Universidade de Vigo

Kernel density estimation with doubly truncated data

Carla Moreira and Jacobo de Uña Álvarez

Report 11/02

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://eioweb.uvigo.es/
E-mail: depc05@uvigo.es

Universidade de Vigo

Kernel density estimation with doubly truncated data

Carla Moreira and Jacobo de Uña Álvarez

Report 11/02

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: 

Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Kernel density estimation with doubly truncated data

C. Moreira
carla@uvigo.es
University of Vigo
Department of Statistics and O.R.
Lagoas - Marcosende, Vigo, 36 310 , Spain

J. de Uña-Álvarez
jacobo@uvigo.es
University of Vigo
Department of Statistics and O.R.
Lagoas - Marcosende, Vigo, 36 310 , Spain

Abstract

In some applications with astronomical and survival data, doubly truncated data are sometimes encountered. In this work we introduce kernel-type density estimation for a random variable which is sampled under random double truncation. Two different estimators are considered. As usual, the estimators are defined as a convolution between a kernel function and an estimator of the cumulative distribution function, which may be the NPMLE (Efron and Petrosian, 1999) or a semiparametric estimator (Moreira and de Uña-Álvarez, 2010b). Asymptotic properties of the introduced estimators are explored. Their finite sample behaviour is investigated through simulations. Real data illustration is included.

1 Introduction

Truncated data play an important role in the statistical analysis of survival times as well as in other fields like astronomy or economy. Double truncation of survival data occurs e.g. when only those individuals whose event time lies within a certain subject-specific observational window are observed. An individual whose event time is not in this interval is not observed and no information on this subject is available to the investigator. Because we are only aware of individuals with event times in the observational window, the inference with truncated data is based on sampling information from a conditional distribution. Hence, suitable corrections to account for the observational bias are needed. This problem goes back to Turnbull (1976).

Among the various existing problems of random truncation, literature has mainly been focused on the left-truncation model or, more generally, in one-sided truncation setups. Woodroofe (1985) investigated the properties of the nonparametric maximum-likelihood estimator (NPMLE) of the distribution function (df) with left-truncated data, see also Lynden-Bell (1971). This estimator was further investigated by Stute (1993), being also extended to the right-censored scenario (see Tsai et al. (1987), Wang (1991) or Zhou and Yip (1999), among many others). However, literature on random double truncation is much scarcer. A possible reason is the absence of closed form estimators; indeed, the existing methods for doubly truncated data are iterative and computationally intensive, and these issues make difficult both the theoretical developments and the practical implementations.

Efron and Petrosian (1999) introduced the NPMLE of the df under double truncation, while Shen (2010a) formally established the uniform strong consistency and the weak convergence of the NPMLE. Bootstrap methods to approximate the finite sample distribution of the NPMLE with doubly truncated data were explored in Moreira and de Uña-Álvarez (2010a). The semiparametric approach, in which the distribution of the truncation times is assumed to belong

to a given parametric family, was investigated in Moreira and de Uña-Álvarez (2010b), see also Shen (2010b). Interestingly, these authors showed that the semiparametric estimator may outperform the NPMLE in the sense of the mean squared error (MSE). An `R` package to compute the NPMLE of a doubly truncated df was presented in Moreira et al. (2010). However, for the best of our knowledge, estimation of a density function observed under random double truncation has not been considered so far.

The rest of the paper is organized as follows. In Section 2 two new estimators of a doubly truncated density function are introduced, and their main asymptotic properties are discussed. As usual in kernel smoothing, these estimators are obtained as a convolution between a kernel function and an appropriate estimator of the cumulative df. The first estimator is purely nonparametric, since it is based on the Efron and Petrosian (1999)'s NPMLE; while the second estimator is semiparametric, being constructed from the semiparametric cumulative df proposed by Moreira and de Uña-Álvarez (2010b). Section 3 provides a simulation study in which the finite-sample properties of the two estimators are investigated. In particular, we explore in much detail the role of the smoothing parameter or bandwidth. Both estimators are critically compared in the sense of the integrated MSE. In Section 4 we give a real data illustration of the proposed methods. To this end, we use data on childhood cancer from Northern region of Portugal (Moreira and de Uña-Álvarez, 2010a). Main conclusions and a final discussion is given in Section 5.

2 The estimators. Asymptotic properties

Let X^* be the random variable of ultimate interest, with df F, and assume that it is doubly truncated by the random pair (U^*, V^*) with joint df T, where U^* and V^* ($U^* \leq V^*$) are the left and right truncation variables respectively. This means that the triplet (U^*, X^*, V^*) is observed if and only if $U^* \leq X^* \leq V^*$, while no information is available when $X^* < U^*$ or $X^* > V^*$. Let (U_i, X_i, V_i), $i = 1, ..., n$, denote the sampling information, these are iid data with the same distribution of (U^*, X^*, V^*) given $U^* \leq X^* \leq V^*$. Introduce $\alpha = P(U^* \leq X^* \leq V^*)$, the probability of no-truncation. For any df W denote the left and right endpoints of its support by $a_W = inf\{t : W(t) > 0\}$ and $b_W = inf\{t : W(t) = 1\}$, respectively. Let $T_1(u) = T(u, \infty)$ and $T_2(v) = T(-\infty, v)$ be the marginal df's of U^* and V^*, respectively. When $a_{T_1} \leq a_F \leq a_{T_2}$ and $b_{T_1} \leq b_F \leq b_{T_2}$, F and T are both identifiable (see Woodroofe, 1985).

In the following two Subsections we introduce respectively the NPMLE and the semiparametric estimator of the df of X^*. Then, in Subsection 2.3 we consider the problem of estimating the density function on the basis of these two cumulative estimators.

2.1 The NPMLE of the cumulative df

Here, we assume without loss of generality, that the NPMLE is a discrete distribution supported by the set of observed data (Turnbull, 1976). Let $\varphi = (\varphi_1, \ldots, \varphi_n)$ be a distribution putting probability φ_i on $X_i, i = 1, \ldots, n$. Similarly, let $\psi = (\psi_1, \ldots, \psi_n)$ be a distribution putting joint probability ψ_i on $(U_i, V_i), i = 1, \ldots, n$. Under the assumption of independence between X^* and (U^*, V^*), the full likelihood, $\mathcal{L}(\varphi, \psi)$, can be decomposed as a product of the conditional likelihood of the X_i's given the (U_i, V_i)'s, say $\mathcal{L}_1(\varphi)$, and the marginal likelihood of the (U_i, V_i)'s, say $\mathcal{L}_2(\varphi, \psi)$:

$$\mathcal{L}(\varphi, \psi) = \prod_{j=1}^{n} \frac{\varphi_j}{\Phi_j} \times \prod_{j=1}^{n} \frac{\Phi_j \psi_j}{\sum_{i=1}^{n} \Phi_i \psi_i} = \mathcal{L}_1(\varphi) \times \mathcal{L}_2(\varphi, \psi) \tag{1}$$

where Φ_i is defined through $\Phi_i = \sum_{m=1}^{n} \varphi_m J_{im}$, $i = 1, \ldots, n$ with $J_{im} = I_{[U_i \leq X_m \leq V_i]} = 1$ if $U_i \leq X_m \leq V_i$ and equal to zero otherwise.

The conditional NPMLE of F (Efron and Petrosian, 1999) is defined as the maximizer of $\mathcal{L}_1(\varphi)$ in equation maximizes indeed the full likelihood, which can be also written as the product

$$\mathcal{L}(\varphi, \psi) = \prod_{j=1}^{n} \frac{\psi_j}{\Psi_j} \times \prod_{j=1}^{n} \frac{\Psi_j \varphi_j}{\sum_{i=1}^{n} \Psi_i \varphi_i} = \mathcal{L}_1^*(\psi) \times \mathcal{L}_2^*(\psi, \varphi)$$

where $\Psi_i = \sum_{m=1}^{n} \psi_m I_{[U_m \leq X_i \leq V_m]} = \sum_{m=1}^{n} \psi_m J_{mi}$, for $i = 1 \ldots, n$. Here, $\mathcal{L}_1^*(\psi)$ denotes the conditional likelihood of the (U_i, V_i)'s given the X_i's and $\mathcal{L}_2^*(\psi, \varphi)$ refers to the marginal likelihood of the X_i's. Introduce $\hat{\psi} = (\hat{\psi}_1, ..., \hat{\psi}_n)$ as the maximizer of $\mathcal{L}_1(\psi)$; then, $T_n(u, v) = \sum_{i=1}^{n} \hat{\psi}_i I_{[U_i \leq u, V_i \leq v]}$ is the NPMLE of T (Shen, 2010a).

The NPMLE of F also admits the representation

$$F_n(x) = \alpha_n \int_{a_F}^{x} \frac{F_n^*(dt)}{G_n(t)}$$

where F_n^* is the ordinary empirical df of the X_i's,

$$G_n(t) = \int_{\{u \leq t \leq v\}} T_n(du, dv)$$

is a nonparametric estimator for the conditional probability of sampling a lifetime $X^* = t$, i.e. $G(t) = P(U^* \leq t \leq V^*)$, and $\alpha_n = (\int_{a_F}^{\infty} G_n^{-1}(t) F_n^*(dt))^{-1}$ is an estimator for α. Shen (2010a) established the uniform strong consistency and the weak convergence of F_n.

2.2 The semiparametric estimator of the cumulative df

In the semiparametric approach it is assumed that T belongs to a parametric family of df's $\{T_\theta\}_{\theta \in \Theta}$, where θ is a vector of parameters and Θ stands for the parametric space. As a consequence, $G(t)$ is parametrized as

$$G_\theta(t) = \int_{\{u \leq t \leq v\}} T_\theta(du, dv).$$

The parameter θ is estimated by the maximizer $\hat{\theta}$ of the conditional likelihood of the (U_i, V_i)'s given the X_i's, that is,

$$\mathcal{L}_1^*(\psi) \equiv \mathcal{L}_1^*(\theta) = \prod_{i=1}^{n} \frac{g_\theta(U_i, V_i)}{G_\theta(X_i)}$$

where $g_\theta(u, v) = \frac{\partial^2}{\partial u \partial v} P(U^* \leq u, V^* \leq v) = K_\theta(du, dv)$ stands for the joint density of (U^*, V^*) (assumed to exist).

Once θ is estimated, a semiparametric estimator for F is introduced through

$$F_{\widehat{\theta}}(x) = \alpha_{\widehat{\theta}} \int_{a_F}^{x} \frac{F_n^*(dt)}{G_{\widehat{\theta}}(t)}, \tag{2}$$

where $\alpha_{\widehat{\theta}} = (\int_{a_F}^{\infty} G_{\widehat{\theta}}^{-1}(t) F_n^*(dt))^{-1}$. Moreira and de Uña-Álvarez (2010b) established the asymptotic normality of both $\hat{\theta}$ and $F_{\widehat{\theta}}$. They also showed by simulations that $F_{\widehat{\theta}}$ may perform much more efficiently than the NPMLE. As a drawback, the semiparametric estimator requires preliminary specification of a parametric family, which may eventually introduce a bias component when it is far away from reality (Moreira and de Uña-Álvarez (2010b)).

2.3 The density estimators

Introduce

$$f_h(x) = \int K_h(x-t) F_n(dt) = \alpha_n \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) G_n(X_i)^{-1} \tag{3}$$

where $K_h(t) = K(t/h)/h$ is the re-scaled kernel function and $h = h_n$ is a deterministic bandwidth sequence with $h_n \to 0$. Note that (3) is a purely nonparametric estimator of f, the density of X^*(assumed to exist). Introduce also the semiparametric kernel density estimator

$$f_{\widehat{\theta},h}(x) = \int K_h(x-t) F_{\widehat{\theta}}(dt) = \alpha_{\widehat{\theta}} \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) G_{\widehat{\theta}}(X_i)^{-1}. \tag{4}$$

Note that both estimators (3) and (4) correct the double truncation by downweighting the X_i's according to an estimation of the sampling probability $G(X_i)$. This is very intuitive, since the values with less probability of being observed are receiving more mass. The case $G(.) = 1$ is possible; for example, this happens whenever the left-truncation time U^* is uniformly distributed in a suitable interval and $V^* - U^*$ is degenerated. See our real data illustration. In such a case, the correction for truncation vanishes and we obtain the usual kernel density estimators.

Both G_n and $G_{\widehat{\theta}}$ are $\sqrt{n}$-consistent estimators of G. For $G_{\widehat{\theta}}$ this follows from the $\sqrt{n}$-consistency of $\widehat{\theta}$, provided that G_θ is a smooth function of θ (Moreira and de Uña-Álvarez (2010b)). For G_n, the result may be obtained by noting that

$$G_n(x) = \alpha_n^{-1} \int \int_{\{u \le x \le v\}} \frac{T_n^*(du, dv)}{\int_{\{u \le t \le v\}} F_n(dt)} \quad \textit{and} \quad \alpha_n = \int \int \frac{T_n^*(du, dv)}{\int_{\{u \le t \le v\}} F_n(dt)},$$

where T_n^* is the ordinary empirical df of the truncation times. Hence, $\sqrt{n}$-consistency of G_n is a consequence of that of F_n (Shen, 2010a) and T_n^*. Since both G_n and $G_{\widehat{\theta}}$ approach to G at a $\sqrt{n}$-rate, which is faster than the nonparametric rate $\sqrt{nh}$, the asymptotic properties of f_h and $f_{\widehat{\theta},h}$ will be the same, and will coincide with those of the estimator based on the true G. However, for the finite sample case, some error improvements are expected when using $f_{\widehat{\theta},h}$ due to the smaller variance associated to $G_{\widehat{\theta}}$. This issue is illustrated in our simulations section.

Introduce the asymptotically equivalent version of f_h and $f_{\widehat{\theta},h}$ through

$$\overline{f}_h(x) = \int K_h(x-t) \overline{F}_n(dt) = \alpha \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) G(X_i)^{-1} \tag{5}$$

where

$$\overline{F}_n(x) = \alpha \frac{1}{n} \sum_{i=1}^{n} G(X_i)^{-1} I_{[X_i \le x]}.$$

In the next result we establish the strong consistency and the asymptotic normality of $\overline{f}_h(x)$. We implicitly assume $G(x) > 0$ throughout this Section.

Theorem 1. (i) If K is bounded on a compact support, h is such that $\sum_{n=1}^{\infty} \exp(-\eta h n) < \infty$ for each $\eta > 0$, G is continuous at x, and x is a Lebesgue point of f, then $\overline{f}_h(x) \to f(x)$ with probability 1.

(ii) If, in addition to the conditions in (i), K is an even function, $h = o(n^{-1/5})$, $G^{-1}f$ has a second derivative which is bounded in a neighbourhood of x, and $f(x) > 0$, then

$$(nh)^{1/2}\left(\overline{f}_h(x) - f(x)\right) \to N(0, \alpha G(x)^{-1} f(x) R(K))$$

in distribution, where $R(K) = \int K(t)^2 dt$.

Proof. For (i) introduce $\widetilde{f}_h(x) = \alpha G(x)^{-1} f_{0,h}(x)$ where

$$f_{0,h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i).$$

By Devroye and Wagner (1979) we have $f_{0,h}(x) \to f_0(x)$ almost surely, where $f_0(x) = \alpha^{-1} G(x) f(x)$. Now, if the support of K is contained in $[-a, a]$,

$$\left|\overline{f}_h(x) - \widetilde{f}_h(x)\right| \leq \alpha f_{0,h}(x) \sup_{x-ah \leq y \leq x+ah} \left|G(y)^{-1} - G(x)^{-1}\right|,$$

and the supremum goes to zero by the continuity of G at x. This ends with the proof to (i). Statement (ii) is proved similarly to Section 2 of Parzen (1962); by following such lines we obtain

$$(nh)^{1/2}\left(\overline{f}_h(x) - E\overline{f}_h(x)\right) \to N(0, \alpha G(x)^{-1} f(x) R(K)).$$

Now, a two-term Taylor expansion (and the fact that K is even) gives $E\overline{f}_h(x) = f(x) + O(h^2)$. Since $nh^5 \to 0$, this implies the claimed result.■

The asymptotic mean and variance of (5) are given in the following result. We refer to the following standard regularity assumptions.

(A1) The kernel function K is a density function with $\int tK(t)dt = 0$, $\mu_2(K) = \int t^2 K(t)dt < \infty$, and $R(K) = \int K(t)^2 dt < \infty$.

(A2) The sequence of bandwidths $h = h_n$ satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

(A3) The functions f and $G^{-1}f$ are twice continuously differentiable around x.

Theorem 2. Under (A1)-(A3) we have, as $n \to \infty$,

$$E\left[\overline{f}_h(x)\right] = f(x) + \frac{1}{2}h^2 f''(x)\mu_2(K) + o(h^2),$$

$$Var\left[\overline{f}_h(x)\right] = (nh)^{-1}\alpha G(x)^{-1} f(x) R(K) + o((nh)^{-1}).$$

Proof. The proof follows standard steps. A second-order Taylor expansion of f around x is used, and the assumptions on the kernel and the bandwidth are enough to conclude. See e.g. Wand and Jones (1995).■

Theorem 2 shows that the double truncation influences the variance of $\overline{f}_h(x)$, the bias being unaffected otherwise. More specifically, the variance of the estimator is large at points x for which the relative probability of getting X_i values around x (i.e. $G(x)$) is small. Usually one will be interested in the global error of $\overline{f}_h$ as an estimator of the entire curve f. This can be measured through the integrated MSE, namely

$$MISE(\overline{f}_h) = \int MSE(\overline{f}_h(x))dx,$$

where

$$MSE(\overline{f}_h(x)) = \left[E\overline{f}_h(x) - f(x)\right]^2 + Var\left(\overline{f}_h(x)\right).$$

Under regularity, we have from the previous results the following asymptotic expression for the $MISE(\overline{f}_h)$:

$$AMISE(\overline{f}_h) = \frac{1}{4}h^4 R\left(f''\right)\mu_2(K)^2 + (nh)^{-1}\alpha R(K)\int G^{-1}f$$

where $R\left(f''\right) = \int \left(f''\right)^2$. Because of the $\sqrt{nh}$-equivalence between G_n,$G_{\widehat{\theta}}$ and G, the same asymptotic expression will hold for f_h and $f_{\widehat{\theta},h}$ under proper conditions.

Interestingly, Hölder's inequality gives $\alpha\int G^{-1}f \geq 1$, which indicates that the global error when estimating the density in the doubly truncated scenario is at least as large as that pertaining to the no truncated situation. This does not mean that for a particular x the MSE of $\overline{f}_h(x)$ may not be smaller than in the i.i.d. situation, since $\alpha G(x)^{-1}f(x) < 1$ may happen. Minimization of $AMISE(\overline{f}_h)$ w.r.t. h leads to the asymptotically optimal bandwidth

$$h_{AMISE} = \left[\frac{\alpha R(K)\int G^{-1}f}{R\left(f''\right)\mu_2(K)^2}\right]^{1/5} n^{-1/5}.$$

Of course, this expression depends on unknown quantities that must be estimated in practice. There exist several criteria to select the bandwidth from the data at hand. Although in this paper we do not propose any particular automatic bandwidth selector, in Section 3 we investigate through simulations the impact of the smoothing parameter in the performance of the two introduced density estimators f_h and $f_{\widehat{\theta},h}$.

3 Simulation study

In this section we illustrate the finite sample behavior of both estimators, the purely nonparametric estimator and the semiparametric estimator, through simulation studies. We analyze the influence of the bandwidth in the estimators' mean integrated squared errors (MISEs), and we measure the amount of efficiency which is gained through the using of the semiparametric information.

We consider two different situations of double truncation, Case 1 and Case 2. In Case 1, U^*, V^* and X^* are mutually independent. In Case 2, we simulate U^* and then we take $V^* = U^* + \tau$ for some fixed constant $\tau > 0$. Case 2 follows the spirit of the childhood cancer data discussed in Section 4, when the recruited observations are those with terminating events (cancer diagnosis) falling between two specific dates. Two different models are simulated for each of the Cases 1 and 2. For Case 1, we take $U^* \sim U(0,1)$, $V^* \sim U(0,1)$, $X^* \sim U(0.25,1)$ (Model 1.1) and $U^* \sim U(0,1)$, $V^* \sim U(0,1)$, $X^* \sim 0.75N(0.5,0.15) + 0.25$ (Model 1.2). For Case 2, we take $\tau = 0.25$ and $U^* \sim U(0,1)$, $X^* \sim 0.75Beta(3/4,1) + 0.25$ (Model 2.1) and $U^* \sim U(0,1)$, $X^* \sim 0.75N(0.5,0.15) + 0.25$ (Model 2.2). Note that when we move from Model 1.1 (resp. 2.1) to Model 1.2 (resp. 2.2) we are changing the lifetime distribution while fixing the distribution of the truncation variables; while when we move from Model 1.2 to Model 2.2 we are maintaining the same lifetime distribution but we change the truncation distribution. This will be interesting when interpreting the simulation results. We also point out that, due to the random truncation, in Models 1.1 and 1.2 relatively small and moderate values of the lifetime are more probably observed, while in Models 2.1 and 2.2 there is no observational bias on X^* (i.e. $G(.) = 1$; see Remark 2.1 in Moreira and de Uña-Álvarez (2010b)). We will recall this issue below.

For the computation of the semiparametric density estimator, as parametric information on (U^*, V^*) we always consider a $Beta(\theta_1, 1)$ for U^*; besides, a $Beta(1,\theta_2)$ is considered for V^*

in Case 1. Note that this parametric model includes the several truncation distributions in the simulations. For each Model, we simulate 1000 Monte Carlo trials with final sample size $n = 50$, 100, 250 or 500. This means that, for each trial, the number of simulated data is much larger than n, actually $N \approx n\alpha^{-1}$ are needed on average, where recall that α stands for the proportion of no truncation. For the simulated models, the proportion of truncation ranges between 44% and 88%. More specifically, the following right and left truncation proportions occur: 37% (right) and 44% (left) for Model 1.1; 38% and 40% for Model 1.2; 53% and 22% for Model 2.1; and 45% and 28% for Model 2.2.

In Table 1 we report the optimal bandwidths (in the sense of the MISE) and the corresponding minimum MISE's for both the nonparametric and the semiparametric estimators. The theoretical MISE function is approximated by the average of the ISEs along the M =1000 trials, namely

$$\overline{ISE}(f_h) = \frac{1}{M}\sum_{m=1}^{M}\int (f_h^m - f)^2 \text{ and } \overline{ISE}(f_{\widehat{\theta},h}) = \frac{1}{M}\sum_{m=1}^{M}\int \left(f_{\widehat{\theta},h}^m - f\right)^2$$

where f_h^m and $f_{\widehat{\theta},h}^m$ are the nonparametric and the semiparametric estimators when based on the m-th Monte Carlo trial.

From Table 1 it is seen that the optimal bandwidths and the MISEs decrease when increasing the sample size; besides, the semiparametric estimator has an error which is smaller than that pertaining the the nonparametric estimator. It is also seen that the optimal bandwidths for the semiparametric estimator are smaller than those of the nonparametric estimator, according to the extra amount of information. As the sample size grows, the relative efficiency of the nonparametric estimator approaches to one; this is in agreement to the asymptotic equivalence of the semiparametric and the nonparametric density estimators discussed in Section 2. Interestingly, for finite sample sizes we see that such relative efficiency may be as poor as 45% (Model 2.2, $n = 50$).

When comparing Models 1.1 and 1.2, one can appreciate that the density corresponding to the first one is not so well approximated by the two estimators; this is because the strong boundary effects of the uniform density (Model 1.1), which disappear when considering a Gaussian model (Model 1.2). Also, the difficulties for estimating the normal density in Case 2 (Model 2.2) are greater than under Model 1.2; this could be explained from the above mentioned fact that Model 1.2 favors the observation of intermediate lifetimes, so there is more sampling information around the density mode (the difficult part to estimate). Model 2.1 is the one presenting the largest MISEs; this Model 2.1 presents difficulties at the left boundary, where the density goes to infinity.

In Figures 1 to 4 we report for each simulated model: (i) the ratio between the MISE's of the semiparametric and the nonparametric estimators along a grid of bandwidths (top row); (ii) the ratio between the MISE of the semiparametric estimator and the minimum MISE of the nonparametric estimator (middle row); and (iii) the target density together with its semiparametric and nonparametric estimators averaged along the 1000 Monte Carlo trials (bottom row). From these Figures 1 to 4 several interesting features are appreciated. First, for each given smoothing degree, the MISE of the semiparametric estimator is less than that of the nonparametric estimator; the relative benefits of using the semiparametric information are more clearly seen when working with relatively smaller bandwidths, when the variance component of the MISE is larger. This illustrates how the semiparametric estimator achieves a variance reduction w.r.t. the NPMLE. The minimum relative efficiency of the nonparametric kernel density estimator varies from about 0.4 to about 0.85, depending on the simulated model and the sample size. Also importantly, we see that the ratios of the MISE's approach to one as the sample size increases. This was expected,

Model	n	h_{opt}		$MISE(h_{opt})$	
		EP	SP	EP	SP
	50	0.173	0.145	0.1449	0.1229
	100	0.130	0.107	0.1174	0.0994
1.1	200	0.091	0.076	0.0890	0.0749
	500	0.051	0.048	0.0537	0.0503
	50	0.062	0.059	0.1281	0.1126
	100	0.052	0.051	0.0817	0.0709
1.2	200	0.044	0.043	0.0480	0.0434
	500	0.037	0.036	0.0240	0.0219
	50	0.216	0.085	0.6940	0.5465
	100	0.126	0.049	0.6142	0.4891
2.1	200	0.039	0.029	0.5181	0.4321
	500	0.015	0.014	0.4054	0.3654
	50	0.074	0.061	0.3091	0.1381
	100	0.056	0.052	0.1587	0.0925
2.2	200	0.046	0.044	0.0748	0.0532
	500	0.037	0.036	0.0385	0.0273

Table 1: Optimal bandwidths (h_{opt}) and minimum MISEs of the density estimators: nonpametric estimator (EP) and semiparametric estimator (SP). Averages along 1000 trials of a sample size n.

since (as discussed in Section 2) both estimators are asymptotically equivalent. However, even when $n = 500$, the relative performance of the nonparametric estimator may be as poor as 70% (Figure 4, top).

Second, from the middle rows of Figures 1 to 4, we see that the semiparametric estimator behaves more efficiently than the nonparametric estimator even when the former uses a suboptimal bandwidth. Indeed, for Models 1.1, 2.1, and 2.2 it becomes clear that there exists a large interval of suboptimal bandwidths which maintain the superiority of the semiparametric density estimator with respect to the nonparametric estimator based on its optimal smoothing parameter. Finally, the averaged estimators depicted in Figures 1- 4 reveal that the semiparametric estimator fits better the target than its nonparametric competitor when the sample size is moderate.

4 Real data illustration

For illustration purposes, in this section we consider data on the age at diagnosis of childhood cancer. These data concern all the cases of childhood cancer diagnosed in North Portugal between 1 January 1999 and 31 December 2003. The age at diagnosis (ranging from 0 to 15 years old) is doubly truncated by (U^*, V^*), where V^* stands for the elapsed time (in years) between birth and end of the study (31 December 2003), and $U^* = V^* - 5$. Information on the 406 diagnosed cases is entirely reported in Moreira and de Uña-Álvarez (2010a).

The semiparametric and the nonparametric kernel estimators for the density of X^* computed from the $n = 406$ cases are given in Figure 5. The scale in the horizontal axis comes from the transformation $(t + 5)/20$, which has been used for the ages at diagnosis and the truncation variables. With this transformation, the U^* is supported on the $(0, 1)$ interval. For the semiparametric estimator, we assume a $Beta(\theta_1, \theta_2)$ model for U^*, and the parameters are estimated by maximizing the conditional likelihood of the truncation times (see Section 2 for details). In this case the pair (U^*, V^*) does not have a density, and the likelihood $\mathcal{L}_1^*(\theta)$ must be properly re-defined by substituting the density of U^* for g_θ in that expression, see Remark 2.1 in (Moreira and de Uña-Álvarez, 2010b) for further details.

Three different bandwidths are used: $h = 0.02$, $h = 0.035$, and $h = 0.06$. As expected, more bumps appear as the smoothing degree decreases. For large bandwidths, only two bumps remain,

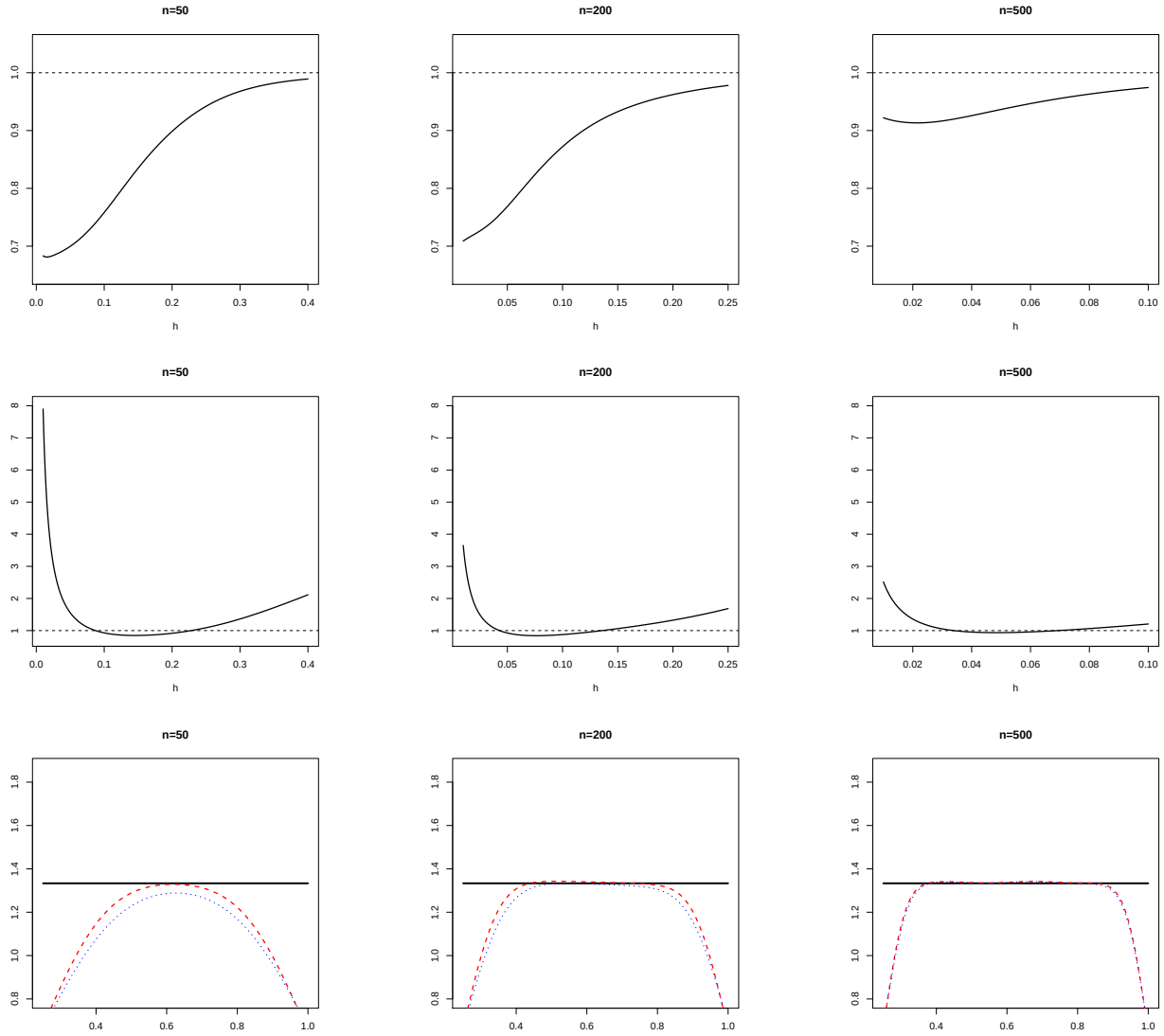

Figure 1: (i) The ratio between the MISE's of the semiparametric and the nonparametric estimators along a grid of bandwidths (top row); (ii) the ratio between the MISE of the semiparametric estimator and the minimum MISE of the nonparametric estimator (middle row); and (iii) the target density (solid line) together with its semiparametric (dashed line) and nonparametric (dotted line) estimators averaged along the 1000 Monte Carlo trials (bottom row) for Model 1.1.

indicating the existence of two subgroups of cases: early cancer detection and late detection (less frequent). For comparison, the naive kernel density estimator which does not correct for the double truncation is also reported. We see that the three estimators are close to each other. This is not surprising, since previous analysis of these data have shown that there is almost no observational bias on the age at diagnosis because of the uniformity of U^* (Moreira and de Uña-Álvarez, 2010a). This fact is also confirmed in Figure 7, left, in which a fairly flat shape of G_n is seen.

For further illustration, in Figure 6 we provide these three estimators for a subgroup of cases. Specifically, we consider the $n = 38$ diagnosed cases of neuroblastoma. For this subgroup the uniformity of U^* is lost, and as a consequence there exists some observational bias (Moreira

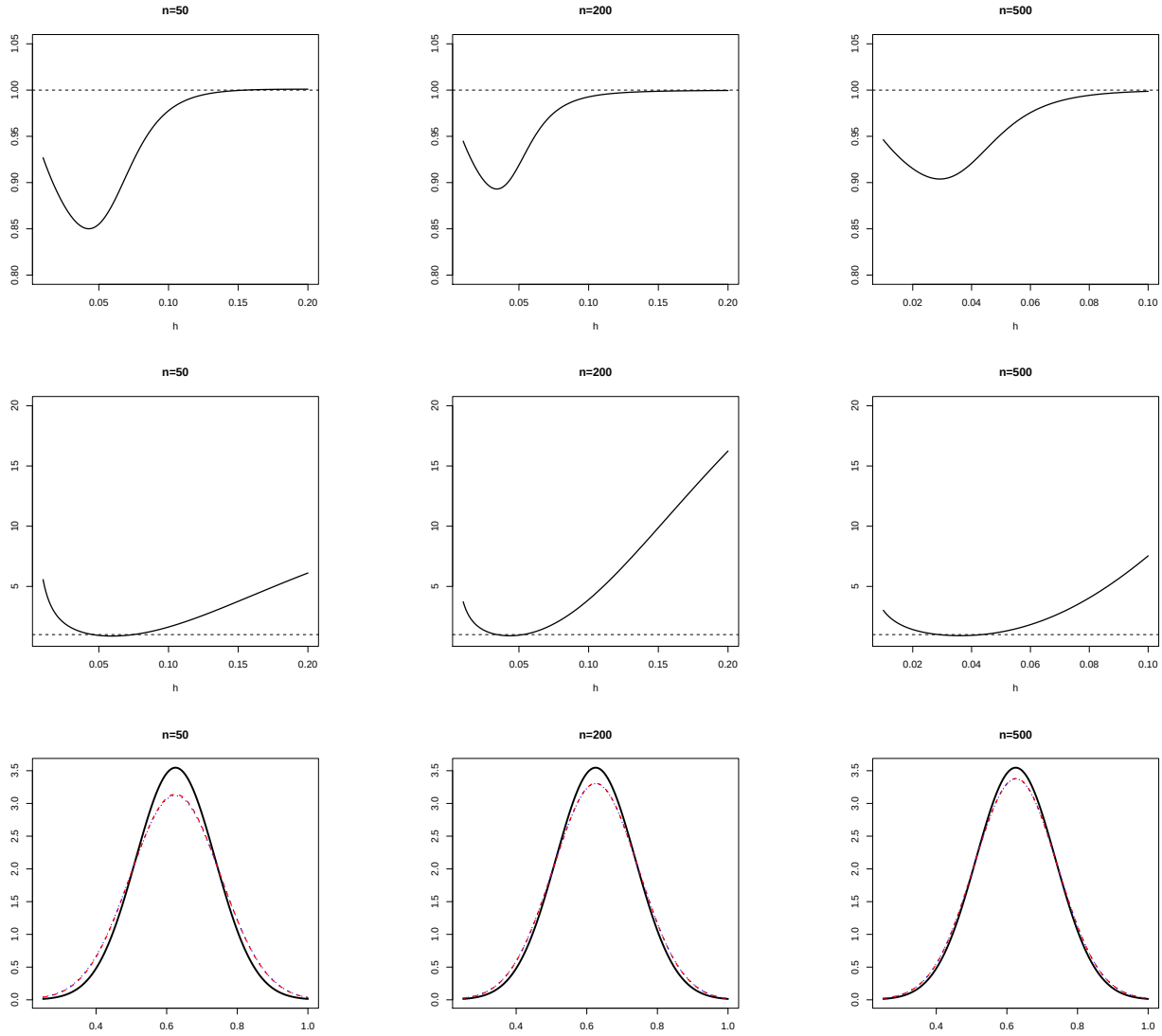

Figure 2: (i) The ratio between the MISE's of the semiparametric and the nonparametric estimators along a grid of bandwidths (top row); (ii) the ratio between the MISE of the semiparametric estimator and the minimum MISE of the nonparametric estimator (middle row); and (iii) the target density (solid line) together with its semiparametric (dashed line) and nonparametric (dotted line) estimators averaged along the 1000 Monte Carlo trials (bottom row) for Model 1.2.

(2010), page 78). Certainly, Figure 7, right, suggests that relatively small ages at diagnosis are more probably observed. This explains the overestimation of the density carried out by the naive estimator at the left tail. Unlike the naive estimator, both the nonparametric and the semiparametric estimators which take the double truncation issue into account declare a second mode at the right tail. These two estimators are similar on the interval $[0.25, 0.45]$ while differences appear from 0.45 on. In order to explain this, we report in Figure 7 the estimators G_n and $G_{\widehat{\theta}}$ for the full data set and for the neuroblastoma cases. Note that the semiparametric estimator is based on a parametric specification of the truncation df; this introduces a bias term which influences the shape of the final density estimator while reducing its variance. Indeed, Figure 7, right, indicates that $G_{\widehat{\theta}}^{-1}$ is smaller than G_n^{-1} at intermediate values of X^*, while the contrary

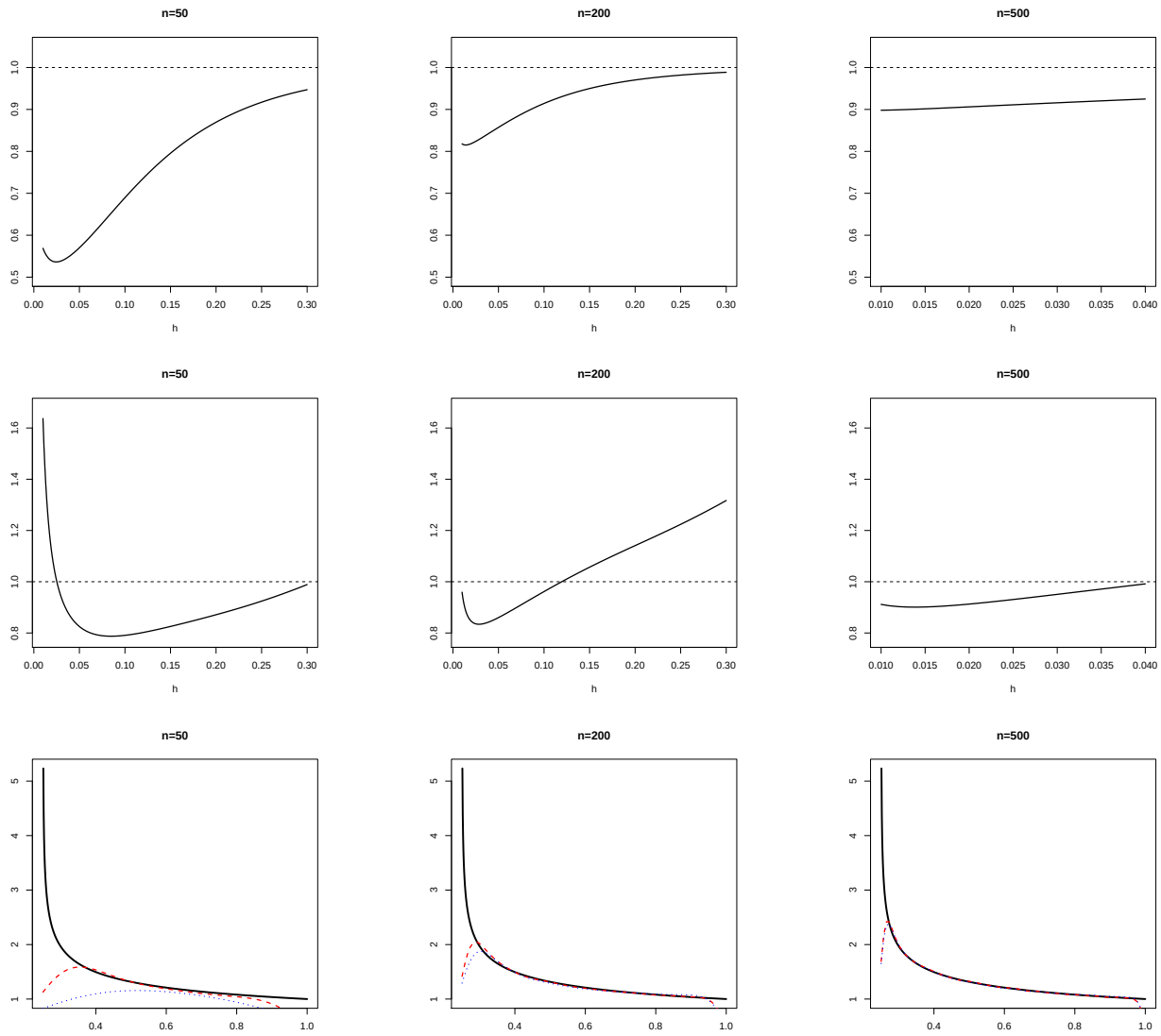

Figure 3: (i) The ratio between the MISE's of the semiparametric and the nonparametric estimators along a grid of bandwidths (top row); (ii) the ratio between the MISE of the semiparametric estimator and the minimum MISE of the nonparametric estimator (middle row); and (iii) the target density (solid line) together with its semiparametric (dashed line) and nonparametric (dotted line) estimators averaged along the 1000 Monte Carlo trials (bottom row) for Model 2.1.

occurs at large times. This explains why the semiparametric estimator locates the second mode more to the right. This biasing effect of the parametric model is not appreciated when analyzing the full data set because G_n and $G_{\widehat{\theta}}$ are close to each other in this case (Figure 7, left).

5 Conclusions and final discussion

In this paper we have introduced kernel density estimation for a variable which is observed under random double truncation. Two estimators have been proposed. The first one is purely nonparametric, and it is defined as a convolution of a kernel function with the NPMLE of the cumulative

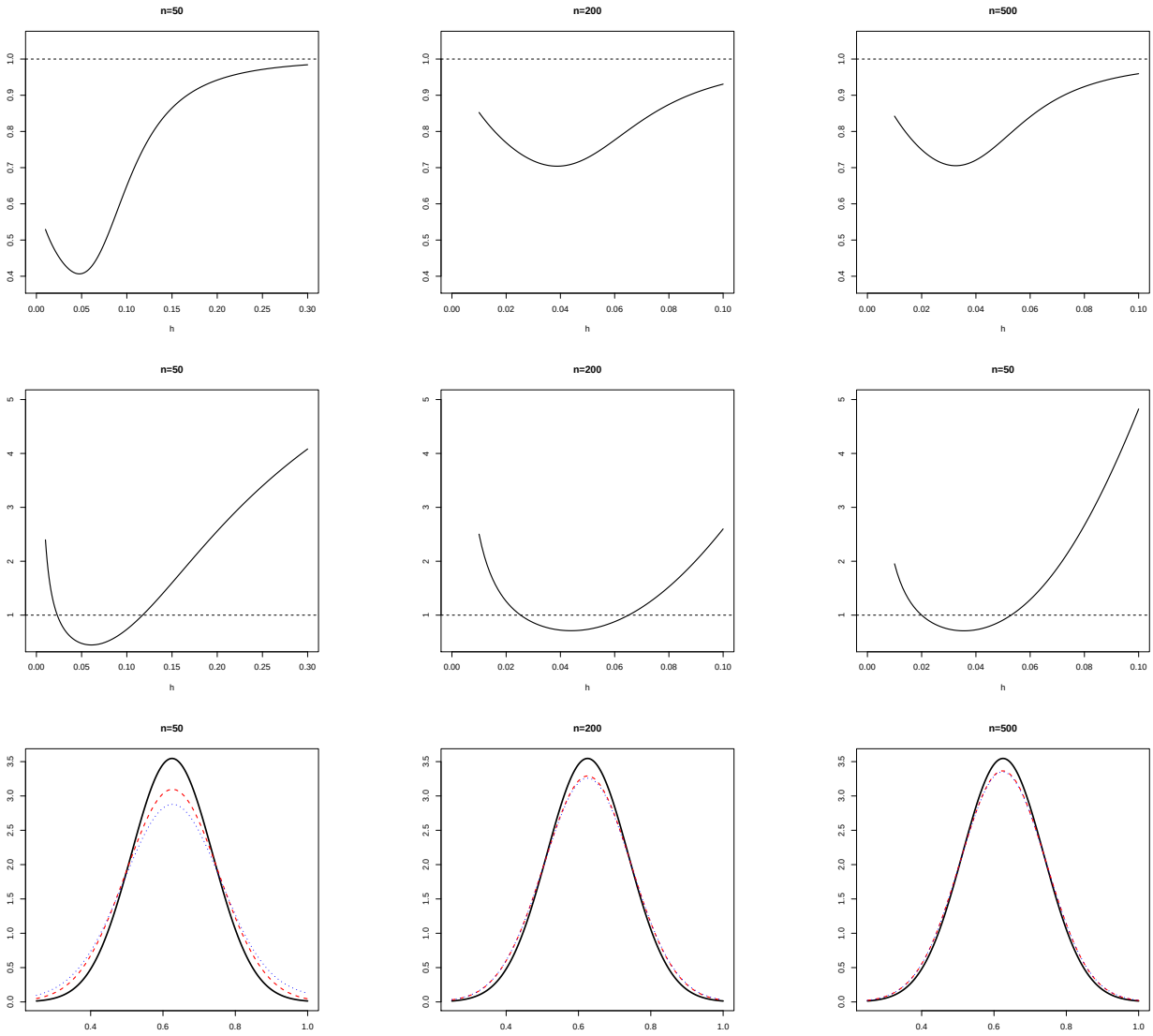

Figure 4: (i) The ratio between the MISE's of the semiparametric and the nonparametric estimators along a grid of bandwidths (top row); (ii) the ratio between the MISE of the semiparametric estimator and the minimum MISE of the nonparametric estimator (middle row); and (iii) the target density (solid line) together with its semiparametric (dashed line) and nonparametric (dotted line) estimators averaged along the 1000 Monte Carlo trials (bottom row) for Model 2.2.

df. The second estimator is semiparametric, since it is based on a parametric specification for the df of the truncation times. Asymptotic properties of the two estimators have been discussed, including a formula for the asymptotic mean integrated squared error (MISE).

Both estimators are asymptotically equivalent in the sense of having the same asymptotic MISE. However, for small and moderate sample sizes, we have seen that the semiparametric estimator may outperform the nonparametric estimator. More explicitly, the relative efficiency of the nonparametric estimator may be as poor as 45% in special situations with small sample sizes. Moreover, in special instances, the relative benefits of using the semiparametric approach are clearly seen even when the sample size is as large as $n = 500$. Finally, our simulation results have

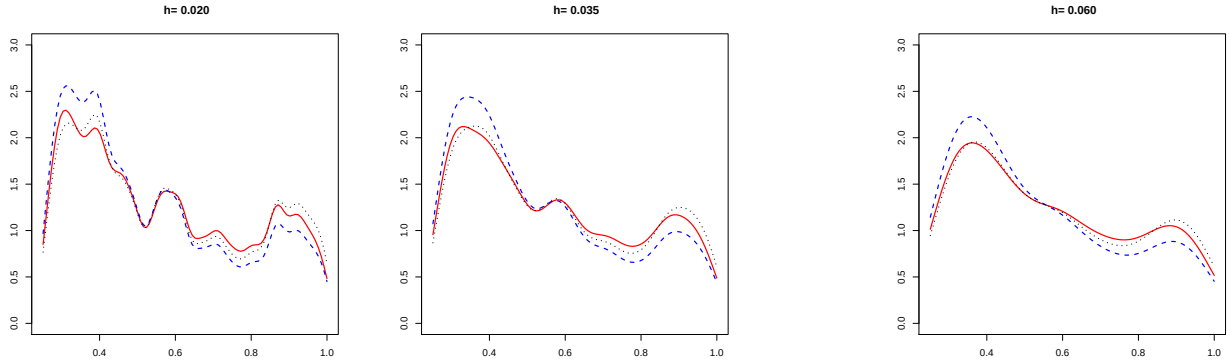

Figure 5: Kernel density estimators for the age at diagnosis, childhood cancer data ($n = 406$). Nonparametric estimator (solid line), semiparametric estimator (dashed line), and naive estimator (dotted line).

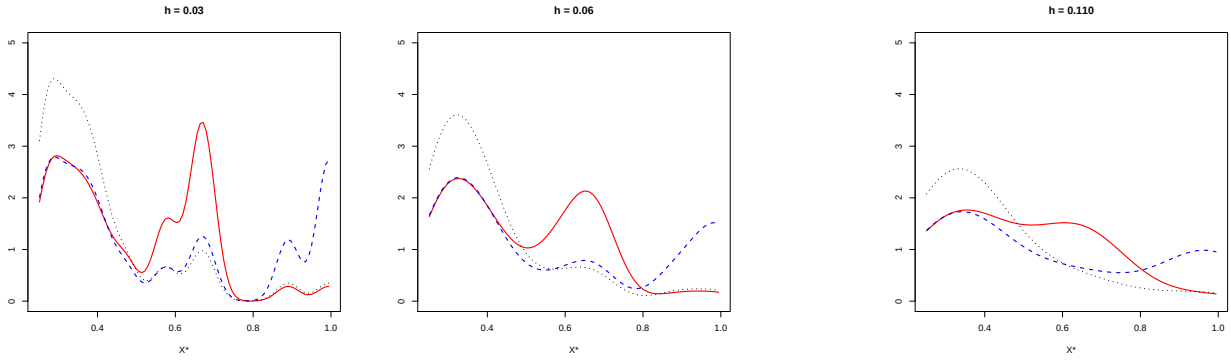

Figure 6: Kernel density estimators for the age at diagnosis, childhood cancer data, for the specific case of neuroblastoma ($n = 38$). Nonparametric estimator (solid line), semiparametric estimator (dashed line), and naive estimator (dotted line).

revealed that the semiparametric estimator may be preferable even when based on a sub-optimal bandwidth. A real data illustration has been provided.

A crucial issue in the construction of the semiparametric estimator is how to choose the parametric model for the truncation distribution. Note that, rather than the truncation distribution itself, the function G influences the shape of the final estimator. Hence, an informal assessment of the parametric family may be performed by plotting the empirical biasing function G_n together with the fitted G_θ. Formal goodness-of-fit tests for a parametric model could be developed too, and this problem is currently under research.

Since the bandwidth h plays a very important role in the performance of the estimators, an interestingly topic for future research is to investigate automatic bandwidth selectors. Also, the application of kernel smoothing to the estimation of the hazard rate function (another important curve in Survival Analysis) in a doubly truncated setup is currently under investigation.

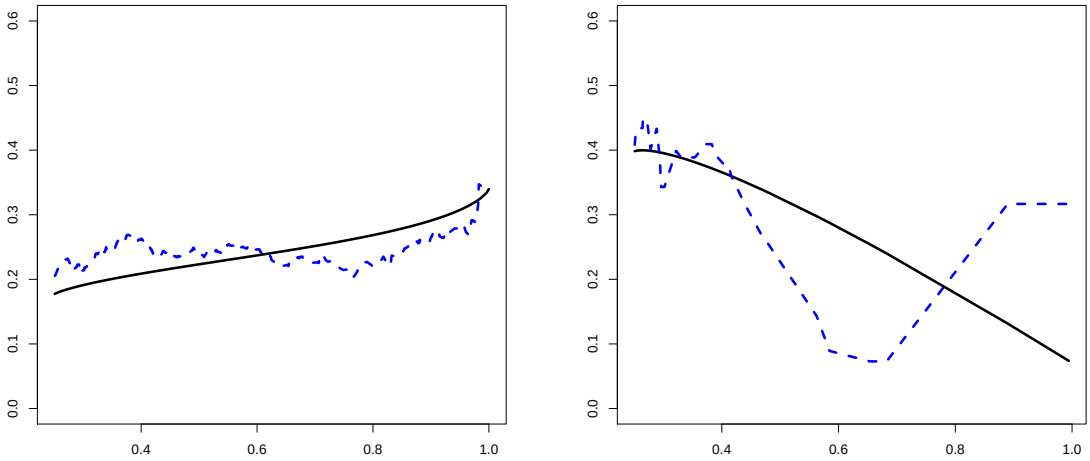

Figure 7: Left: Estimators G_n (dashed line) and $G_{\widehat{\theta}}$ (dotted line) for the childhood cancer data. Left: full sample. Right: neuroblastoma cases.

Acknowledgements

Work supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación, by the Grant 10PXIB300068PR of the Xunta de Galicia and SFRH/BPD/68328/2010 Grant of Portuguese Fundação Ciência e Tecnologia. We also thank the IPO (the Portuguese Institute of Cancer at Porto) for kindly providing the childhood cancer data.

References

Devroye, L. and T. Wagner (1979). The l^1 convergence of kernel density estimates. *The Annals of Statistics 7*, 1136–1139.

Efron, B. and V. Petrosian (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association 94*, 824–834.

Lynden-Bell, D. (1971). A method for allowing for known observational selection in small samples applied to 3cr quasars. *Monthly Notices of the Royal Astronomical Society 155*, 95–118.

Moreira, C. (2010). *The Statistical Analysis of Doubly Truncated Data: new Methods, Software Development, and Biomedical Applications. PhD Dissertation*. PhD in statistics, Departamento de Estadística e I. O. – Universidade de Vigo, Lagoas –Marcosende, Vigo–Spain. ISBN 978–84–95046–30–7.

Moreira, C. and J. de Uña-Álvarez (2010a). Bootstrappping the npmle for doubly truncated data. *Journal of Nonparametric Statistics 22*, 567–583.

Moreira, C. and J. de Uña-Álvarez (2010b). A semiparametric estimator of survival for doubly truncated data. *Statistics in Medicine 29*, 3147–3159.

Moreira, C., J. de Uña-Álvarez, and R. Crujeiras (2010). Dtda: an r package to analyze randomly truncated data. *Journal of Statistical Software 37*, 1–20.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*, 1065–1076.

Shen, P. (2010a). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics 62*, 835–853.

Shen, P. (2010b). Semiparametric analysis of doubly truncated data. *Communications in Statistics – Theory and Methods 39*, 3178–3190.

Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *The Annals of Statistics 21*, 146–156.

Tsai, W., N. Jewell, and M. Wang (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika 74*, 883–886.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B. Methodological 38*, 290–295.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*, Volume 60 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall Ltd.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association 86*, 130–143.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics 13*, 163–177.

Zhou, Y. and P. S. F. Yip (1999). A strong representation of the product-limit estimator for left truncated and right censored data. *Journal of Multivariate Analysis 69*, 261–280.