Guided Censored Regression

Majda Talamakrouni, Anouar El Ghouch and Ingrid Van Keilegom

Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain

ABSTRACT. Parametrically guided nonparametric regression is an appealing method that can reduce the bias of a nonparametric regression function estimator without increasing the variance. In this paper, we adapt this method to the censored data case using an unbiased transformation of the data and a local linear fit. The asymptotic properties of the proposed estimator are established and its performance is evaluated via finite sample simulations.

Key words: Least squares estimation, local linear regression, nonparametric regression, parametric regression, right censoring, synthetic data, U-statistics

1 Introduction

In the area of regression analysis, a lot of research has been carried out on completely parametric or completely nonparametric regression. Both approaches have, however, a number of important drawbacks in practice. Parametric models, have the advantage of being powerful and precise when the chosen model is the correct one, but can otherwise give a completely wrong picture of the underlying regression function. Nonparametric methods, on the other hand, have in general a slower rate of convergence, but need no explicit specification of the form of the regression function.

This motivates the consideration of an approach called parametrically guided nonparametric regression, that contains both a parametric and a nonparametric component, and is therefore a good compromise between the two extreme approaches described above. Unlike the classical semi-parametric models (additive model, partially linear model, etc) a guided nonparametric estimator is completely nonparametric in the sense that it does not rely on any assumed global structure. On the other hand, a guided nonparametric estimator takes advantage of both parametric and nonparametric methods: It will always converge to the true regression function no matter if the parametric part is correct or not, and it will adapt automatically to the parametric model if the latter is locally or globally close to the true underlying curve.

In the context of completely observed i.i.d. data, many techniques and papers are available in the literature. These include Glad (1998), Fan & Ullah (1999), Gozalo & Linton (2000), Mays *et al.* (2001), Rahman & Ullah (2002) and Martins-Filho *et al.* (2008).
In this paper we consider the additive approach as discussed in Martins-Filho *et al.* (2008). Suppose for the moment that we have completely observed data $(Y_i, X_i), i = 1, \ldots, n$, let $m(x) = E(Y|X = x)$ be the true regression function and let $m(x, \hat{\theta})$ be any parametric estimator of m, where $\hat{\theta}$ is an estimator of the least false value θ^* according to a certain distance measure between m and the parametric regression model $m(., \theta)$ (see Assumption 2, below). The basic idea is based on the identity

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

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

$$\hat{m}(x) = m(x, \hat{\theta}) + \hat{r}_{\hat{\theta}}(x). \tag{1}$$

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

A common problem in practice is the presence of censoring, i.e. a competing event C that makes the variable of interest Y unobservable. As for the non-censored case, there is a vast literature on parametric and nonparametric regression with censored data, see for example Lai & Ying (1991), Delecroix *et al.* (2008) and more recently Ding & Nan (2011) for the case of parametric regression, and Fan & Gijbels (1994) and El Ghouch & Van Keilegom (2008) for the case of nonparametric regression. Unfortunately, the parametrically guided nonparametric regression technique is not directly applicable to censored data. Indeed, some of the responses may not be observed due to censoring. In this framework, an important preoccupation is the adaptation and extension of the procedures existing for completely observed data. The aim of this paper is to propose a new estimator of the mean regression function, which is an extension of the guided nonparametric estimator to the situation where the response is randomly right censored. To deal with censoring we use a synthetic data approach. Thus, we first transform the observed data in an unbiased way before making any inference. Such a technique is largely used in the literature, see for example Koul *et*

al. (1981), among many others. We establish the asymptotic normality of the proposed estimator and illustrate its performance via finite sample simulations.

The paper is organized as follows. Section 2 explains in detail the proposed methodology and shows how the observed data can be transformed in an appropriate way. Section 3 provides some asymptotic results for the proposed method, while Section 4 illustrates the performance of the proposed estimator via a simulation study. Finally, some general conclusions are drawn in Section 5. All the proofs are given in the Appendix.

2 Model and estimation procedure

2.1 Definitions and notations

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

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

where $\mathbf{I}(\cdot)$ is the indicator function and C is the censoring variable, supposed to be independent of the response Y given the covariate X.

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

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

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

2.2 Parametrically guided nonparametric regression with censored data

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

using local linear regression techniques by $\hat{b}_0$, where $(\hat{b}_0, \hat{b}_1)$ minimize

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

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

$$\hat{m}(x) = \hat{b}_0 = \sum_{i=1}^{n} W_i(x,h) Z_i(x,\hat{\theta}), \tag{2}$$

where

$$W_i(x,h) = \frac{s_{n,2}(x) - (X_i - x)s_{n,1}(x)}{s_{n,2}(x)s_{n,0}(x) - s_{n,1}^2(x)} K\left(\frac{X_i - x}{h}\right), \tag{3}$$

and

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

Note that, since $\sum_{i=1}^{n} W_i(x,h) = 1$, the estimator defined in (2) has the form given in (1) with $r_\theta(x) = \sum_{i=1}^{n} W_i(x,h)(Y_i - m(X_i,\theta))$.

Since, in the case of censored data, the observed variable T does not have the same conditional expectation as the variable of interest Y, we will use some unbiased transformation based on (T, δ, X). This procedure is known in the literature as the synthetic data approach. Several transformations exist in the literature, see for instance Leurgans (1987) and Zheng (1987). In this paper, we will use the transformation proposed by Koul *et al.* (1981), which is given by

$$Y^* = \frac{\delta \phi(T)}{1 - G(T|X)}. \tag{4}$$

Since Y and C are independent given X we have

$$E(Y^*|X = x) = E(\phi(Y)|X = x). \tag{5}$$

Unfortunately, in real data analysis, $G(y|x)$ is typically unknown and so the transformation (4) can not be computed without estimating G. In the uncensored case, the estimation of the conditional distribution function has been considered by Stone (1977) among many others. In the censored data framework, $G(y|x)$ can be estimated using Beran's estimator, see Beran (1981), defined by

$$1 - \hat{G}(y|x) = \prod_{i=1}^{n} \left(1 - \frac{(1 - \delta_i)\mathbf{1}_{\{\phi(T_i) \le y\}} W_{0i}(x, h_0)}{\sum_{j=1}^{n} \mathbf{1}_{\{\phi(T_i) \le \phi(T_j)\}} W_{0j}(x, h_0)}\right), \tag{6}$$

for y less than $T_{(n)}$, the largest observation of the sample $(T_i)_{1 \le i \le n}$. Here, W_0's are

Nadaraya-Watson weights defined by

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

where K_0 is a kernel function and $0 < h_0 \equiv h_{0,n}$ is a bandwidth parameter. Note that Beran's estimator is a conditional version of the well known Kaplan-Meier estimator and reduces to the latter when all weights $W_{0i}(x, h_0)$ are equal to n^{-1}. The asymptotic properties of this estimator have been studied by Dabrowska (1987), Gonzalez-Manteiga & Cadarso-Suarez (1994), Van Keilegom & Veraverbeke (1997 b), among others. Plugging-in Beran's estimator $\hat{G}(y|x)$ in the transformation (4), provides an estimator of the unknown transformed response Y^*, that will be denoted by $\hat{Y}^*$. Starting from the transformed data $(\hat{Y}_i^*, X_i),\ i = 1, \ldots, n$, we now apply the guided local linear method (GLL) to estimate the function $m(x) = E(\phi(Y)|X = x)$. Define

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

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

$$\hat{m}_{\hat{G},\hat{\theta}}(x) = \hat{b}_0 = \sum_{i=1}^{n} W_i(x, h) Z_{\hat{G},i}(x, \hat{\theta}), \tag{7}$$

where $W_i(x, h), i = 1, \ldots, n$, are defined by expression (3).

Note that if $m(x, \hat{\theta})$ is linear in x, then $\hat{m}_{\hat{G},\hat{\theta}}(x)$ reduces to the classical local linear estimator, i.e. $\sum_{i=1}^{n} W_i(x, h)\hat{Y}_i^*$, which means that our new estimator $\hat{m}_{\hat{G},\hat{\theta}}(x)$ is a generalization of the classical one. In principle, the parametric guide can be obtained using any parametric technique adapted to censored data. Often, even a simple parametric guide can improve the regression estimator compared to the purely non parametric version. However, using a completely non correct parametric pilot will not improve or may even harm the quality of the nonparametric estimator. Without any prior knowledge about the structure under study, one can use any of the well known model selection criterions such as AIC adapted to censored data; see Collett (1994) and more recently Liang & Zou (2008). However, the problem of parametric model selection is not in the scope of this paper.

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

3 Asymptotic results

The presence of censorship adds considerable complexity to the method described in the previous sections. In fact, the quantities $Z_{\hat{G},i}(x,\hat{\theta})$ in expression (7) depend on Beran's estimator $\hat{G}(y|x)$ defined by (6), which is computed from the whole sample and this makes the estimators (7) more difficult to study than (2).

To establish the asymptotic normality of $\hat{m}_{\hat{G},\hat{\theta}}(x)$ we follow the traditional approach of breaking the problem into two parts. First, we establish in Theorem 1 the asymptotic normality of $\tilde{m}_{\hat{G}}(x)$, an estimator of $m(x)$ constructed using a given non random approximation $\tilde{m}(x)$. Then, in Theorem 2 we provide sufficient conditions for the asymptotic equivalence of $\hat{m}_{\hat{G},\theta^*}(x)$ and $\hat{m}_{\hat{G},\hat{\theta}}(x)$, where θ^* is a fixed parameter that will be defined later. Throughout, c will be a positive constant that may take different values. We now provide a list of regularity conditions that are required in these theorems.

Assumption 1.

(A.1) *X belongs to a compact subset $S \subset \Re$ with marginal density $f_X(.)$, which is continuously differentiable and $\inf_{x\in S} f_X(x) > 0$.*

(A.2) *1. ϕ is a bounded function that vanishes outside the interval $[0,\tau]$ for some $\tau < \inf_{x\in S}\tau_x$ with $\tau_x = \sup\{y : H(y|x) < 1\}$, i.e. the right endpoint of the support of $H(y|x) = P(T \le y|X = x) = 1 - (1 - F(y|x))(1 - G(y|x))$.*

2. The functions $H_j(y|x) = P(T \le y, \delta = j|X = x),\ j = 0, 1$, have four derivatives with respect to x. Furthermore, the derivatives are bounded uniformly for all $y \le \tau$ and $x \in S$.

(A.3) *1. $f_{Y|X}(y|x)$, the conditional density of Y given $X = x$, is differentiable in y for all $x \in S$.*

2. $f_{\varepsilon^*|X}(y|x)$, the density of the synthetic errors $\varepsilon^* = Y^* - m(X)$ given $X = x$, is uniformly bounded and continuous in y for all $x \in S$.

3. For all $x \in S$, $\sigma^2_*(x) = Var(\varepsilon^*|X = x) \neq 0$ and is continuous in x.

4. $E(\phi(Y)^2) < \infty$.

(A.4) *The kernel K is a symmetric, continuously differentiable probability density function with compact support S_K, and $\int x^2K(x)dx = \mu^2_K < \infty$. The kernel K_0 is a symmetric, continuously differentiable function of order four with compact support S_{K_0}, $\int K_0(x)dx = 1$ and $\int x^4K_0(x)dx = \mu^4_{K_0} < \infty$.*

(A.5) *$(\log n)^3/nh_0^3 = O(1)$, $nh_0^8 = O(1)$ and $nh^3 \to \infty$ as $n \to \infty$.*

(A.6) *$m(x)$ is twice continuously differentiable.*

3.1 The estimator with a fixed guide

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

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

where $\tilde{Z}_{\hat{G},i}(x) = \tilde{m}(x) + (\hat{Y}_i^* - \tilde{m}(X_i)),\ i = 1, \ldots,\ n$, and $W_i(x,h)$ are the local linear weight functions defined in (3). The following Theorem establishes the asymptotic normality of $\tilde{m}_{\hat{G}}(.)$.

Theorem 1. *Suppose that Assumption 1 holds. Then,*

$$(nh)^{1/2}\left(\tilde{m}_{\hat{G}}(x) - m(x) - \tilde{B}(x) + O_p\left(\frac{(\log n)^{1/2}}{(nh_0)^{1/2}}\right) + o_p(h^2)\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma_*^2(x)}{f_X(x)}\int K^2(u)du\right), \tag{8}$$

with

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

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

3.2 The estimator with an estimated guide

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

Assumption 2.

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

1. $\theta \in \Theta$, Θ a compact subset of $\Re^p$.

2. The function $(x,\theta) \mapsto m(x,\theta)$ is twice continuously differentiable with respect to x and θ.

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

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

Note that from equality (5) one can prove that $\theta^* = \arg\min_{\theta\in\Theta} M(\theta)$ with $M(\theta) = E\big(Y^* - m(X,\theta)\big)^2$.

Proposition 1. *Under Assumptions 1 and 2,*

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

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

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then,*

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

with

$$B(x,\theta^*) = \frac{1}{2}h^2\mu_K^2\big(m^{(2)}(x) - \ m^{(2)}(x,\theta^*)\big). \tag{9}$$

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

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

It is clear from expression (9) that the choice of the parametric guide $m(x,\theta)$ will have an impact on the bias of the GLL estimator. If $m(x,\theta) = \theta_1 + \theta_2 x$ is a linear guide, then $B(x,\theta^*)$ is the asymptotic bias of the local linear estimator, which means as said before that the GLL estimator is a generalization of the classical local linear estimator. On the other hand, if the parametric function $m(x,\theta)$ is twice differentiable with respect to x and $m^{(2)}(x) = m^{(2)}(x,\theta^*)$, then the asymptotic bias $B(x,\theta^*) = 0$, which means that one can increase the bandwidth to reduce the variance without increasing bias. Finally, if the second derivatives

of $m(x,\theta)$ and $m(x)$ are close to each other, then the bias will be reduced in absolute value compared to that of the local linear estimator when $|m^{(2)}(x)-m^{(2)}(x,\theta^*)| < |m^{(2)}(x)|$. Note that the expression of the asymptotic variance is equal to that of the local linear estimator, which means that with a suitable parametric guide one can reduce the bias of the local linear estimator without any increase in the variance, and thus the bias-variance problem can be bypassed. The condition $h_0(h\log n)^{-1} \to \infty$ means that the selected bandwidth for Beran's estimator must be asymptotically larger than the bandwidth of the GLL estimator. In this case ($h << h_0$) and under the condition $nh_0^8 = O(1)$, the asymptotic bias from the estimation of $G(y|x)$ will have no asymptotic effect on the GLL estimator. Finally, the condition $nh^5 = O(1)$ is exactly as in the classical nonparametric approach.

Remark 1. In practice, choosing the bandwidth parameters is crucial. To select h_0, the bandwidth used in the estimation of the conditional cumulative distribution $G(.|x)$, one can use, for example, the least squares cross validation method proposed by Dabrowska (1992) or the bootstrap method developed by Van Keilegom & Veraverbeke (1997 a). As for the regression's bandwidth h, which is more important in our setting, the theoretical optimal value that minimizes the asymptotic mean integrated squared error (see Theorem 3.2) is given by

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

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

Remark 2. With completely observed data Martins-Filho *et al.* (2008) proposed a class of guided local linear estimators that generalizes the estimator (1). Their approach can be easily extended to censored data. The starting point is the following identity :

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

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

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

where $\hat{r}_{\hat{\theta},\alpha}(x)$ is now a nonparametric fit based on the pseudo-estimated data $\big(X_i, (\hat{Y}_i^* - m(X_i,\hat{\theta}))/m(X_i,\hat{\theta})^{\alpha}\big)$. It is easy to verify that the case $\alpha = 0$ leads back to the estimator (1). The generalized parametrically guided local linear estimator is now given by

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

where $Z_{\hat{G},i}(x,\theta,\alpha) = m(x,\theta) + (\hat{Y}_i^* - m(X_i,\theta))\rho_{x,\theta}(X_i),\ i = 1,\ldots,n$, and
$\rho_{x,\theta}(v) = \big(m(x,\theta)/m(v,\theta)\big)^{\alpha}$.
All the results proved before can be extended to the generalized guided local linear estimator, the generalization of the proof is straightforward and is omitted here. The most important result is the following theorem, which is a generalization of Theorem 2 :

Theorem 3. *Suppose that Assumptions 1 and 2 hold and suppose that* $m(x,\theta^*) \neq 0$. *Then,*

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

with

$$B(x,\theta^*,\alpha) = \frac{1}{2}h^2\mu_K^2\big(\Gamma^{(2)}_{\theta^*,\alpha}(x) - \Lambda^{(2)}_{\theta^*,\alpha}(x)\big), \tag{10}$$

and

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

Theorem 3 shows that, in addition to $m(x,\theta^*)$ the asymptotic bias $B(x,\theta^*,\alpha)$ is now also depending on the parameter α. In practice, the parameter α needs also to be selected in order to minimize the bias, Martins-Filho *et al.* (2008) proposed some guidelines for choosing α. Their method can be extended to censored data, however, adding another tuning parameter may not be very practical in the present context. For simplicity, in the following simulations studies we will limit ourselves to the case where $\alpha = 0$.

4 Simulations

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

4.1 Exponential proportional hazards model

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

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

where $r(x) = \exp(-\theta x) + \lambda(\sin(2\pi x))^2$, X_i and U_i are independent uniform variables on $[0, 1]$, and $\theta = 2$. The bandwidths are selected over a grid on $[0.1, 0.8]$ with steps of length 0.1. The censoring variable C_i is independent of the response Y_i given X_i, and is defined by $C_i = -\log(V_i)\exp(\tilde{r}(X_i))$, where $\tilde{r}(x) = \log(r(x)) + \log(b^2)$ and V_i is a uniform variable on $[0, 1]$ independent of X_i and U_i. The parameter b allows to control the rate of censoring (RC) which is given by $RC(x) = P(Y_i > C_i|X_i = x) = (b^2 + 1)^{-1}$. The value $b = 2$ corresponds to a fixed censoring rate of 20% and $b = 1.22$ corresponds to a fixed censoring rate of 40%. Note that the model defined above is equivalent to the case where conditional on $X_i = x$, the variables Y_i and C_i are independent and exponentially distributed with rates $1/r(x)$ and $\exp(-\tilde{r}(x))$. We aim to compare the GLL estimator with the fully nonparametric estimator. For the parametric guide we consider an exponential proportional hazards model (PPH) $m(x, \theta) = \exp(-\theta x)$ and for the nonparametric one we use the local linear (LL) estimator. The parameter $\lambda \geq 0$ in the expression of $r(x)$ allows to control the deviation of the parametric guide from the true regression function. The case where $\lambda = 0$ corresponds to the situation where the parametric guide is correct, which is the ideal situation for the GLL estimator. Note that this case corresponds to the situation where the response $Y_i \sim \exp(1)$ is the survival time of an exponential proportional hazards model; see Bender *et al.* (2005). If $\lambda \neq 0$ then the parametric pilot is incorrect and deviates increasingly from the true model when the value of λ increases. We consider five values of λ : 0, 0.04, 0.2, 1, 2, the graphs of the truncated conditional mean are plotted in Figure 1 for each value of λ. We compare the performance of the estimators in the situations of

: correct, approximately correct and incorrect parametric pilots. We conducted $N = 200$ simulations for samples of size $n = 100$. For each model and for each set of data, we evaluate the different estimators at 30 equally spaced points, x_i, taken from 0 to 1. At every data point $x_i, i = 1, ..., 30$, we approximate the bias and the variance by $B(x_i) = N^{-1}\sum_{k=1}^{N}\left[\hat{m}^k_{\hat{G},\hat{\theta}}(x_i) - m(x_i)\right]$, $S^2(x_i) = N^{-1}\sum_{k=1}^{N}\left[\hat{m}^k_{\hat{G},\hat{\theta}}(x_i) - N^{-1}\sum_{k=1}^{N}\hat{m}^k_{\hat{G},\hat{\theta}}(x_i)\right]^2$, where $\hat{m}^k_{\hat{G},\hat{\theta}}(x_i)$ is the GLL estimator for the k^{th} replication. Let $Bias^2 = 30^{-1}\sum_{i=1}^{30}B^2(x_i)$, $Var = 30^{-1}\sum_{i=1}^{30}S^2(x_i)$ and $MSE = Bias^2 + Var$ be the average squared bias, the average variance and the average mean squared error of the estimates, respectively. The results are summarized in Table 1 and show that with a correct parametric (PPH) guide ($\lambda = 0$) we get the best results for the GLL estimator, the observed bias of GLL estimator is approximately 10^{-1} times that of the LL estimator. With roughly correct parametric guide ($\lambda = 0.04$), we observe that the GLL estimator remains as good as before and the observed bias is reduced (by a factor 10) compared to the LL estimator. For $\lambda = 0.2$ the bias is still reduced by a factor 2 compared to the LL estimator. In the two last cases the parametric guide is incorrect, the bias of the GLL estimator is very close to that of the LL estimator. Regarding the variance, as expected, the LL and GLL estimator behave similarly. The MSE is not obviously reduced, this is because in this case the variance dominates the bias. Finally, it seems that increasing the censoring rate does not harm the performance of the GLL estimator, since the estimated transformed data depend on Beran's estimator $\hat{G}$ of which the performance increases with the rate of censorship.

INSERT TABLE 1 HERE

INSERT FIGURE 1 HERE

4.2 Sinusoidal model

The data are generated from the following model:

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

where X_i is drawn from a uniform density on $[0, 1]$ and ε_i is normally distributed with mean 0 and variance 1. The bandwidths are selected over grid on $[0.2, 0.8]$ with steps of length 0.1. The censoring variable C_i is defined as $C_i|_{X_i=x} \sim N(\mu(x), 1)$ with $\mu(x) = \sin(2\pi x)+(a_1+a_2x)^2$. The variables X_i and ε_i are independent. As before, the parameters a_1 and a_2 control the rate of censoring (RC), which is given by $RC(x) = 1-\Phi((a_1+a_2x)^2/\sqrt{2})$, where Φ is the cumulative distribution function of the standard normal variable. We selected

$a_1 = 0.84$, $a_2 = 0$. This leads to a constant rate of 30% censoring. We present results from a sinusoidal regression curve with three various forms for the parametric guide. The first guide is a linear model $m(x, \theta) = \theta_0 + \theta_1 x$, the second guide is a cubic model $m(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$ and the last guide is guessed correctly and belongs to the true parametric family $m(x, \theta) = \theta \sin(2\pi x)$. The parameters $\hat{\theta}$ are obtained using least squares estimation. We carried out the simulations for various sample sizes (25, 50, 100, 200) and for 30 equidistant data points in $[0, 1]$. The results are obtained by using 200 simulations and are presented in Table 2. As expected in Section 3.2, with a linear guide, the GLL estimator (denoted L-GLL) behaves exactly as the LL estimator, and is therefore only calculated for $n = 25$. Regarding the bias, the GLL estimator with cubic (C-GLL) or sinus guide (S-GLL) outperforms the LL estimator even for small sample sizes. It is clear from the results that the bias is substantially reduced for all sample sizes. The variance decreases when the sample size increases, which confirms the asymptotic results of the previous section. The MSE is also reduced compared to the local linear estimator, except for the GLL estimator with cubic guide, which has a slightly larger variance especially with few data. Finally, we note that throughout the simulations the bias is generally smaller when the bandwidth of Beran's estimator is larger than the bandwidth used for the LL estimator or the GLL estimator ($h << h_0$), which is in concordance with the results of Theorem 2. Figure 2 illustrates this point. We also notice that the bandwidths h_0, used in Beran's estimator, and h, used in the local linear approximation, do not affect the results in the same manner. In fact, h seems to have a more significant effect on the MSE than h_0. Figure 3 illustrates this remark.

INSERT TABLE 2 HERE

INSERT FIGURE 2 HERE

INSERT FIGURE 3 HERE

5 Conclusion

In this work we have adapted the parametrically guided nonparametric regression to censored data. The new estimator is obtained by local linear smoothing based on the estimated transformed data. We have proved the asymptotic normality of the proposed estimator with $\sqrt{nh}$ rate of convergence. Under certain conditions, we found that the bias of the GLL estimator can be reduced compared to that of the LL estimator, while the variance remains unchanged. Simulations confirm the theoretical results and provide the following conclusions: as in the uncensored framework, the GLL estimator with censored data outperforms the local linear estimator if the parametric guide is equal or close to the true regression

curve and performs as the local linear estimator if the guide is linear or misspecified.

Acknowledgements

We thank the associate editor and the referees for their valuable comments. This research was supported by IAP research network grant nr. P07/06 of the Belgian government (Belgian Science Policy), and by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'. The research of the third author was also supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

References

Bender, R., Augustin T. & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **24**, 1713-1723.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report, Univ. California, Berkeley.*

Collett, D. (1998). *Modeling survival data in medical research.* Chapman & Hall, New York.

Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Stat.* **14**, 181-197.

Dabrowska, D. M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scand. J. Stat.* **19**, 351-361.

Delecroix, M., Lopez, O. & Patilea, V. (2008). Nonlinear censored regression using synthetic data. *Scand. J. Stat.* **35**, 248-265.

Ding, Y. & Nan, B. (2011). A sieve M-theorem for bundled parameters in semi-parametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.* **39**, 3032- 3061.

Du, Y. & Akritas, M. G. (2002). I.i.d. representations of the conditional Kaplan-Meier process for arbitrary distributions. *Math. Methods Statist.* **11**, 152-182.

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

El Ghouch, A. & Van Keilegom, I. (2008). Nonparametric regression with dependent censored data. *Scand. J. Stat.* **35**, 228-247.

Fan, J. & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann.*

Statist. **20**, 2008-2036.

Fan, J. & Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89**, 560-570.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall, London.

Fan, Y. & Ullah, A. (1999). Asymptotic normality of a combined regression estimator. *J. Multivariate Anal.* **71**, 191-240.

Glad, I. K. (1998). Parametrically guided nonparametric regression. *Scand. J. Stat.* **25**, 649-668.

Gonzalez-Manteiga, W. & Cadarso-Suarez, C. (1994). Asymptotic properties of generalized Kaplan-Meier estimator with some applications. *J. Nonparametr. Statist.* **4**, 65-78.

Gozalo, P. & Linton, O. (2000). Local nonlinear least squares: using parametric information in nonparametric regression. *J. Econometrics* **99**, 63-106.

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

Lai, T. L. & Ying, Z. (1991). Large-sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist.* **19**, 1370-1402.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* **74**, 301-309.

Liang, H. & Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Comput. Statist. Data Anal.* **52**, 2538-2548.

Lopez, O. & Patilea, V. (2009). Nonparametric lack-of-fit tests for parametric mean-regression models with censored data. *J. Multivariate. Anal.* **100**, 210-230.

Martins-Filho, C., Mishra, S. & Ullah, A. (2008). A class of improved parametrically guided nonparametric regression estimators. *Econometric Reviews* **27**, 542-573.

Martins-Filho, C. & Yao, F. (2006). A note on the use of V and U statistics in nonparametric models of regression. *Ann. Inst. Statist. Math.* **58**, 389-406.

Mays, J. E., Birch, J. B. & Starnes, B. A. (2001). Model robust regression: combining parametric, nonparametric and semiparametric methods. *J. Nonparametr. Stat.* **13**, 245-277.

Newey, W. K. & McFadden, D. (1999). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*. Vol. **4**. Edited by D. McFadden & R. Engle, Amsterdam, The Netherlands.

Rahman, M. & Ullah, A. (2002). Improved combined parametric and nonparametric regression: estimation and hypothesis testing. In *Handbook of Applied Econometrics and Statistical Inference*, Edited by Ullah, A., Wan, A., Chaturvedi, A. Marcel Dekker, New

York.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595-645.

Van Keilegom, I. & Akritas, M.G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.* **27**, 1745-1784.

Van Keilegom, I. & Veraverbeke, N. (1997 a). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Statist. Math.* **49**, 467-491.

Van Keilegom, I. & Veraverbeke, N. (1997 b). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles. *Comm. Statist. Theory Methods* **25**, 2251-2265.

Yao, F. & Martins-Filho, C. (2011). An asymptotic characterization of finite order U-statistics with sample size dependent kernels: applications to nonparametric estimators and test statistics. *Comm. Statist. Theory Methods* (Forthcoming).

Zheng, Z. (1987). A class of estimators of the parameters in linear regression with censored data. *Acta Mathematicae Applicatae Sinica* **3**, 231-241.

Corresponding author

I. Van Keilegom, Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium (ingrid.vankeilegom@uclouvain.be).

Appendix: Proofs

Proof of Theorem 1. We start with the case when G is known. The guided local linear estimator based on the non random parametric model $\tilde{m}(x)$ and transformation (4) is $\tilde{m}_G(x) = \sum_{i=1}^n W_i(x,h)\tilde{Z}_{G,i}(x)$, where $\tilde{Z}_{G,i}(x) = \tilde{m}(x) + (Y_i^* - \tilde{m}(X_i)),\ i = 1, \ldots, n$. We have the following result.

Lemma 1. *Suppose that the assumptions of Theorem 1 hold. Then,*

$$\sqrt{nh}\Big(\tilde{m}_G(x) - m(x) - \tilde{B}(x) + o_p(h^2)\Big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\sigma_*^2(x)}{f_X(x)}\int K^2(u)du\Big), \tag{11}$$

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

Proof. Write

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

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

$$\sqrt{nh}\left(I_{1,n}(x) - \frac{h^2}{2}m^{(2)}(x)\mu_K^2 + o(h^2)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_*^2(x)}{f_X(x)}\int K^2(u)du\right). \tag{12}$$

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

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

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

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

Hence, $I_{2,n}(x) = \frac{h^2}{2}\tilde{m}^{(2)}(x)\mu_K^2 + o_p(h^2)$. The result of Lemma 1 now follows from this last expression and (12). □

Now we turn to the case when G is unknown.

Lemma 2. *Suppose that the assumptions of Theorem 1 hold. Then,*

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

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

$$\max_i |\hat{Y}_i^* - Y_i^*| \le O_p(1) \times \sup_{t\le\tau, x\in S}|\hat{G}(t^-|x) - G(t|x)|. \tag{13}$$

The result of Lemma 2 now follows from (13). □

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

$$\sup_{t\le\tau, x\in S}|\hat{G}(t^-|x) - G(t|x)| = O_p((nh_0)^{-1/2}(\log n)^{1/2}), \tag{14}$$

(note that their condition $nh_0^5(\log n)^{-1} = O(1)$ is replaced by $nh_0^9(\log n)^{-1} = O(1)$ since we

work with a kernel of order 4), and from Lemma 2 we have,

$$\sqrt{nh}\big(\tilde{m}_{\hat{G}}(x) - \tilde{m}_G(x)\big) = O_p\bigg(\big(\frac{h\log n}{h_0}\big)^{1/2}\bigg). \tag{15}$$

The result of Theorem 1 is now a direct consequence of Lemma 6.1 and equation (15). □

For the proof of Proposition 1, the following Lemmas are needed.

Lemma 3. *Under the assumptions of Proposition 1, we have,*

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

Proof. Define the following functions:
$M(\theta) = E\big(Y^* - m(X,\theta)\big)^2$, $M_n(\theta) = n^{-1}\sum_{i=1}^n \big(Y_i^* - m(X_i,\theta)\big)^2$ and $\hat{M}_n(\theta) = n^{-1}\sum_{i=1}^n \big(\hat{Y}_i^* - m(X_i,\theta)\big)^2$. To prove the result of Lemma 3 we have to show that $\sup_{\theta\in\Theta}|\hat{M}_n(\theta) - M(\theta)| = o_p(1)$. Note that $|\hat{M}_n(\theta) - M(\theta)| \le |\hat{M}_n(\theta) - M_n(\theta)| + |M_n(\theta) - M(\theta)|$. First consider

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

Then,

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

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

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

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

Now, under assumptions (A.3) and (B.1) and by Lemma 2.4 in Newey & McFadden

(1994), we have $\sup_{\theta\in\Theta}|M_n(\theta)- M(\theta)| = o_p(1)$. Then the result of Lemma 3 is a consequence of Theorem 2.1 in Newey & McFadden (1994). □

Define the following functions:

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

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

Lemma 4. *Under the assumptions of Proposition 1, we have,* (1). $\sup_{\theta\in\Theta}\|\hat{H}_n(\theta)- H_n(\theta)\|_2 = o_p(1)$, (2). $\sup_{\theta\in\Theta}\|H_n(\theta) - H(\theta)\|_2 = o_p(1)$, (3). $\hat{D}_n(\theta^*) - D_n(\theta^*) = O_p(n^{-1/2})$, *and* (4). $D_n(\theta^*) = O_p(n^{-1/2})$.

Proof.

1. We have

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

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

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

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

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

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

$$\hat{G}(t^-|x) - G(t|x) = \sum_{j=1}^{n} W_{0j}(x,h_0)\eta_j(t^-,x) + O_p\left(\left(\frac{\log n}{nh_0}\right)^{3/4}\right), \tag{16}$$

where

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

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

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

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

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

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

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

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

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

We start with U_{11n}. We have

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

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

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

where $\tilde{U}_{12n} = \frac{2}{n(n-1)}\sum_{i<j}\psi_n(Z_i,Z_j)$ is a second order U-statistic. By Theorem 1 in Yao & Martins-Filho (2011),

$$\tilde{U}_{12n} = \theta_n + O_p\big((n^{-1}\sigma_{1n}^2)^{1/2}\big) + O_p\big((n^{-2}\sigma_{2n}^2)^{1/2}\big), \tag{17}$$

where $\theta_n = E(\psi_n(Z_i,Z_j))$, $\sigma_{1n}^2 = Var\Big[E\big(\psi_n(Z_i,Z_j)|Z_i\big)\Big]$ and $\sigma_{2n}^2 = Var\big[\psi_n(Z_i,Z_j)\big]$. Next, write

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

Now consider

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

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

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

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

$$\sigma_{1n}^2 \le 2\Big\{E\big[E\big(h_{ij}|Z_i\big)^2\big] + E\big[E\big(h_{ji}|Z_i\big)^2\big]\Big\},$$

and

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

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

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

From these two equations, it follows that

$$\sigma_{1n}^2 = O(1). \tag{18}$$

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

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

and

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

Hence,

$$\sigma_{2n}^2 = O(h_0^{-1}). \tag{19}$$

From equations (17), (18) and (19) we finally get $U_{1n} = O_p(n^{-1/2})$ under the conditions $nh_0^2 \to \infty$ and $nh_0^8 = O(1)$. Similarly, we can prove that the second term $U_{2n} = O_p(n^{-1/2})$ using a third order U-statistic and the condition $(\log n)^3/nh_0^3 = O(1)$; see Martins-Filho & Yao (2006) and Yao & Martins-Filho (2011).

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

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

4. This is obvious, since

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

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

Proof of Proposition 1. This proposition is an immediate consequence of Lemmas 3 and 4, and of Theroem 3.1 in Newey & McFadden (1999). □

Proof of Theorem 2. Consider the following decomposition:

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

where $\tilde{m}_{\hat{G},\theta^*}(x)$ is the guided local linear estimator of $m(x)$ using the nonrandom function $\tilde{m}(x, \theta^*)$. The second term in the sum is asymptotically normal by Theorem 1. To prove the result of Theorem 2 it is sufficient to establish the following:

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

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

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

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

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

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

Fig 1: Truncated conditional mean for different values of λ.

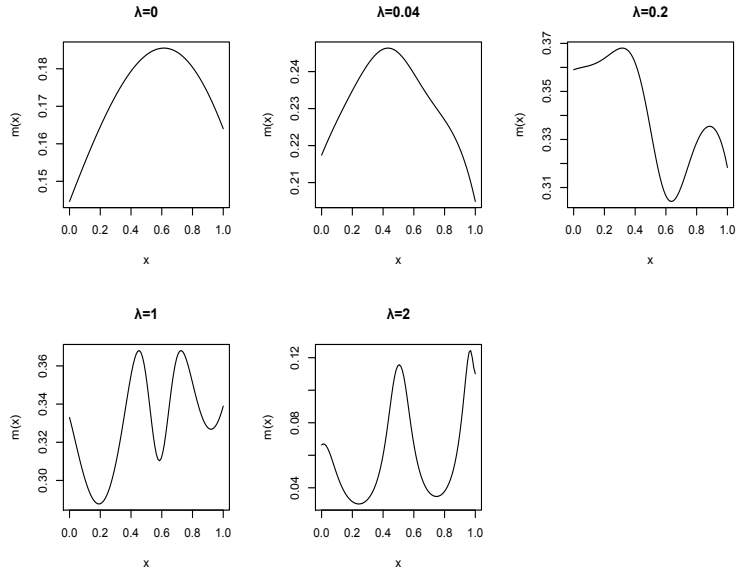

Table 2: Average squared bias (Bias^2 $\times 10^2$), average variance (Var $\times 10^2$) and average MSE ($\times 10^2$) for samples of size $n = 25, 50, 100, 200$, with a censoring rate of 30%, and $N = 200$ replications.

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

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

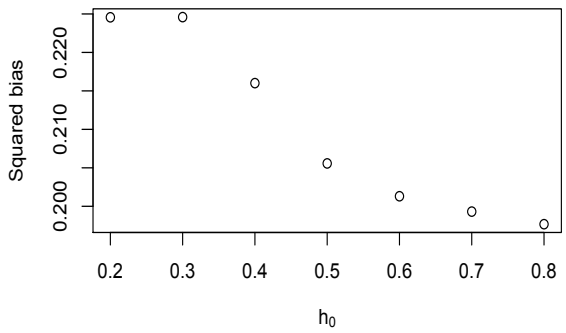

Fig 3: MSE ($\times 10^2$) of the S-GLL estimator, for $n = 200$, $B = 200$, h varies from 0.2 to 0.8 and two values of h_0: $h_0 = 0.3$ on the left and $h_0 = 0.7$ on the right.