Bernoulli **19**(3), 2013, 721–747
DOI: 10.3150/12-BEJ464

Single index regression models in the presence of censoring depending on the covariates

OLIVIER LOPEZ[1], VALENTIN PATILEA[2] and INGRID VAN KEILEGOM[3]

[1]*Université Pierre et Marie Curie, Laboratoire de Statistique Théorique et Appliquée, 4 place Jussieu Barre 15-25 Etage 2, 75005 Paris, France. E-mail: olivier.lopez0@upmc.fr*
[2]*CREST (Ensai) & IRMAR, Campus de Ker-Lann, Rue Blaise Pascal BP 37203, 35172 Bruz cedex, France. E-mail: patilea@ensai.fr*
[3]*Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium. E-mail: ingrid.vankeilegom@uclouvain.be*

Consider a random vector $(X', Y)'$, where X is d-dimensional and Y is one-dimensional. We assume that Y is subject to random right censoring. The aim of this paper is twofold. First, we propose a new estimator of the joint distribution of $(X', Y)'$. This estimator overcomes the common curse-of-dimensionality problem, by using a new dimension reduction technique. Second, we assume that the relation between X and Y is given by a mean regression single index model, and propose a new estimator of the parameters in this model. The asymptotic properties of all proposed estimators are obtained.

Keywords: curse-of-dimensionality; dimension reduction; multivariate distribution; right censoring; semiparametric regression; survival analysis

1. Introduction and model

Consider a random vector $(X', Y)'$, where $X = (X^{(1)}, \ldots, X^{(d)})'$ is d-dimensional and Y is one-dimensional. We assume that Y is subject to random right censoring, that is, instead of observing $(X', Y)'$, we observe the triplet $(X', T, \delta)'$, where $T = Y \wedge C$, $\delta = \mathbf{1}_{Y \leq C}$, and the random variable C is the censoring variable. Typically, Y is (a transformation of) the survival time (whose range can span the whole real line), and X is a vector of characteristics. The data consist of n i.i.d. replications $(X_i', T_i, \delta_i)'$ of $(X', T, \delta)'$.

Under this setting, the purpose of this paper is twofold. First, we propose a new estimator of the joint distribution $F(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ of X and Y (where $X \leq x$ means that $X^{(j)} \leq x^{(j)}$ for $j = 1, \ldots, d$). Second, we assume that the relation between X and Y is given by a single index mean regression model (as in, e.g., Härdle and Stoker [12], Powell, Stock and Stoker [25], Ichimura [16], Härdle, Hall and Ichimura [11], Klein and Spady [17], Horowitz and Härdle [14], Hristache, Juditsky and Spokoiny [15]), and we propose new estimators of the parameters under this model. These estimators will be constructed under the following fundamental model assumption on the relation between Y and C, which we impose throughout this paper:

1350-7265

(A0) There exists a function $g : \mathbb{R}^d \to \mathbb{R}$, such that:
 (i) Y and C are independent, conditionally on $g(X)$
 (ii) $\mathbb{P}(Y \leq C|X, Y) = \mathbb{P}(Y \leq C|g(X), Y)$.

Note that assumption (A0) holds in the particular case where $\mathcal{L}(C|X, Y) = \mathcal{L}(C|g(X))$. By assuming that the censoring variable depends on X only through a one-dimensional variable $g(X)$, we avoid the curse-of-dimensionality problems which strike regression approaches where X is multivariate and Y is independent of C conditionally on X, and at the same time the dependence of C on X is not too restrictive. A related dimension reduction model assumption for the censoring time has been considered in Section 4 of Li, Wang and Chen [18].

The function g will be unknown in general. When g is known, assumption (A0) has been proposed by Lopez [19]. The assumption is needed for identifying the model. In the literature on nonparametric censored regression, alternatives to assumption (A0) have been proposed. There are basically two alternatives, which can be regarded as limiting cases of assumption (A0), and in that sense our assumption is a trade-off between these two. The first alternative has been used by, for example, Akritas [1] and Van Keilegom and Akritas [33], among many others. They assume that Y is independent of C, conditionally on X, and propose kernel type estimators of the distribution $F(x, y)$ under this assumption. This assumption is a particular case of (A0) by taking $g(X) \equiv X$. Their estimators are however restricted to the case where $d = 1$. Although they could in principle be extended to higher dimensions, this is not recommended in practice, since they will suffer from the curse-of-dimensionality and higher order kernels will need to be used. The second alternative to assumption (A0) has been proposed by Stute [28,29]. He assumes that Y is independent of C, and that $\mathbb{P}(Y \leq C|X, Y) = \mathbb{P}(Y \leq C|Y)$. This is again a particular case of (A0), by taking $g(X) \equiv 1$. Although his estimator can be used for any $d \geq 1$, it has the drawback that it assumes that the censoring variable C depends on X in a very particular way. This type of dependence might hold true when the censoring is purely 'administrative' (censoring at the end of the study), but when the censoring can be caused by other factors (like death due to another disease, change of treatment, ...), then less restrictive assumptions on the censoring mechanism are required.

Our assumption (A0) balances somewhere in between these two extreme assumptions. By imposing assumption (A0), we propose a new dimension reduction technique, which overcomes the drawbacks of these two classical sets of assumptions, by allowing for $d \geq 1$ without assuming the complete independence between Y and C.

In some cases, the function g will be known exactly from some a priori information. For example, we might know that the censoring only depends on one component of X, for example, $g(X) = X^{(1)}$. Lopez [19] proposed an estimator of the joint distribution $F(x, y)$ when g is supposed to be known. However, in many other cases, g will be unknown and needs to be estimated. Throughout this paper, we will assume that

$$g \in \mathcal{G}, \qquad \text{where } \mathcal{G} = \{x \to \lambda(\theta, x) \colon \theta \in \Theta\}, \tag{1.1}$$

where λ is a known function, and Θ is a compact parameter set in $\mathbb{R}^k$. The true (but unknown) value of θ will be denoted by θ_0. This semiparametric assumption on the conditional distribution of C allows to avoid the curse of dimensionality that would have stroke our approach if no restriction on the censoring time would have been made.

Throughout the paper, we will assume that we know some root-n consistent estimator $\hat{\theta}$ of θ_0, that satisfies the following:

(C0) The estimator $\hat{\theta}$ satisfies:

$$\hat{\theta} - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} \mu(T_i, \delta_i, X_i) + o_P\bigl(n^{-1/2}\bigr),$$

with $E[\mu(T, \delta, X)] = 0$ and $E[\mu(T, \delta, X)^2] < \infty$.

Hence, the set Θ can from now on be an arbitrarily small environment of θ_0.

To illustrate the nature of assumptions (A0) and (C0), consider the function $g(x) = \theta_0' x$, and the case where C follows a Cox regression model given X, in the sense that the conditional hazard $h(\cdot|x, y)$ of C given $X = x$ and $Y = y$ satisfies

$$h(c|x, y) = h_0(c) \exp\bigl(\theta_0' x\bigr)$$

for some baseline function h_0 only depending on c. This model assumption on C seems realistic since often the censoring variable C represents itself a lifetime, like the time until a patient dies from a disease other than the disease under study. Under this model, we clearly have $\mathcal{L}(C|X, Y) = \mathcal{L}(C|\theta_0' X)$, and the estimator $\hat{\theta}$ proposed by Andersen and Gill [3] satisfies condition (C0), with

$$\mu(t, \delta, x) = \Sigma^{-1}\left((1-\delta)\phi(x, t) - \int \phi(x, u)\mathbf{1}_{t>u}\bigl[1 - G(u - |x)\bigr]^{-1} \mathrm{d}G(u|x)\right),$$

where the matrix Σ is defined by condition D in Andersen and Gill [3],

$$\phi(x, t) = x - \frac{E[X \mathrm{e}^{\theta_0' X}(1 - H(t|X))]}{E[\mathrm{e}^{\theta_0' X}(1 - H(t|X))]},$$

with $H(t|x) = \mathbb{P}(T \le t | X = x)$ and $G(c|x) = \mathbb{P}(C \le c | X = x)$. See also Gorgens and Horowitz [10] for regression models more general than Cox in which $\mathcal{L}(C|X, Y) = \mathcal{L}(C|\theta_0' X)$. Alternatively, one could also assume that $C = r(\theta_0' X) + U$, where $r(\cdot)$ is given, $E(U) = 0$, and U is independent of X and Y. For the estimation of θ_0 and the verification of condition (C0) under this model, see, for example, Akritas and Van Keilegom [2] and Heuchenne and Van Keilegom [13].

The purpose of this paper is twofold. The first contribution of this paper consists in proposing and studying a new nonparametric estimator of the joint distribution of X and Y under assumption (A0). Under different sets of assumptions on the relation between X, Y and C, this distribution has been the object of study of many papers in the past. See, for example, Akritas [1], Stute [28,29], Van Keilegom and Akritas [33], among others. As mentioned before, assumption (A0) allows to avoid the curse-of-dimensionality problem present in some of these contributions, and the heavy assumptions on the relation between C and X, which are present in many others.

The second contribution of this paper is the estimation of a semiparametric single index regression model for the censored response Y given X under assumption (A0). The proposed estimator is based on a two-step procedure, in which first a preliminary (consistent) estimator is obtained, which is then used to build a least squares criterion that defines our new semiparametric estimator in order to achieve $n^{1/2}$-consistency. Both steps heavily rely on the estimator of $F(x, y)$ studied before. Note that in this second contribution two dimension reduction techniques are used: the first one comes from assumption (A0), which is concerned with the relation between Y and C, and the second one comes from the single index model, which is making an hypothesis on the relation between Y and X.

Single index regression models are now a common semiparametric multivariate explanatory approach, see for instance Delecroix, Hristache and Patilea [5] for a review. However, the literature on single index models with a censored response variable is rather poor. To the best of our knowledge, the only contribution that allows for a general relationship between the censoring variable and the covariates is Li, Wang and Chen [18] and it is based on sliced inverse regression (SIR). However, it is well known that the SIR approach requires a linear conditional expectation condition among the covariates, which may be restrictive in applications, see equation (2.3) in Li, Wang and Chen [18].

Lopez [20] proposed a semiparametric least squares estimator for the single index regression in the particular case where $g(X) \equiv 1$ in assumption (A0). A similar procedure was introduced by Wang *et al.* [34] under the stronger assumption that C is independent of $(X', Y)'$. See also Lu and Cheng [23]. Lu and Burke [22] used the same more restrictive condition to define an average derivative estimator of the index. It is worthwhile to notice that these three contributions involve a Kaplan–Meier estimate of the censoring distribution, while in general assumption (A0) requires a nonparametric estimate of the *conditional* distribution of C given $g(X)$.

This paper is organized as follows. In the next section, the estimators of the joint distribution and of the parameters in the single index model are explained in detail. Section 3 is devoted to the presentation of the asymptotic results of the proposed estimators, while in Section 4 we compare our estimator with an existing estimator in the literature. Finally, Appendix A contains the assumptions under which the results of Section 3 are valid, while Appendix B contains some technical lemmas and the proofs of the main results.

2. The estimators

2.1. Estimation of the distribution $F(x, y)$

We first explain how to estimate the joint distribution $F(x, y)$ of X and Y. For an arbitrary value of θ, let

$$G_\theta(t|z) = \mathbb{P}\big(C \le t|\lambda(\theta, X) = z\big), \tag{2.1}$$

and define

$$\hat{G}_\theta(t|z) = 1 - \prod_{T_i \le t}\left(1 - \frac{w_{in}^\theta(z)}{\sum_{j=1}^n w_{jn}^\theta(z)\mathbf{1}_{T_j \ge T_i}}\right)^{1-\delta_i}, \tag{2.2}$$

where

$$w_{in}^{\theta}(z) = K\left(\frac{\lambda(\theta, X_i) - z}{a_n}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{\lambda(\theta, X_j) - z}{a_n}\right).$$

Here, a_n is a bandwidth sequence converging to zero as n tends to infinity, and K is a probability density function (kernel). Note that $\hat{G}_\theta(t|z)$ reduces to the estimator proposed by Beran [4] when $\lambda(\theta, X)$ is equal to X.

With at hand the estimator $\hat{\theta}$ introduced in condition (C0), and the corresponding estimator $\hat{g}(x) = \lambda(\hat{\theta}, x)$ of $g(x)$, we now define the following estimator of $F(x, y)$:

$$\hat{F}_{\hat{g}}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \mathbf{1}_{T_i \le y, X_i \le x}}{1 - \hat{G}_{\hat{\theta}}(T_i - |\hat{g}(X_i))}. \tag{2.3}$$

Note that this estimator is in the same spirit as the estimator proposed by Stute [28,29], but the denominators of the two estimators are different, because of the different sets of underlying assumptions. See also Fan and Gijbels [9] for a similar weighting scheme in a nonparametric regression framework. Also note that when g would be known, this estimator equals the estimator proposed and studied in Lopez [21].

In Section 3.1, we will study the asymptotic properties of the estimator $\hat{F}_{\hat{g}}(x, y)$.

2.2. Estimation of the single index model

We first need to introduce some notations. For $\theta \in \Theta$, let $Z_\theta = \lambda(\theta, X)$, and let $\mathcal{Z}_\theta \subset \mathbb{R}$ be the support of the variable Z_θ. We assume that $\mathcal{Z}_\theta$ is compact for all $\theta \in \Theta$. Also, define $H_\theta(t|z) = \mathbb{P}(T \le t | Z_\theta = z)$ and let $\tau_{H_\theta, z} = \inf\{t\colon\ H_\theta(t|z) = 1\}$.

We assume that the following single index mean regression model is valid: for some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^d$, with, say, first component $\beta_0^{(1)} = 1$,

$$E[Y|X, Y \le \tau] = E\left[Y|\beta_0' X, Y \le \tau\right] = m\left(\beta_0' X\right), \tag{2.4}$$

where m is an unknown function, and where τ is some fixed truncation point, satisfying

$$\tau < \inf_{\theta \in \Theta} \inf_{z \in \mathcal{Z}_\theta} \tau_{H_\theta, z}.$$

Let $f(t; \beta) = E[Y|\beta' X = t, Y \le \tau]$. Then, $f(\cdot; \beta_0) = m(\cdot)$. Also, let $\mathcal{B} = \{1\} \times \tilde{\mathcal{B}}$, where $\tilde{\mathcal{B}}$ is a compact subset of $\mathbb{R}^{d-1}$, and denote by $\mathcal{X}$ the support of the covariate vector X, which is a compact subset of $\mathbb{R}^d$.

The truncation at τ in model (2.4) is very common in the context of regression with right censored observations, and is caused by the lack of information in the right tail of the conditional distribution of Y given X. See, for example, Akritas [1] and Akritas and Van Keilegom [2] for similar truncation mechanisms. Note that when $\mathcal{L}(Y|X) = \mathcal{L}(Y|\beta_0' X)$, that is, when the whole distribution of Y given X only depends on X via $\beta_0' X$, then model (2.4) is satisfied for any value of τ.

The estimation of β_0 consists of several steps. We first explain these steps in an informal, intuitive way to outline the main ideas behind the proposed method, and we next work out each of these steps in a rigorous way.

1. Estimate $f(t;\beta)$ using some nonparametric estimator $\hat{f}(t;\beta)$.
2. Construct a preliminary consistent estimator β_n of β_0.
3. Use β_n to compute a trimming function that helps to avoid technical problems caused by denominators close to zero in the nonparametric estimation of $f(t;\beta)$.
4. Construct a second semi-parametric estimator $\hat{\beta}$ of β_0 by using the trimming function of the preceding step.

2.2.1. *Estimation of $f(t;\beta)$*

One possible estimator of $f(t;\beta)$ is

$$\hat{f}(t;\beta) = \int \tilde{K}\left(\frac{\beta' x - t}{h}\right) y \mathbf{1}_{y\le\tau}\, \mathrm{d}\hat{F}_{\hat{g}}(x,y) \Big/ \left(\int \tilde{K}\left(\frac{\beta' x - t}{h}\right) \mathbf{1}_{y\le\tau}\, \mathrm{d}\hat{F}_{\hat{g}}(x,y)\right), \tag{2.5}$$

where $h = h_n$ is a second bandwidth sequence, possibly different from the bandwidth a_n used to estimate the joint distribution $F(x,y)$, and where $\tilde{K}$ is a kernel function. However, other estimators may be used, for example, $[\hat{F}_\beta(\tau|t)]^{-1}\int y\mathbf{1}_{y\le\tau}\,\mathrm{d}\hat{F}_\beta(y|t)$, where $\hat{F}_\beta(y|t)$ denotes Beran's [4] estimator of $\mathbb{P}(Y\le y|\beta' X = t)$.

In what follows, we do not specify the choice of estimator of $f(t;\beta)$. Instead we will work with a generic estimator $\hat{f}(t;\beta)$ that satisfies certain conditions that need to be fulfilled in order to obtain the asymptotic normality of $\hat{\beta}$, and we will prove in Section 3.2 that the estimator in (2.5) satisfies these conditions.

2.2.2. *Preliminary estimation of β_0*

We assume that we know some set B such that

$$\inf_{\beta\in\mathcal{B}, x\in B} f_\beta^\tau(\beta' x) = c > 0,$$

where the function f_β^τ denotes the density of $\beta' X$, conditionally on $Y\le\tau$. Define the following preliminary trimming function:

$$\tilde{J}(x) = \mathbf{1}_{x\in B}. \tag{2.6}$$

Let $M(\beta, f, \tilde{J}) = E[(Y - f(\beta' X;\beta))^2\mathbf{1}_{Y\le\tau}\tilde{J}(X)]$, and note that this is minimized as a function of β when $\beta=\beta_0$. Motivated by this fact, we define the preliminary estimator β_n of β_0 by replacing all unknown quantities in $M(\beta, f, \tilde{J})$ by appropriate estimators, that is,

$$\begin{aligned}\beta_n &= \arg\min_{\beta\in\mathcal{B}} \int \big(y - \hat{f}(\beta' x;\beta)\big)^2 \mathbf{1}_{y\le\tau}\tilde{J}(x)\,\mathrm{d}\hat{F}_{\hat{g}}(x,y)\\ &= \arg\min_{\beta\in\mathcal{B}} M_n(\beta, \hat{f}, \tilde{J}).\end{aligned} \tag{2.7}$$

Note that other criterion functions can be used, based on M or L-estimating functions. We do not consider them here, since their analysis is very similar to the one for the least squares criterion function.

2.2.3. *New trimming function*

We will now refine the definition of the trimming function, by using the preliminary estimator β_n. Define

$$J(x) = \mathbf{1}_{f^{\tau}_{\beta_n}(\beta_n' x) > c}, \tag{2.8}$$

so instead of requiring that $f^{\tau}_{\beta}(\beta' x) > c$ for all β, we now only consider $\beta = \beta_n$, which will be satisfied for many more x-values, and hence this new function $J(x)$ is trimming much less than the preliminary naive trimming function $\tilde{J}(x)$.

To simplify our discussion, we will directly consider that the true function $f^{\tau}_{\beta_n}$ is used in the definition of J. In practice, the trimming function can be estimated by $\mathbf{1}_{\hat{f}^{\tau}_{\beta_n}(\beta_n' x) > c}$, where

$$\hat{f}^{\tau}_{\beta}(t) = \frac{1}{nb_n \mathbb{P}(Y \leq \tau)} \sum_{i=1}^{n} \frac{\delta_i \mathbf{1}_{T_i \leq \tau}}{1 - \hat{G}_{\hat{\theta}}(T_i - |\hat{g}(X_i))} K\left(\frac{\beta' X_i - t}{b_n}\right),$$

and where $b_n \to 0$ is a bandwidth parameter. In applications, $c_1 = c\mathbb{P}(Y \leq \tau)$ can be chosen arbitrarily small by the statistician. Considering $f^{\tau}_{\beta_n}$ or $\hat{f}^{\tau}_{\beta_n}$ does not change anything asymptotically speaking, see the arguments in Delecroix, Hristache and Patilea [5], see also Step 0 in the proof of Theorem 3.5 below. By similar arguments, the estimator of β_0 obtained with $\mathbf{1}_{\hat{f}^{\tau}_{\beta_n}(\beta_n' x) > c}$ is asymptotically equivalent to the 'ideal' estimator obtained with the trimming function

$$J_0(x) = \mathbf{1}_{f^{\tau}_{\beta_0}(\beta_0' x) > c}, \tag{2.9}$$

as long as β_n is a consistent estimator of β_0. Let us point out that J_0 only depends on $\beta_0' x$ and, in view of equation (A.14) in the proof of Theorem 3.5, this property will be essential for achieving $\sqrt{n}$-asymptotic normality of our estimator $\hat{\beta}$ defined below.

2.2.4. *Estimation of β_0*

With at hand this new trimming function, we can now define a new semi-parametric least squares estimator of β_0:

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta \in \mathcal{B}_n} \int \big(y - \hat{f}(\beta' x; \beta)\big)^2 \mathbf{1}_{y \leq \tau} J(x) \, \mathrm{d}\hat{F}_{\hat{g}}(x, y) \\ &= \arg\min_{\beta \in \mathcal{B}_n} M_n(\beta, \hat{f}, J), \end{aligned} \tag{2.10}$$

where $\mathcal{B}_n$ is a set shrinking to $\{\beta_0\}$, which is computed from the preliminary step. The proof of the asymptotic normality of $\hat{\beta}$ will be carried out in two steps. We will first show that minimizing $M_n(\beta, \hat{f}, J)$ is asymptotically equivalent to minimizing $M_n(\beta, f, J_0)$. This then brings back the minimization problem to a fully parametric one.

3. Asymptotic properties

3.1. Estimation of the distribution $F(x, y)$

Let us first introduce a few notations. Denote $H(t) = \mathbb{P}(T \le t)$, $H_\theta(t|z) = \mathbb{P}(T \le t|Z_\theta = z)$, $H_{\theta,0}(t|z) = \mathbb{P}(T \le t, \delta = 0|Z_\theta = z)$, and $H_{\theta,1}(t|z) = \mathbb{P}(T \le t, \delta = 1|Z_\theta = z)$. For any function $L(u)$, let $\nabla_u L(u)$ (resp., $\nabla^2_{u,u} L(u)$) denote the vector (resp., matrix) of partial derivatives of order 1 (resp., order 2) of $L(u)$ with respect to u. In particular, denote by $\nabla_\theta G_\theta(t|\lambda(\theta, x))$ the vector of partial derivatives of the function $G_\theta(t|\lambda(\theta, x))$ with respect to all occurrences of θ. Let us point out that, in general, the vector valued function $\nabla_\theta G_\theta(t|\lambda(\theta, x))$ depends on x, and not only on $\lambda(\theta, x)$. Finally, for any matrix A of dimensions $k \times \ell$ (where $k, \ell \ge 1$) we denote $|A| = [\mathrm{trace}(A'A)]^{1/2}$.

We further need to introduce two (intermediate) estimators of $F(x, y)$:

$$\tilde{F}_g(x, y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \mathbf{1}_{T_i \le y, X_i \le x}}{1 - G_{\theta_0}(T_i - |g(X_i))}, \tag{3.1}$$

$$\hat{F}_g(x, y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \mathbf{1}_{T_i \le y, X_i \le x}}{1 - \hat{G}_{\theta_0}(T_i - |g(X_i))}. \tag{3.2}$$

In the following result, we consider integrals of the form $\int \phi(x, y)\,\mathrm{d}\hat{F}_g(x, y)$ with ϕ belonging to some class of functions $\mathcal{F}$, and we state that this class of integrals is Glivenko–Cantelli and admits an i.i.d. representation uniformly over all $\phi \in \mathcal{F}$. The proof can be found in Lopez [21]. For a completely nonparametric estimator of $F(x, y)$ that is not based on model assumption (A0), Sánchez-Sellero, González-Manteiga and Van Keilegom [26] obtained a similar uniform consistency and convergence result. The assumptions mentioned below can be found in Appendix A.

Theorem 3.1. (i) *Under Assumptions 1 and 3, for $a_n \to 0$ and $na_n \to \infty$, and for a class $\mathcal{F}$ satisfying Condition 1, we have*

$$\sup_{\phi \in \mathcal{F}} \left| \int \phi(x, y)\,\mathrm{d}[\hat{F}_g - F](x, y) \right| \to_{a.s.} 0.$$

(ii) *For $Z_i = \lambda(\theta_0, X_i)$, define*

$$M_i(t) = (1 - \delta_i)\mathbf{1}_{T_i \le t} - \int_{-\infty}^{t} \frac{\mathbf{1}_{T_i \ge y}\,\mathrm{d}G_{\theta_0}(y|Z_i)}{1 - G_{\theta_0}(y - |Z_i)},$$

which is a continuous time martingale with respect to the natural filtration $\sigma(\{Z_i \mathbf{1}_{T_i \le t}, T_i \mathbf{1}_{T_i \le t}, \delta_i \mathbf{1}_{T_i \le t}, i = 1, \ldots, n\})$. Under Assumptions 1–4 and for a class $\mathcal{F}$ satisfying Conditions 2 and 3,

$$\int \phi(x, y)\,\mathrm{d}[\hat{F}_g - \tilde{F}_g](x, y) = \frac{1}{n}\sum_{i=1}^{n} \int \frac{\bar{\phi}(Z_i, s)\,\mathrm{d}M_i(s)}{[1 - F(s - |Z_i)][1 - G_{\theta_0}(s|Z_i)]} + R_n(\phi),$$

where $\sup_{\phi\in\mathcal{F}}|R_n(\phi)| = o_P(n^{-1/2})$, $\bar\phi$ *is defined above Condition* 3, *and* $F(s|z) = \mathbb{P}(Y \le s|Z_{\theta_0} = z)$.

The following theorem provides the behavior of the difference between integrals with respect to $\hat F_{\hat g}$ and integrals with respect to $\hat F_g$.

Theorem 3.2. (i) *Under Assumptions* 1, 3 *and* 5, *for* $a_n \to 0$ *and* $na_n \to \infty$, *and for a class* $\mathcal{F}$ *satisfying Condition* 1, *we have*

$$\sup_{\phi\in\mathcal{F}}\left|\int \phi(x,y)\,\mathrm{d}[\hat F_{\hat g} - \hat F_g](x,y)\right| = o_P(1).$$

(ii) *Under Assumptions* 1–3 *and* 5, *for* $a_n \to 0$ *and* $na_n^3(\log n)^{-1} \to \infty$, *and for a class* $\mathcal{F}$ *whose envelope is as in Condition* 1,

$$\begin{aligned}&\int \phi(x,y)\,\mathrm{d}[\hat F_{\hat g} - \hat F_g](x,y)\\ &\quad = -E\left(\frac{\phi(X,Y)\{\nabla_\theta G_{\theta_0}(Y-|\lambda(\theta_0,X))\}'}{1-G_{\theta_0}(Y-|g(X))}\right)\frac{1}{n}\sum_{i=1}^n \mu(T_i,\delta_i,X_i) + \tilde R_n(\phi),\end{aligned}$$

where the function μ *is defined in* (C0), *and where* $\sup_{\phi\in\mathcal{F}}|\tilde R_n(\phi)| = o_P(n^{-1/2})$.

3.2. Estimation of the single index model

We now return to the single index model (2.4) and to the estimators β_n and $\hat\beta$ defined in (2.7) and (2.10). We start with stating the asymptotic consistency of the estimator β_n. Note that the estimator $\hat\beta$ is by construction consistent, since it is defined on a shrinking neighborhood of β_0.

Theorem 3.3. *Let* $\tilde J$ *be defined as in* (2.6). *Under Assumptions* 1, 3, 5, 7, *and* 9 – (A.1), *and for* $a_n \to 0$ *and* $na_n \to \infty$, *we have*

$$\sup_{\beta\in\mathcal{B}}\left|M_n(\beta,\hat f,\tilde J) - M(\beta,f,\tilde J)\right| \to 0,$$

in probability. Consequently, $\beta_n \to \beta_0$ *in probability.*

The next lemma is an important property in the literature on single index models. In the classical uncensored single index regression model, the property $E[\nabla_\beta f(\beta_0'X;\beta_0)|\beta_0'X] = 0$ plays a major role in proving the asymptotic normality of M-estimators. See Delecroix, Hristache and Patilea [5]. The next lemma shows that in our context, where we have to truncate at τ because of censoring in the data, the analogous truncated version of this property holds true without any further model conditions.

Lemma 3.4. *Assume that the derivative $\nabla_\beta f(\beta_0'\cdot;\beta_0)$ exists and is bounded. Then, for any β_0 satisfying condition* (2.4),

$$E\big[\nabla_\beta f\big(\beta_0'X;\beta_0\big)\mathbf{1}_{Y\le\tau}|\beta_0'X\big]=0.$$

This lemma is crucial for obtaining our i.i.d. representation and the asymptotic normality of $\hat{\beta}$, which we state in the next theorem. We denote by $\nabla_{\tilde{\beta}} f(\beta_0'\cdot;\beta_0)$ the vector of partial derivatives with respect to the last $d-1$ components of β.

Theorem 3.5. *Let $\phi(x,y)=(y-f(\beta_0'x;\beta_0))\nabla_{\tilde{\beta}} f(\beta_0'x;\beta_0)\mathbf{1}_{y\le\tau}J_0(x)$. Under Assumptions* 1–11, *we have*

$$\begin{aligned}\hat{\tilde{\beta}}-\tilde{\beta}_0=\Omega^{-1}\Bigg[&\int\phi(x,y)\,\mathrm{d}\big(\tilde{F}_g(x,y)-F(x,y)\big)\\&+\frac{1}{n}\sum_{i=1}^n\int\frac{\bar{\phi}(g(X_i),s)\,\mathrm{d}M_i(s)}{[1-F(s-|g(X_i))][1-G_{\theta_0}(s|g(X_i))]}\\&-E\left(\frac{\phi(X,Y)\{\nabla_\theta G_{\theta_0}(Y-|\lambda(\theta_0,X))\}'}{1-G_{\theta_0}(Y-|g(X))}\right)\frac{1}{n}\sum_{i=1}^n\mu(T_i,\delta_i,X_i)\Bigg]\qquad(3.3)\\&+\mathrm{o}_P\big(n^{-1/2}\big)\\=\Omega^{-1}&\left[\frac{1}{n}\sum_{i=1}^n\eta(T_i,\delta_i,X_i)\right]+\mathrm{o}_P\big(n^{-1/2}\big),\end{aligned}$$

where the function μ is defined in (C0), *and where*

$$\Omega=E\big[\mathbf{1}_{Y\le\tau}J_0(X)\nabla_{\tilde{\beta}} f\big(\beta_0'X;\beta_0\big)\nabla_{\tilde{\beta}} f\big(\beta_0'X;\beta_0\big)'\big].$$

Hence,

$$n^{1/2}(\hat{\tilde{\beta}}-\tilde{\beta}_0)\xrightarrow{d}N\big(0,\Omega^{-1}E\big[\eta(T,\delta,X)\eta(T,\delta,X)'\big]\Omega^{-1}\big).$$

If we wish to estimate the asymptotic variance in Theorem 3.5, we see that we need to estimate the variance of $\Omega^{-1}\eta$. However, one can consistently estimate Ω by

$$\hat{\Omega}=\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{Y_i\le\tau}J(X_i)\nabla_{\tilde{\beta}}\hat{f}\big(\hat{\beta}'X_i;\hat{\beta}\big)\nabla_{\tilde{\beta}}\hat{f}\big(\hat{\beta}'X_i;\hat{\beta}\big)'.$$

Similarly, when it comes to estimate the covariance matrix of η, one can proceed by taking the empirical variance of a random vector $(\hat{\eta}(T_i,\delta_i,X_i))_{1\le i\le n}$, where $\hat{\eta}$ denotes an estimated version of η in which we replaced each unknown quantity by its empirical counterpart (f replaced by $\hat{f}$, β_0 by $\hat{\beta}$, F by $\hat{F},\ldots$).

We end this section with the verification of Assumptions 9–11 for the estimator $\hat{f}(t;\beta)$ defined in (2.5). Define the (uncomputable) kernel estimator based on $\tilde{F}_g$,

$$f^*(t;\beta) = \int \tilde{K}\left(\frac{\beta' x - t}{h}\right) y \mathbf{1}_{y\le\tau}\, \mathrm{d}\tilde{F}_g(x,y) \bigg/ \left(\int \tilde{K}\left(\frac{\beta' x - t}{h}\right) \mathbf{1}_{y\le\tau}\, \mathrm{d}\tilde{F}_g(x,y)\right). \tag{3.4}$$

The advantage of $\tilde{F}_g$, and hence of f^*, is that it is composed of sums of i.i.d. terms. Classical arguments show that f^* satisfies Assumptions 9 to 11. This is shown in Proposition 3.6 below. On the other hand, Proposition 3.7 shows that the difference between $\hat{f}$ and f^* is sufficiently small so that $\hat{f}$ also satisfies these assumptions.

Proposition 3.6. *Assume that*

(i) *$\tilde{K}$ is a symmetric density function with compact support, and with two continuous derivatives of bounded variation*;
(ii) *$f(\cdot;\beta_0) \in \mathcal{H}_1^0$ and $\nabla_\beta f(\beta_0'\cdot;\beta_0) \in \mathcal{H}_2^0$, with $\mathcal{H}_1^0$ and $\mathcal{H}_2^0$ defined in* (A.4) *and* (A.5);
(iii) *$nh^5(\log n)^{-1/2} \to \infty$, and $nh^8 \to 0$.*

Then, f^* satisfies Assumptions 9–11.

Proposition 3.7. *Under the assumptions of Theorem* 3.2, *we have*

$$\sup_{\beta\in\mathcal{B},x\in\mathcal{X}} \left| f^*(\beta' x;\beta) - \hat{f}(\beta' x;\beta)\right| = \mathrm{O}_P\big((\log n)^{1/2} n^{-1/2} a_n^{-1/2}\big),$$

$$\sup_{\beta\in\mathcal{B},x\in\mathcal{X}} \left| \nabla_\beta f^*(\beta' x;\beta) - \nabla_\beta \hat{f}(\beta' x;\beta)\right| = \mathrm{O}_P\big((\log n)^{1/2} h^{-1} n^{-1/2} a_n^{-1/2}\big),$$

$$\sup_{\beta\in\mathcal{B},x\in\mathcal{X}} \left| \nabla^2_{\beta,\beta} f^*(\beta' x;\beta) - \nabla^2_{\beta,\beta} \hat{f}(\beta' x;\beta)\right| = \mathrm{O}_P\big((\log n)^{1/2} h^{-2} n^{-1/2} a_n^{-1/2}\big),$$

where $\hat{f}$ is the estimator defined in (2.5). *Moreover, $\nabla_\beta \hat{f}(\beta_0' x;\beta_0) = x\hat{m}_1(\beta_0' x) + \hat{m}_2(\beta_0' x)$, with, for $j = 1, 2$,*

$$\sup_{x\in\mathcal{X}} \left|\hat{m}_j(\beta_0' x) - m_j^*(\beta_0' x)\right| = \mathrm{O}_P\big((\log n)^{1/2} h^{-1} n^{-1/2} a_n^{-1/2}\big),$$

$$\sup_{u\in\beta_0'\mathcal{X}} \left|\hat{m}_j'(u) - m_j^{*\prime}(u)\right| = \mathrm{O}_P\big((\log n)^{1/2} h^{-2} n^{-1/2} a_n^{-1/2}\big),$$

where the functions m_j^* are defined in (A.15), *and where m' denotes the derivative of the univariate function $\beta_0'\mathcal{X} \ni u \to m(u)$.*

Note that $\hat{f}'(u;\beta_0) = \hat{m}_1(u)$ (resp., $f^{*\prime}(u;\beta_0) = m_1^*(u)$). Combine Propositions 3.6 and 3.7 and deduce that $\hat{f}$ satisfies Assumptions 9–11 if $nh^8 \to 0$, $na_n h^4(\log n)^{-1} \to \infty$ and $ha_n^{-1/2}(\log n)^{1/2} \to 0$. In the case where $a_n = n^{-1/[4-\delta]}$ for some $\delta \in (0,1)$, these conditions are satisfied if $nh^{4(4-\delta)/(3-\delta)}(\log n)^{-(4-\delta)/(3-\delta)} \to \infty$ and $nh^{8-2\delta}(\log n)^{4-\delta} \to 0$.

4. Simulation study

To investigate the small sample behavior of our procedure, we carry out a small simulation study in which we consider two models. In the first model, the regression function is given by

$$m_1(\beta_0' x) = \beta_0' x - 0.5(\beta_0' x)^2,$$

and in the second

$$m_2(\beta_0' x) = \log(1 + 0.5\beta_0' x),$$

with $\beta_0 = (1, 0.75, 0.25, -0.5)$. We consider residuals $\varepsilon = Y - m_j(\beta_0' X)$ (for $j = 1, 2$) that are Gaussian variables $\mathcal{N}(0, 1)$ independent of X. The covariates are composed of 4 independent components, following an uniform distribution on [0, 1].

The censoring variable C follows an exponential distribution with mean $\gamma \exp(\theta_0' X)$ conditional on the covariate X_i, where $\theta_0 = (-0.1, -0.2, 0.1, -0.3)$, and γ is a parameter that allows us to modify the average proportion of censored responses. The parameter θ_0 is estimated by maximizing the Cox pseudo-likelihood, since the regression model on C is a proportional hazards model.

We consider 10,000 replications of this simulation scheme for $n = 200$. For each simulated sample j, we compute the resulting estimator $\hat{\beta}^{(j)}$ of β_0 and compute $\|\hat{\beta}^{(j)} - \beta_0\|_2^2$. We then deduce an estimator of the mean squared error (MSE) $E[\|\hat{\beta} - \beta_0\|_2^2]$. We take $a_n = 2$ for the bandwidth involved in Beran's estimator. Since the procedure is more sensitive to the choice of the second bandwidth h, we consider a set of bandwidths $h_j = 0.5 + j0.1$, for $j = 1, \ldots, 10$, and for each sample, we take the bandwidth that gives the lowest value of $M_n(\beta, \hat{f}, J)$ defined in (2.10). In Table 1, we compare the MSE of the estimator that we propose to the MSE of an estimator based on Kaplan–Meier weights, that is if we replace Beran's estimator in our approach by a standard Kaplan–Meier estimator. This alternative estimator is the one defined in Lopez [20]. As for our approach, this estimator puts more weights to the largest uncensored observations caused by censoring. Nevertheless this alternative procedure is not adapted to Assumption (A0) that we use herein. Hence, the estimator of Lopez [20] is expected to fail in our simulation setting.

As expected, our estimator based on the conditional Kaplan–Meier weighting outperforms the estimator of Lopez [20] in the different situations we consider. It is also natural to observe that the MSE of our $\hat{\beta}$ increases with the proportion of censoring.

Table 1. Comparison of the MSE of the proposed estimator $\hat{\beta}$ (columns CKM) with the MSE of the estimator based on Kaplan–Meier weights (columns KM) for different proportions of censoring

	Proportion of censoring					
	15%		30%		50%	
Regression model	CKM	KM	CKM	KM	CKM	KM
m_1	1.022	1.463	1.147	1.279	1.619	1.728
m_2	0.580	1.480	1.290	1.613	1.407	1.633

Appendix A: Assumptions and conditions

We split the assumptions in three parts, namely those required for the estimation of $F(x, y)$, the estimation of β_0, and the estimation of $f(\cdot; \beta)$.

Assumptions needed for the estimation of $F(x, y)$

The asymptotic results related to the estimator $\hat{F}_{\hat{g}}(x, y)$ will be valid under the following assumptions and conditions.

Assumption 1. *The distribution $\mathbb{P}(Z_\theta \le z)$ has three uniformly bounded derivatives for $z \in \mathcal{Z}_\theta$ and $\theta \in \Theta$, and the densities $f_{Z_\theta}(z)$ satisfy* $\inf_{\theta\in\Theta} \inf_{z\in\mathcal{Z}_\theta} f_{Z_\theta}(z) > 0$.

For any function $J(t|z)$ we will denote by $J_c(t|z)$ the continuous part, and $J_d(t|z) = J(t|z) - J_c(t|z)$. Assumption 2 below has been introduced by Du and Akritas [7] to obtain their asymptotic i.i.d. representation of the conditional Kaplan–Meier estimator.

Assumption 2. (i) *Let $L(y|z)$ denote $H_{\theta_0}(y|z)$ or $H_{\theta_0,0}(y|z)$. Then, $\nabla_z L(y|z)$ and $\nabla^2_{z,z} L(y|z)$ exist, are continuous with respect to z, and are uniformly bounded as functions of (z, y).*

(ii) *For some positive nondecreasing bounded (on $[-\infty; \tau]$) functions L_1, L_2, L_3, we have, for all $z \in \mathcal{Z}_{\theta_0}$,*

$$\begin{aligned} \left|H_{\theta_0 c}(t_1|z) - H_{\theta_0 c}(t_2|z)\right| &\le \left|L_1(t_1) - L_1(t_2)\right|, \\ \left|\nabla_z H_{\theta_0 c}(t_1|z) - \nabla_z H_{\theta_0 c}(t_2|z)\right| &\le \left|L_2(t_1) - L_2(t_2)\right|, \\ \left|\nabla_z H_{\theta_0,0c}(t_1|z) - \nabla_z H_{\theta_0,0c}(t_2|z)\right| &\le \left|L_3(t_1) - L_3(t_2)\right|, \end{aligned}$$

the last two assumptions implying the same kind for $\nabla_z H_{1c}$.

(iii) *The jumps of $F_g(\cdot|z)$ and $G_{\theta_0}(\cdot|z)$ are the same for all $z \in \mathcal{Z}_{\theta_0}$. Let $(d_1, d_2, \ldots)$ be the atoms of G.*

(iv) *$F_g(\cdot|z)$ and $G_{\theta_0}(\cdot|z)$ have two derivatives with respect to z, with the first derivatives uniformly bounded (on $[-\infty; \tau]$). The variation of the functions $\nabla_z F_g(\cdot|z)$ and $\nabla^2_{z,z} F_g(\cdot|z)$ on $[-\infty; \tau]$ is bounded by a constant not depending on z.*

(v) *For all d_i, define*

$$\begin{aligned} s_i &= \sup_{z\in\mathcal{Z}_{\theta_0}} \left|F_g(d_i - |z) - F_g(d_i|z)\right|, \\ s_i' &= \sup_{z\in\mathcal{Z}_{\theta_0}} \left|\nabla_z F_g(d_i - |z) - \nabla_z F_g(d_i|z)\right|, \\ r_i &= \sup_{z\in\mathcal{Z}_{\theta_0}} \left|G_{\theta_0}(d_i - |z) - G_{\theta_0}(d_i|z)\right|, \\ r_i' &= \sup_{z\in\mathcal{Z}_{\theta_0}} \left|\nabla_z G_{\theta_0}(d_i - |z) - \nabla_z G_{\theta_0}(d_i|z)\right|. \end{aligned}$$

Then, $\sum_{d_i \le \tau}(s_i + s_i' + r_i + r_i') < \infty$.

Assumption 3. *The kernel K is a symmetric probability density function with compact support, and K has bounded second derivative.*

Assumption 4. *The bandwidth a_n satisfies* $(\log n)n^{-1}a_n^{-3} \to 0$ *and* $na_n^4 \to 0$.

Assumption 5. *The function $(x,t,\theta) \mapsto G_\theta(t|\lambda(\theta,x))$ is differentiable with respect to θ, and the vector $\nabla_\theta G_\theta(t|\lambda(\theta,x))$ is uniformly bounded in (x,t,θ).*

The class of functions $\mathcal{F}$ considered in Section 3.1 should satisfy the following conditions, which are taken over from Lopez [21]. The conditions make use of concepts from the context of empirical processes, which can be found, for example, in Van der Vaart and Wellner [32].

Condition 1. *Let $p_0(x,y,c) = \mathbf{1}_{y\le c}[1 - G_{\theta_0}(y - |g(x))]^{-1}$. The class $p_0\mathcal{F}$ is $\mathbb{P}_{(X,Y,C)}$-Glivenko–Cantelli, and has an integrable envelope Φ_0 satisfying $\Phi_0(x,y,c) = 0$ for $y > \tau$.*

Condition 2. *The covering number $N(\varepsilon, \mathcal{F}, L^2(\mathbb{P}_{(X,Y)}))$ is bounded by $A\varepsilon^{-V}$ for $\varepsilon > 0$ and for some $A, V > 0$, and $\mathcal{F}$ has a square integrable envelope Φ satisfying $\Phi(x,y) = 0$ for $y > \tau$.*

Let $Z = Z_{\theta_0} = g(X)$, let $F_z(x,y) = \mathbb{P}(X \le x, Y \le y | Z = z)$, and for any function $\phi(x,y)$, define $\bar\phi(z,s) = \int \mathbf{1}_{s\le y}\phi(x,y)\,\mathrm{d}F_z(x,y)$. Let $\mathcal{Z}_{\theta_0,\eta}$ be the set of all points at a distance at least $\eta > 0$ from the complementary of $\mathcal{Z}_{\theta_0}$.

Condition 3. *For all $\phi \in \mathcal{F}$, $\bar\phi$ is twice differentiable with respect to z, and*

$$\sup_{s\le\tau, z\in\mathcal{Z}_{\theta_0,\eta}} \left\{\left|\nabla_z\bar\phi(z,s)\right| + \left|\nabla^2_{z,z}\bar\phi(z,s)\right|\right\} \le M < \infty$$

for some constant M not depending on ϕ. Moreover, $\bar\Phi$ is bounded on $\mathcal{Z}_{\theta_0,\eta} \times]-\infty;\tau]$, and has bounded partial derivatives with respect to z, where Φ is the envelope function of Condition 2.

The reason for introducing the set $\mathcal{Z}_{\theta_0,\eta}$ is to prevent us from boundary effects coming from kernel estimators. See Lopez [21] for a detailed discussion on this issue.

Assumptions needed for the estimation of β_0

We next state the additional assumptions needed for the asymptotic results concerning the estimation of the parameters in the single index model.

Assumption 6. *There exist $0 < c_0 < c_1 < \infty$ and $\eta > 0$ such that, for each $c \in [c_0, c_1]$ and $x \in \mathcal{X}$,*

$$\mathbf{1}_{f^\tau_{\beta_0}(\beta_0'x) > c} = 1 \quad \Longrightarrow \quad g(x) \in \mathcal{Z}_{\theta_0,\eta}.$$

Moreover, assume that

$$\left|f_{\beta_1}^{\tau}\left(\beta_1' x\right) - f_{\beta_2}^{\tau}\left(\beta_2' x\right)\right| \le C\|\beta_1 - \beta_2\|^{\alpha}$$

for some positive constant C and some $\alpha > 0$.

Assumption 7. (i) $E(|Y|^3) < \infty$;

(ii) $E[\{f(\beta' X;\beta) - f(\beta_0' X;\beta_0)\}^2 \mathbf{1}_{Y\le\tau}] = 0 \Longrightarrow \beta = \beta_0$;

(iii) $\beta_0 = (1, \tilde{\beta}_0')'$ *with $\tilde{\beta}_0$ an interior point of $\tilde{\mathcal{B}}$;*

(iv) *The class $\{(x, y) \to f(\beta' x;\beta)\mathbf{1}_{y\le\tau}\colon\ \beta\in\mathcal{B}\}$ satisfies Condition 1 for a continuous integrable envelope Ψ.*

Assumption 8. *The classes $\{x \to \nabla_\beta f(\beta' x;\beta)\colon\ \beta\in\mathcal{B}\}$ and $\{x \to \nabla^2_{\beta,\beta} f(\beta' x;\beta)\colon\ \beta\in\mathcal{B}\}$ are VC-classes of continuous functions for a uniformly bounded envelope.*

Assumptions needed for the estimation of $f(\cdot;\beta)$

The last group of assumptions is required for the generic estimator $\hat{f}(\cdot;\beta)$. They are verified in Section 3.2 for the estimator defined in (2.5).

Assumption 9. *For all $c > 0$,*

$$\sup_{\beta\in\mathcal{B},x\in\mathcal{X}} \left|\hat{f}\left(\beta' x;\beta\right) - f\left(\beta' x;\beta\right)\right| \mathbf{1}_{f_\beta^\tau(\beta' x) > c} = o_P(1), \tag{A.1}$$

$$\sup_{\beta\in\mathcal{B},x\in\mathcal{X}} \left|\nabla_\beta\hat{f}\left(\beta' x;\beta\right) - \nabla_\beta f\left(\beta' x;\beta\right)\right| \mathbf{1}_{f_\beta^\tau(\beta' x) > c} = o_P(1), \tag{A.2}$$

$$\sup_{\beta\in\mathcal{B},x\in\mathcal{X}} \left|\nabla^2_{\beta,\beta}\hat{f}\left(\beta' x;\beta\right) - \nabla^2_{\beta,\beta} f\left(\beta' x;\beta\right)\right| \mathbf{1}_{f_\beta^\tau(\beta' x) > c} = o_P(1). \tag{A.3}$$

Assumption 10. *There exist Donsker classes $\mathcal{H}_1$ and $\mathcal{H}_2$ such that $f(\cdot;\beta_0)\in\mathcal{H}_1$ and $\nabla_\beta f(\beta_0'\cdot;\beta_0)\in\mathcal{H}_2$, and such that with probability tending to one, $\hat{f}(\cdot;\beta_0)\in\mathcal{H}_1$ and $\nabla_\beta\hat{f}(\beta_0'\cdot;\beta_0)\in\mathcal{H}_2$.*

Typical examples of such kind of Donsker classes are classes of regular functions. Let $\mathcal{T} = \{\beta_0' x\colon\ x\in\mathcal{X}\}\subset\mathbb{R}$ and let $\mathcal{C}^1_\ell(\mathcal{T}, M) = \{h : \mathcal{T}\mapsto\mathbb{R}^\ell\colon\ \sup_{t\in\mathcal{T}}\{|h(t)| + |h'(t)|\}\le M\}$ for $\ell\ge 1$ and for some $M < \infty$. Define

$$\mathcal{H}_1^0 = \mathcal{C}_1^1(\mathcal{T}, M), \tag{A.4}$$

$$\mathcal{H}_2^0 = \left\{h : \mathcal{X}\mapsto\mathbb{R}^d\colon\ x\mapsto x h_1\left(\beta_0' x\right) + h_2\left(\beta_0' x\right)\colon\ h_1\in\mathcal{C}_1^1(\mathcal{T}, M), h_2\in\mathcal{C}_d^1(\mathcal{T}, M)\right\}. \tag{A.5}$$

The class $\mathcal{H}_2^0$ is a Donsker class, which follows from stability properties of Donsker classes (see, e.g., Examples 2.10.7 and 2.10.10 in Van der Vaart and Wellner [32]).

Assumption 11. *For all $c > 0$,*

$$\sup_{x\in\mathcal{X}} \big|\hat{f}(\beta_0' x;\beta_0) - f(\beta_0' x;\beta_0)\big|\mathbf{1}_{f^{\tau}_{\beta_0}(\beta_0' x)>c} = \mathrm{O}_P(\varepsilon_n),$$

$$\sup_{x\in\mathcal{X}} \big|\nabla_\beta \hat{f}(\beta_0' x;\beta_0) - \nabla_\beta f(\beta_0' x;\beta_0)\big|\mathbf{1}_{f^{\tau}_{\beta_0}(\beta_0' x)>c} = \mathrm{O}_P(\varepsilon_n'),$$

where ε_n and ε_n' satisfy $\varepsilon_n\varepsilon_n' = \mathrm{o}(n^{-1/2})$, $a_n^{-1/2}(\log n)^{1/2}\varepsilon_n \to 0$ and $a_n^{-1/2}(\log n)^{1/2}\varepsilon_n' \to 0$.

Appendix B: Technical lemmas and proofs

We start this Appendix with two technical lemmas, needed in the proofs of the main results. The first technical lemma gives a concentration inequality for the convergence rate of semi-parametric estimators.

Let b_n be a sequence of real numbers tending to zero, and let $\{\zeta_\alpha\colon \alpha\in\mathcal{A}\}$ be a family of uniformly bounded functions, where $\mathcal{A}$ is a compact subset of $\mathbb{R}^p$ (with $p \geq 1$). Consider the class of functions

$$\begin{aligned}\mathcal{G} = \bigg\{&(u,z,t,\delta) \mapsto g_{\alpha,x,v}(u,z,t,\delta)\\ &= K^0\bigg(\frac{\psi(\alpha,u)-\psi(\alpha,x)}{b_n}\bigg)\zeta_\alpha(x,u,z,t,\delta)\xi(t)\mathbf{1}_{t\leq v}\colon \alpha\in A, x\in\mathcal{X}, v\in\mathbb{R}\bigg\},\end{aligned} \tag{A.6}$$

where K^0, ψ and ξ are fixed functions, $\mathcal{X}\subset\mathbb{R}^d$ is a compact set, and $t\in\mathbb{R}$, and consider the process (in α, x and v)

$$\nu_n(g_{\alpha,x,v}) = \sum_{i=1}^n \big(g_{\alpha,x,v}(X_i,Z_i,T_i,\delta_i) - E\big[g_{\alpha,x,v}(X,Z,T,\delta)\big]\big).$$

Typically, K^0 denotes either a kernel or its derivative of order 1 or 2.

Lemma A.1. *Assume that the class of functions*

$$\bigg\{(u,z,t,\delta)\to K^0\bigg(\frac{\psi(\alpha,u)-\psi(\alpha,x)}{b_n}\bigg)\zeta_\alpha(x,u,z,t,\delta)\colon \alpha\in\mathcal{A}, x\in\mathcal{X}\bigg\} \tag{A.7}$$

is a VC-class of functions for a constant envelope, assume that $E[|\xi(T)|^3]<\infty$, and that $nb_n^3/(\log n)\to\infty$. Then,

$$n^{-1/2}b_n^{-1/2}\big[\log(1/b_n)\big]^{-1}\|\nu_n\|_{\mathcal{G}} = \mathrm{O}_P(1),$$

where $\|\cdot\|_{\mathcal{G}}$ denotes the uniform norm over all maps in $\mathcal{G}$.

The proof of Lemma A.1 is a consequence of Proposition 1 in Einmahl and Mason [8] and Talagrand's inequality [30], and it is available from the long version of this paper, see arXiv:1111.6232.

Remark. Note that if K^0 is of bounded variation with compact support, and if $\psi(\alpha, x) = \alpha' x$, then (A.7) holds, see Nolan and Pollard [24].

The second technical lemma shows the consistency of the estimator $\hat{G}_\theta(t|\lambda(\theta, x))$ and its vector of partial derivatives, uniformly in t, θ and x, and it also establishes the rate of convergence of the estimator $\hat{G}_{\hat{\theta}}(t|\hat{g}(x))$, uniformly in t and x.

Lemma A.2. *Under the assumptions of Theorem* 3.2, *we have*

$$\sup_{t\le\tau,\theta\in\Theta,x\in\mathcal{X}} \bigl|\hat{G}_\theta\bigl(t|\lambda(\theta, x)\bigr) - G_\theta\bigl(t|\lambda(\theta, x)\bigr)\bigr| = \mathrm{o}_P(1), \tag{A.8}$$

$$\sup_{t\le\tau,\theta\in\Theta,x\in\mathcal{X}} \bigl|\nabla_\theta\hat{G}_\theta\bigl(t|\lambda(\theta, x)\bigr) - \nabla_\theta G_\theta\bigl(t|\lambda(\theta, x)\bigr)\bigr| = \mathrm{o}_P(1), \tag{A.9}$$

$$\sup_{t\le\tau}\ \sup_{x:g(x)\in\mathcal{Z}_{\theta_0,\eta}} \bigl|\hat{G}_{\hat{\theta}}\bigl(t|\hat{g}(x)\bigr) - G_{\theta_0}\bigl(t|g(x)\bigr)\bigr| = \mathrm{O}_P\bigl(n^{-1/2}a_n^{-1/2}(\log n)^{1/2}\bigr). \tag{A.10}$$

Proof. For the first part, with probability tending to 1, for $t \le \tau$, $1 - \hat{G}(t|\lambda(\theta, x)) > 0$. Taking the logarithm, one obtains

$$\log\bigl(1 - \hat{G}\bigl(t|\lambda(\theta, x)\bigr)\bigr) = \sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i\le t}\log\bigl(1 - W_{n,i}(x,\theta)\bigr),$$

where

$$\begin{aligned} W_{n,i}(x,\theta) &= W_n(X_i, T_i; x, \theta) \\ &= K\biggl(\frac{\lambda(\theta, X_i) - \lambda(\theta, x)}{a_n}\biggr)\bigg/\biggl(\sum_{j=1}^n \mathbf{1}_{T_j\ge T_i} K\biggl(\frac{\lambda(\theta, X_j) - \lambda(\theta, x)}{a_n}\biggr)\biggr). \end{aligned}$$

A Taylor expansion leads to

$$\log\bigl(1 - \hat{G}\bigl(t|\lambda(\theta, x)\bigr)\bigr) = -\sum_{i=1}^n (1-\delta_i) W_{n,i}(x,\theta)\mathbf{1}_{T_i\le t} + \mathrm{O}_P\bigl(n^{-1}a_n^{-2}\bigr),$$

where the order of the remainder term is uniform in t, θ, x, as

$$\sup_{i:T_i\le\tau}\ \sup_{x,\theta}\bigl|W_{n,i}(x,\theta)\bigr| = \mathrm{O}_P\bigl(n^{-1}a_n^{-1}\bigr).$$

The remainder term is $o_P(1)$ if $na_n^2 \to \infty$. Rewrite

$$\sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i \le t} W_{n,i}(x,\theta) = \frac{1}{na_n}\sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i\le t} K\left(\frac{\lambda(\theta,X_i)-\lambda(\theta,x)}{a_n}\right) S_\theta\big(\lambda(\theta,x),T_i\big)^{-1}$$
$$+ \frac{1}{na_n}\sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i\le t} K\left(\frac{\lambda(\theta,X_i)-\lambda(\theta,x)}{a_n}\right)$$
$$\times \frac{\hat{S}_\theta(\lambda(\theta,x),T_i) - S_\theta(\lambda(\theta,x),T_i)}{S_\theta(\lambda(\theta,x),T_i)\hat{S}_\theta(\lambda(\theta,x),T_i)},$$

where

$$S_\theta\big(\lambda(\theta,x),y\big) = \big[1 - H_\theta\big(y|\lambda(\theta,x)\big)\big] f_{Z_\theta}\big(\lambda(\theta,x)\big),$$
$$\hat{S}_\theta\big(\lambda(\theta,x),y\big) = \frac{1}{na_n}\sum_{j=1}^n \mathbf{1}_{T_j\ge y} K\left(\frac{\lambda(\theta,X_j)-\lambda(\theta,x)}{a_n}\right).$$

Apply Lemma A.1 to obtain the uniform convergence of $\hat{S}_\theta$ towards S_θ, and to show that

$$\sup_{x,\theta\in\Theta,t\le\tau}\left|\frac{1}{na_n}\sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i\le t} K\left(\frac{\lambda(\theta,X_i)-\lambda(\theta,x)}{a_n}\right) S_\theta\big(\lambda(\theta,x),T_i\big)^{-1} \right.$$
$$\left. - \int_{-\infty}^t \frac{\mathrm{d}H_{\theta,0}(s|\lambda(\theta,x))}{1-H_\theta(s-|\lambda(\theta,x))}\right| = o_P(1).$$

Since S_θ is uniformly bounded away from zero for $y \le \tau$, see Assumption 1, the result follows from

$$\exp\left[-\int_{-\infty}^t \frac{\mathrm{d}H_{\theta,0}(s|\lambda(\theta,x))}{1-H_\theta(s-|\lambda(\theta,x))}\right] = 1 - G_\theta\big(t|\lambda(\theta,x)\big).$$

For the gradient, we have

$$\nabla_\theta \hat{G}_\theta\big(t|\lambda(\theta,x)\big) = \big(1-\hat{G}_\theta\big(t|\lambda(\theta,x)\big)\big)\sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i\le t}\frac{\nabla_\theta W_{n,i}(x,\theta)}{1-W_{n,i}(x,\theta)}.$$

From this, we deduce that the convergence of $\nabla_\theta \hat{G}_\theta$ follows from the convergence of $\hat{G}_\theta$, of $\hat{S}_\theta$ and of

$$\frac{1}{na_n^2}\sum_{i=1}^n (1-\delta_i)\mathbf{1}_{T_i\le t}\nabla_\theta\lambda(\theta,x) K'\left(\frac{\lambda(\theta,X_i)-\lambda(\theta,x)}{a_n}\right),$$

and

$$\frac{1}{na_n^2}\sum_{i=1}^{n}\mathbf{1}_{T_i\leq t}\nabla_\theta\lambda(\theta,x)K'\left(\frac{\lambda(\theta,X_i)-\lambda(\theta,x)}{a_n}\right).$$

These two quantities can be studied using Lemma A.1, which shows that their centered versions converge uniformly with rate $(na_n^3)^{-1/2}\log n$, while the bias term is of order a_n^2.

The third result can be deduced from a Taylor expansion, Assumption 5 and Proposition 4.3 in Van Keilegom and Akritas [33]. Indeed, we can deduce that

$$\begin{aligned}
&\sup_{t\leq\tau}\sup_{x:g(x)\in\mathcal{Z}_{\theta_0,\eta}}\left|\hat{G}_{\hat{\theta}}\big(t|\hat{g}(x)\big)-G_{\theta_0}\big(t|g(x)\big)\right| \\
&\quad\leq\sup_{t\leq\tau}\sup_{x:g(x)\in\mathcal{Z}_{\theta_0,\eta}}\left|\hat{G}_{\theta_0}\big(t|g(x)\big)-G_{\theta_0}\big(t|g(x)\big)\right|+\mathrm{O}_P\big(\|\hat{\theta}-\theta_0\|\big).
\end{aligned}$$

□

We are now ready to give the proofs of the main results.

Proof of Theorem 3.2. Part (i) of the theorem can be easily derived by replacing the differentiability condition in Assumption 5 by a uniform continuity condition on G_θ with respect to θ, and equation (A.8) in Lemma A.2.

For part (ii), a Taylor expansion with respect to θ leads to

$$\int\phi(x,y)\,\mathrm{d}[\hat{F}_{\hat{g}}-\hat{F}_g](x,y)=-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(X_i,T_i)\nabla_\theta\hat{G}_{\theta_n}(T_i-|\lambda(\theta_n,X_i))(\hat{\theta}-\theta_0)}{[1-\hat{G}_{\theta_n}(T_i-|\lambda(\theta_n,X_i))]^2}$$

for some θ_n between $\hat{\theta}$ and θ_0. From the convergence of $\hat{\theta}$ towards θ_0, it follows that θ_n tends to θ_0. Moreover, applying equation (A.8) and (A.9) in Lemma A.2, we obtain that

$$\begin{aligned}
\int\phi(x,y)\,\mathrm{d}[\hat{F}_{\hat{g}}-\hat{F}_g](x,y)&=-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(X_i,T_i)\nabla_\theta G_{\theta_0}(T_i-|\lambda(\theta_0,X_i))(\hat{\theta}-\theta_0)}{[1-G_{\theta_0}(T_i-|g(X_i))]^2}\\
&\quad+R_n(\phi),\\
&=U_n(\phi)+R_n(\phi),
\end{aligned}$$

with $\sup_\phi|R_n(\phi)|\leq|R_n(\Phi)|=\mathrm{o}_P(n^{-1/2})$, and

$$U_n(\phi)=\left\{-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(X_i,T_i)\nabla_\theta G_{\theta_0}(T_i-|\lambda(\theta_0,X_i))}{[1-G_{\theta_0}(T_i-|g(X_i))]^2}\right\}\left\{\frac{1}{n}\sum_{j=1}^{n}\mu(T_j,\delta_j,X_j)\right\}+R_n'(\phi),$$

with $\sup_\phi|R_n'(\phi)|\leq|R_n'(\Phi)|=\mathrm{o}_P(n^{-1/2})$. Centering the first sum in $U_n(\phi)$ and applying a uniform central limit theorem (see, e.g., Van der Vaart and Wellner [32]), we obtain the stated representation. □

Proof of Theorem 3.3. Consider the difference

$$\begin{aligned}
&\left|M_n(\beta, \hat{f}, \tilde{J}) - M_n(\beta, f, \tilde{J})\right| \\
&\quad \le 2\int |y|\mathbf{1}_{y\le\tau}\,\mathrm{d}\hat{F}_{\hat{g}}(x,y) \sup_{x:\tilde{J}(x)=1,\beta\in\mathcal{B}} \left|\hat{f}\left(\beta'x;\beta\right) - f\left(\beta'x;\beta\right)\right| \\
&\qquad + \int \mathbf{1}_{y\le\tau}\left|\hat{f}\left(\beta'x;\beta\right) + f\left(\beta'x;\beta\right)\right|\mathrm{d}\hat{F}_{\hat{g}}(x,y) \sup_{x:\tilde{J}(x)=1,\beta\in\mathcal{B}} \left|\hat{f}\left(\beta'x;\beta\right) - f\left(\beta'x;\beta\right)\right|.
\end{aligned}$$

The first term on the right-hand side converges uniformly to zero by Assumption 9 and the law of large numbers for $\hat{F}_{\hat{g}}$ (see Theorem 3.1 and Theorem 3.2). The integral in the second term can be bounded by

$$\left(1+\mathrm{o}_P(1)\right) \times \int 2\Psi(x)\,\mathrm{d}\hat{F}_{\hat{g}}(x,y),$$

where $\mathrm{o}_P(1)$ is uniform in β, by Assumption 7 and 9 – (A.1). Now we have to show that $M_n(\beta, f, J^*)$ converges to $M(\beta, f, J^*)$ uniformly in β. For this, apply Theorem 3.1 and Theorem 3.2 using Assumption 7. By usual arguments for proving consistency (see, e.g., Van der Vaart [31], Theorem 5.7), the consistency of β_n follows. □

Proof of Lemma 3.4. The proof is somewhat similar to the proof of Lemma 5A in Dominitz and Sherman [6]. First, observe that

$$f\left(\beta'X;\beta\right) = E\left[Y|\beta'X, Y\le\tau\right] = E\left[f\left(\beta_0'X;\beta_0\right)|\beta'X, Y\le\tau\right] = \frac{E[f(\beta_0'X;\beta_0)\mathbf{1}_{Y\le\tau}|\beta'X]}{\mathbb{P}(Y\le\tau|\beta'X)}.$$

Let $\alpha(X,\beta) = \beta_0'X - \beta'X$. Define

$$\Gamma_X(\beta_1,\beta_2) = E\left[f\left(\alpha(X,\beta_1) + \beta_2'X;\beta_0\right)\mathbf{1}_{Y\le\tau}|\beta_2'X\right],$$

and note that $f(\beta'X;\beta) = \Gamma_X(\beta,\beta)/\mathbb{P}(Y\le\tau|\beta'X)$. Then,

$$\begin{aligned}
\nabla_{\beta_1}\Gamma_X(\beta_0,\beta_0) &= -f'\left(\beta_0'X;\beta_0\right)E\left[X\mathbb{P}(Y\le\tau|X)|\beta_0'X\right], \\
\nabla_{\beta_2}\Gamma_X(\beta_0,\beta_0) &= f'\left(\beta_0'X;\beta_0\right)X\mathbb{P}\left(Y\le\tau|\beta_0'X\right) + f\left(\beta_0'X;\beta_0\right)\nabla_\beta h(X,\beta_0),
\end{aligned}$$

where $h(x,\beta) = \mathbb{P}(Y\le\tau|\beta'X = \beta'x)$. It follows that

$$\begin{aligned}
&\nabla_\beta f\left(\beta_0'x;\beta_0\right) \\
&\quad = \frac{f'(\beta_0'x;\beta_0)\{x\mathbb{P}(Y\le\tau|\beta_0'X=\beta_0'x) - E[X\mathbb{P}(Y\le\tau|X)|\beta_0'X=\beta_0'x]\}}{\mathbb{P}(Y\le\tau|\beta_0'X=\beta_0'x)} \\
&\qquad + \frac{\nabla_\beta h(x,\beta_0)f(\beta_0'x;\beta_0)}{\mathbb{P}(Y\le\tau|\beta_0'X=\beta_0'x)} - \frac{\nabla_\beta h(x,\beta_0)f(\beta_0'x;\beta_0)E[\mathbf{1}_{Y\le\tau}|\beta_0'X=\beta_0'x]}{\mathbb{P}(Y\le\tau|\beta_0'X=\beta_0'x)^2} \qquad \text{(A.11)} \\
&\quad := xm_1\left(\beta_0'x\right) + m_2\left(\beta_0'x\right).
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&E\bigl[\nabla_\beta f\bigl(\beta_0'X;\beta_0\bigr)\mathbf{1}_{Y\le\tau}|\beta_0'X\bigr]\\
&\quad=\frac{E[f'(\beta_0'X;\beta_0)\{X\mathbb{P}(Y\le\tau|\beta_0'X)-E[X\mathbb{P}(Y\le\tau|X)|\beta_0'X]\}\mathbf{1}_{Y\le\tau}|\beta_0'X]}{\mathbb{P}(Y\le\tau|\beta_0'X)}\\
&\quad=0. \qquad\square
\end{aligned}
$$

Proof of Theorem 3.5. The proof consists of three steps:

Step 0: Replace J by J_0. For any $\mathcal{B}_n$ a sequence of shrinking neighborhoods of β_0,

$$\sup_{\beta\in\mathcal{B}_n}\bigl|M_n(\beta,\hat f,J)-M_n(\beta,\hat f,J_0)\bigr|\le \mathrm{o}_P\bigl(M_n(\beta,\hat f,J_0)+n^{-1}\bigr).$$

See Delecroix, Hristache and Patilea [5], page 738. Similar arguments apply also when the trimming J is defined with $\hat f^{\tau}_{\beta_n}(\beta_n'x)$ justifying the practical implementation of the trimming function.

Step 1: Bring the problem back to the parametric case.

For notational simplicity, we work with $\nabla_\beta f$ instead of $\nabla_{\tilde\beta} f$. Note that $\nabla_\beta f=(0,\nabla_{\tilde\beta}' f)'$. We will show that, on $\mathcal{B}_n$,

$$M_n(\beta,\hat f,J_0)=M_n(\beta,f,J_0)+\mathrm{o}_P\biggl(\frac{\|\beta-\beta_0\|}{\sqrt n}\biggr)+\mathrm{o}_P\bigl(\|\beta-\beta_0\|^2\bigr)+C_n',$$

where C_n' does not depend on β. Decompose

$$
\begin{aligned}
M_n(\beta,\hat f,J_0)&=M_n(\beta,f,J_0)\\
&\qquad-\frac{2}{n}\sum_{i=1}^n\frac{\delta_iJ_0(X_i)(T_i-f(\beta'X_i;\beta))\mathbf{1}_{T_i\le\tau}}{1-\hat G_{\hat\theta}(T_i-|\hat g(X_i))}\bigl[\hat f\bigl(\beta'X_i;\beta\bigr)-f\bigl(\beta'X_i;\beta\bigr)\bigr]\\
&\qquad+\frac{1}{n}\sum_{i=1}^n\frac{\delta_iJ_0(X_i)\mathbf{1}_{T_i\le\tau}}{1-\hat G_{\hat\theta}(T_i-|\hat g(X_i))}\bigl[\hat f\bigl(\beta'X_i;\beta\bigr)-f\bigl(\beta'X_i;\beta\bigr)\bigr]^2\\
&=M_n(\beta,f,J_0)-2A_{1n}+B_{1n}.
\end{aligned}
$$

Step 1.1: Study of A_{1n}.

A_{1n} can be expressed as

$$
\begin{aligned}
A_{1n}&=\frac{1}{n}\sum_{i=1}^n\frac{\delta_iJ_0(X_i)(T_i-f(\beta_0'X_i;\beta_0))\mathbf{1}_{T_i\le\tau}}{1-\hat G_{\hat\theta}(T_i-|\hat g(X_i))}\bigl[\hat f\bigl(\beta_0'X_i;\beta_0\bigr)-f\bigl(\beta_0'X_i;\beta_0\bigr)\bigr]\\
&\quad+\frac{1}{n}\sum_{i=1}^n\frac{\delta_iJ_0(X_i)\mathbf{1}_{T_i\le\tau}(f(\beta_0'X_i;\beta_0)-f(\beta'X_i;\beta))}{1-\hat G_{\hat\theta}(T_i-|\hat g(X_i))}\\
&\qquad\times\bigl[\hat f\bigl(\beta'X_i;\beta\bigr)-f\bigl(\beta'X_i;\beta\bigr)-\hat f\bigl(\beta_0'X_i;\beta_0\bigr)+f\bigl(\beta_0'X_i;\beta_0\bigr)\bigr]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}(f(\beta_0' X_i;\beta_0) - f(\beta' X_i;\beta))}{1-\hat G_{\hat\theta}(T_i - |\hat g(X_i))}\big[\hat f(\beta_0' X_i;\beta_0) - f(\beta_0' X_i;\beta_0)\big] \\
&+ \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i J_0(X_i)(T_i - f(\beta_0' X_i;\beta_0))\mathbf{1}_{T_i\le\tau}}{1-\hat G_{\hat\theta}(T_i - |\hat g(X_i))} \\
&\qquad \times \big[\hat f(\beta' X_i;\beta) - f(\beta' X_i;\beta) - \hat f(\beta_0' X_i;\beta_0) + f(\beta_0' X_i;\beta_0)\big] \\
&= A_{2n} + A_{3n} + A_{4n} + A_{5n}.
\end{aligned}
$$

A_{2n} does not depend on β. For A_{3n}, observe that, for any $\beta \in \mathcal{B}_n$, we can replace $J_0(X_i)$ by $\mathbf{1}_{f_\beta^\tau(\beta' X_i)>c/2}$ using Assumption 6. As $\nabla_\beta f(\beta' x;\beta)$ is a bounded function of x and β (Assumption 8, since the class of functions has a bounded envelope), and using the uniform convergence of $\nabla_\beta \hat f(\beta' x;\beta)$ (Assumption 9), we can obtain from a first order Taylor expansion applied twice (for $f(\beta' x;\beta)$ and for $\hat f(\beta' x;\beta) - f(\beta' x;\beta)$ around β_0), that $A_{3n} = o_P(\|\beta-\beta_0\|^2)$.

For A_{4n}, first replace $\hat G_{\hat\theta}$ with G_{θ_0}. For this, note that $[1 - G_{\theta_0}(T_i - |g(X_i))]$ is bounded away from zero with probability tending to 1 for $T_i \le \tau$, and that

$$
\sup_{t\le\tau, x: J_0(x)=1} \big|\hat G_{\hat\theta}\big(t|\hat g(x)\big) - G_{\theta_0}\big(t|g(x)\big)\big|\big|\hat f\big(\beta_0' x;\beta_0\big) - f\big(\beta_0' x;\beta_0\big)\big| = o_P\big(n^{-1/2}\big) \tag{A.12}
$$

using part 2 of Assumption 11, and Lemma A.2. A first order Taylor expansion for $f(\beta' x;\beta) - f(\beta_0' x;\beta_0)$ and property (A.12) lead to

$$
\begin{aligned}
A_{4n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}(f(\beta_0' X_i;\beta_0) - f(\beta' X_i;\beta))}{1-G_{\theta_0}(T_i - |g(X_i))}\big[\hat f(\beta_0' X_i;\beta_0) - f(\beta_0' X_i;\beta_0)\big] \\
&\quad + o_P\left(\frac{\|\beta-\beta_0\|}{\sqrt{n}}\right).
\end{aligned}
$$

Next, a second order Taylor development shows that the first term above can be rewritten as

$$
\begin{aligned}
&\frac{(\beta-\beta_0)'}{n}\sum_{i=1}^{n} \frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}\nabla_\beta f(\beta_0' X_i;\beta_0)}{1-G_{\theta_0}(T_i - |g(X_i))}\big[f(\beta_0' X_i;\beta_0) - \hat f(\beta_0' X_i;\beta_0)\big] \\
&\quad + o_P\big(\|\beta-\beta_0\|^2\big).
\end{aligned}
\tag{A.13}
$$

To show that this term is negligible, we will use empirical process theory. We have that $f \in \mathcal{H}_1$, where $\mathcal{H}_1$ is the Donsker class defined in Assumption 10, and $\hat f \in \mathcal{H}_1$ with probability tending to 1. Consequently, the class of functions

$$
\mathcal{H}_1' = \left\{(y,c,x,t) \to \frac{\mathbf{1}_{y\le c}\mathbf{1}_{y\le\tau}\nabla_\beta f(\beta_0' x;\beta_0)J_0(t)\phi(\beta_0' t)}{1-G_{\theta_0}(y\wedge c - |g(x))} : \phi \in \mathcal{H}_1\right\}
$$

is a Donsker class, see Example 2.10.8 in Van der Vaart and Wellner [32]. Furthermore, for all $\phi \in \mathcal{H}_1$,

$$E\left[\frac{\delta J_0(X)\nabla_\beta f(\beta_0'X;\beta_0)\phi(\beta_0'X)\mathbf{1}_{T\le\tau}}{1-G_{\theta_0}(T-|g(X))}\right] = E\left[\nabla_\beta f\left(\beta_0'X;\beta_0\right)\phi\left(\beta_0'X\right)J_0(X)\mathbf{1}_{Y\le\tau}\right] \\ = 0, \tag{A.14}$$

since $E[\nabla_\beta f(\beta_0'X,\beta_0)\mathbf{1}_{Y\le\tau}|\beta_0'X]=0$ (see Lemma 3.4), and since $J_0(X)$ is a function of $\beta_0'X$ alone. Deduce that, since $\mathcal{H}_1'$ is a Donsker class, and since $\hat{f}$ tends uniformly to f, that the first term in (A.13) is of order $o_P(\|\beta-\beta_0\|n^{-1/2})$. See the asymptotic equicontinuity of Donsker classes, cf. Van der Vaart and Wellner [32], Section 2.1.2.

For A_{5n}, apply a second order Taylor expansion. Using that $\nabla^2_{\beta,\beta}f$ is bounded, and that $\nabla^2_{\beta,\beta}\hat{f}$ converges uniformly to $\nabla^2_{\beta,\beta}f$, we obtain

$$A_{5n} = \frac{(\beta-\beta_0)'}{n}\sum_{i=1}^n \frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}(T_i-f(\beta_0'X_i;\beta_0))[\nabla_\beta f(\beta_0'X_i;\beta_0)-\nabla_\beta\hat{f}(\beta_0'X_i;\beta_0)]}{1-\hat{G}_{\hat{\theta}}(T_i-|\hat{g}(X_i))} \\ + o_P\left(\|\beta-\beta_0\|^2\right).$$

Proceed as for A_{4n} to replace $\hat{G}$ and $\hat{g}$ by G and g, using part 3 of Assumption 11. The same arguments as for A_{4n} can then be used, but considering instead the Donsker class

$$\mathcal{H}_2' = \left\{(y,c,x)\to\frac{\mathbf{1}_{y\le c}J_0(x)\mathbf{1}_{y\le\tau}(y-f(\beta_0'x;\beta_0))\phi(x)}{1-G_{\theta_0}(y-|g(x))}\colon \phi\in\mathcal{H}_2\right\},$$

where $\mathcal{H}_2$ is defined in Assumption 10, and observing that, for any function ϕ,

$$E\left[\frac{\delta J_0(X)\phi(X)(Y-f(\beta_0'X;\beta_0))\mathbf{1}_{T\le\tau}}{1-G_{\theta_0}(T-|g(X))}\right] \\ = E\left[E\left[\left(Y-f\left(\beta_0'X;\beta_0\right)\right)\mathbf{1}_{Y\le\tau}|X\right]J_0(X)\phi(X)\right]=0,$$

by the definition of our regression model. Deduce that $A_{5n}=o_P(\|\beta-\beta_0\|n^{-1/2}+\|\beta-\beta_0\|^2)$.

Step 1.2: Study of B_{1n}.

Rewrite B_{1n} as

$$B_{1n} = \frac{1}{n}\sum_{i=1}^n\frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}}{1-\hat{G}_{\hat{\theta}}(T_i-|\hat{g}(X_i))} \\ \times\left[\hat{f}\left(\beta'X_i;\beta\right)-f\left(\beta'X_i;\beta\right)-\hat{f}\left(\beta_0'X_i;\beta_0\right)+f\left(\beta_0'X_i;\beta_0\right)\right]^2 \\ +\frac{1}{n}\sum_{i=1}^n\frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}}{1-\hat{G}_{\hat{\theta}}(T_i-|\hat{g}(X_i))}\left[\hat{f}\left(\beta_0'X_i;\beta_0\right)-f\left(\beta_0'X_i;\beta_0\right)\right]^2$$

$$
\begin{aligned}
&+\frac{2}{n}\sum_{i=1}^{n}\frac{\delta_i J_0(X_i)\mathbf{1}_{T_i\le\tau}}{1-\hat G_{\hat\theta}(T_i-|\hat g(X_i))}\bigl[\hat f\bigl(\beta_0'X_i;\beta_0\bigr)-f\bigl(\beta_0'X_i;\beta_0\bigr)\bigr]\\
&\qquad\times\bigl[\hat f\bigl(\beta'X_i;\beta\bigr)-f\bigl(\beta'X_i;\beta\bigr)-\hat f\bigl(\beta_0'X_i;\beta_0\bigr)+f\bigl(\beta_0'X_i;\beta_0\bigr)\bigr]\\
&=B_{2n}+B_{3n}+2B_{4n}.
\end{aligned}
$$

Observe that, for any $\beta\in\mathcal{B}_n$, we can replace $J_0(X_i)$ by $\mathbf{1}_{f_\beta^\tau(\beta'X_i)>c/2}$ using Assumption 8. Next, by a Taylor expansion and the uniform convergence of $\nabla_\beta\hat f$, we have that $B_{2n}=\mathrm{o}_P(\|\beta-\beta_0\|^2)$. The term B_{3n} does not depend on β. For B_{4n}, a second order Taylor expansion leads to

$$
\begin{aligned}
B_{4n}=\frac{(\beta-\beta_0)'}{n}\sum_{i=1}^{n}&\frac{\delta_i J_0(X_i)[\hat f(\beta_0'X_i;\beta_0)-f(\beta_0'X_i;\beta_0)]}{1-\hat G_{\hat\theta}(T_i-|\hat g(X_i))}\\
&\times\bigl[\nabla_\beta\hat f\bigl(\beta_0'X_i;\beta_0\bigr)-\nabla_\beta f\bigl(\beta_0'X_i;\beta_0\bigr)\bigr]+\mathrm{o}_P\bigl(\|\beta-\beta_0\|^2\bigr).
\end{aligned}
$$

Replace $\hat G$ by G and use Assumption 11, part 1, to conclude.

Step 2: Study of $M_n(\beta,f,J_0)$.

Observe that, on $\mathrm{o}_P(1)$-neighborhoods of β_0, from a Taylor expansion,

$$
\begin{aligned}
&M_n(\tilde\beta,f,J_0)-M_n(\tilde\beta_0,f,J_0)\\
&\quad=(\tilde\beta-\tilde\beta_0)'\nabla_{\tilde\beta}M_n(\tilde\beta_0,f,J_0)+(\tilde\beta-\tilde\beta_0)'\nabla^2_{\tilde\beta,\tilde\beta}M_n(\tilde\beta_0,f,J_0)(\tilde\beta-\tilde\beta_0)+\mathrm{o}_P\bigl(\|\tilde\beta-\tilde\beta_0\|^2\bigr),
\end{aligned}
$$

and apply Theorem 1 and 2 of Sherman [27] to conclude. □

Proof of Proposition 3.6. The uniform convergence results in Assumptions 9 and 11 can be deduced from studying the uniform convergence rate of the numerator and the denominator in (3.4) (and their derivatives) separately. This is a consequence of Lemma A.1. Since the other terms can be studied in a similar way, we only consider the case of the denominator and its derivatives in (3.4). In each case, the bias part can be dealt with uniformly with classical kernel arguments, and is of order h^2. For the centered version of f^*, the result can be deduced from the study of the uniform convergence rate of empirical processes indexed by some class of functions as the one defined in (A.6), with

$$
\zeta_\beta(x,X,Z,T,\delta)=\frac{\delta(x-X)^j}{1-G_{\theta_0}(T-|Z)},
$$

where $j=0$ (resp., 1, 2) for f^* (resp., $\nabla_\beta f^*$, $\nabla^2_{\beta,\beta}f^*$), and $\xi(T)=T$. The kernel K^0 in (A.6) is either $\tilde K$ or $\tilde K'$ or $\tilde K''$, and $\psi(\beta,x)=\beta'x$. It follows from the conditions on $\tilde K$ and from Nolan and Pollard [24] that the class of functions

$$
\biggl\{x\to K^0\biggl(\frac{\beta'x-\beta'u}{h}\biggr):\ u\in\mathcal{X},h>0,\beta\in\mathcal{B}\biggr\}
$$

is a VC-class of functions. Moreover, $u \to (x-u)^j$ ($j = 0, 1, 2$) is also a VC-class of bounded functions using permanence properties of VC-classes, see Lemma 2.6.18 in Van der Vaart and Wellner [32]. Finally, since $1 - G_{\theta_0}(T - |Z)$ is bounded away from zero, (A.7) holds. Now applying Lemma A.1, we get

$$\sup_{\beta,x}\bigl|f^*(\beta'x;\beta) - f(\beta'x;\beta)\bigr| = \mathrm{O}_P\bigl((\log n)^{1/2}n^{-1/2}h^{-1/2} + h^2\bigr),$$

$$\sup_{\beta,x}\bigl|\nabla_\beta f^*(\beta'x;\beta) - \nabla_\beta f(\beta'x;\beta)\bigr| = \mathrm{O}_P\bigl((\log n)^{1/2}n^{-1/2}h^{-3/2} + h^2\bigr),$$

$$\sup_{\beta,x}\bigl|\nabla^2_{\beta,\beta} f^*(\beta'x;\beta) - \nabla^2_{\beta,\beta} f(\beta'x;\beta)\bigr| = \mathrm{O}_P\bigl((\log n)^{1/2}n^{-1/2}h^{-5/2} + h^2\bigr),$$

where h^2 comes from the bias term. Hence, Assumption 9 holds if $h \to 0$ and $nh^5(\log n)^{-1/2} \to \infty$. Assumption 11 holds if $(\log n)^{-1}n^{1/2}a_n^{1/2}h \to \infty$, and $nh^8 \to 0$.

The first part of Assumption 10 follows directly from the uniform convergence of f^*. Elementary algebra shows that the gradient of f^* can be written as

$$\nabla_\beta f^*(\beta_0'x;\beta_0) = xm_1^*(\beta_0'x) + m_2^*(\beta_0'x). \tag{A.15}$$

Using the same arguments as above, these two functions converge uniformly to $m_1(\beta_0'x)$ and $m_2(\beta_0'x)$, respectively, where $\nabla_\beta f(\beta_0'x;\beta_0) = xm_1(\beta_0'x) + m_2(\beta_0'x)$, see equation (A.11), and Assumption 10 follows. □

The proof of Proposition 3.7 is a direct consequence of Lemma A.2, equation (A.10), and the fact that

$$\sup_\beta\left[(nh)^{-1}\sum_{i=1}^n\bigl|\tilde K^{(j)}\bigr|\left(\frac{\beta'x - \beta'X_i}{h}\right)|T_i|^k\right] = \mathrm{O}_P(1),$$

and hence will be omitted.

Acknowledgements

The third author acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement No. 203650.

References

[1] Akritas, M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Statist.* **22** 1299–1327. MR1311977

[2] Akritas, M.G. and Van Keilegom, I. (2000). The least squares method in heteroscedastic censored regression models. In *Asymptotics in Statistics and Probability* (M.L. Puri, ed.) 379–391. Utrecht: VSP.
[3] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646
[4] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.
[5] Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136** 730–769. MR2181975
[6] 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
[7] Du, Y. and Akritas, M.G. (2002). Uniform strong representation of the conditional Kaplan–Meier process. *Math. Methods Statist.* **11** 152–182. MR1941314
[8] Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380–1403. MR2195639
[9] Fan, J. and Gijbels, I. (1994). Censored regression: Local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89** 560–570. MR1294083
[10] Gørgens, T. and Horowitz, J.L. (1999). Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *J. Econometrics* **90** 155–191. MR1703340
[11] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171
[12] Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488
[13] Heuchenne, C. and Van Keilegom, I. (2007). Polynomial regression with censored data based on preliminary nonparametric estimation. *Ann. Inst. Statist. Math.* **59** 273–297. MR2405280
[14] Horowitz, J.L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104
[15] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333
[16] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981
[17] Klein, R.W. and Spady, R.H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387–421. MR1209737
[18] Li, K.C., Wang, J.L. and Chen, C.H. (1999). Dimension reduction for censored regression data. *Ann. Statist.* **27** 1–23. MR1701098
[19] Lopez, O. (2007). Réduction de dimension en présence de données censurées. Ph.D. thesis, CREST-ENSAI. Available at http://tel.archives-ouvertes.fr/tel-00195261/fr.
[20] Lopez, O. (2009). Single-index regression models with right-censored responses. *J. Statist. Plann. Inference* **139** 1082–1097. MR2479851
[21] Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Comm. Statist. Theory Methods* **40** 2639–2660. MR2860770
[22] Lu, X. and Burke, M.D. (2005). Censored multiple regression by the method of average derivatives. *J. Multivariate Anal.* **95** 182–205. MR2164128
[23] Lu, X. and Cheng, T.L. (2007). Randomly censored partially linear single-index models. *J. Multivariate Anal.* **98** 1895–1922. MR2396946
[24] Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Statist.* **15** 780–799. MR0888439

[25] Powell, J.L., Stock, J.H. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117
[26] 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
[27] Sherman, R.P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Ann. Statist.* **22** 439–459. MR1272092
[28] Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45** 89–103. MR1222607
[29] Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.* **23** 461–471. MR1439707
[30] Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28–76. MR1258865
[31] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247
[32] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671
[33] Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.* **27** 1745–1784. MR1742508
[34] Wang, Y., He, S., Zhu, L. and Yuen, K.C. (2007). Asymptotics for a censored generalized linear model with unknown link function. *Probab. Theory Related Fields* **138** 235–267. MR2288070

Received February 2010 and revised July 2012