.

2.2.1. *Estimation of $F_{X,Y}$*

In the case where there is no censoring (as in Delecroix *et al.* (2003)), $F_{X,Y}$ can be estimated by the empirical distribution function. In our censoring framework, the empirical distribution function of (X, Y) is unavailable since it relies on the true Y_i's, which are not observed. A convenient way to proceed involves replacing it by some Kaplan–Meier estimator, such as the one proposed by Stute (1993). Let us define the Kaplan–Meier estimator (Kaplan and Meier (1958)) of F_Y,

$$\hat{F}_Y(y) = 1 - \prod_{i: Z_i \leq t} \left(1 - \frac{1}{\sum_{j=1}^n \mathbb{1}_{Z_j \geq Z_i}}\right)^{\delta_i} = \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i \leq y},$$

where W_{in} denotes the "jump" of the Kaplan–Meier estimator at observation i (see Stute (1993)). To estimate $F_{X,Y}$, Stute proposes the use of

$$\hat{F}(x, y) = \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i \leq y, X_i \leq x}.$$

Let us also define the following (uncomputable) estimator of the distribution function:

$$\tilde{F}(x, y) = \sum_{i=1}^{n} \delta_i W_i^* \mathbb{1}_{Z_i \le y, X_i \le x},$$

where $W_i^* = n^{-1}[1 - G(Z_i-)]^{-1}$. The link between $\hat{F}$ and $\tilde{F}$ comes from the fact that, in the case where $\mathbb{P}(Y = C) = 0$,

$$W_{in} = n^{-1}[1 - \hat{G}(Z_i-)]^{-1}, \tag{3}$$

where $\hat{G}$ denotes the Kaplan–Meier estimator of G (see Satten and Datta (2001)). Asymptotic properties of $\hat{F}$ can be deduced by studying the difference with the simplest, but uncomputable, estimator $\tilde{F}$.

If we know the family of regression functions f_θ^τ, it is possible to compute the empirical version of $L^\tau(\theta, J)$ using $\hat{F}$, that is,

$$\begin{aligned} L_n^\tau(\theta, f^\tau, J) &= \int \log f_\theta^\tau(y, \theta' x) J(x) \mathbb{1}_{y \in A_\tau} \, \mathrm{d}\hat{F}(x, y) \\ &= \sum_{i=1}^{n} \delta_i W_{in} \log f_\theta^\tau(Z_i, \theta' X_i) J(X_i) \mathbb{1}_{Z_i \in A_\tau}. \end{aligned}$$

In the case $J \equiv 1$, the estimator of θ_0 obtained by maximizing L_n^τ would turn out to be an extension of the maximum likelihood estimator of θ_0, used in the presence of censoring.

2.3. Estimation of f_θ^τ

In our regression model (2), the family $\{f_\theta^\tau, \theta \in \Theta\}$ is actually unknown. As in Delecroix *et al.* (2003), we propose to use nonparametric kernel smoothing to estimate f_θ^τ. Introducing a kernel function K and a sequence of bandwidths h, define

$$\hat{f}_\theta^{h,\tau}(z, \theta' x) = \frac{\int K_h(\theta' x - \theta' u) K_h(z - y) \mathbb{1}_{y \in A_\tau} \, \mathrm{d}\hat{F}(u, y)}{\int K_h(\theta' x - \theta' u) \mathbb{1}_{y \in A_\tau} \, \mathrm{d}\hat{F}(u, y)}, \tag{4}$$

where $K_h(\cdot) = h^{-1} K(\cdot / h)$. Also, define $f^{*h,\tau}$, the kernel estimator based on the function $\tilde{F}$, that is,

$$f_\theta^{*h,\tau}(z, \theta' x) = \frac{\int K_h(\theta' x - \theta' u) K_h(z - y) \mathbb{1}_{y \in A_\tau} \, \mathrm{d}\tilde{F}(u, y)}{\int K_h(\theta' x - \theta' u) \mathbb{1}_{y \in A_\tau} \, \mathrm{d}\tilde{F}(u, y)}.$$

$f^{*h,\tau}$ will play an important role in studying the asymptotic behavior of $\hat{f}^{h,\tau}$. Indeed, $f^{*h,\tau}$ is theoretically easier to handle since it relies on sums of i.i.d. quantities, which is not the case for $\hat{F}$. Since $f^{*h,\tau}$ can be studied by standard kernel arguments, the most important difficulty will arise from studying the difference between $\hat{f}^{h,\tau}$ and $f^{*h,\tau}$.

In what follows, we will impose the following conditions on the kernel function:

Assumption 1. *Assume that*:

(A1) K *is a twice differentiable and fourth order kernel with derivatives of order* 0, 1 *and* 2 *of bounded variation, its support contained in* $[-1/2, 1/2]$ *and* $\int_{\mathbb{R}} K(s)\,\mathrm{d}s = 1$;
(A2) $\kappa := \|K\|_\infty := \sup_{x\in\mathbb{R}} |K(x)| < \infty$;
(A3) $\mathcal{K} := \{K((x-\cdot)/h) : h > 0, x \in \mathbb{R}^d\}$ *is a pointwise measurable class of functions*;
(A4) $h \in \mathcal{H}_n \subset [an^{-\alpha}, bn^{-\alpha}]$ *with* $a, b \in \mathbb{R}$, $1/8 < \alpha < 1/6$ *and where* $\mathcal{H}_n$ *is of cardinality* k_n *satisfying* $k_n n^{-4\alpha} \to 0$.

2.4. The trimming function J

The reason for introducing function J must be related to the need to avoid denominators close to zero in the definition (4). Ideally, we would need to use the trimming function

$$J_0(x, c) = \tilde{J}(f_{\theta_0' X}, \theta_0' x, c), \tag{5}$$

where c is a strictly positive constant, $f_{\theta_0' X}$ denotes the density of $\theta_0' X$ and $\tilde{J}(g, u, c) = \mathbb{1}_{g(u)>c}$. Unfortunately, this function relies on the knowledge of the parameter θ_0 and $f_{\theta_0' X}$. Therefore, we will have to proceed in two steps; that is, we first obtain a preliminary consistent estimator of θ_0 and then use it to estimate the trimming function J_0 which will be needed to achieve asymptotic normality of our estimators of θ_0.

We will assume that we know some set B on which $\inf\{f_{\theta' X}(\theta' x) : x \in B, \theta \in \Theta\} > c$, where c is a strictly positive constant. In a preliminary step, we can use this set B to compute the preliminary trimming $J_B(x) = \mathbb{1}_{x\in B}$. Using this trimming function and a deterministic sequence of bandwidth h_0 satisfying (A4) in Assumption 1, we define a preliminary estimator θ_n of θ_0,

$$\theta_n = \arg\max_{\theta\in\Theta} L_n^\tau(\theta, \hat{f}^{h_0,\tau}, J_B). \tag{6}$$

Let us stress the fact that B is assumed to be known. This is a classical assumption in single-index regression (see Delecroix *et al.* (2006)). However, in practice, the procedure does not seem very sensitive to the choice of B. The bandwidth h_0 we consider in the preliminary step can be any sequence decreasing to zero slower than $n^{-1/2}$. An adaptive choice of h_0 could be considered (using, for instance, the same choice as in the final estimation step; see below). However, since we will only need θ_n to be a preliminary consistent estimator and the final estimator will not be very sensitive to an adaptive choice of h_0 while computing θ_n, we do not consider this case in what follows.

With this preliminary estimator θ_n to hand, we can compute an estimated version of J_0 which will happen to be equivalent to J_0 (see Delecroix *et al.* (2006), page 738), that is,

$$\hat{J}_0(x, c) = \tilde{J}(\hat{f}^{h_0,\tau}_{\theta_n' X}, \theta_n' x, c). \tag{7}$$

For each sequence of bandwidths satisfying (A4) in Assumption 1 and for each truncation bound τ, we can define an estimator of θ_0,

$$\hat{\theta}^\tau(h) = \arg\max_{\theta\in\Theta_n} L_n^\tau(\theta, \hat{f}^{h,\tau}, \hat{J}_0), \tag{8}$$

where Θ_n is a shrinking sequence of neighborhoods accordingly to the preliminary estimation. However, as for any smoothing approach, the performance of this procedure strongly depends on the bandwidth sequence. Therefore, it becomes particularly relevant to provide an approach which automatically selects the most adapted bandwidth according to the data. The new question which then arises from the censored framework concerns the adaptive choice of the truncation parameter τ.

2.5. Adaptive choice of the bandwidth

Our procedure consists of choosing from the data, for each θ, a bandwidth which is adapted to the computation of $f_\theta^\tau(z,u)$. For this, we use an adaptation of the cross-validation technique of Fan and Yim (2004), that is,

$$\hat{h}^\tau(\theta) = \arg\min_{h\in\mathcal{H}_n} \sum_{i=1}^n W_{in} \mathbb{1}_{Z_i\in A_\tau} \left\{ \int_{A_\tau} \hat{f}_\theta^{h,\tau}(z,\theta' X_i)^2 \,\mathrm{d}z - 2\hat{f}_\theta^{h,\tau}(Z_i,\theta' X_i) \right\}.$$

This criterion is (up to a quantity which does not depend on h) an empirical version of the ISE criterion defined in equation (3.3) of Fan and Yim (2004) (in a censored framework), that is, $\int_{A_\tau}\int \{\hat{f}_\theta^{h,\tau}(z,\theta' x) - f_\theta^\tau(z,\theta' x)\}^2 f_{\theta' X}(\theta' x)\,\mathrm{d}x\,\mathrm{d}z$.

The estimator of θ_0 with an adaptive bandwidth is now defined as

$$\hat{\theta}^\tau = \arg\max_{\theta\in\Theta_n} L_n^\tau(\theta, \hat{f}^{\hat{h},\tau}, \hat{J}_0). \tag{9}$$

In the above notation, $\hat{h}$ depends on θ and τ, which was not emphasized in order to simplify the notation.

2.6. Adaptive choice of τ

As we have already mentioned, the Kaplan–Meier estimator does not behave well in the tail of the distribution. For example, if some moment conditions are not satisfied, it is not even $n^{1/2}$-consistent. Moreover, even in the case where an appropriate moment condition holds, it may happen (at least for a finite sample size) that the weights corresponding to the large observations are too important and exert considerable influence on the estimation procedure. For this reason, we introduced a truncation by a bound τ. However, a large number of existing procedures which also rely on this type of truncation do not consider the problem of choosing τ from the data. We propose that τ is selected from the data in the following way. Suppose that we have a consistent estimator of the asymptotic mean square error,

$$E^2(\tau) = \limsup_n E[\|\hat{\theta}^\tau(\hat{h}^\tau) - \theta_0\|^2],$$

say $\hat{E}^2(\tau)$, satisfying

$$\sup_{\tau_1\le\tau\le\tau_0} |\hat{E}^2(\tau) - E^2(\tau)| \to 0 \qquad \text{in probability.} \tag{10}$$

Such an estimator will be proposed in Section 4. Using this empirical estimator, we propose to choose τ in the following way, that is,

$$\hat{\tau} = \underset{\tau_1 \leq \tau \leq \tau_0}{\arg\min}\, \hat{E}^2(\tau).$$

Our final estimator of θ_0 is based on an adaptive bandwidth and an adaptive choice of truncation parameter τ, that is,

$$\hat{\theta} = \hat{\theta}^{\hat{\tau}}.$$

As we have already said, truncating the data does not introduce additional bias in the estimation of θ_0. On the other hand, removing too many data points could strongly increase the variance and removing some of the largest data points will decrease it. Our selection procedure $\hat{\tau}$ is then based on estimating the variance of $\hat{\theta}$ and consists of taking from the data the truncation parameter τ that seems to be the best compromise between these two aspects.

3. Asymptotic results

3.1. Consistency

The assumptions needed for consistency can basically be split into three categories, that is, identifiability assumptions, assumptions on the regression model (2) itself and, finally, assumptions on the censoring model.

Identifiability assumption and assumption on the regression model

Assumption 2. *Assume that for all $\tau_1 \leq \tau \leq \tau_0$ and all $\theta \in \Theta - \{\theta_0\}$,*

$$L_\tau(\theta_0, J_B) - L_\tau(\theta, J_B) > 0.$$

Assumption 3. *Assume that for $\theta_1, \theta_2 \in \Theta$, a bounded function $\Phi(X)$ and some $\gamma > 0$, we have*

$$\sup_\tau \| f^\tau_{\theta_1}(y, \theta_1' x) - f^\tau_{\theta_2}(y, \theta_2' x) \|_\infty \leq \|\theta_1 - \theta_2\|^\gamma \Phi(X).$$

Assumptions on the censoring model

Assumption 4. $\mathbb{P}(Y = C) = 0$.

This classical assumption in a censored framework avoids problems caused by the lack of symmetry between C and Y in the case where there are ties.

Assumption 5. *Identifiability assumption*: *we assume that*:

- *Y and C are independent*;
- $\mathbb{P}(Y \leq C | X, Y) = \mathbb{P}(Y \leq C | Y)$.

This last assumption was initially introduced by Stute (1993). An important particular case in which Assumption 5 holds is when C is independent from (X, Y). However, Assumption 5 is a more general and widely accepted assumption, which allows the censoring variables to depend on the covariates.

Theorem 1. *Under Assumptions 2–5,*

$$\sup_{\theta\in\Theta,\tau_1\le\tau\le\tau_0} |L_n^\tau(\theta, \hat{f}^{h_0,\tau}, J_B) - L^\tau(\theta, J_B)| = o_P(1) \tag{11}$$

and, consequently,

$$\theta_n \to_{\mathbb{P}} \theta_0.$$

Proof. To show (11), we will proceed in two steps. First, we consider $L_n^\tau(\theta, f^\tau, J_B) - L^\tau(\theta, J_B)$ (parametric problem) and then $L_n^\tau(\theta, \hat{f}^{h_0,\tau}, J_B) - L_n^\tau(\theta, f^\tau, J_B)$.

Step 1. From Assumption 3, the family $\{\log(f_\theta^\tau(\cdot, \theta'\cdot)), \theta \in \Theta, \tau_1 \le \tau \le \tau_0\}$ is seen to be P-Glivenko–Cantelli. Using a uniform version of Stute (1993) leads to $\sup_\theta |L_n^\tau(\theta, f^\tau, J_B) - L^\tau(\theta, J_B)| \to_{\mathbb{P}} 0$.

Step 2. We have, on the set $\Theta' B$,

$$|\log \hat{f}_\theta^{h_0,\tau}(y, u) - \log f_\theta^\tau(y, u)| \le c^{-1}[\hat{f}_\theta^{h_0,\tau}(y, u) - f_\theta^\tau(y, u)].$$

Hence,

$$\begin{aligned}
&\sup_{\theta,\tau} |L_n^\tau(\theta, \hat{f}^{h_0,\tau}, J_B) - L_n(\theta, f^\tau, J_B)| \\
&\quad\le c^{-1} \sup_{\theta,y,u,\tau} |\hat{f}_\theta^{h_0,\tau}(y, u) - f_\theta^\tau(y, u)| \mathbb{1}_{u\in\Theta' B, y\le\tau} \int \mathrm{d}\hat{F}(x, y) \\
&\quad\le c^{-1} \sup_{\theta,y,u,\tau} |\hat{f}_\theta^{h_0,\tau}(y, u) - f_\theta^\tau(y, u)| \mathbb{1}_{u\in\Theta' B, y\le\tau}.
\end{aligned}$$

Using the uniform convergence of $\hat{f}_\theta^{h,\tau}$ (see Proposition 2 and Lemma 2), we deduce that $\sup_{\theta,\tau} |L_n^\tau(\theta, \hat{f}^{h_0,\tau}, J_B) - L_n^\tau(\theta, f^\tau, J_B)| \to_{\mathbb{P}} 0$. □

3.2. Asymptotic normality

To obtain the asymptotic normality of our estimator, we need to introduce some regularity assumptions to the regression model.

Assumption 6. *Denote by $\nabla_\theta f_\theta^\tau(y, x)$ (resp., $\nabla_\theta^2 f_\theta^\tau(y, x)$) the vector of partial derivatives (resp., the matrix of second derivatives with respect to θ) of f_θ^τ with respect to θ and computed at point (θ, x, y). Assume that for $\theta_1, \theta_2 \in \Theta$, a bounded function $\Phi(X)$ and some $\gamma > 0$, we have*

$$\sup_\tau \|\nabla_\theta^2 f_{\theta_1}^\tau(y, x) - \nabla_\theta^2 f_{\theta_2}^\tau(y, x)\|_\infty \le \|\theta_1 - \theta_2\|^\gamma \Phi(X).$$

Assumption 7. *Using the notation of Van der Vaart and Wellner* (1996) *in Section* 2.7, *define*

$$\mathcal{H}_1 = \mathcal{C}^{1+\delta}(\theta_0'\mathcal{X} \times A_\tau, M),$$
$$\mathcal{H}_2 = x\mathcal{C}^{1+\delta}(\theta_0'\mathcal{X} \times A_\tau, M) + \mathcal{C}^{1+\delta}(\theta_0'\mathcal{X} \times A_\tau, M).$$

Assume that $f_{\theta_0}^\tau(\cdot,\cdot) \in \mathcal{H}_1$ (*as a function of* $\theta_0'x$ *and* y) *and* $\nabla_\theta f_{\theta_0}^\tau(\cdot,\cdot) \in \mathcal{H}_2$.

If the family of functions f^τ was known (parametric problem), the asymptotic normality of $\hat\theta$ could be deduced from elementary results on Kaplan–Meier integrals (see the Appendix for a brief review of these results), as in Stute (1999) or in Delecroix *et al.* (2008). Using results of this kind, we can derive the following lemma (see the Appendix for the proof), which is sufficient to obtain the asymptotic law of $\hat\theta$ in the parametric case, from Sherman (1994), Theorems 1 and 2.

Lemma 1. *Under Assumptions* 6 *and* 7, *we have the following representations*:

1. *on* $\mathrm{o}_P(1)$*-neighborhoods of* θ_0,

$$L_n^\tau(\theta, f^\tau, J_0) = L^\tau(\theta, J_0) + (\theta-\theta_0)'T_{1n}(\theta) + (\theta-\theta_0)'T_{2n}(\theta)(\theta-\theta_0) + T_{3n}(\theta) + T_{4n}(\theta_0),$$

with $\sup_{\theta,\tau}|T_{1n}| = \mathrm{O}_P(n^{-1/2})$, $\sup_{\theta,\tau}|T_{2n}| = \mathrm{o}_P(1)$, $\sup_{\theta,\tau}|T_{3n}| = \mathrm{O}_P(n^{-1})$ *and* $T_{4n}(\theta_0) = L_n^\tau(\theta_0, f^\tau, J_0)$;

2. *on* $\mathrm{O}_P(n^{-1/2})$*-neighborhoods of* θ_0,

$$L_n^\tau(\theta, f^\tau, J_0) = n^{-1/2}(\theta-\theta_0)'W_{n,\tau} - \tfrac{1}{2}(\theta-\theta_0)'V_\tau(\theta-\theta_0) + T_{4n}(\theta_0) + T_{5n}(\theta),$$

with $\sup_{\theta,\tau}|T_{5n}| = \mathrm{o}_P(n^{-1})$, *and defining* $f_1(x,y) = f_{\theta_0}^{\tau^{-1}}(y, \theta_0'x)J_0(x,c)\nabla_\theta f_{\theta_0}^\tau(y,x)$,

$$W_{n,\tau} = \frac{1}{n^{1/2}}\sum_{i=1}^n \psi(Z_i, \delta_i, X_i; f_1 \mathbb{1}_{A_\tau}),$$
$$V_\tau = E[f_{\theta_0}^{\tau^{-2}}(Y, \theta_0'X)J_0(X,c)\nabla_\theta f_{\theta_0}^\tau(Y,X)\nabla_\theta f_{\theta_0}^\tau(Y,X)'\mathbb{1}_{Y\in A_\tau}],$$

where ψ *is defined as in Theorem* 3.

In the following theorem, we show that the semi-parametric estimator proposed in Section 2 has the same asymptotic law as in the fully parametric case.

Theorem 2. *Define* $\tau^* = \arg\min_\tau E^2(\tau)$. *Under Assumptions* 1–7, *we have the following asymptotic i.i.d. representation*:

$$\hat\theta - \theta_0 = -\frac{1}{n^{1/2}}V_{\tau^*}^{-1}W_{n,\tau^*} + \mathrm{o}_P(n^{-1/2}), \tag{12}$$

where V_τ *and* $W_{n,\tau}$ *are defined in Lemma* 1. *As a consequence*,

$$n^{1/2}(\hat\theta - \theta_0) \Longrightarrow \mathcal{N}(0, \Sigma_{\tau^*}),$$

where $\Sigma_{\tau^*} = V_{\tau^*}^{-1}\Delta_{\tau^*}(f_1)V_{\tau^*}^{-1}$, $\Delta_{\tau^*}(f_1) = \mathrm{Var}(\psi(Z,\delta,X; f_1\mathbb{1}_{A_{\tau^*}}))$ *and* f_1 *is defined as in Lemma* 1.

This theorem is a consequence of the following lemma. This result shows that, asymptotically speaking, maximizing $L_n^\tau(\theta, \hat{f}^{h,\tau}, J)$ is equivalent to maximizing $L_n^\tau(\theta, f^\tau, J)$.

Main Lemma. *Under Assumptions* 1–7,

$$L_n^\tau(\theta, \hat{f}^{h,\tau}, \hat{J}_0) = L_n^\tau(\theta, f^\tau, J_0) + (\theta-\theta_0)' R_{1n}(\theta,h,\tau) + (\theta-\theta_0)' R_{2n}(\theta,h,\tau)(\theta-\theta_0) + \tilde{L}_n^\tau(\theta_0),$$

where

$$\sup_{\theta\in\Theta_n, h\in\mathcal{H}_n, \tau_1\le\tau\le\tau_0} R_{1n}(\theta,h,\tau) = \mathrm{o}_P(n^{-1/2}),$$
$$\sup_{\theta\in\Theta_n, h\in\mathcal{H}_n, \tau_1\le\tau\le\tau_0} R_{2n}(\theta,h,\tau) = \mathrm{o}_P(1)$$

and

$$\tilde{L}_n^\tau(\theta_0) = A_{1n}^\tau(\theta_0, \hat{f}^{h,\tau}) - B_{4n}^\tau(\theta_0, \hat{f}^{h,\tau}),$$

$A_{1n}^\tau(\theta_0, \hat{f}^{h,\tau})$ *and* $B_{4n}^\tau(\theta_0, \hat{f}^{h,\tau})$ *being defined in the proof of this lemma.*

In view of Sherman (1994), Theorems 1 and 2, this result will allow us to obtain the rate of convergence of our estimators and the asymptotic law is then the same as the asymptotic law in the parametric problem.

Proof of Theorem 2. Define

$$\Gamma_{0n}(\theta,\tau,h) = L_n^\tau(\theta, \hat{f}^{h,\tau}, \hat{J}_0),$$
$$\Gamma_{1n}(\theta,\tau) = L_n^\tau(\theta, \hat{f}^{\hat{h},\tau}, \hat{J}_0),$$
$$\Gamma_{2n}(\theta) = L_n^{\hat{\tau}}(\theta, \hat{f}^{\hat{h},\hat{\tau}}, \hat{J}_0).$$

We now apply Sherman (1994), Theorems 1 and 2, to Γ_{in}, for $i = 0, 1, 2$. From our Main Lemma and Lemma 1, we deduce that the representation (11) in Sherman (1994), Theorem 2, holds for $i = 0, 1, 2$ on $\mathrm{O}_P(n^{-1/2})$-neighborhoods of θ_0, with W_n and V defined as in Lemma 1. The asymptotic representation (12) is a by-product of the proof of Sherman (1994), Theorem 2, and of the i.i.d. representations of Kaplan–Meier integrals (see Theorem 3). □

4. Simulation study and real data analysis

4.1. Practical implementation of the adaptive choice of τ

From the proof of Theorem 2, we have the representation

$$\hat{\theta} - \theta_0 = -\frac{1}{n}\sum_{i=1}^{n} V_\tau^{-1}\psi(Z_i, \delta_i, X_i; f_1 \mathbb{1}_{A_\tau}) + o_P(n^{-1/2}).$$

As in Stute (1995), the function ψ of Theorem 3 can be estimated from the data by

$$\hat{\psi}(Z, \delta, X; \hat{f}_1 \mathbb{1}_{A_\tau}) = \frac{\delta \hat{f}_1(X, Z)}{1 - \hat{G}(Z-)} + \int \frac{\int_y^{\tau_0}\int_{\mathcal{X}} \hat{f}_1(x, t)\, \mathrm{d}\hat{F}(x, t)\, \mathrm{d}M^{\hat{G}}(y)}{1 - \hat{H}(y)},$$

where $\hat{f}_1$ is our kernel estimator of f_1 and $\hat{H}$ is the empirical estimator of H. To consistently estimate $\Delta(f_1)$, we use the general technique proposed by Stute (1996), that is,

$$\hat{\Delta}_\tau(f_1) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\psi}(Z_i, \delta_i, X_i; \hat{f}_1) - \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}(Z_i, \delta_i, X_i; \hat{f}_1)\right]^{\otimes 2}, \tag{13}$$

where $\otimes 2$ denotes the product of the matrix with its transpose. A consistent estimator of V_τ can then be computed as

$$\hat{V}_\tau = \int \hat{f}_{\hat{\theta}}^{h,\tau^{-2}}(y, \hat{\theta}' x)\hat{J}_0(x, c)\nabla_\theta \hat{f}_{\hat{\theta}}^{h,\tau}(y, x)\nabla_\theta \hat{f}_{\hat{\theta}}^{h,\tau}(y, x)' \mathbb{1}_{y \in A_\tau}\, \mathrm{d}\hat{F}(x, y).$$

To estimate the asymptotic mean square error, we use

$$\hat{E}_\tau^2 = \frac{1}{n}\hat{W}_{n,\tau}' \hat{V}_\tau^{-1}\hat{V}_\tau^{-1}\hat{W}_{n,\tau}.$$

4.2. Simulation study

In order to check the finite-sample behavior of our estimators of θ_0, we conducted some simulations using a model similar to the one in Delecroix *et al.* (2003). We considered the regression model

$$Y_i = \theta_0' X_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $Y_i \in \mathbb{R}$, $\theta_0 = (1, 0.5, 1.4, 0.2)'$ and $X_i \sim \otimes^4\{0.2\mathcal{N}(0, 1) + 0.8\mathcal{N}(0.25, 2)\}$. The errors are centered and normally distributed with conditional variance equal to $|\theta_0' X|$. We used the kernel

$$K(u) = 2k(u) - k * k(u),$$

Table 1. Biases, variances and mean square errors for 25% of censoring and sampling of size 100

$p = 25\%, n = 100$	Bias	Variance	MSE
$\hat{\theta}^{ADE}$	$\begin{pmatrix} -0.112 \\ -0.551 \\ -0.155 \end{pmatrix}$	$\begin{pmatrix} 0.14 & 0.005 & -0.022 \\ 0.005 & 0.075 & 0.016 \\ -0.022 & 0.016 & 0.116 \end{pmatrix}$	0.6714181
$\hat{\theta}^{\infty}$	$\begin{pmatrix} 0.057 \\ 0.215 \\ 0.048 \end{pmatrix}$	$\begin{pmatrix} 0.033 & 0.012 & 0.001 \\ 0.012 & 0.073 & -0.004 \\ 0.001 & -0.004 & 0.027 \end{pmatrix}$	0.1841227
$\hat{\theta}^{\hat{\tau}}$	$\begin{pmatrix} 0.07 \\ 0.221 \\ 0.028 \end{pmatrix}$	$\begin{pmatrix} 0.034 & 0.002 & 0.002 \\ 0.002 & 0.074 & 0 \\ 0.002 & 0 & 0.02 \end{pmatrix}$	0.1825980

where $*$ denotes the convolution product and

$$k(u) = \tfrac{3}{4}(1 - u^2)\mathbb{1}_{|u|\leq 1}$$

is the classical Epanechnikov kernel. The censoring distribution was selected to be exponential with parameter λ, which allows us to fix the proportion of censored responses ($p = 25\%$ and $p = 40\%$ in our simulations). $\hat{h}$ was chosen using a regular grid between 1 and 1.5.

Our estimator $\hat{\theta}^{\hat{\tau}}$ was compared with two other estimators, that is, $\hat{\theta}^{\infty}$, which does not rely on an adaptive choice of τ, and $\hat{\theta}^{ADE}$, which is obtained using the average derivative method of Lu and Burke (2005). In the tables below, we report our results over 100 simulations from samples of size 100 and 200 for two different rates of censoring. Recalling that the first component of θ_0 is fixed as 1, we need only to estimate the three other components. For each estimator, the mean square error $E(\|\hat{\theta} - \theta_0\|^2)$ is decomposed into bias and covariance.

To give a precise idea of the number of observations which are removed from the study by choosing τ adaptively, we introduce $N = \sharp\{1 \leq i \leq n, Z_i \leq \hat{\tau}\}$. In the following Table 5, we evaluate $E[N]$ in the different cases which we considered in the simulation study. We also include

Table 2. Biases, variances and mean square errors for 40% of censoring and sampling of size 100

$p = 40\%, n = 100$	Bias	Variance	MSE
$\hat{\theta}^{ADE}$	$\begin{pmatrix} -0.334 \\ -0.743 \\ -0.158 \end{pmatrix}$	$\begin{pmatrix} 0.159 & 0.009 & -0.014 \\ 0.009 & 0.268 & 0.048 \\ -0.014 & 0.048 & 0.165 \end{pmatrix}$	1.280163
$\hat{\theta}^{\infty}$	$\begin{pmatrix} 0.127 \\ 0.296 \\ 0.096 \end{pmatrix}$	$\begin{pmatrix} 0.11 & -0.034 & -0.01 \\ -0.034 & 0.101 & 0.021 \\ -0.01 & 0.021 & 0.059 \end{pmatrix}$	0.3829797
$\hat{\theta}^{\hat{\tau}}$	$\begin{pmatrix} 0.074 \\ 0.176 \\ 0.061 \end{pmatrix}$	$\begin{pmatrix} 0.064 & -0.005 & -0.004 \\ -0.005 & 0.051 & 0.014 \\ -0.004 & 0.014 & 0.069 \end{pmatrix}$	0.2239023

Table 3. Biases, variances and mean square errors for 25% of censoring and sampling of size 200

$p = 25\%, n = 200$	Bias	Variance	MSE
$\hat{\theta}^{ADE}$	$\begin{pmatrix} -0.189 \\ -0.578 \\ -0.133 \end{pmatrix}$	$\begin{pmatrix} 0.096 & 0.003 & 0.006 \\ 0.003 & 0.148 & -0.016 \\ 0.006 & -0.016 & 0.131 \end{pmatrix}$	0.7620268
$\hat{\theta}^{\infty}$	$\begin{pmatrix} 0.073 \\ 0.133 \\ 0.015 \end{pmatrix}$	$\begin{pmatrix} 0.033 & 0.004 & -0.004 \\ 0.004 & 0.023 & 0.002 \\ -0.004 & 0.002 & 0.012 \end{pmatrix}$	0.0910719
$\hat{\theta}^{\hat{\tau}}$	$\begin{pmatrix} 0.034 \\ 0.107 \\ 0.014 \end{pmatrix}$	$\begin{pmatrix} 0.007 & 0.001 & 0.004 \\ 0.001 & 0.011 & 0 \\ 0 & 0 & 0.006 \end{pmatrix}$	0.0364064

the average weight allocated to the largest (uncensored) data point, first in the case where we consider the whole data set (we denote it Weight$^{\infty}$), then in the truncated data set where we removed all data points with $Z_i \geq \hat{\tau}$ (we denote it Weight$^{\hat{\tau}}$).

Clearly, the MSE deteriorates when the percentage of censoring increases. According to the simulations, $\hat{\theta}^{\hat{\tau}}$ and $\hat{\theta}^{\infty}$ outperform $\hat{\theta}^{ADE}$, while, as expected, choosing τ adaptively improves the quality of the estimation. This is not obvious in the case where there is only 25% censoring. However, in the case where the level of censoring is high, estimation of the tail of the distribution by Kaplan–Meier estimators becomes more erratic and the importance of choosing a proper truncation appears in the significant difference between the MSEs of $\hat{\theta}^{\hat{\tau}}$ and $\hat{\theta}^{\infty}$. Moreover, the importance of truncation becomes obvious if we look at Table 5. In the case where there is 40% censoring, we see that if we do not use truncation, the weight allocated to the largest data point can be up to (approximately) ten times the weight allocated to the largest observation in the truncated data set. The ratio is less important in the case where there is 25% censoring, but still significant (in this case, the ratio is approximately 3). Therefore, it seems that, considering the whole data set, the weight allocated to the largest observation can have an overly strong influence

Table 4. Biases, variances and mean square errors for 40% of censoring and sampling of size 200

$p = 40\%, n = 200$	Bias	Variance	MSE
$\hat{\theta}^{ADE}$	$\begin{pmatrix} -0.109 \\ -0.763 \\ -0.053 \end{pmatrix}$	$\begin{pmatrix} 0.146 & -0.02 & 0.056 \\ -0.02 & 0.143 & -0.014 \\ 0.056 & -0.014 & 0.192 \end{pmatrix}$	1.078027
$\hat{\theta}^{\infty}$	$\begin{pmatrix} 0.104 \\ 0.151 \\ 0.077 \end{pmatrix}$	$\begin{pmatrix} 0.109 & 0.008 & 0.042 \\ 0.008 & 0.049 & 0.003 \\ 0.042 & 0.003 & 0.055 \end{pmatrix}$	0.2521227
$\hat{\theta}^{\hat{\tau}}$	$\begin{pmatrix} 0.043 \\ 0.14 \\ 0.021 \end{pmatrix}$	$\begin{pmatrix} 0.018 & -0.001 & 0.002 \\ -0.001 & 0.022 & 0.002 \\ 0.002 & 0.002 & 0.014 \end{pmatrix}$	0.07533921

Table 5. Last observed data in the truncating model and weight allocated to the largest observation in each model for different sample sizes and censoring rates

	$\hat{\mathbb{E}}(N)$	Weight^{∞}	$\text{Weight}^{\hat{\tau}}$
$n = 100$, $p = 25\%$	90	0.0667	0.0204
$n = 100$, $p = 40\%$	87	0.124	0.0236
$n = 200$, $p = 25\%$	185	0.0402	0.0119
$n = 200$, $p = 40\%$	172	0.0997	0.0122

on the estimation procedure, which explains the difference in performance of the estimators with or without truncation.

4.3. Example: Stanford Heart Transplant data

We now illustrate our method using data from the Stanford Heart Transplant program. This data set was initially studied by Miller and Halpern (1982). 184 of 249 patients in this program received a heart transplant between October 1967 and February 1980. From this data, we considered the survival time as the response variable Z, age as the first component of X and the square of age as the second component. Patients alive beyond February 1980 were considered censored. For easier comparison to previous work on this data set, we concentrated our analysis on the 157 patients out of 184 who had complete tissue typing. Among these 157 cases, 55 were censored.

Several methods of estimation have already been applied to this data set to estimate the following regression model:

$$Z = \alpha + \beta' X + \varepsilon(X), \tag{14}$$

where $\beta = (\beta_1, \beta_2)'$, $E[\varepsilon(X)|X] = 0$; see Miller and Halpern (1982), Wei *et al.* (1990) and Stute *et al.* (2000). Furthermore, nonparametric lack-of-fit tests have shown that the regression model (14) seems reasonable; see Stute *et al.* (2000) and Lopez and Patilea (2009). Therefore, it seems appropriate to experiment with our model on this data set. This strengthens the assumption on the residual, by assuming that $\varepsilon(X) = \varepsilon(\theta_0' X)$, where $\theta_0 = (1, \beta_2/\beta_1)'$, but allows more flexibility on the regression function.

In the following table, we present our estimators and recall the values of the estimators of β_2/β_1 for the linear regression model (14). We first computed $\hat{\theta}^{\infty}$, which is our estimator using the whole data set, that is, with $\tau = +\infty$, and compared it to the one obtained by choosing τ from the data, as in Section 4.1. In this last case, $\hat{\tau} = Z_{(90)}$, where $Z_{(i)}$ denotes the ith order statistic, which means that it required us to remove the 67 largest observations to estimate θ_0 (but not to estimate Kaplan–Meier weights, which were computed using the whole data set). We computed $\text{Weight}^{\infty} = 0.0397$ and $\text{Weight}^{\hat{\tau}} = 0.0076$ for the truncated data. The adaptive bandwidth was 1.7 for $\hat{\theta}^{\infty}$ and 1.3 for $\hat{\theta}^{\hat{\tau}}$. The estimated values of the mean square error were $E^2_{\infty} = 0.1089375$ and $E^2_{\hat{\tau}} = 0.01212701$ for $\hat{\theta}^{\infty}$ and $\hat{\theta}^{\hat{\tau}}$, respectively.

Table 6. Comparison of different estimators of $\theta_{0,2}$

	Estimator of $\theta_{0,2} = \beta_2/\beta_1$
Miller and Halpern	−0.01588785
Wei *et al.*	63.75
Stute *et al.*	−0.01367034
$\hat{\theta}^{\infty}$ (without adaptive choice of τ)	−0.07351351
$\hat{\theta}^{\hat{\tau}}$ (with adaptive choice of τ)	−0.0421508

Our estimators seem relatively close to the ones obtained by Miller and Halpern (1982) and Stute *et al.* (2000) using, respectively, the Buckley–James method and the Kaplan–Meier integrals method for the linear regression model.

5. Proof of Main Lemma

First, the same arguments as in Delecroix *et al.* (2006) apply to replace $\hat{J}_0$ by J_0. Define $J_\theta(x,c) = \mathbb{1}_{f_{\theta'X}(\theta'x)\geq c}$. From Assumption 3 on the density of $\theta'x$, we deduce that, on shrinking neighborhoods of θ_0, $J_0(x,c)$ can be replaced by $J_\theta(x,c/2)$. Using a Taylor expansion, we write

$$\begin{aligned}
&L_n^\tau(\theta, \hat{f}^{h,\tau}, J_0) - L_n^\tau(\theta, f^\tau, J_0)\\
&\quad= \sum_{i=1}^n \delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau} \log\left(\frac{\hat{f}_\theta^{h,\tau}(Z_i,\theta'X_i)}{f_\theta^\tau(Z_i,\theta'X_i)}\right) J_0(X_i,c)\\
&\quad= \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau}(\hat{f}_\theta^{h,\tau}(Z_i,\theta'X_i) - f_\theta^\tau(Z_i,\theta'X_i))J_0(X_i,c)}{f_\theta^\tau(Z_i,\theta'X_i)}\\
&\qquad- \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau}[\hat{f}_\theta^{h,\tau}(Z_i,\theta'X_i) - f_\theta^\tau(Z_i,\theta'X_i)]^2 J_0(X_i,c)}{\phi(f_\theta^\tau(Z_i,\theta'X_i), \hat{f}_\theta^{h,\tau}(Z_i,\theta'X_i))^2}\\
&\quad= A_{1n}^\tau(\theta, \hat{f}^{h,\tau}) - B_{1n}^\tau(\theta, \hat{f}^{h,\tau}),
\end{aligned}$$

where $\phi(f_\theta^\tau(Z_i,\theta'X_i), \hat{f}_\theta^{h,\tau}(Z_i,\theta'X_i))$ is between $\hat{f}_\theta^{h,\tau}(Z_i,\theta'X_i)$ and $f_\theta^\tau(Z_i,\theta'X_i)$.

Step 1. We first study A_{1n}. A Taylor expansion leads to the following decomposition,

$$\begin{aligned}
A_{1n}^\tau &= (\theta-\theta_0)' \sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau}(\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}(Z_i,X_i) - \nabla_\theta f_{\theta_0}^\tau(Z_i,X_i))J_\theta(X_i,c/2)}{f_\theta^\tau(Z_i,\theta'X_i)}\\
&\quad+ (\theta-\theta_0)' \left[\sum_{i=1}^n \frac{\delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau}(\nabla_\theta^2 \hat{f}_{\tilde{\theta}}^{h,\tau}(Z_i,X_i) - \nabla_\theta^2 f_{\tilde{\theta}}^\tau(Z_i,X_i))J_\theta(X_i,c/2)}{2f_\theta^\tau(Z_i,\theta'X_i)}\right](\theta-\theta_0)
\end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i W_{in}\mathbb{1}_{Z_i\in A_\tau}(\hat{f}^{h,\tau}_{\theta_0}(Z_i,\theta_0'X_i)-f^{\tau}_{\theta_0}(Z_i,\theta_0'X_i))}{f^{\tau}_{\theta}(Z_i,\theta'X_i)f^{\tau}_{\theta_0}(Z_i,\theta_0'X_i)}$$

$$\times\big(f^{\tau}_{\theta_0}(Z_i,\theta_0'X_i)-f^{\tau}_{\theta}(Z_i,\theta'X_i)\big)J_0(X_i,c)J_\theta(X_i,c/2)+A^{\tau}_{1n}(\theta_0,\hat{f}^{h,\tau})$$

$$=A^{\tau}_{1n}(\theta_0,\hat{f}^{h,\tau})+(\theta-\theta_0)'A^{\tau}_{2n}(\theta_0,\hat{f}^{h,\tau})+(\theta-\theta_0)'A^{\tau}_{3n}(\tilde{\theta},\hat{f}^{h,\tau})(\theta-\theta_0)+A^{\tau}_{4n}(\theta,\hat{f}^{h,\tau})$$

for some $\tilde{\theta}$ between θ and θ_0. Observe that, using the uniform consistency of $\nabla^2_\theta\hat{f}^{h,\tau}_\theta$ (deduced from Proposition 2 and Lemma 2), we obtain $\sup_{\tilde{\theta}\in\Theta_n,\tau\le\tau_0,h\in\mathcal{H}_n}A^{\tau}_{3n}(\tilde{\theta},\hat{f}^{h,\tau})=\mathrm{o}_P(1)$. We now study $A^{\tau}_{2n}(\theta_0,\hat{f}^{h,\tau})$. Using the expression (3) for the jumps of the Kaplan–Meier estimator, observe that

$$A^{\tau}_{2n}(\theta,\hat{f}^{h,\tau})=\sum_{i=1}^{n}\frac{W_i^*\mathbb{1}_{Z_i\in A_\tau}(\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}(Z_i,X_i)-\nabla_\theta f^{\tau}_{\theta_0}(Z_i,X_i))J_\theta(X_i,c/2)}{f^{\tau}_{\theta}(Z_i,\theta'X_i)}$$

$$+\frac{1}{n}\sum_{i=1}^{n}W_i^*Z_G(Z_i-)\frac{\delta_i\mathbb{1}_{Z_i\in A_\tau}(\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}(Z_i,X_i)-\nabla_\theta f^{\tau}_{\theta_0}(Z_i,X_i))J_\theta(X_i,c/2)}{f^{\tau}_{\theta}(Z_i,\theta'X_i)}$$

$$=A^{\tau}_{21n}(\theta,\hat{f}^{h,\tau})+A^{\tau}_{22n}(\theta,\hat{f}^{h,\tau}),$$

where

$$Z_G(t)=\frac{\hat{G}(t)-G(t)}{1-\hat{G}(t)}.$$

The term A^{τ}_{22n} can be bounded using (21), (22) and Lemma 2, by

$$\sup_{\tau\le\tau_0}|A_{22n}(\theta,\hat{f}^{h,\tau})|\le\mathrm{o}_P(n^{-1/2})\times n^{-1}\sum_{i=1}^{n}\delta_i[1-G(Z_i-)]^{-1}$$

and the last term is $\mathrm{O}_P(1)$ since it has finite expectation. Now, for A^{τ}_{21n}, first replace θ in the denominator by θ_0. We have

$$A^{\tau}_{21n}(\theta,\hat{f}^{h,\tau})=\sum_{i=1}^{n}\frac{W_i^*\mathbb{1}_{Z_i\in A_\tau}(\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}(Z_i,X_i)-\nabla_\theta f^{\tau}_{\theta_0}(Z_i,X_i))J_0(X_i,c/4)}{f^{\tau}_{\theta_0}(Z_i,\theta_0'X_i)}$$

$$+R^{\tau}_n(\theta,h)(\theta-\theta_0),$$

with $\sup_{\theta\in\Theta_n,\tau\le\tau_0,h\in\mathcal{H}_n}|R^{\tau}_n(\theta,h)|=\mathrm{o}_P(1)$ from Assumption 3 and the uniform consistency of $\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}$ deduced from Proposition 2 and Lemma 2. Now, use Assumption 7 and Proposition 3. Using the equicontinuity property of Donsker classes (see, for example, Van der Vaart and Wellner (1996) or Van der Vaart (1998)), we obtain that

$$A^{\tau}_{2n}(\theta,\hat{f}^{h,\tau})=\iint\frac{[\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}(y,x)-\nabla_\theta f^{\tau}_{\theta_0}(y,x)]\mathbb{1}_{y\in A_\tau}J_0(x,c/4)\,\mathrm{d}\mathbb{P}(x,y)}{f^{\tau}_{\theta_0}(y,u)}+\mathrm{o}_P(n^{-1/2}),$$

where the o_P-rate does not depend on θ, h or τ. From classical kernel arguments, $\sup_{y,x,\tau} |\int (\nabla_\theta f^{*h,\tau}_{\theta_0}(y,x) - \nabla_\theta f^{\tau}_{\theta_0}(y,x)) \mathbb{1}_{y\in A_\tau} J_0(x, c/4)\, d\mathbb{P}(x,y)| = O_{\mathbb{P}}(h^4) = o_{\mathbb{P}}(n^{-1/2})$ since $nh^8 \to 0$. Lemma 3 then completes the proof for $A^{\tau}_{2n}(\theta, \hat{f}^{h,\tau})$. $A^{\tau}_{4n}(\theta, \hat{f}^{h,\tau})$ can be handled similarly.

Step 2. B^{τ}_{1n} can be rewritten as

$$
\begin{aligned}
&B^{\tau}_{1n}(\theta, \hat{f}^{h,\tau}) \\
&\quad = \sum_{i=1}^{n} \delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau} J_\theta(X_i, c/2) \frac{\{(\theta-\theta_0)'[\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}(Z_i, X_i) - \nabla_\theta f^{\tau}_{\theta_0}(Z_i, X_i)]\}^2}{\phi(f^{\tau}_{\theta}(Z_i, \theta' X_i), \hat{f}^{h,\tau}_{\theta}(Z_i, \theta' X_i))^2} \\
&\qquad + 2\sum_{i=1}^{n} \delta_i W_{in} J_\theta(X_i, c/2) \mathbb{1}_{Z_i\in A_\tau} [\hat{f}^{h,\tau}_{\theta_0}(Z_i, \theta_0' X_i) - f^{\tau}_{\theta_0}(Z_i, \theta_0' X_i)](\theta-\theta_0)' \\
&\qquad\qquad \times [\nabla_\theta \hat{f}^{h,\tau}_{\tilde{\theta}}(Z_i, X_i) - \nabla_\theta f^{\tau}_{\tilde{\theta}}(Z_i, X_i)][\phi(f^{\tau}_{\theta}(Z_i, \theta' X_i), \hat{f}^{h,\tau}_{\theta}(Z_i, \theta' X_i))^2]^{-1} \\
&\qquad + B^{\tau}_{4n}(\theta_0, \hat{f}^{h,\tau}) + o_P(\|\theta-\theta_0\|^2) \\
&\quad = (\theta-\theta_0)' B^{\tau}_{2n}(\theta_0, \hat{f}^{h,\tau})(\theta-\theta_0) + (\theta-\theta_0)' B^{\tau}_{3n}(\theta, \hat{f}^{h,\tau}) + B^{\tau}_{4n}(\theta_0, \hat{f}^{h,\tau}) + o_P(\|\theta-\theta_0\|^2)
\end{aligned}
$$

for some $\tilde{\theta}$ between θ and θ_0. The third term does not depend on θ. For B^{τ}_{2n}, use the uniform consistency of $\nabla_{\theta_0} \hat{f}^{h,\tau}_{\theta_0}$ (Proposition 2 and Lemma 2) to obtain $\sup_{\tau\le\tau_0, h\in\mathcal{H}_n} |B^{\tau}_{2n}(\theta, \hat{f}^{h,\tau})| = o_P(n^{-1/2})$. Finally, for $B^{\tau}_{3n}(\theta, \hat{f}^{h,\tau})$, from a Taylor expansion,

$$
\begin{aligned}
B^{\tau}_{3n}(\theta, \hat{f}^{h,\tau}) = 2\sum_{i=1}^{n} &\frac{\delta_i W_{in} \mathbb{1}_{Z_i\in A_\tau} J_\theta(X_i, c/2)}{\phi(f^{\tau}_{\theta}(Z_i, \theta' X_i), \hat{f}^{h,\tau}_{\theta}(Z_i, \theta' X_i))^2} \\
&\times [\hat{f}^{h,\tau}_{\theta_0}(Z_i, \theta_0' X_i) - f^{\tau}_{\theta_0}(Z_i, \theta_0' X_i)][\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}(Z_i, X_i) - \nabla_\theta f^{\tau}_{\theta_0}(Z_i, X_i)] \\
&+ (\theta-\theta_0)' R^{\tau}_n(\theta, \hat{f}^{h,\tau}),
\end{aligned}
$$

with $\sup_{\theta\in\Theta_n, \tau\le\tau_0, h\in\mathcal{H}_n} R^{\tau}_n(\theta, \hat{f}^{h,\tau}) = o_P(1)$, from Proposition 2 and Lemma 2. For the main term, the product of the uniform convergence rates of $\hat{f}^{h,\tau}_{\theta_0}$ and $\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}$ obtained from Proposition 2 and Lemma 2 is $o_P(n^{-1/2})$ for $h \in \mathcal{H}_n$.

6. Conclusion

We proposed a new estimation procedure for a conditional density under a single-index assumption and random censoring. This procedure is an extension of the approach of Delecroix *et al.* (2003) in the case of a censored response. One of the advantages of this model is that it relies on fewer assumptions than a Cox regression model, in the case where the random variables of the model are absolutely continuous. By showing that estimating in this semi-parametric model is asymptotically equivalent to estimating in a parametric one (unknown in practice), we

obtain an $n^{-1/2}$-rate for the estimator of the index. This estimator can then be used to estimate the conditional density or the conditional distribution function by using a traditional nonparametric estimator under censoring. A new feature of our procedure is that it provides an adaptively driven choice of the bandwidth involved in the kernel estimators we use and that it also provides an adaptive choice of truncation parameter needed to avoid problems caused by the bad behavior of Kaplan–Meier estimators in the tail of the distribution. In this specific problem, this truncation does not introduce an additional bias into the procedure and seems, according to our simulations, to increase the quality of the estimator, especially in the case where the proportion of censored responses is important. Our way of choosing τ was motivated by minimizing the MSE in the estimation of $\hat{\theta}$. However, our method could easily be adapted to other kinds of criteria which, for example, focus more on the error in estimating one specific direction or on the error in the estimation of the conditional density itself.

Appendix

A.1. Kaplan–Meier integrals for the parametric case

We first recall a classical asymptotic representation of integrals with respect to $\hat{F}$; see Stute (1995), Stute (1996) and Sánchez Sellero *et al.* (2005).

Theorem 3. *Let $\mathcal{F}$ be a VC-class of functions with envelope Φ, such as*

$$\Phi(x, y) = 0 \qquad \text{for all } y \geq \tau_0, \tag{15}$$

where $\tau_0 \leq \tau_H$. We have the following asymptotic i.i.d. representation for all $\phi \in \mathcal{F}$:

$$\int \phi(x, y)\, \mathrm{d}\hat{F}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \psi(Z_i, \delta_i, X_i; \phi) + R(\phi),$$

where $\sup_{\phi \in \mathcal{F}} |R(\phi)| = \mathrm{O}_{a.s.}([\log n]^3 n^{-1})$, and

$$\psi(Z_i, \delta_i, X_i; \phi) = \frac{\delta \phi(X_i, Z_i)}{1 - G(Z_i-)} + \int \frac{\int_y^{\tau_0} \int_{\mathcal{X}} \phi(x, t)\, \mathrm{d}F(x, t)\, \mathrm{d}M_i^G(y)}{1 - H(y)},$$

where $M_i^G(y) = (1 - \delta_i)\mathbb{1}_{Z_i \leq y} - \int_{-\infty}^{y} \mathbb{1}_{Z_i \geq t}[1 - G(t-)]^{-1}\, \mathrm{d}G(t)$ is a martingale with respect to the filtration $\mathcal{G}_y = \{(Z_i, \delta_i, X_i)\mathbb{1}_{Z_i \leq y}\}$. Define $\Delta(\phi) = \mathrm{Var}(\psi(Z, \delta, X; \phi))$. It then follows that

$$\sqrt{n} \int \phi(x, y)\, \mathrm{d}[\hat{F} - F](x, y) \Longrightarrow \mathcal{N}(0, \Delta(\phi)).$$

Initially, the result of Stute was derived for a single function ϕ. Furthermore, Theorem 1.1 in Stute (1996) gives a convergence rate which is only $\mathrm{o}_P(n^{-1/2})$ for the remainder term; however, a higher convergence rate is obtained in his proof of Theorem 1.1 for functions satisfying (15),

which is the only case considered in our work. To obtain uniformity on a VC-class of functions, see Sánchez Sellero *et al.* (2005) who provided a more general representation that extends the one of Stute in the case where Y is right-censored and left-truncated. Their result is very useful since it provides, as a corollary, uniform law of large numbers results and a uniform central limit theorem. The representation we present in our Theorem 3 is a simple rewriting of Stute's representation. Theorem 3 is then a key ingredient to prove Lemma 1.

Proof of Lemma 1. We directly show the second part of the lemma since the first can be approached using similar techniques. By means of a Taylor expansion,

$$\begin{aligned} L_n^\tau(\theta, f^\tau, J_0) &= (\theta-\theta_0)' \sum_{i=1}^n \delta_i W_{in} J_0(X_i, c) \mathbb{1}_{Z_i \in A_\tau} \frac{\nabla_\theta f_{\theta_0}^\tau(Z_i, X_i)}{f_{\theta_0}^\tau(Z_i, \theta_0' X_i)} \\ &\quad + \frac{1}{2}(\theta-\theta_0)' \sum_{i=1}^n \delta_i W_{in} J_0(X_i, c) \mathbb{1}_{Z_i \in A_\tau} \nabla_\theta^2 [\log f_{\tilde\theta}^\tau](Z_i, X_i)(\theta-\theta_0) \qquad (16) \\ &\quad + T_{4n}(\theta_0) \end{aligned}$$

for some $\tilde\theta$ between θ_0 and θ. Theorem 3 provides an i.i.d. representation for the first term (which corresponds to $W_{n,\tau}$ in Lemma 1), while, from Assumption 6, the family of functions $\nabla_\theta^2[\log f_{\tilde\theta}^\tau](y,x)\mathbb{1}_{y\in A_\tau}$ is a VC-class of functions satisfying (15). Hence, the sum in the second term of (16) tends to V almost surely using a uniform law of large numbers property. □

A.2. The gradient of f

In the following, for any function φ, we will denote by $\varphi_h^{(n)}(\cdot)$ the expression $h^{-n}\varphi^{(n)}(\cdot/h)$, such as, for example, $K_h'(\cdot) = h^{-1}K'(\cdot/h)$.

Proposition 1. *Let*

$$f_\tau'(y,u) = \partial_u f_{\theta_0}^\tau(y,u).$$

We have

$$\nabla_\theta f_{\theta_0}^\tau(y', x) = x f_{1,\tau}(y, \theta_0' x) + f_{2,\tau}(y, \theta_0' x),$$

with

$$\begin{aligned} f_{1,\tau}(y, \theta_0' x) &= f_\tau'(y, \theta_0' x), \\ f_{2,\tau}(y, \theta_0' x) &= -f_\tau'(y, \theta_0' x) E[X | \theta_0' X]. \end{aligned}$$

In particular, $E[\nabla_\theta f_{\theta_0}^\tau(Y, X) | \theta_0' X] = 0$.

Proof. This follows by direct adaptation of Dominitz and Sherman (2005), Lemma 5A. □

A.3. Convergence properties of $f^{*h,\tau}$

We first recall some classical properties of kernel estimators. Consider the class of functions $\mathcal{K}$ introduced in Assumption 1. Let $N(\varepsilon, \mathcal{K}, d_Q)$ be the minimal number of balls $\{g : d_Q(g, g') < \varepsilon\}$ of d_Q-radius ε needed to cover $\mathcal{K}$. For $\varepsilon > 0$, let $N(\varepsilon, \mathcal{K}) = \sup_Q N(\kappa\varepsilon, \mathcal{K}, d_Q)$, where the supremum is taken over all probability measures Q on $(\mathbb{R}^d, \mathcal{B})$ and d_Q is the $L_2(Q)$-metric. From Nolan and Pollard (1987), it can easily be seen that, using a kernel K satisfying Assumption 1, for some $C > 0$ and $\nu > 0$, $N(\varepsilon, \mathcal{K}) \le C\varepsilon^{-\nu}$, $0 < \varepsilon < 1$.

Proposition 2. *Under Assumption 1, we have, for some $c > 0$,*

$$\sup_{x,y,h,\tau} |f_{\theta_0}^{*h,\tau}(y, \theta_0' x) - f_{\theta_0}^{\tau}(y, \theta_0' x)| \mathbb{1}_{y \in A_\tau} J_0(x, c) = \mathrm{O}_P(n^{-1/2} h^{-1} [\log n]^{1/2}), \tag{17}$$

$$\sup_{x,y,h,\tau} |\nabla_\theta f_{\theta_0}^{*h,\tau}(y, x) - \nabla_\theta f_{\theta_0}^{\tau}(y, x)| \mathbb{1}_{y \in A_\tau} J_0(x, c) = \mathrm{O}_P(n^{-1/2} h^{-2} [\log n]^{1/2}), \tag{18}$$

$$\sup_{x,y,h,\tau,\theta} |\nabla_\theta^2 f_{\theta}^{*h,\tau}(y, x) - \nabla_\theta^2 f_{\theta}^{\tau}(y, x)| \mathbb{1}_{y \in A_\tau} J_\theta(x, c) = \mathrm{O}_P(n^{-1/2} h^{-3} [\log n]^{1/2}). \tag{19}$$

Proof. (17) is a direct application of Einmahl and Mason (2005), Theorems 1 and 4. For (18), we only show convergence for the term

$$\hat{r}_{\theta_0}^{h,\tau}(x, y) := \frac{1}{h} \sum_{i=1}^{n} \delta_i W_i^* \mathbb{1}_{Z_i \in A_\tau} J_0(x, c)(X_i - x) K_h'(X_i' \theta_0 - x' \theta_0) K_h(Z_i - y).$$

Define

$$\bar{r}_{\theta_0}^{h,\tau}(x, y) = \frac{1}{h} \mathbb{E}[\mathbb{1}_{Y \in A_\tau} J_0(x, c)(X - x) K_h'(X' \theta_0 - x' \theta_0) K_h(Y - y)]$$

and

$$r_{\theta_0}^{\tau}(x, y) = \frac{\partial}{\partial u} \{\mathbb{E}[(X - x) | \theta_0' X = u, Y = y] \mathbb{1}_{y \in A_\tau} J_0(x, c) f_{\theta_0' X, Y}(u, y)\} \Big|_{u = \theta_0' x}.$$

Note that, from our assumptions, $r_{\theta_0}^{\tau}$ is a finite quantity. Next, Einmahl and Mason (2005), Theorem 4, yields

$$\sup_{x,y,h,\tau} |\hat{r}_{\theta_0}^{h,\tau}(x, y) - \bar{r}_{\theta_0}^{h,\tau}(x, y)| \mathbb{1}_{y \in A_\tau} = \mathrm{O}_P(n^{-1/2} h^{-2} [\log n]^{1/2}).$$

For the bias term, $\sup_{x,y,h,\tau} |\bar{r}_{\theta_0}^{h,\tau}(x, y) - r_{\theta_0}^{\tau}(x, y)| \mathbb{1}_{y \in A_\tau} = \mathrm{O}(h^4) = \mathrm{o}(n^{-1/2})$; see, for example, Bosq and Lecoutre (1997). As a consequence,

$$\sup_{x,y,h,\tau} |\hat{r}_{\theta_0}^{h,\tau}(x, y) - r_{\theta_0}^{\tau}(x, y)| \mathbb{1}_{y \in A_\tau} = \mathrm{O}_P(n^{-1/2} h^{-2} [\log n]^{1/2}).$$

For (19), we also need uniformity with respect to θ. The result can be deduced from the uniform convergence (with respect to θ, x, u) of quantities such as

$$S_n^{h,\tau}(\theta, x, y, \beta) = \frac{1}{h^2} \sum_{i=1}^{n} \delta_i W_i^* \phi(Z_i, X_i, \theta) \nabla_\theta^\beta K\left(\frac{\theta' X_i - \theta' x}{h}\right) K\left(\frac{Z_i - y}{h}\right), \tag{20}$$

where $\nabla_\theta^\beta K([\theta' X_i - \theta' x]h^{-1})$ for $\beta = 1$ (resp., for $\beta = 2$) denotes the gradient vector of function $K([\theta' X_i - \theta' x]/h)$ (resp., Hessian matrix) with respect to θ and evaluated at θ, and where ϕ is a bounded function with respect to θ and x. The function ϕ we consider is $\phi(Z, X, \theta) = f_{\theta' X}^{\tau}(\theta' X)^{-1} \mathbb{1}_{Z \in A_\tau} J_0(x, c)$, with the convention $0/0 = 0$ and where $f_{\theta' X}^{\tau}(\theta' X)$ is the conditional density of $\theta' X$ given $Y \in A_\tau$. (20) can be treated using the same method as Einmahl and Mason (2005). For this, observe that the family of functions $\{(X, Z) \to \nabla_\theta^\beta K([\theta' X - \theta' x]h^{-1}) K([Z - y]h^{-1}), \theta \in \Theta, x, y\}$ satisfies the assumptions of Einmahl and Mason (2005), Proposition 1 (see Nolan and Pollard (1987), Lemma 22(ii)). Hence, we apply Talagrand's inequality (Talagrand (1994); see also Einmahl and Mason (2005)) to obtain that

$$\sup_{\theta, x, y, h, \tau} |S_n^{h,\tau}(\theta, x, y, \alpha) - E[S_n^{h,\tau}(\theta, x, y, \alpha)]| \mathbb{1}_{y \in A_\tau} = \mathrm{O}_P(n^{-1/2} [\log n]^{1/2} h^{-1-\beta}).$$

Again, the bias term converges uniformly at rate $\mathrm{O}(h^4)$. □

A.4. The difference between f^* and $\hat{f}$

A.4.1. *Convergence rate of* $\hat{f}$

In this section, we show that replacing $f^{*h,\tau}$ by $\hat{f}^{h,\tau}$ (which is the estimator used in practice) does not modify the rate of convergence. For an intuitive understanding of these results, observe that $f^{*h,\tau}$ was obtained from $\hat{f}^{h,\tau}$ by replacing $\hat{G}$ by G. Let us recall some convergence properties of $\hat{G}$. We have

$$\sup_{t \le \tau_0} |\hat{G}(t) - G(t)| = \mathrm{O}_P(n^{-1/2}), \tag{21}$$

$$\sup_{t \le \tau_0} \frac{1 - G(t)}{1 - \hat{G}(t)} = \mathrm{O}_P(1); \tag{22}$$

see Gill (1983) for (21) and Zhou (1992) for (22). From (21), we see that the convergence rate of $\hat{G}$ is faster than the convergence rate of $f^{*h,\tau}$, which explains the asymptotic equivalence of $\hat{f}^{h,\tau}$ and $f^{*h,\tau}$. Lemma 2 makes things more precise and also provides a representation of the difference between $\nabla_\theta f_{\theta_0}^{*h,\tau}$ and $\nabla_\theta \hat{f}_{\theta_0}^{h,\tau}$ which is needed in the proof of the Main Lemma. Also required to prove our Main Lemma, Lemma 3 below supplies a technical result on the integral of this difference.

Lemma 2. *Under the assumption of Lemma 1, we have, for some* $c > 0$,

$$\sup_{x, y, h, \tau} |\hat{f}_{\theta_0}^{h,\tau}(y, \theta' x) - f_{\theta_0}^{*h,\tau}(y, \theta' x)| \mathbb{1}_{y \in A_\tau} J_0(x, c) = \mathrm{O}_P(n^{-1/2}), \tag{23}$$

$$\sup_{x,y,h,\tau} |\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}(y,x) - \nabla_\theta f^{*h,\tau}_{\theta_0}(y,x)| \mathbb{1}_{y\in A_\tau} J_0(x,c) = \mathrm{O}_P(n^{-1/2}h^{-1}), \tag{24}$$

$$\sup_{x,y,h,\tau,\theta} |\nabla^2_\theta \hat{f}^{h,\tau}_{\theta}(y,x) - \nabla^2_\theta f^{*h,\tau}_{\theta}(y,x)| \mathbb{1}_{y\in A_\tau} J_\theta(x,c) = \mathrm{O}_P(n^{-1/2}h^{-2}). \tag{25}$$

Furthermore, for x such as $J_0(x,c)\neq 0$,

$$\begin{aligned}\left(\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}(y,x) - \nabla_\theta f^{*h,\tau}_{\theta_0}(y,x)\right) = \int \frac{\int_{\mathcal{X}}\int_t^{\tau_0} g^{h,\tau}_{f,x,y}(x_2,y_2)\,\mathrm{d}\mathbb{P}(x_2,y_2)\,\mathrm{d}\bar{M}^G(t)}{1-H(t)} \\ + R_n(\tau,h,x,y),\end{aligned} \tag{26}$$

where $\bar{M}^G(y) = n^{-1}\sum_{i=1}^n M^G_i(y)$, M^G_i is defined in Theorem 3, $\sup_{x,y,\tau,h}|R_n(\tau,h,x,y)| = \mathrm{O}_P((\log n)^{1/2}n^{-1}h^{-3})$ and $g^h_{f,x,y}$ is defined by

$$\begin{aligned} g^{h,\tau}_{f,x_1,y_1}(x_2,y_2) = \frac{1}{h}\frac{(x_1-x_2)K_h'(\theta_0'x_1-\theta_0'x_2)K_h(y_1-y_2)}{f^\tau_{\theta_0'X}(\theta_0'x_1)} \\ - \frac{K_h(\theta_0'x_1-\theta_0'x_2)K_h(y_1-y_2)f'^\tau_{\theta_0'X}(\theta_0'x_1)}{f^\tau_{\theta_0'X}(\theta_0'x_1)^2},\end{aligned}$$

where $f'^\tau_{\theta_0'X}$ denotes the derivative of $u\to f^\tau_{\theta_0'X}(u)$, the conditional density of $\theta_0'X$ given $Y\in A_\tau$.

Lemma 3. *Under the assumptions of Lemma 1,*

$$\sup_{h,\tau}\int \frac{[\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}(y,x) - \nabla_\theta f^{*h,\tau}_{\theta_0}(y,x)]\mathbb{1}_{y\in A_\tau} J_0(x,c/4)\,\mathrm{d}\mathbb{P}(x,y)}{f^\tau_{\theta_0}(y,\theta_0'x)} = \mathrm{o}_P(n^{-1/2}).$$

Proof of Lemma 2. To prove (23)–(25), we just prove (25) since the others are similar. To prove (25), we consider only the terms in which the second derivative is involved, the others being treated analogously. Consider

$$\begin{aligned}&\frac{1}{h}\sum_{i=1}^n \delta_i W_{in}(X_i-x)K_h''(\theta'X_i-\theta'x)K_h(Z_i-y)(X_i-x)'\mathbb{1}_{Z_i\in A_\tau} f^\tau_{\theta'X}(\theta'x)^{-1}\\ &\quad= \frac{1}{h}\sum_{i=1}^n \delta_i W_i^* J_\theta(X_i,c)(X_i-x)K_h''(\theta'X_i-\theta'x)K_h(Z_i-y)(X_i-x)'\mathbb{1}_{Z_i\in A_\tau} f^\tau_{\theta'X}(\theta'x)^{-1}\\ &\qquad+ \frac{1}{h}\sum_{i=1}^n \delta_i W_i^* Z_G(Z_i-)(X_i-x)K_h''(\theta'X_i-\theta'x)K_h(Z_i-y)(X_i-x)'\mathbb{1}_{Z_i\in A_\tau} f^\tau_{\theta'X}(\theta'x)^{-1},\end{aligned}$$

where the first term is contained in $\nabla_\theta^2 f_\theta^{*h,\tau}$, while the second can be bounded by

$$O_P(n^{-1/2}h^{-2})\left[\frac{1}{nh^2}\sum_{i=1}^n \delta_i \mathbb{1}_{Z_i\le\tau_0}|K''|\left(\frac{\theta' X_i-\theta' x}{h}\right)|K|\left(\frac{Z_i-y}{h}\right)\right].$$

Using the results of Sherman (1994), the term inside the brackets is $O_P(1)$, uniformly in x, y, θ and h.

Now, for the representation (26), observe that

$$\begin{aligned}
&\nabla_\theta[\hat f_{\theta_0}^{h,\tau}-f_{\theta_0}^{*h,\tau}](y,x)\\
&\quad= h^{-1}\sum_{i=1}^n \delta_i W_i^* Z_G(Z_i-)(x-X_i)K_h'(\theta_0' x-\theta_0' X_i)K_h(y-Z_i)f_{\theta_0' X}^{\tau}(\theta_0' x)^{-1}\mathbb{1}_{Z_i\in A_\tau}\\
&\qquad-\sum_{i=1}^n \delta_i W_i^* Z_G(Z_i-)J_0(x,c)K(\theta_0' X_i-\theta_0' x)K_h(Z_i-y)f_{\theta_0' X}^{\prime\tau}(\theta_0' x)f_{\theta_0' X}^{\tau}(\theta_0' x)^{-2}\mathbb{1}_{Z_i\in A_\tau}\\
&\qquad+R_n'(\tau,h,x,y),
\end{aligned}\tag{27}$$

with $\sup_{x,y,h,\tau}|R_n'(\tau,h,x,y)| = O_P(n^{-1}h^{-3/2}[\log n]^{1/2})$, from the convergence rate of Z_G (see (21) and (22)) and the convergence rate of the denominator in (4) and its derivative, say $(\hat f_{\theta_0' X}^{\tau}-f_{\theta_0' X}^{\tau})$ and $(\hat f_{\theta_0' X}^{\prime\tau}-f_{\theta_0' X}^{\prime\tau})$ (which are of uniform rate $O_P(n^{-1/2}h^{-1/2}[\log n]^{1/2})$ and $O_P(n^{-1/2}h^{-3/2}[\log n]^{1/2}$, from arguments similar to those used in the proofs of (17)–(19) and (23)–(25)). An i.i.d. representation of the main term in (27) can be deduced from Theorem 3 since the class $\{h^3 g_{f,x,y}^{h,\tau}, x, y, h\}$ is a VC-class; see Nolan and Pollard (1987). □

Proof of Lemma 3. Observe that, from classical kernel arguments,

$$\begin{aligned}
&\sup_t\left|\iint_{x_2,t\le y_2\le\tau_0} g_{f,x,y}^{h,\tau}(x_2,y_2)J_0(x,c/4)\,\mathrm{d}\mathbb{P}(x_2,y_2)\,\mathrm{d}\mathbb{P}(x,y)-E[\nabla_\theta f_{\theta_0}^{\tau}(Y,X)J_0(X,c/4)]\right|\\
&\quad= O(h^4)
\end{aligned}$$

since K is of order 4. From the representation (26) in Lemma 2,

$$\begin{aligned}
&\int[\nabla_\theta \hat f_{\theta_0}^{h,\tau}(y,x)-\nabla_\theta f_{\theta_0}^{*h,\tau}(y,x)]J_0(x,c/4)\,\mathrm{d}\mathbb{P}(x,y)\\
&\quad=\int[1-H(t)]^{-1}E[\nabla_\theta f_{\theta_0}^{\tau}(Y,X)J_0(X,c/4)]\,\mathrm{d}\bar M^G(t)\\
&\qquad+\int[1-H(t)]^{-1}\left[\iint_{x_2,t\le y_2\le\tau_0} g_{f,x,y}^{h,\tau}(x_2,y_2)J_0(x,c/4)\,\mathrm{d}\mathbb{P}(x_2,y_2)\,\mathrm{d}\mathbb{P}(x,y)\right.\\
&\qquad\qquad\left.-E[\nabla_\theta f_{\theta_0}^{\tau}(Y,X)J_0(X,c/4)]\right]\mathrm{d}\bar M^G(t)+\int R_n(\tau,h,x,y)\,\mathrm{d}\mathbb{P}(x,y),
\end{aligned}\tag{28}$$

where the last term is $o_P(n^{-1/2})$, uniformly in τ and h. The first term is zero by Proposition 1 and because J_0 depends only on $\theta_0' X$.

For the second term, let $\phi_n(t,h,\tau) = [1 - H(t)]^{-1}\{\iint_{x_2,t\le y_2\le\tau_0} g^{h,\tau}_{f,x,y}(x_2,y_2)J_0(x, c/4)\,\mathrm{d}\mathbb{P}(x_2,y_2)\,\mathrm{d}\mathbb{P}(x,y) - E[\nabla_\theta f^\tau_{\theta_0}(Y,X)J_0(X,c/4)]\}$. Using the fact that $\mathcal{H}_n$ is of cardinality k_n, we have, for the second term in (28),

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}_n}\left|\int \phi_n(t,h,\tau)\,\mathrm{d}\bar{M}^G(t)\right|\ge\varepsilon\right)\le k_n\sup_{h\in\mathcal{H}_n}\mathbb{P}\left(\left|\int\phi_n(t,h,\tau)\,\mathrm{d}\bar{M}^G(t)\right|\ge\varepsilon\right).$$

Now, apply Lenglart's inequality; see Lenglart (1977) or Fleming and Harrington (1991), Theorem 3.4.1. This shows that, for all $\varepsilon>0$ and all $\eta>0$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\tau\le s\le\tau_0}\left\{\int_0^s\phi_n(t,h,\tau)\,\mathrm{d}\bar{M}^G(t)\right\}^2\ge\varepsilon^2\right)\\&\quad\le\frac{\eta}{\varepsilon^2}+\mathbb{P}\left(n^{-1}\int_0^{\tau_0}\phi_n^2(t,h,\tau)\frac{[1-\hat{H}(t-)]\,\mathrm{d}G(t)}{1-G(t-)}\ge\eta\right).\end{aligned}\tag{29}$$

As mentioned before, $\sup_t|\phi_n(t,h,\tau)| = \mathrm{O}(h^4)$. From (29) and the condition on k_n in Assumption 1, the lemma follows. □

A.4.2. Donsker classes

As stated in Assumption 7, to obtain an $n^{-1/2}$-convergence of $\hat{\theta}$, we need the regression function (and its gradient) to be sufficiently regular. In the lemma below, we first show that the classes of functions defined in Assumption 7 are Donsker and that $\hat{f}^{h,\tau}_{\theta_0}$ also belongs to the same regular class as $f^\tau_{\theta_0}$ with probability tending to one.

Proposition 3. *Consider the classes $\mathcal{H}_1$ and $\mathcal{H}_2$ defined in Assumption 7. $\mathcal{H}_1$ and $\mathcal{H}_2$ are Donsker classes. Furthermore, $\hat{f}^{h,\tau}_{\theta_0}$ and $\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}$ belong, respectively, to $\mathcal{H}_1$ and $\mathcal{H}_2$ with probability tending to one for some sufficiently large constant M.*

Proof. The class $\mathcal{H}_1$ is Donsker from Van der Vaart and Wellner (1996), Corollary 2.7.4. The class $\mathcal{H}_2$ is Donsker from a permanence property of Donsker classes; see Van der Vaart and Wellner (1996), Examples 2.10.10 and 2.10.7. We only show the proof for $\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}$ since the proof for $\hat{f}^{h,\tau}_{\theta_0}$ is similar. We write

$$\begin{aligned}&\nabla_\theta\hat{f}^{h,\tau}_{\theta_0}(z,x)\\&\quad=\frac{1}{nh}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{Z_i\in A_\tau}(X_i-x)K_h'(\theta_0'X_i-\theta_0'x)K_h(Z_i-z)}{[1-\hat{G}(Z_i-)]f^\tau_{\theta_0'X}(\theta_0'x)}J_0(X_i,c/2)\\&\qquad+\frac{1}{nh}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{Z_i\in A_\tau}(X_i-x)K_h'(\theta_0'X_i-\theta_0'x)K_h(Z_i-z)[\hat{f}^\tau_{\theta_0'X}(\theta_0'x)-f^\tau_{\theta_0'X}(\theta_0'x)]}{[1-\hat{G}(Z_i-)]\hat{f}^\tau_{\theta_0'X}(\theta_0'x)f^\tau_{\theta_0'X}(\theta_0'x)}\end{aligned}$$

$$
\begin{aligned}
&\times J_0(X_i, c/2) \\
&- \left[\frac{1}{nh} \sum_{i=1}^{n} \frac{(X_i - x) K_h'(\theta_0' X_i - \theta_0' x) J_0(X_i, c/2)}{(f^{\tau}_{\theta_0' X}(\theta_0' x))^2} \right] \\
&\quad \times \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i K_h(\theta_0' X_i - \theta_0' x) K_h(Z_i - z) \mathbb{1}_{Z_i \in A_\tau}}{[1 - \hat{G}(Z_i -)]} \right] \\
&+ \left[\frac{1}{nh} \sum_{i=1}^{n} \frac{(X_i - x) K_h'(\theta_0' X_i - \theta_0' x) [(\hat{f}^{\tau}_{\theta_0' X}(\theta_0' x))^2 - (f^{\tau}_{\theta_0' X}(\theta_0' x))^2] J_0(X_i, c/2)}{(\hat{f}^{\tau}_{\theta_0' X}(\theta_0' x) f^{\tau}_{\theta_0' X}(\theta_0' x))^2} \right] \\
&\quad \times \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{Z_i \in A_\tau} K_h(\theta_0' X_i - \theta_0' x) K_h(Z_i - z)}{[1 - G(Z_i -)]} \right].
\end{aligned}
$$

From this expression, we clearly see that $\nabla_\theta \hat{f}^{h,\tau}_{\theta_0}(y, x) = x\phi_1(x'\theta_0, y) + \phi_2(x'\theta_0, y)$. We must now check that ϕ_1 and ϕ_2 are in $\mathcal{H}_1$ with probability tending to one. Since the functions are twice continuously differentiable (from the assumptions on K), we only have to check their boundedness. By Lemma 2, this can be done by replacing $\hat{f}^{h,\tau}$ by $f^{*h,\tau}$ (that is, $\hat{G}$ by the true function G). Among the several terms in the decomposition of $\nabla_\theta f^{*h,\tau}$, we will only study

$$
\phi(u, y) = \frac{1}{nh} \sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{Z_i \in A_\tau} X_i K_h'(\theta_0' X_i - u) K_h(Z_i - z) J_0(X_i, c/2)}{[1 - G(Z_i -)] f^{\tau}_{\theta_0' X}(u)}
$$

since the others are similar. We will show that the derivatives of order 0, 1 and $1 + \delta$ of this function are uniformly bounded by some constant M with probability tending to one.

Now, a centered version of ϕ converges to zero at rate $\mathrm{O}_P([\log n]^{1/2} n^{-1/2} h^{-1})$ (see Einmahl and Mason (2005)), which tends to zero as long as $nh^2 \to \infty$. Furthermore, $E[\phi]$ is uniformly bounded from our Assumption 7 on the regression function. For the derivative,

$$
\begin{aligned}
\partial_u \phi(u, y) = &- \frac{1}{nh} \sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{Z_i \in A_\tau} X_i K_h''(\theta_0' X_i - u) K_h(Z_i - z) J_0(X_i, c/4)}{[1 - G(Z_i -)] f^{\tau}_{\theta_0' X}(u)} \\
&- \frac{1}{nh} \sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{Z_i \in A_\tau} X_i K_h'(\theta_0' X_i - u) K_h(Z_i - z) J_0(X_i, c/4) f'^{\tau}_{\theta_0' X}(u)}{[1 - G(Z_i -)](f^{\tau}_{\theta_0' X}(u))^2}.
\end{aligned}
$$

Again, $E[\partial_u \phi]$ is uniformly bounded from our Assumption 7. Now, using the results of Einmahl and Mason (2005), the centered version of $\partial_u \phi$ tends to zero provided that $nh^6 \to \infty$. The same arguments apply for $\partial_y \phi$. Hence, with $f_i(u, y) = E(\phi_i(u, y))$, we have proven that $\sup_{u,y} |\partial_u^j \partial_y^k \phi_i(u, y) - \partial_u^j \partial_y^k f_i(u, y)|$ tends to zero in probability for $i = 1, 2$, $k + j \leq 1$. We must now show that $\partial_u \phi_j$ and $\partial_y \phi_j$ are δ-Hölder for $j = 1, 2$ with a Hölderian constant bounded

by some M with probability tending to one. We only prove the result for $\partial_u \phi_1$. We have

$$\begin{aligned}
&\sup_{u',y',x,y} \frac{|\partial_u \phi_1(u,y) - \partial_u \phi_1(u',y')|}{\|(u,y)-(u',y')\|^{\delta}} \\
&\quad = \max\Bigg(\sup_{|u-u'|\geq n^{-1},y,y'} \frac{|\partial_u \phi_1(u,y) - \partial_u \phi_2(u',y')|}{\|(u,y)-(u',y')\|^{\delta}}, \\
&\qquad\qquad \sup_{|u-u'|\leq n^{-1},y,y'} \frac{|\partial_u \phi_1(u,y) - \partial_u \phi_1(u',y')|}{\|(u,y)-(u',y')\|^{\delta}} \Bigg) \\
&\quad = \max(S_1, S_2).
\end{aligned}$$

We have

$$S_1 \leq \sup_{u,y,u',y'} \frac{|\partial_u f_1(u',y') - \partial_u f_1(u,y)|}{\|(u',y')-(u,y)\|^{\delta}} + 2n^{\delta} \sup_{u,y,u',y'} |\partial_u \phi_1(u,y) - \partial_u f_1(u,y)|.$$

From our assumptions, the first supremum is bounded, while the last is $\mathrm{O}_P(n^{-1/2+\delta} \times [\log n]^{1/2} h^{-3})$ from the convergence rate of $\partial_u \phi_2$. It tends to zero provided that $nh^{6+\delta} \to \infty$. For S_2, since K is $\mathcal{C}^3$ with bounded derivatives, for some positive constant M,

$$\begin{aligned}
&\sup_{\|(u,y)-(u',y')\|\leq n^{-1},y,y'} \frac{|\partial_u \phi_1(u,y) - \partial_u \phi_1(u',y')|}{\|(u,y)-(u',y')\|^{\delta}} \\
&\quad \leq M \times \mathrm{O}_P(1) \left\| \sum_{i=1}^{3} \left| K^{(i)} \right| \right\|_{\infty} \sup_{\|(u,y)-(u',y')\|\leq n^{-1}} \|(u,y)-(u',y')\|^{1-\delta} \\
&\qquad \times h^{-1} \frac{1}{nh^4} \sum_{i=1}^{n} \frac{\delta_i}{1-G(Z_i-)}.
\end{aligned}$$

The last supremum is bounded by $\mathrm{O}_P(1) \times n^{-1+\delta} h^{-5}$ and it tends to zero when $nh^6 \to \infty$ (and the $\mathrm{O}_P(1)$ term does not depend on u, y). □

References

Bashtannyk, D.M. and Hyndman, R.J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.* **36** 279–298. MR1836204

Bosq, D. and Lecoutre, J.P. (1997). *Théorie de l'estimation fonctionnelle*. *Economie et statistiques avancées* **3**. Paris: Economica.

Brunel, E. and Comte, F. (2006). Adaptive nonparametric regression estimation in presence of right censoring. *Math. Methods Statist.* **15** 233–255. MR2278288

Cox, D.R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.* **86** 213–226. MR1997761

Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136** 730–769. MR2181975

Delecroix, M., Lopez, O. and Patilea, V. (2008). Nonlinear censored regression using synthetic data. *Scand. J. Statist.* **35** 248–265. MR2418739

Dominitz, J. and Sherman, R.P. (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory* **21** 838–863. MR2189497

Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380–1403. MR2195639

Fan, J. and Yim, T.H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91** 819–834. MR2126035

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley. MR1100924

Gannoun, A., Saracco, J., Yuan, A. and Bonney, G.E. (2005). Non-parametric quantile regression with censored data. *Scand. J. Statist.* **32** 527–550. MR2232341

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.* **11** 49–58. MR0684862

Heuchenne, C. and van Keilegom, I. (2007). Nonlinear regression with censored data. *Technometrics* **49** 34–44. MR2345450

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867

Koul, H., Susarla, V. and van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9** 1276–1288. MR0630110

Lenglart, E. (1977). Relation de domination entre deux processus. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **13** 171–179.

Lopez, O. (2009). Single-index regression models with right-censored responses. *J. Statist. Plann. Inference* **139** 1082–1097. MR2479851

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. MR2460488

Lu, X. and Burke, M.D. (2005). Censored multiple regression by the method of average derivatives. *J. Multivariate Anal.* **95** 182–205. MR2164128

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69** 521–531. MR0695199

Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Statist.* **15** 780–799. MR0888439

Sánchez Sellero, C., González Manteiga, W. and Van Keilegom, I. (2005). Uniform representation of product-limit integrals with applications. *Scand. J. Statist.* **32** 563–581. MR2232343

Satten, G.A. and Datta, S. (2001). The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *Amer. Statist.* **55** 207–210. MR1947266

Sherman, R.P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Ann. Statist.* **22** 439–459. MR1272092

Stute, W., González Manteiga, W. and Sánchez Sellero, C. (2000). Nonparametric model checks in censored regression. *Comm. Statist. Theory Methods* **29** 1611–1629. MR1793304

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45** 89–103. MR1222607

Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* **23** 422–439. MR1332574

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.* **23** 461–471. MR1439707

Stute, W. (1999). Nonlinear censored regression. *Statist. Sinica* **9** 1089–1102. MR1744826

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28–76. MR1258865

Van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer. MR1385671

Wei, L.J., Ying, Z. and Lin, D.Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77** 845–851. MR1086694

Zhou, M. (1992). Asymptotic normality of the "synthetic data" regression estimator for censored survival data. *Ann. Statist.* **20** 1002–1021. MR1165603

Received July 2008 and revised March 2009