Universidade de Vigo

Nonparametric regression with doubly truncated data

Carla Moreira, Jacobo de Uña-Álvarez and Luís Meira-Machado

Report 12/04

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://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade de Vigo

Nonparametric regression with doubly truncated data

Carla Moreira, Jacobo de Uña-Álvarez and Luís Meira-Machado

Report 12/04

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

Nonparametric regression with doubly truncated data

Carla Moreira * Jacobo de Uña - Álvarez § Luís Meira- Machado ††

Abstract

In this paper nonparametric regression with a doubly truncated response is introduced. Local constant and local linear kernel-type estimators are proposed. Asymptotic expressions for the bias and the variance of the estimators are obtained, showing the deterioration provoked by the random truncation. To solve the crucial problem of bandwidth choice, two different bandwidth selectors based on plug-in and cross-validation ideas are introduced. The performance of both the estimators and the bandwidth selectors is investigated through simulations. A real data illustration is included. The main conclusion is that the introduced regression methods perform satisfactorily in the complicated scenario of random double truncation.

Key Words: Local polynomial regression; Kernel smoothing; bandwidth selection; random truncation; biased data; mean squared error.

*University of Vigo, Department of Statistics and O.R., Lagoas - Marcosende, 36 310, Vigo, Spain. E-mail address: carlamgmm@gmail.com. University of Minho, Department of Mathematics for Science and Technology, Campus de Azurém, 4800-058 Guimarães, Portugal. Research is supported by research Grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación and SFRH/BPD/68328/2010 Grant of Portuguese Fundação Ciência e Tecnologia.

§University of Vigo, Department of Statistics and O.R., Lagoas - Marcosende, 36 310, Vigo, Spain. E-mail address: jacobo@uvigo.es. Research is supported by research Grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación and by the Grant PGIDIT07PXIB300191PR of the Xunta de Galicia.

††University of Minho, Department of Mathematics for Science and Technology, Campus de Azurém, 4800-058 Guimarães, Portugal.

1 Introduction

Random truncation is a well-known phenomenon which may be present when observing lifetime data. For example, recruitment of lifetimes through a cross-section induces left-truncation, so larger event times are observed with higher probability. Another example is found when analyzing data which correspond to events taking place before some specific date; in this case, the time-to-event is right-truncated and, therefore, small lifetimes are over-sampled. These two forms of truncation are one-sided, and relatively simple estimators exist. See e.g. Klein and Moeshberger (2003). Nonparametric estimation methods suitable for one-sided random truncation were developed in the last three decades, see for example Woodroofe (1985), Tsai et al. (1987) or Stute (1993) for the estimation of a cumulative distribution function, and for nonparametric regression, Gross and Lai (1996), Iglesias-Pérez and González-Manteiga (1999), Akritas and LaValley (2005) or Ould-Saïd and Lemdani (2006).

In some applications, two-sided (rather than one-sided) random truncation appears. This occurs, for example, when the sample restricts to those individuals with event falling between two particular dates. This is the case of the sample provided by Moreira and de Uña-Álvarez (2010a), who reported data corresponding to children diagnosed from cancer between 1999 and 2003; in this case, the age at cancer diagnosis is doubly truncated, the truncation times been determined by the two specific limiting dates of observation. The AIDS Blood Transfusion data in Kalbfleisch and Lawless (1989) is another example of such a situation. These data are restricted to those cases diagnosed from AIDS prior January 1987. For this data set, the induction times are doubly truncated because HIV was unknown before 1982, so any case of transfusion-related AIDS before this time would not have been properly classified. Efron and Petrosian (1999) investigated quasar luminosities which were doubly truncated by some detection limits. See Section 4 for more details about the AIDS Blood Transfusion data and quasar luminosities data. Under double truncation, the observational bias is not so evident as in the one-sided truncated setup. Generally speaking, one may say that, under double truncation, large and small inter-event times will be less probably observed. Unlike for one-sided truncation, the nonparametric maximum-likelihood estimator (NPMLE) of the lifetime distribution has no explicit form under double truncation; this complicates the practice and the theoretical developments. We mention that censoring is a problem different to random truncation, because with censored data the researcher as at least some partial information on the censored lifetimes.

Compared to the huge literature devoted to one-sided truncation, there are only few papers devoted to the random double truncation model. Efron and Petrosian (1999) introduced the NPMLE of a cumulative distribution function (df) under double truncation. The asymptotic properties of this NPMLE were further investigated by Shen (2010). Moreira and de Uña-Álvarez (2010b) introduced a semiparametric estimator of a doubly truncated df, while Moreira et al. (2010) presented an `R` package to compute the NPMLE and confidence bands. Methods for testing a quasi-independence assumption between the lifetime of interest and the truncation times were investigated by Martin and Betensky (2005). Despite of the existence of these papers, random double truncation is a phenomenon which is still quite unknown

nowadays. In some applications, the goal is the estimation of a smooth curve such as the density function, the hazard rate function, or the regression function. The estimation of these curves crucially depends on the selected bandwidth or smoothing parameter (Wand and Jones, 1995). For the best of our knowledge, the only paper dealing with smoothing methods under double truncation is Moreira and de Uña-Álvarez (2012), who considered kernel density estimation. In this paper we rather focus in nonparametric kernel regression.

Let (X^*, Y^*) be the two-dimensional variable of interest, where Y^* is the lifetime or the inter-event time of main interest, and X^* is a one-dimensional continuous covariate. The goal is the estimation of the regression function $m(x) = E[Y^*|X^* = x]$. Due to the presence of random double truncation, we are only able to observe (X^*, Y^*) when $U^* \leq Y^* \leq V^*$, where (U^*, V^*) are the truncation times; in that case, (U^*, V^*) are also observed. On the contrary, when $U^* \leq Y^* \leq V^*$ is violated, nothing is observed. As usual with random truncation, we assume that the truncation times are independent of (X^*, Y^*). Let $(U_1, V_1, X_1, Y_1), ..., (U_n, V_n, X_n, Y_n)$ be the observed sample, these are iid data with the same distribution as (U^*, V^*, X^*, Y^*) given $U^* \leq Y^* \leq V^*$, and let $m^T(x) = E[Y_1|X_1 = x]$ be the observed regression function. In general, $m^T(x)$ and the target $m(x)$ will differ; see e.g. Figure 8, in which these two curves are estimated for the AIDS Blood Transfusion data. This is because of the truncating condition which introduces an observational bias. Similar features were reported in the context of length-biasing, in which the relative probability of sampling a given value of (X^*, Y^*) is proportional to the length of Y^*, see e.g. Cristóbal and Alcalá (2000). In the doubly truncated setup, this relative probability of observing $(X^*, Y^*) = (x, y)$ is given by $G(y) = P(U^* \leq y \leq V^*)$. This function G can be estimated from the data by maximum likelihood principles, see the iterative algorithm in Section 2.

The rest of the paper is organized as follows. In Section 2 we introduce the relationship between the observed conditional distribution and that of interest. As it will be seen, by downweighting the (X_i, Y_i)'s with the largest values of $G_n(Y_i)$ (where G_n is an estimator for G), we are able to obtain a consistent estimator of $m(x)$. Weighted local polynomial type estimators are considered to this end. We give the asymptotic bias and variance of the weighted Nadaraya-Watson (i.e. local constant) estimator and the weighted local linear kernel estimator. We also propose two different methods to choose the bandwidth for these estimators in practice. In Section 3 we investigate the finite-sample performance of the estimators and the bandwidth selectors through simulations. Section 4 illustrates all the proposed methods by considering AIDS Blood Transfusion data of Kalbfleisch and Lawless (1989) and also the quasar luminosities of Efron and Petrosian (1999). Finally, in Section 5 we report the main conclusions of our investigation. The technical proofs and details are deferred to the Appendix.

2 The estimators.

In this Section we introduce the proposed estimators. We also include the asymptotic results (Section 2.1) and the bandwidth selection algorithms (Section 2.2). Firstly we introduce the needed notations.

Let $F(.|x)$ be the conditional df of Y^* given $X^* = x$, so $m(x) = \int_{-\infty}^{\infty} tF(dt|x)$, and let $\alpha(x) = P(U^* \leq Y^* \leq V^*|X^* = x) = \int_{-\infty}^{\infty} G(t)F(dt|x)$ be the conditional probability of no truncation. It is assumed that $\alpha(x) > 0$. Let $F^*(.|x)$ be the observable conditional df, that is $F^*(y|x) = P(Y_1 \leq y|X_1 = x)$. We have

$$F^*(y|x) = \alpha(x)^{-1} \int_{-\infty}^{y} G(t)F(dt|x), y \geq 0.$$

This means that, for a fixed value of the covariate, the response Y^* is observed with a relative probability proportional to $G(Y^*)$. Conversely, provided that $G(t) > 0$ for all t, one may write $F(y|x) = \alpha(x) \int_{-\infty}^{y} G(t)^{-1} F^*(dt|x)$, where $\alpha(x) = 1/\alpha^*(x)$ with $\alpha^*(x) = \int_{-\infty}^{\infty} G(t)^{-1} F^*(dt|x) = E\left[G(Y_1)^{-1}|X_1 = x\right]$. Therefore, the target $m(x)$ is written as $m(x) = m^*(x)/\alpha^*(x)$ where $m^*(x) = E\left[Y_1 G(Y_1)^{-1}|X_1 = x\right]$.

Note that the functions $m^*(x)$ and $\alpha^*(x)$ are conditional means of observable variables (once G is replaced by a proper estimator) and, consequently, they can be estimated by standard methods as, e.g., Nadaraya-Watson. Hence, a consistent estimator of $m(x)$ may be introduced as $\hat{m}_{NW}(x) = \hat{m}^*_{NW}(x)/\hat{\alpha}^*_{NW}(x)$, where

$$\widehat{m}^*_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i G_n(Y_i)^{-1}}{\sum_{i=1}^{n} K_h(x - X_i)}$$

and

$$\widehat{\alpha}^*_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) G_n(Y_i)^{-1}}{\sum_{i=1}^{n} K_h(x - X_i)}$$

are the Nadaraya-Watson estimators of $m^*(x)$ and $\alpha^*(x)$ respectively. In these expressions, G_n stands for a nonparametric estimator of the biasing function G, namely $G_n(y) = \int_{u \leq y \leq v} T_n(du, dv)$ where T_n is the NPMLE of the joint df of the truncation times (Shen, 2010). Also, K and h are the kernel function and the bandwidth respectively, while $K_h(.) = K(./h)/h$ is the re-scaled kernel; see e.g. Wand and Jones (1995) for more on kernel regression.

As mentioned, for the computation of T_n (and hence of G_n) an iterative algorithm is needed. To this end, we have followed the algorithm proposed by Shen (2010), see also Moreira et al. (2010). Explicitly, the steps are as follows:

Step S_0 Compute $\Phi_i^{(0)} = \sum_{m=1}^{n} \varphi_m^{(0)} I(U_i \leq Y_m \leq V_i)$ where $\varphi_m^{(0)} = \frac{1}{n}$, $1 \leq m \leq n$, is the initial solution for F.

Step S_1 Compute the associated solution for T: $\psi_j^{(1)} = \left[\sum_{i=1}^{n} \frac{1}{\Phi_i^{(0)}}\right]^{-1} \frac{1}{\Phi_j^{(0)}}$, $1 \leq j \leq n$, and $\Psi_i^{(1)} = \sum_{m=1}^{n} \psi_m^{(1)} I(U_m \leq Y_i \leq V_m)$.

Step S_2 Compute the improved solution for F: $\varphi_j^{(1)} = \left[\sum_{i=1}^{n} \frac{1}{\Psi_i^{(1)}}\right]^{-1} \frac{1}{\Psi_j^{(1)}}$, $1 \leq j \leq n$, and $\Phi_i^{(1)} = \sum_{m=1}^{n} \varphi_m^{(1)} I(U_i \leq Y_m \leq V_i)$.

Step S_3 Repeat Steps S_1 and S_2 until a convergence criterion is reached.

As convergence criterion, we have used $\max_{1\leq j\leq n} |\varphi_j^{(k-1)} - \varphi_j^{(k)}| \leq 1e-06$. Then, the estimated T_n is constructed from the k-th solution $\psi_j^{(k)}$, $1 \leq j \leq n$, as $T_n(u,v) = \sum_{i=1}^{n} \psi_j^{(k)} I(U_i \leq u, V_i \leq v)$. Accordingly, G_n is computed as $G_n(y) = \int_{u\leq y\leq v} T_n(du,dv) = \sum_{i=1}^{n} \psi_j^{(k)} I(U_i \leq y \leq V_i)$.

An alternative way of introducing a consistent estimator for $m(x)$ is through weighted local least squares. Introduce the criterion function

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \ldots - \beta_p (X_i - x)^p)\}^2 K_h(x - X_i) G_n(Y_i)^{-1}.$$

Let $(\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$ be the minimizer of this criterion. Then, $\hat{m}_{(p)}(x) = \widehat{\beta}_0$ is an estimator for $m(x)$. Under length-bias ($G(y) = y$), this is the (weighted) local polynomial estimator introduced by Cristóbal and Alcalá (2000). For $p = 0$, we obtain the Nadaraya-Watson (i.e. local constant) type estimator $\widehat{m}_{NW}(x)$ introduced above. For $p = 1$, we obtain the weighted local linear kernel regression estimator, say $\widehat{m}_{LLK}(x)$, which (in the ordinary setting with no truncation) has been recommended in applications due to its smaller bias and boundary adaptability when compared to the local constant estimator.

2.1 Asymptotic performance

Theorem below gives an asymptotic expression for the bias and the variance of $\hat{m}_{NW}(x)$ and $\hat{m}_{LLK}(x)$. Some further notation is needed. We put $f(x)$ for the density of the covariate X^*, and $\alpha = P(U^* \leq Y^* \leq V^*)(> 0)$ for the (unconditional) probability of no truncation, and we introduce

$$\sigma^2(x) = E[(Y^* - m(X^*))^2 G(Y^*)^{-1} | X^* = x].$$

We also put $\mu_2(K) = \int t^2 K(t) dt$ and $R(K) = \int K(t)^2 dt$. The following conditions are assumed:

(C1) The function m'' is continuous and bounded in a neighborhood of x

(C2) The kernel function K is a density function symmetrical about zero, and with compact support

(C3) As $n \to \infty$, $h \to 0$ and $nh \to \infty$

(C4) The conditional expectations $E[(Y^* - m(X^*))^d G(Y^*)^{-1} | X^* = x]$, $d = 0, 1, 2$, are finite

Theorem 2.1. *Assume (C1)-(C4). The asymptotic bias and variance of $\widehat{m}_{NW}(x)$ and $\widehat{m}_{LLK}(x)$ are given by*

$$Bias(\widehat{m}_{NW}(x)) \sim \frac{1}{2}\mu_2(K)h^2 \left[m''(x)f(x) + 2m'(x)f'(x)\right] / f(x) \equiv h^2 B_{NW}(x), \tag{2.1}$$

$$Bias(\widehat{m}_{LLK}(x)) \sim \frac{1}{2}\mu_2(K)h^2 m''(x) \equiv h^2 B_{LLK}(x), \tag{2.2}$$

and

$$Var(\widehat{m}_{NW}(x)) \sim Var(\widehat{m}_{LLK})(x) \sim (nh)^{-1} R(K)\alpha\sigma^2(x)/f(x) \equiv (nh)^{-1} V(x). \tag{2.3}$$

Proof. See the Appendix. □

Theorem 2.1 shows that the asymptotic bias of the proposed estimators is the same as that corresponding to the iid case and, therefore, it is unaffected by the double truncation issue. However, truncation may influence the variance; the same happens under one-sided truncation (see e.g. Ould-Saïd and Lemdani, 2006, or Liang et al., 2011) and for length-biased data (Cristóbal and Alcalá, 2000). In the untruncated situation, the asymptotic variance of both NW and LLK estimators is $(nh)^{-1}R(K)\tau^2(x)/f(x)$ where $\tau^2(x) = E[(Y^* - m(X^*))^2 | X^* = x]$. This quantity $\tau^2(x)$ may be greater or smaller than $\alpha\sigma^2(x)$, so the estimation of $m(x)$ could be performed with less variance under double truncation than under iid sampling at particular points x. This can be explained by the fact that, because of truncation, specific parts of the support of the variable of interest are over-sampled (while others not), thus introducing extra information in some areas. However, by Hölder's inequality we have $\int \alpha\sigma^2(x)dx \geq \int \tau^2(x)dx$, and hence the doubly truncated scenario is 'more difficult' (in the sense of having more variance in estimation) in average. A similar feature was found when estimating a density function under double truncation, see

Moreira and de Uña-Álvarez (2012).

Theorem 2.1 also informs about the relative asymptotic performance of Nadaraya-Watson and local linear kernel smoothers. As usual when comparing both methods, it is seen that the NW involves and extra bias term $(2m'(x)f'(x))$ which is missing in the LLK. In practice, this will result in a poorer relative performance.

2.2 Bandwidth choice

As always with kernel methods, the choice of the bandwidth sequence h is very important for the accuracy of the proposed estimators. One possible criterion for choosing the bandwidth is to minimize the mean integrated squared error (MISE), which for any estimator $\hat{m}(x)$ is defined as

$$MISE(\widehat{m}) = E\int(\widehat{m}(x) - m(x))^2 dx = \int Bias(\widehat{m}(x))^2 dx + \int Var(\widehat{m}(x))dx.$$

Theorem 2.1 suggests the following asymptotic approximation to the MISE of the NW and LLK estimators:

$$MISE(\widehat{m}_{NW}) \sim h^4 \int B_{NW}(x)^2 dx + (nh)^{-1} \int V(x)dx$$

and

$$MISE(\widehat{m}_{LLK}) \sim h^4 \int B_{LLK}(x)^2 dx + (nh)^{-1} \int V(x)dx$$

respectively. The bandwidth minimizing these two functions is given by

$$h_{opt} = n^{-1/5} \left[\frac{\int V(x)dx}{4\int B^2(x)dx} \right]^{1/5} \tag{2.4}$$

where $V(x)$ is defined in (2.3) and $B(x)$ is $B_{NW}(x)$ in (2.1) for the NW estimator and $B_{LLK}(x)$ in (2.2) for the LLK estimator. Practical usage of (2.4) requires estimation of $\int V(x)dx$ and $\int B^2(x)dx$, which can be performed on the basis of formulas (2.1), (2.2) and (2.3). To this end, we follow the direct plug-in (DPI) method in Härdle and Marron (1995), which makes use of polynomials and histograms for the preliminary estimation of the density and regression functions involved in $B(x)$ and $V(x)$. However, these estimators must be adapted to the doubly truncated scenario. In the Appendix we give the most relevant details behind this modified DPI bandwidth h_{DPI}.

Alternatively, one may use a cross-validation (CV) criterion to compute the bandwidth in a data-driven way. For any estimator $\hat{m}(x)$ of $m(x)$ depending on a bandwidth h, a cross-validation function is given by $CV(h) = \sum_{i=1}^{n} (Y_i - \hat{m}_{-i}(X_i))^2$, where $\hat{m}_{-i}(x)$ is the leave-one-out version of $\hat{m}(x)$ computed by

removing the i-th datum from the initial sample (cfr. Härdle et al., 2004). Both the DPI and the CV bandwidths are investigated in the simulations of the next section.

3 Simulation study

In this section, we illustrate the finite sample behaviour of the proposed estimators, through a simulation study. We compare the global mean squared errors of NW and LLK estimators, and we explore their graphical fit to the true underlying curve. We also investigate the performance of the two bandwidth selectors introduced in Section 2. We generate the data according to three different patterns of truncation:

- Model 1:
  1. Draw X^* from $Exp(\lambda)$, with $\lambda = 4$, and truncated to the support $(0, 1)$;
  2. Given $X^* = x$, set $Y^* = m(x) + \epsilon$ with $\epsilon \sim N(0, \tau)$ independent of X^*.
  3. Draw independently $U^* \sim U(0, b_U)$ and $V^* \sim U(a_V, 1)$, with $b_U = 0.5$ and $a_V = 0.5$.
  4. We accept the (U^*, V^*, X^*, Y^*) satisfying the condition $U^* \leq Y^* \leq V^*$.

Models 2 and 3 are similar to Model 1 but with a different Step 3 for the simulation of the truncating variables. Specifically, for Model 2 we take $U^* \sim b_U Beta(3, 1)$ and $V^* \sim U(a_V, 3)$, with $b_U = 0.5$ and $a_V = 0.5$; while for Model 3 we take $U^* \sim U(a_U, b_U)$ and $V^* = U^* + k$, with $a_U = -0.25$, $b_U = 0.5$ and $k = 0.5$. Given the covariate X_i $(i = 1, \ldots, n)$, the response Y_i is generated in Step 2 by considering as true regression function $m(x) = \mathbb{E}(Y|X = x) = \frac{2+sin(2\pi x)}{3}$. The whole procedure Step 1-Step 4 is repeated until a final sample with size n is obtained, with $n = 50$, 100, 250 and 500. We consider two different values for the standard deviation τ of ϵ in each model, $\tau = 0.01$ and $\tau = 0.1$.

In Figure 1 the observational bias behind the simulated models is depicted. For Model 1, the biasing function $G(t)$ is symmetrical about 1/2, taking smaller values as t approaches to 0 and 1, and being zero beyond 1 (top-left figure). As a consequence, the observable regression function separates from the target for X around 1/4 and 3/4, corresponding to the maximum and minimum values of the sinus function. This departure is hardly noticed for $\tau = 0.01$ (middle-left figure), since in this case the response is roughly constant and therefore the double truncation does not induce any bias; however, the separation becomes very clear for $\tau = 0.1$ (bottom-left figure), particularly around $x = 1/4$, when the observation of the response occurs with a smaller probability. On the other hand, for Models 2 and 3 the biasing function is asymmetrical (Figure 1, top-center and top-right, respectively). For Model 2 the main problems are in the observation of small responses; this provokes the departure between the target and the observed regression function around 3/4 (bottom-center figure). While for Model 3 the situation is the opposite, with a large observational bias for the large responses (bottom-right figure). As for Model 1, the observational bias is almost negligible when $\tau = 0.01$.

We mention that there is no hope that the proposed estimators will solve the biasing problem around the maximum of the sinus function in bottom-left (Model 1) and bottom-right (Model 3) figures. This is because the zero value of $G(t)$ for $t > 1$ does not allow for the observation of the response to the right of 1 and, consequently, no information on the conditional distribution of Y^* around $X^* = 1/4$ will be available. However, the biases around the minimum of $m(x)$ (bottom-left and bottom-center figures) will be corrected by the estimators up to a certain extent. See Figure 7.

The global performance of $\hat{m}_{NW}(x)$ and $\hat{m}_{LLK}(x)$, both based on a Gaussian kernel, was assessed along $M = 500$ Monte Carlo trials. To this end we used the global mean squared error (GMSE) as a measure of fit, which for any given estimator $\hat{m}_h$ is defined as

$$GMSE(h) = \frac{1}{Mn} \sum_{l=1}^{M} \sum_{k=1}^{n} [\hat{m}_{h,l}(X_{k,l}) - m(X_{k,l})]^2.$$

where $\hat{m}_{h,l}(x)$ is the estimator $\hat{m}_h(x)$ based on the l-th trial, and $X_{k,l}$ is the k-th covariate value in the l-th trial. In Tables 1 to 3 we report the optimal bandwidth h_{GMSE} (defined as the bandwidth leading to the smallest GMSE) and the minimum GMSE for NW and LLK estimators and Models 1 to 3. The functions which are minimized are depicted in Figure 2 ($\tau = 0.01$) and Figure 3 ($\tau = 0.1$) for $n = 250$ (the other sample sizes report similar figures). As expected, the optimal bandwidths and the GMSE decrease as the sample size increases, and they increase with the noise (τ). On the other hand, the error of the LLK estimator is always smaller than that of NW, so the local linear smoother will be preferred in practice. Interestingly, by comparing the GMSE for the three different biasing functions, it is seen that the error of the estimator under Models 1 and 3 are larger than under Model 2. These relative difficulties agree with the deterioration level of the regression function shown in Figure 1 for the largest value of τ.

The median and the interquartile range (IQR) of CV and DPI bandwidths h_{CV} and h_{DPI} along the Monte Carlo simulations are also provided in Tables 1 to 3. Generally speaking, it is seen that the CV bandwidth tends to oversmooth. On the other hand, the DPI method tends to choose a bandwidth smaller than optimal for LLK and n=50, 100. However, for LLK and n=250, 500 and for NW, the DPI bandwidth is in general greater than h_{GMSE}, the distance between h_{DPI} and h_{GMSE} being larger for NW than for LLK. It is also seen from the Tables that CV is more variable than DPI; this agrees with previous comparative studies on both bandwidth selectors.

In Figures 4 and 5 the attained MSE's (for each simulated sample) of the NW and LLK estimators when based on CV and DPI bandwidths are described by using boxplots. These Figures correspond to $n = 250$, other values of n reported similar results (not shown). It is clear that the error of the NW estimator is smaller than that of LLK when using the CV algorithm, while the contrary occurs for DPI. In this sense, one may say that DPI provides results in agreement with the better performance of the LLK estimator. Interestingly, the error of the LLK estimator is smaller when using the DPI bandwidth; therefore, one practical conclusion from our simulations is that one should use the LLK estimator rather than NW, and that one should take the DPI bandwidth for the computation of the LLK.

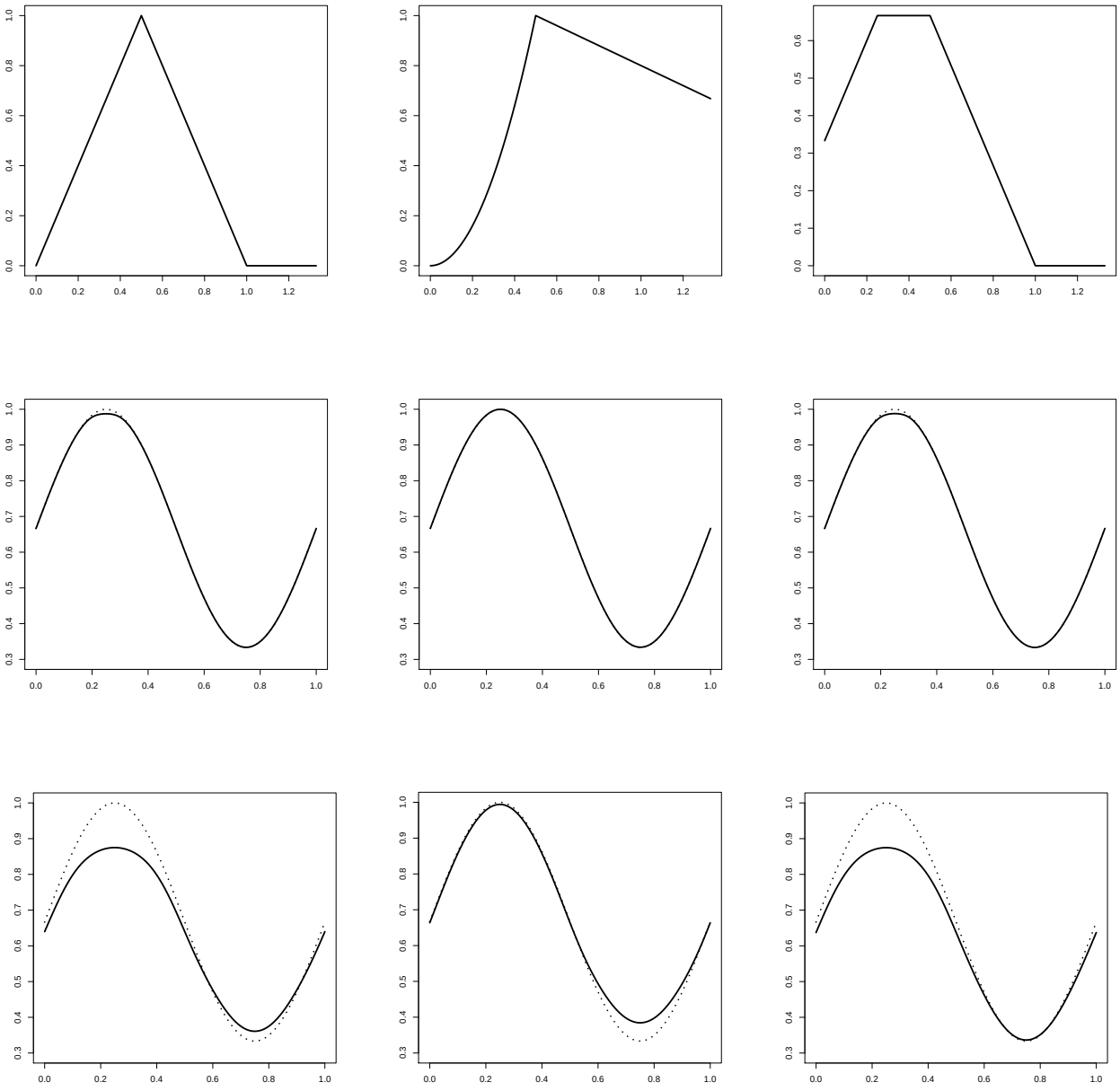

Figure 1: Biasing function G for Model 1 (top-left figure), for Model 2 (top-center figure) and for Model 3 (top-right figure). Observational bias behind the simulated models 1 to 3; Model 1 (bottom-left figure); Model 2 (bottom-center figure) and Model 3 (bottom-right figure). Target (dashed line) and observed regression function (solid line).

The relative behaviour of NW and LLK estimators is further illustrated in Figures 6 and 7, where we report for each model the estimators averaged along the 500 Monte Carlo trials, in the cases $n = 250$ and $n = 500$. For the computation of the estimators we used the optimal bandwidths (h_{GMSE}) in Tables 1 to 3. The quality of fit tends to improve when increasing the sample size. However, a systematic bias is visible in the case $\tau = 0.1$ for Models 1 and 3, even with $n = 500$. This is because of the absence of sampling information on the response around $x = 1/4$, as discussed above, see Figure 1. We also note that values of X close to 1 are less represented in the sample because of the exponential model taken for the covariate. Therefore, the bias around $x = 3/4$ under Model 2, case $\tau = 0.1$, is still visible for $n = 500$. In this case, however, this bias will decrease when considering larger sample sizes (not shown).

Table 1: Minimum $GMSE$'s of the NW and LLK estimators, and corresponding optimal bandwidths for Model 1. The median and the interquartile range of CV and DPI bandwidth selectors are also reported.

		$\widehat{m}_{NW}$					
τ	n	h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0060	5.2959×10^{-5}	0.0110	7.0000×10^{-3}	0.0178	4.2103×10^{-3}
	100	0.0060	3.9501×10^{-5}	0.0090	4.0000×10^{-3}	0.0183	3.0197×10^{-3}
	250	0.0060	2.2629×10^{-5}	0.0070	2.0000×10^{-3}	0.0164	1.8303×10^{-3}
	500	0.0050	1.3223×10^{-5}	0.0060	1.0000×10^{-3}	0.0146	1.7519×10^{-3}
0.01		$\widehat{m}_{LLK}$					
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0270	2.6223×10^{-5}	0.0280	1.4000×10^{-2}	0.0244	3.8596×10^{-3}
	100	0.0240	1.6791×10^{-5}	0.0260	8.0000×10^{-3}	0.0232	2.4707×10^{-3}
	250	0.0190	8.6523×10^{-6}	0.0220	4.0000×10^{-3}	0.0204	1.3200×10^{-3}
	500	0.0170	4.7123×10^{-6}	0.0190	3.0000×10^{-3}	0.0181	1.2630×10^{-3}
		$\widehat{m}_{NW}$					
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0310	3.3139×10^{-3}	0.0310	2.5000×10^{-2}	0.0381	1.3787×10^{-2}
	100	0.0280	2.3607×10^{-3}	0.0260	1.4250×10^{-2}	0.0436	1.1005×10^{-2}
	250	0.0240	1.5297×10^{-3}	0.0220	8.2500×10^{-3}	0.0405	7.9349×10^{-3}
	500	0.0200	1.1792×10^{-3}	0.0190	5.0000×10^{-3}	0.0364	4.7742×10^{-3}
0.1		$\widehat{m}_{LLK}$					
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0610	2.6875×10^{-3}	0.0790	3.5500×10^{-2}	0.0459	1.5335×10^{-2}
	100	0.0520	1.9447×10^{-3}	0.0750	2.8250×10^{-2}	0.0508	1.0983×10^{-2}
	250	0.0410	1.2994×10^{-3}	0.0770	1.7000×10^{-2}	0.0484	8.1262×10^{-3}
	500	0.0340	1.0249×10^{-3}	0.0700	0.0000×10^{-2}	0.0435	5.3420×10^{-3}

Table 2: Minimum $GMSE$'s of the NW and LLK estimators, and corresponding optimal bandwidths for Model 2. The median and the interquartile range of CV and DPI bandwidth selectors are also reported.

				$\widehat{m}_{NW}$			
τ	n	h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0080	4.9537×10^{-5}	0.0130	9.0000×10^{-3}	0.0171	6.6149×10^{-3}
	100	0.0070	3.4174×10^{-5}	0.0110	6.0000×10^{-3}	0.0168	5.1992×10^{-3}
	250	0.0070	1.8963×10^{-5}	0.0080	2.0000×10^{-3}	0.0176	4.1057×10^{-3}
	500	0.0060	1.1874×10^{-5}	0.0070	2.0000×10^{-3}	0.0168	2.7829×10^{-3}
0.01				$\widehat{m}_{LLK}$			
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0230	2.5103×10^{-5}	0.2400	1.2250×10^{-2}	0.0241	3.8750×10^{-3}
	100	0.0200	1.5843×10^{-5}	0.0220	9.0000×10^{-3}	0.0224	3.5226×10^{-3}
	250	0.0170	8.3326×10^{-6}	0.0190	8.0000×10^{-3}	0.0213	3.2622×10^{-3}
	500	0.0150	5.0942×10^{-6}	0.0165	3.0000×10^{-3}	0.0201	2.0562×10^{-3}
				$\widehat{m}_{NW}$			
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0350	1.7077×10^{-3}	0.0390	1.8000×10^{-2}	0.0395	1.7868×10^{-2}
	100	0.0310	1.1060×10^{-3}	0.0320	1.3000×10^{-2}	0.0377	1.4368×10^{-2}
	250	0.0260	6.3582×10^{-4}	0.0270	9.0000×10^{-3}	0.0397	1.2556×10^{-2}
	500	0.0230	3.9048×10^{-4}	0.0230	6.2500×10^{-3}	0.0397	8.1604×10^{-3}
0.1				$\widehat{m}_{LLK}$			
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0600	1.3355×10^{-3}	0.0680	3.8000×10^{-2}	0.0489	1.5610×10^{-2}
	100	0.0530	8.1070×10^{-4}	0.0600	2.5000×10^{-2}	0.0466	1.4155×10^{-2}
	250	0.0450	4.4170×10^{-4}	0.0520	1.4000×10^{-2}	0.04677	1.1123×10^{-2}
	500	0.0400	2.6572×10^{-4}	0.0470	1.0000×10^{-2}	0.0471	8.8781×10^{-3}

Table 3: Minimum $GMSE$'s of the NW and LLK estimators, and corresponding optimal bandwidths for Model 3. The median and the interquartile range of CV and DPI bandwidth selectors are also reported.

				$\widehat{m}_{NW}$			
τ	n	h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0060	5.3897×10^{-5}	0.0120	8.0000×10^{-3}	0.0179	4.1690×10^{-3}
	100	0.0060	3.9785×10^{-5}	0.0090	4.0000×10^{-3}	0.0178	3.0842×10^{-3}
	250	0.0060	2.2461×10^{-5}	0.0070	2.0000×10^{-3}	0.0164	1.9815×10^{-3}
	500	0.0060	1.4003×10^{-5}	0.0060	2.0000×10^{-3}	0.0148	1.3820×10^{-3}
0.01				$\widehat{m}_{LLK}$			
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0270	2.6537×10^{-5}	0.0290	1.4000×10^{-2}	0.0241	3.4288×10^{-3}
	100	0.0230	1.6545×10^{-5}	0.0250	8.0000×10^{-3}	0.0229	2.2151×10^{-3}
	250	0.0190	8.6117×10^{-6}	0.0210	4.0000×10^{-3}	0.0202	1.3066×10^{-3}
	500	0.0160	5.1441×10^{-6}	0.0180	2.0000×10^{-3}	0.0179	8.9016×10^{-4}
				$\widehat{m}_{NW}$			
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0310	3.2685×10^{-2}	0.0310	2.4000×10^{-2}	0.0409	1.5634×10^{-2}
	100	0.0290	2.3049×10^{-3}	0.0270	1.4000×10^{-2}	0.0426	1.1020×10^{-2}
	250	0.0230	1.5559×10^{-3}	0.0210	9.0000×10^{-3}	0.0392	5.8629×10^{-3}
	500	0.0200	1.1263×10^{-3}	0.0180	6.0000×10^{-3}	0.0353	4.6045×10^{-3}
0.1				$\widehat{m}_{LLK}$			
		h_{GMSE}	$GMSE$	h_{CV}	$IQR(h_{CV})$	h_{DPI}	$IQR(h_{DPI})$
	50	0.0590	2.6569×10^{-3}	0.0730	4.0000×10^{-2}	0.0479	1.4737×10^{-2}
	100	0.0510	1.8975×10^{-3}	0.0670	3.0250×10^{-2}	0.0509	1.1565×10^{-2}
	250	0.0410	1.2763×10^{-3}	0.0650	1.9000×10^{-2}	0.0470	7.1083×10^{-3}
	500	0.0340	9.7927×10^{-4}	0.0640	1.2000×10^{-2}	0.0421	5.1815×10^{-3}

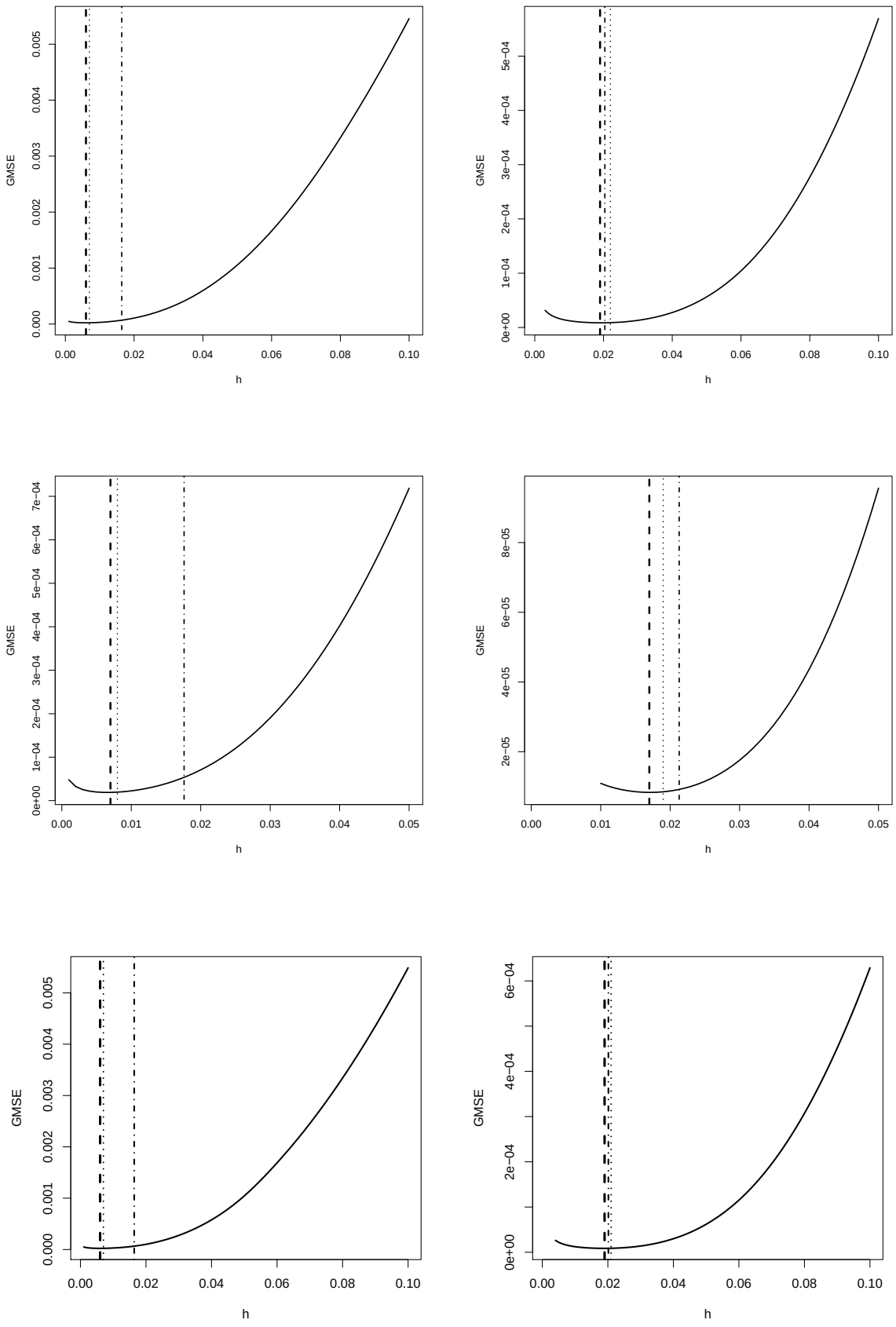

Figure 2: GMSE's for Nadaraya-Watson estimator (left) and local linear kernel estimator (right), with vertical lines representing h_{GMSE} (dashed line), h_{CV} (dotted line) and h_{DPI} (dashed-dotted line), for Models 1 to 3 (from top to bottom), sample size is n=250 and $\tau = 0.01$.

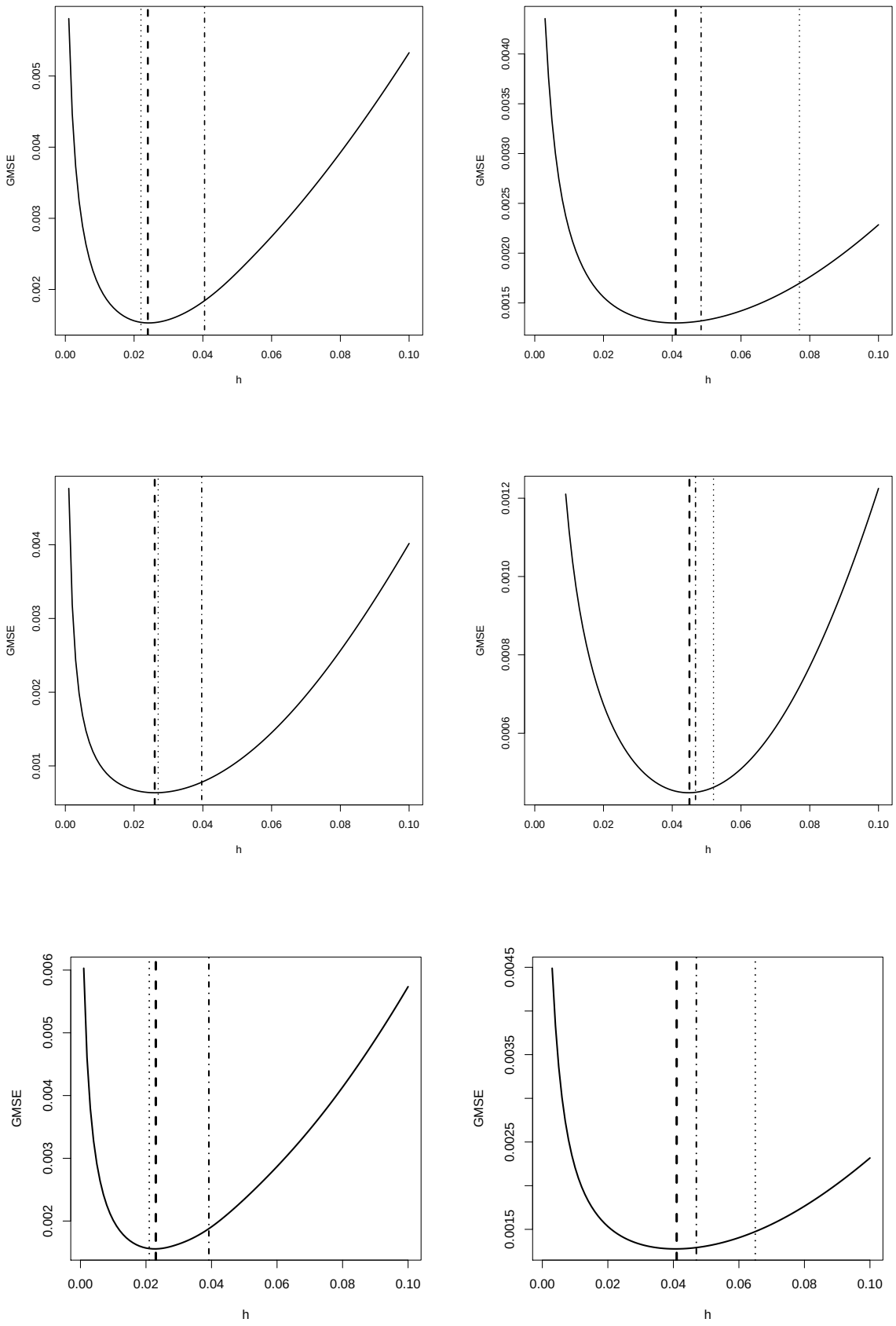

Figure 3: GMSE's for Nadaraya-Watson estimator (left) and local linear kernel estimator (right), with vertical lines representing h_{GMSE} (dashed line), h_{CV} (dotted line) and h_{DPI} (dashed-dotted line), for Models 1 to 3 (from top to bottom), sample size is n=250 and $\tau = 0.1$.

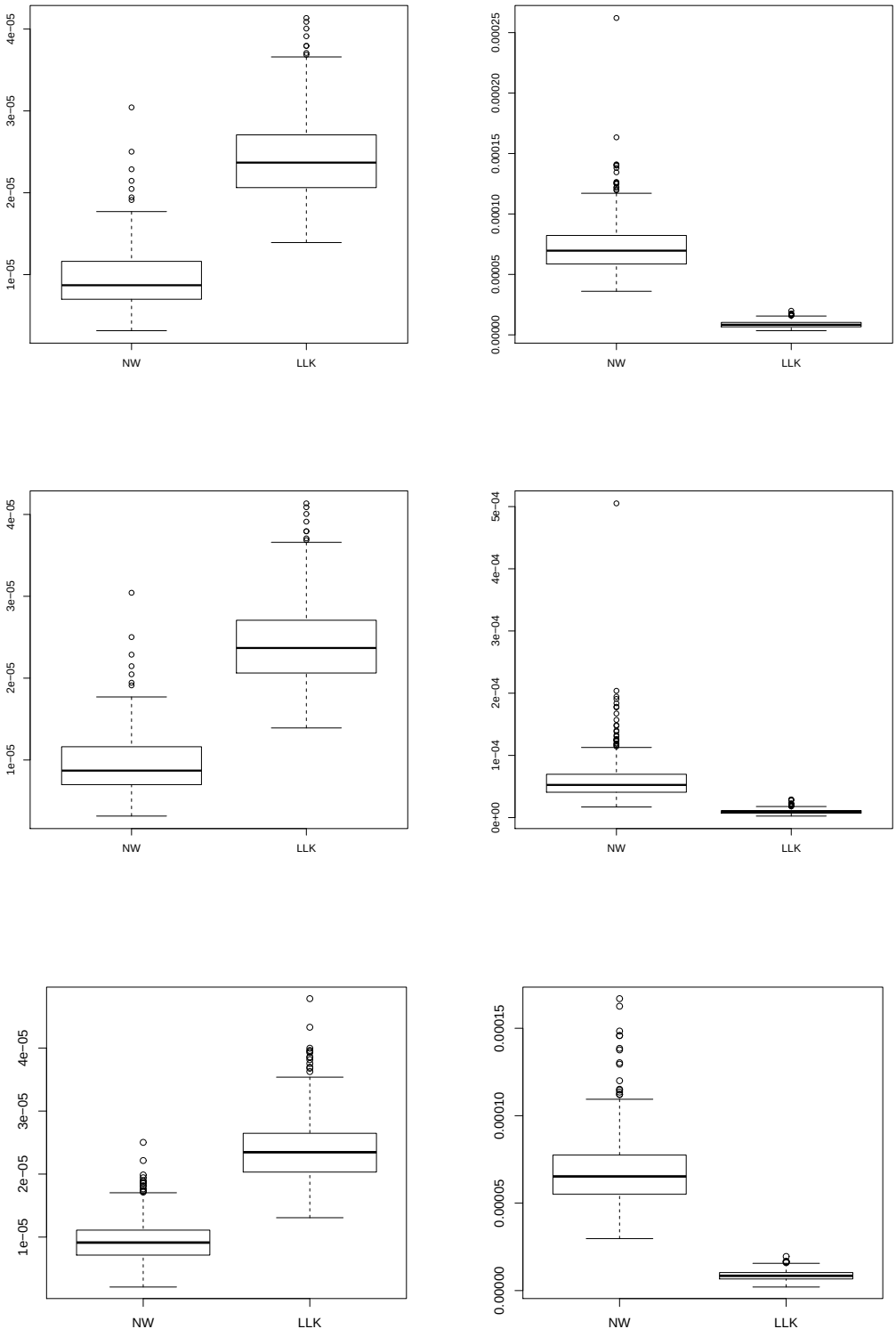

Figure 4: Box-plot of the M=500 mean squared errors obtained using the cross-validation selector (left panels) and direct plug-in selector (right panels) for each estimator (Nadaraya-Watson and local linear kernel), for Models 1 to 3 (from top to bottom), with sample size $n = 250$ and $\tau = 0.01$.

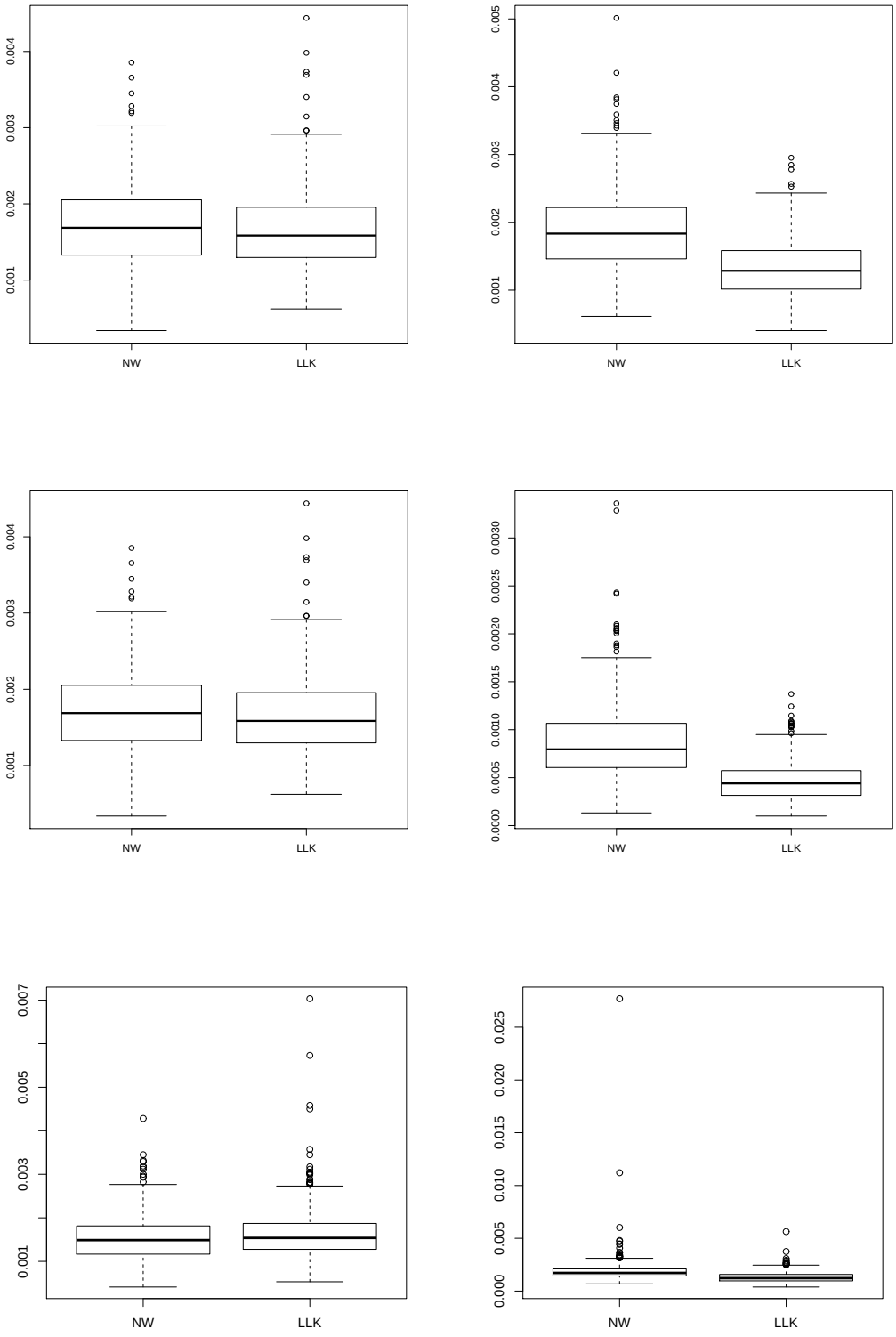

Figure 5: Box-plot of the M=500 mean squared errors obtained using the cross-validation selector (left panels) and direct plug-in selector (right panels) for each estimator (Nadaraya-Watson and local linear kernel), for Models 1 to 3 (from top to bottom), with sample size $n = 250$ and $\tau = 0.1$.

4 Data analysis

For illustration purposes, in this section we consider the analysis of AIDS Blood Transfusion Data and also the quasar luminosities data set.

4.1 AIDS incubation time analysis

In this Subsection we consider epidemiological data on transfusion-related Acquired Immune Deficiency Syndrome (AIDS). The AIDS Blood Transfusion Data are collected by the Centers for Disease Control (CDC), which is from a registry data base, a common source of medical data (see Kalbfleisch and Lawless, 1989; Bilker and Wang, 1986). The variable of interest (X^*) is the induction or incubation time, which is defined as the time elapsed from Human Immunodeficiency virus (HIV) infection to the clinical manifestation of full-blown AIDS. The CDC AIDS Blood Transfusion Data can be viewed as being doubly truncated. The data were retrospectively ascertained for all transfusion-associated AIDS cases in which the diagnosis of AIDS occurred prior to the end of the study, thus leading to right-truncation. Besides, because HIV was unknown prior to 1982, any cases of transfusion-related AIDS before this time would not have been properly classified and thus would have been missed. Thus, in addition to right-truncation, the observed data are also truncated from the left. See Bilker and Wang (1986), section 5.2, for further discussions.

The data include 494 cases reported to the CDC prior to January 1, 1987, and diagnosed prior to July 1, 1986. Of the 494 cases, 295 had consistent data, and the infection could be attributed to a single transfusion or short series of transfusions. Our analyses are restricted to this subset, which is entirely reported in Kalbfleisch and Lawless (1989), Table 1. Values of U^* were obtained by measuring the time from HIV infection to January 1, 1982; while V^* was defined as time from HIV infection to the end of study (July 1, 1986). Note that the difference between V^* and its respective U^* is always 4.5 years.

More specifically, our goal is to study the relationship between AIDS incubation time and age at infection. In Figure 8 we depict the scatterplot of the CDC AIDS blood transfusion data (age at the infection vs. time of incubation) together with the regression function estimators. Figure 8, left, gives the NW estimator computed from the CV and DPI automatic bandwidth selectors. For the CDC AIDS blood transfusion data these bandwidths gave the values 14.4 and 6.12 respectively. Figure 8, right, gives the LLK estimator, for which CV and DPI algorithms gave bandwidths 20 and 7.59 respectively. Overall, the four estimators coincide in that the mean incubation (or induction) time is an increasing-decreasing function of age at infection; this agrees with previous analysis reported for this data set, see e.g. Table 4 in Kalbfleisch and Lawless, 1989.

For comparison purposes, we include in Figure 8 the ordinary Nadaraya-Watson and LLK estimators, which ignore the truncation problem. These estimators were computed by using the respective DPI bandwidth values indicated before. The first thing one sees is that these naive curves underestimate the regression function all along its domain. This is explained by the fact that large incubation times are

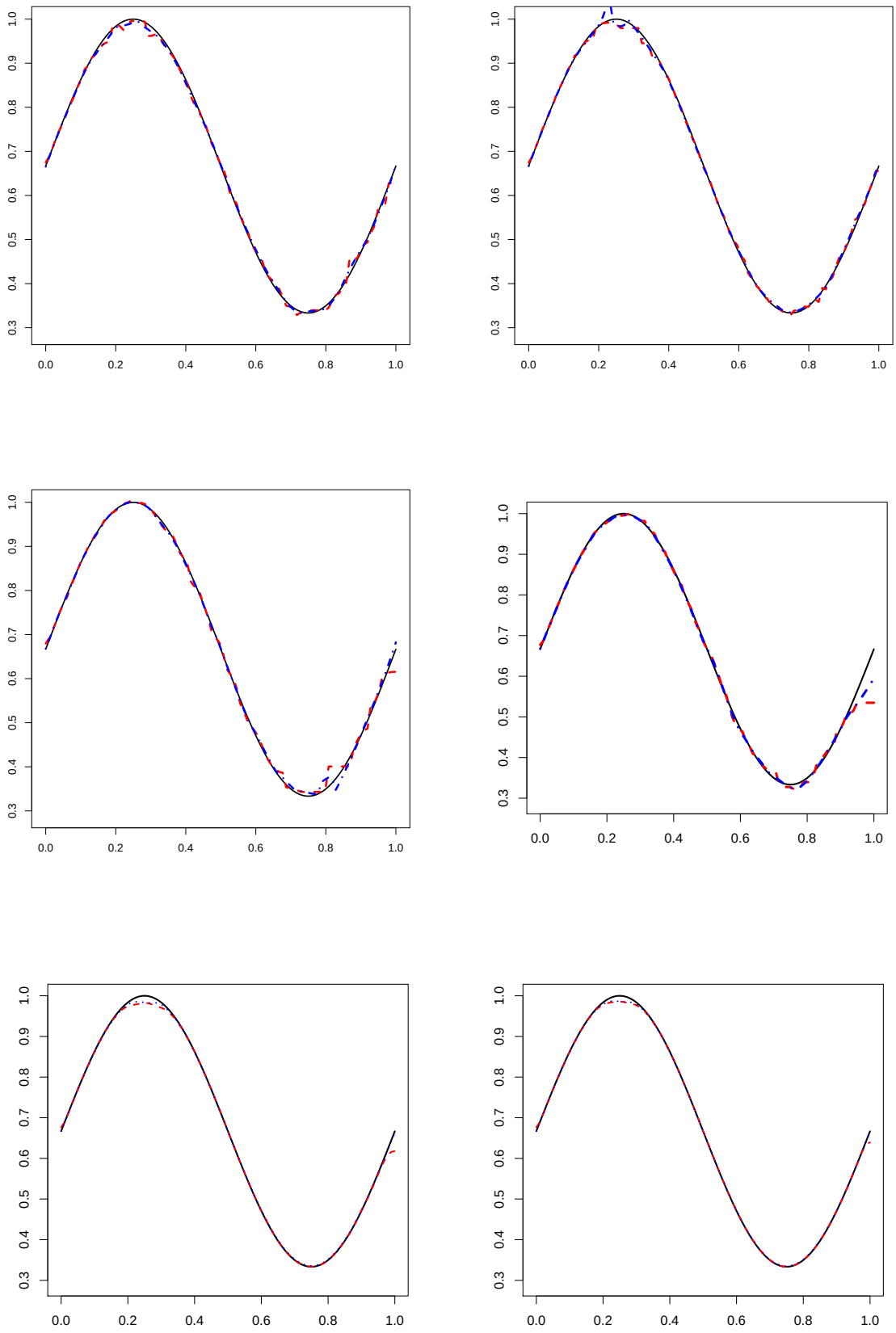

Figure 6: Target regression function (solid line), Nadaraya-Watson estimator (dashed line) and local linear kernel estimator (dashed-dotted line), averaged along 500 Monte Carlo trials, for Models 1 to 3 (top from bottom) and sample sizes $n = 250$ and $n = 500$ (from left to right), with $\tau = 0.01$.

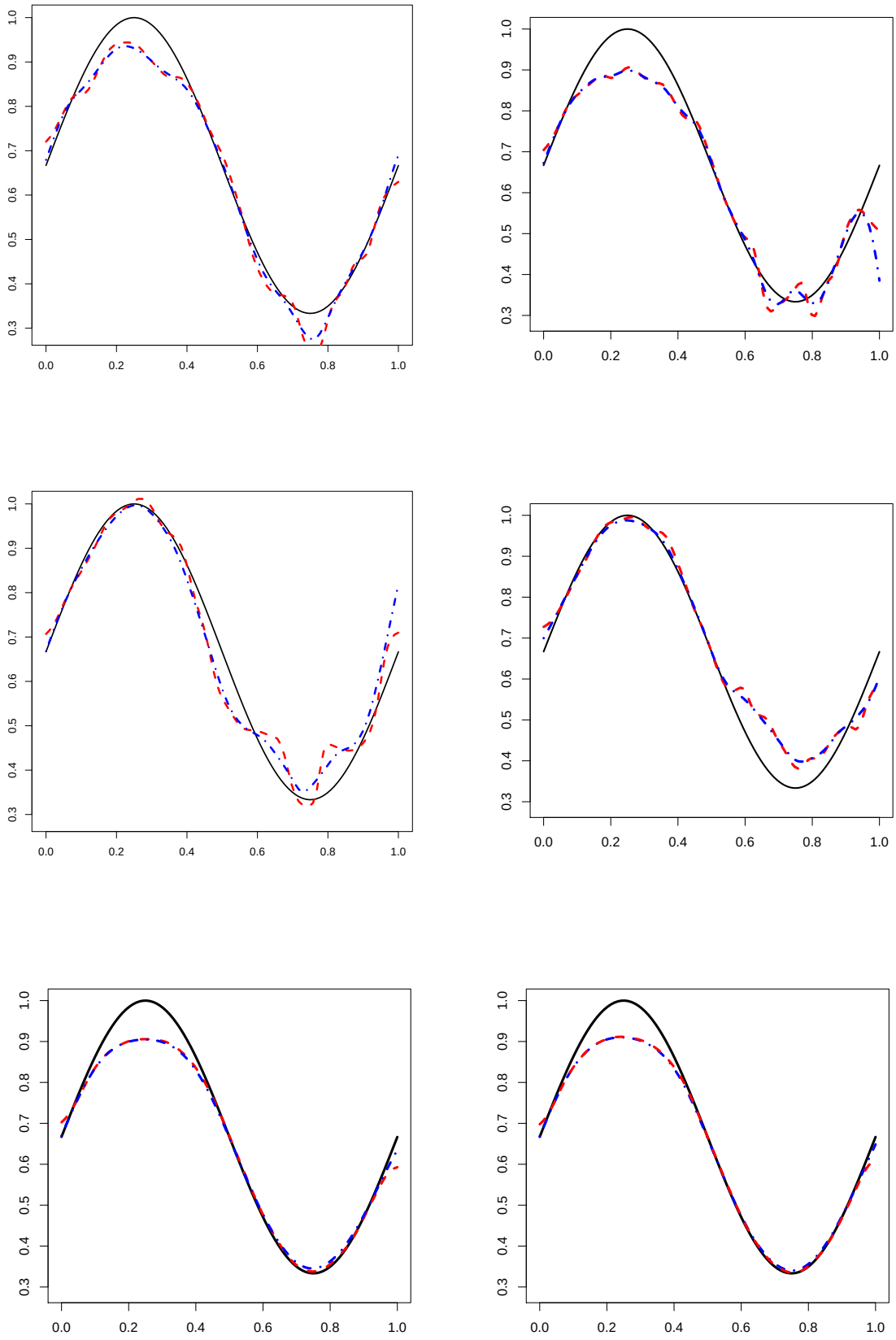

Figure 7: Target regression function (solid line), Nadaraya-Watson estimator (dashed line) and local linear kernel estimator (dashed-dotted line), averaged along 500 Monte Carlo trials, for Models 1 to 3 (top from bottom) and sample sizes $n = 250$ and $n = 500$ (from left to right), with $\tau = 0.1$.

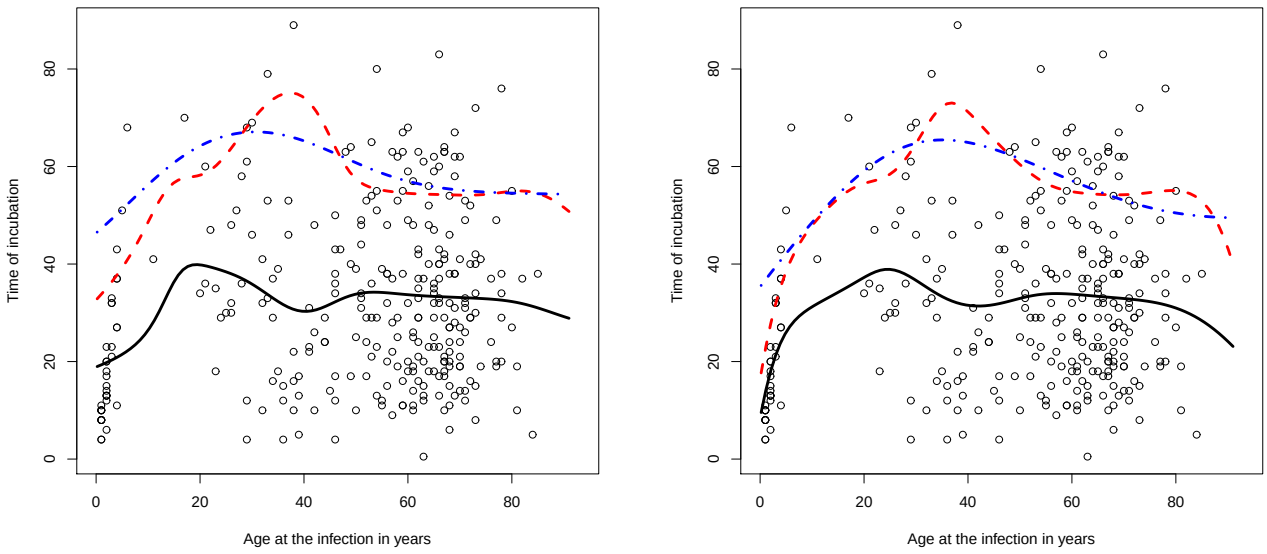

Figure 8: Left panel: NW estimator with different bandwidth selectors: CV bandwidth (h=14.4) (dashed-dotted line), DPI bandwidth (h=6.115472) (dashed line) and its ordinary version (solid line) for random sampling with the same bandwidth as DPI bandwidth. Right panel: LLK estimator with different bandwidth selectors: CV bandwidth (h=20) (dashed-dotted line), DPI bandwidth (h=7.592714) (dashed line) and its ordinary version (solid line) for random sampling with the same bandwidth as DPI bandwidth. AIDS Blood Transfusion data.

observed with smaller probability; see Figure 9, in which the function G_n for the AIDS Blood Transfusion data is depicted. Secondly, the ordinary estimators of $m(x)$ suggest a flat (or even a slightly decreasing) shape of the regression function for $20 < x < 80$, where the true $m(x)$ is concave (according to the corrected estimators). An explanation for this is that intermediate ages are associated to the largest incubation times, which are under-sampled, and therefore the scatterplot gets empty at its top-central part, where the pick of $m(x)$ is located. Summarizing, it is very important to perform a correction for the double truncation issue in regression analysis.

4.2 Quasar luminosities

In this Subsection we consider astronomical data on the quasars luminosity. In Astronomy, one of the main goals of the quasar investigations is to study luminosity evolution. The motivating example presented in the paper of Efron and Petrosian (1999) concerns a set of measurements on quasars in which there is double truncation because the quasars are observed only if their luminosity occurs within a certain finite interval, bounded at both ends, determined by limits of detection.

The original data set studied by Efron and Petrosian (1999), comprised independently collected quadruplets (z_i, m_i, a_i, b_i), $i = 1, \ldots, n$, where z_i is the redshift of the ith quasar and m_i is the ap-

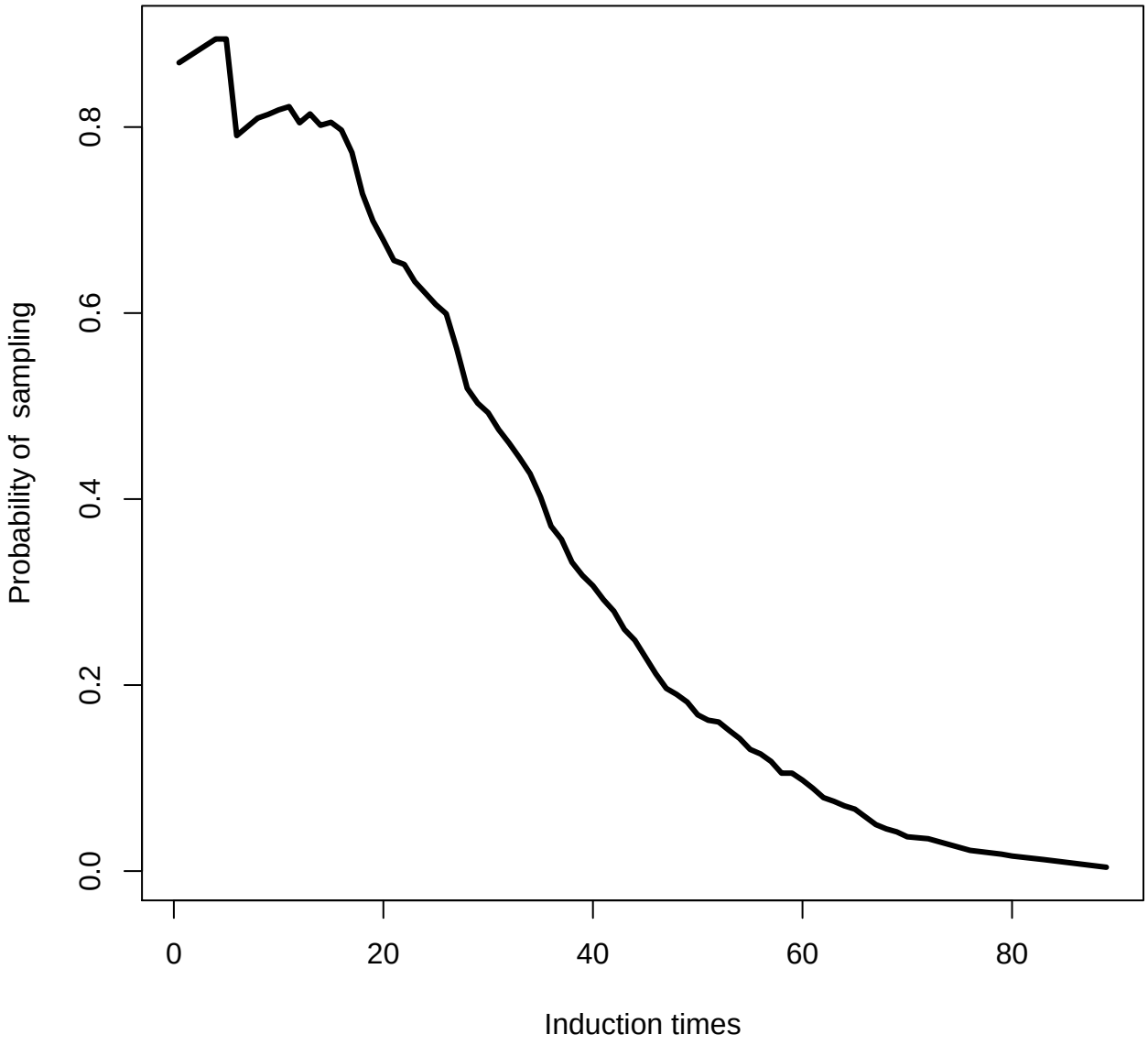

Figure 9: Biasing function for the CDC AIDS blood transfusion data.

parent magnitude. Due to experimental constraints, the distribution of each luminosity in the log-scale ($y_i = t(z_i, m_i)$) is truncated to a known interval $[a_i, b_i]$, where t represents a transformation which depends on the cosmological model assumed (see Efron and Petrosian (1999) for details). Quasars with apparent magnitude above b_i were too dim to yield dependent redshifts, and hence they were excluded from the study. The lower limit a_i was used to avoid confusion with non quasar stellar objects. The $n = 210$ quadruplets investigated by Efron and Petrosian (1999) are included in `DTDA R` package presented in Moreira et al. (2010).

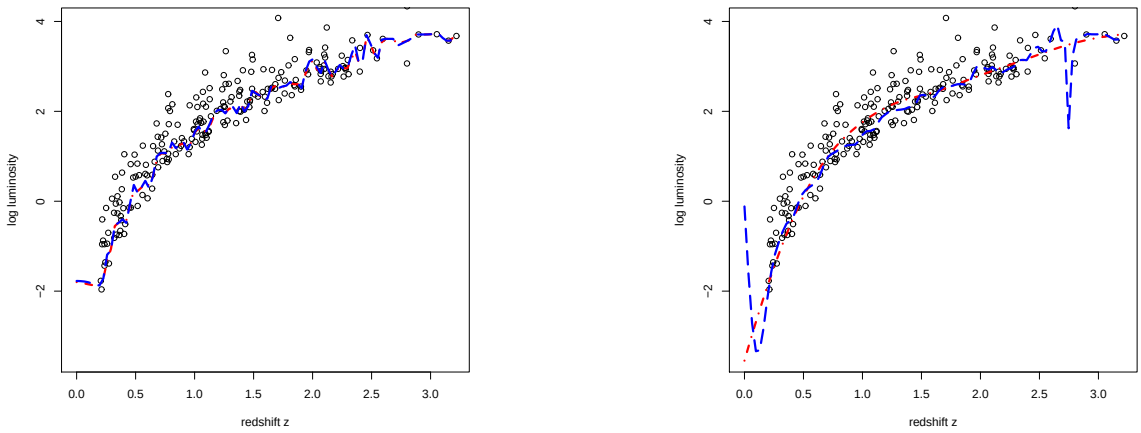

Figure 10: Left panel: NW estimator with different bandwidth selectors: CV bandwidth (h=0.02) (dashed-dotted line) and DPI bandwidth (h=0.01621) (dashed line). Right panel: LLK estimator with different bandwidth selectors: CV bandwidth (h=0.3) (dashed-dotted line) and direct plug-in bandwidth (h=0.03613) (dashed line). Log-luminosity quasar data.

In Figure 10 we depict the scatterplot of the quasar data (redshift vs. log-luminosity) together with the regression function estimators. Figure 10, left, gives the NW estimator computed from the CV and DPI automatic bandwidth selectors. For the quasar data these bandwidths gave the values 0.02 and 0.01621. Figure 10, right, gives the LLK estimator for which CV and DPI algorithms gave bandwidths 0.3 and 0.03613 respectively. In this application the more reasonable estimator seems to be the LLK based on the cross-validation bandwidth, the other estimators providing wiggly curves.

In Figure 11 we report the estimated biasing function G_n for que quasar data, while in Figure 12 we compare the LLK estimator adapted to double truncation and its ordinary version for random sampling. For this comparison two bandwidths were used: the CV bandwidth ($h = 0.3$) and a slightly smoother estimator ($h = 0.4$). The biasing function in Figure 11 indicates that small quasar luminosities are observed with a very small relative probability. This results in a biased observation of the true regression function at the left corner of the plot visible in Figure 12 for ($h = 0.3, 0.4$), where the luminosity is expected to be small. At this corner, the observed scatterplot is shifted up, with respect to the scatterplot one would have observed under random sampling. By the same reason, since the intermediate values of

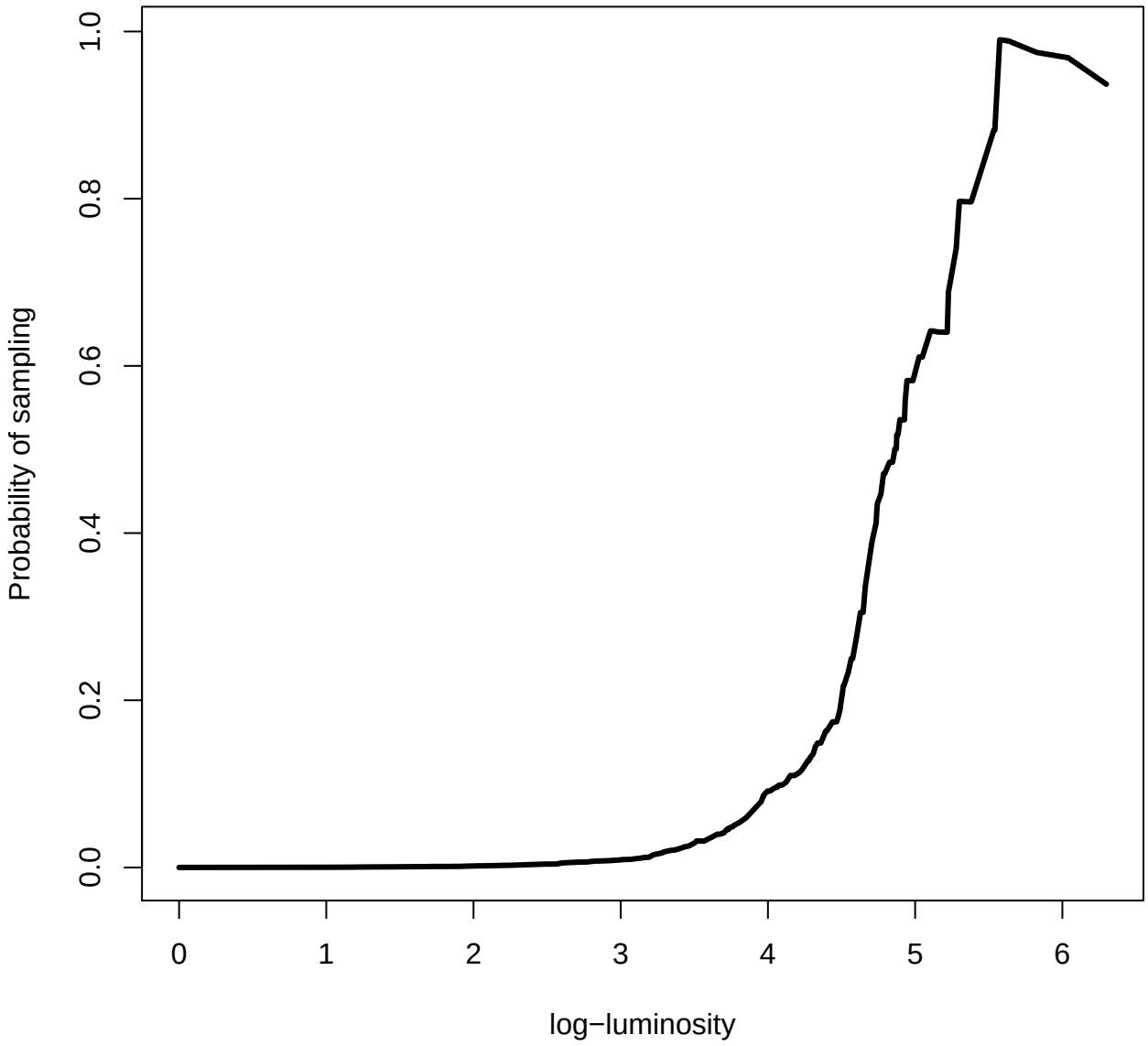

Figure 11: Biasing function for the quasar data.

the response are observed with a relatively larger probability, the scatterplot is shifted down at its central part, something which becomes evident when using the largest bandwidth ($h = 0.4$). In sum, one may say that it is important to use a correction for double truncation when performing a regression analysis.

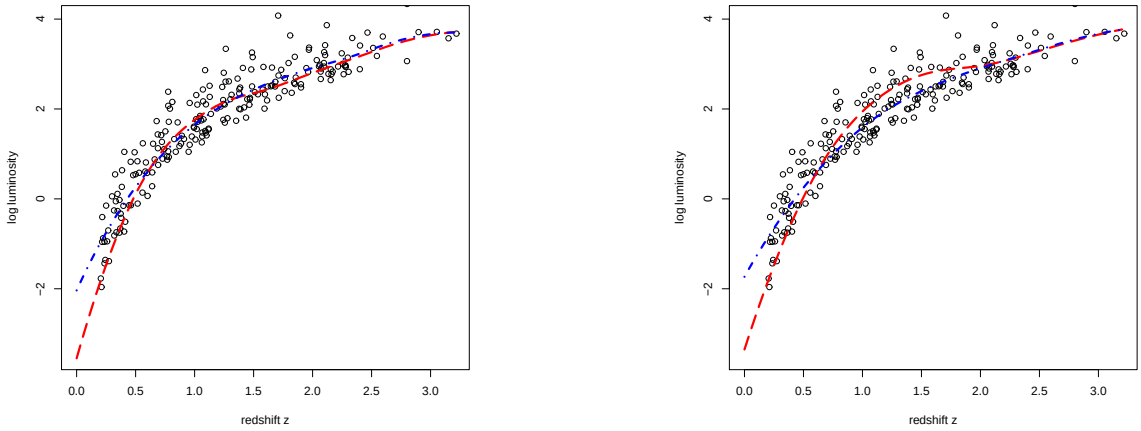

Figure 12: Local linear kernel regression estimator adapted to double truncation (dashed line) and its ordinary version (dashed-dotted line) for random sampling with two bandwidths: the CV bandwidth $h = 0.3$ (left) and a slightly smoother estimator $h = 0.4$ (right).

5 Conclusions

In this paper we have proposed two different nonparametric estimators for a regression function when the response is subject to random double truncation. The proposed estimators are proper adaptations of the Nadaraya-Watson and local linear kernel smoothers to the truncated setup. Without such an adaptation, the ordinary kernel smoothers would be systematically biased.

Asymptotic expressions for the bias and the variance of the estimators have been obtained, and two different bandwidth selectors based on cross-validation and plug-in ideas have been introduced. The practical performance of the estimators and the bandwidth selectors has been investigated through simulations, and a real data analysis has been provided for further illustration. It has been demonstrated that both estimators and both bandwidths selectors perform well, approaching to their targets as the sample size increases. Besides, when comparing the several estimators, the local linear kernel estimator based on the direct plug-in bandwidth seems to be the best one. The estimators, however, may be inconsistent around covariate values for which full observation of the response is not possible. This happens when the truncation skips a relevant part of the support of the response, so its conditional distribution can not be reconstructed. These possible inconsistencies have been illustrated in the simulation study too.

Although the results have been established for a single covariate, similar results can be provided for the multivariate setting. However, as always with nonparametric regression, the practical performance of the estimators will become poorer as the dimension grows. To this regard, the application of semiparametric regression techniques (such as e.g. additive regression) to the randomly truncated scenario would be interesting. Another issue which is left for future research is the construction of confidence bands for the

regression function.

6 Appendix: Technical proofs and details

Proof to Theorem 2.1

In this section we prove the asymptotic expressions given in Theorem 2.1. Note that these expressions refer to the local polynomial kernel estimator $\hat{m}_{(p)}(x)$, where the case $p = 0$ corresponds to the Nadaraya-Watson estimator (NW), and $p = 1$ corresponds to the local linear kernel estimator (LLK). Note that Theorem 1 is similar to Theorem 2.1 in Sköld (1999), where the biasing function here (G_n) is random but independent of the covariate. Note also that conditions (C1)-(C4) imply those in Sköld (1999). As mentioned in that paper, the result for NW follows by a linearization and standard Taylor-arguments (once G_n is replaced by the true biasing function G), while for LLK one may apply techniques as for no truncated data (e.g. Wand and Jones, 1995). For illustrative purposes, we give here a sketch of the proof for the NW estimator.

Consider the Nadaraya-Watson estimator:

$\widehat{m}_{NW}(x) = \frac{\widehat{m}^*(x)}{\widehat{\alpha}^*(x)}$

where $\widehat{m}^*(x) = \widehat{m}^*_{NW}(x)$ and $\widehat{\alpha}^*(x) = \widehat{\alpha}^*_{NW}(x)$. Put for simplicity $\widehat{m}(x) = \widehat{m}_{NW}(x)$. We have:

$$\begin{aligned}
\widehat{m}(x) - m(x) &= \frac{\widehat{m}^*(x)}{\widehat{\alpha}^*(x)} - \frac{m^*(x)}{\alpha^*(x)} = \\
&= \frac{1}{\widehat{\alpha}^*(x)} \left(\widehat{m}^*(x) - m^*(x)\right) + m^*(x) \left(\frac{1}{\widehat{\alpha}^*(x)} - \frac{1}{\alpha^*(x)}\right) = \\
&= \frac{1}{\alpha^*(x)} \left(\widehat{m}^*(x) - m^*(x)\right) + \left(\frac{1}{\widehat{\alpha}^*(x)} - \frac{1}{\alpha^*(x)}\right) \left(\widehat{m}^*(x) - m^*(x)\right) - \\
&- \frac{m^*(x)}{\widehat{\alpha}^*(x)\alpha^*(x)} \left(\widehat{\alpha}^*(x) - \alpha^*(x)\right) = \\
&= \frac{1}{\alpha^*(x)} \left(\widehat{m}^*(x) - m^*(x)\right) - \frac{m^*(x)}{\alpha^*(x)^2} \left(\widehat{\alpha}^*(x) - \alpha^*(x)\right) + R(x)
\end{aligned}$$

where

$$\begin{aligned}
R(x) &= \left(\frac{1}{\widehat{\alpha}^*(x)} - \frac{1}{\alpha^*(x)}\right) \left(\widehat{m}^*(x) - m^*(x)\right) + \frac{m^*(x)}{\alpha^*(x)} \left(\widehat{\alpha}^*(x) - \alpha^*(x)\right) \left(\frac{1}{\alpha^*(x)} - \frac{1}{\widehat{\alpha}^*(x)}\right) = \\
&= \left(\frac{1}{\widehat{\alpha}^*(x)} - \frac{1}{\alpha^*(x)}\right) \left[\widehat{m}^*(x) - m^*(x) + \frac{m^*(x)}{\alpha^*(x)} \left(\widehat{\alpha}^*(x) - \alpha^*(x)\right)\right].
\end{aligned}$$

This shows the asymptotic equivalence $\widehat{m}(x) - m(x) \sim \hat{\phi}(x)$ where

$$\widehat{\phi}(x) = \frac{1}{\alpha^*(x)} \left(\widehat{m}^*(x) - m^*(x)\right) - \frac{m^*(x)}{\alpha^*(x)^2} \left(\widehat{\alpha}^*(x) - \alpha^*(x)\right).$$

Since G_n is a $\sqrt{n}-$consistent estimator of G (Shen, 2010; Moreira and de Uña-Álvarez, 2012), and since both $\widehat{m}^*(x)$ and $\widehat{\alpha}^*(x)$ are only $\sqrt{nh}-$consistent estimators, one may replace G_n by G in the formulae to compute the asymptotic bias and variance in a simple way. This leads to the well-known asymptotics for the NW-type estimators $\widehat{m}^*(x)$ and $\widehat{\alpha}^*(x)$, namely (cfr. Härdle et al., 2004)

$$\begin{aligned} E\widehat{m}^*(x) - m^*(x) &\sim \frac{1}{2}\mu_2(K)h^2 \frac{1}{f^*(x)} \left[m^{*\prime\prime}(x)f^*(x) + 2m^{*\prime}(x)f^{*\prime}(x)\right], \\ E\widehat{\alpha}^*(x) - \alpha^*(x) &\sim \frac{1}{2}\mu_2(K)h^2 \frac{1}{f^*(x)} \left[\alpha^{*\prime\prime}(x)f^*(x) + 2\alpha^{*\prime}(x)f^{*\prime}(x)\right] \end{aligned}$$

(for the biases), and

$$\begin{aligned} &Var(\widehat{m}^*(x)) \sim (nh)^{-1}R(K)\tfrac{Var[Y_1^2 G(Y_1)^{-1}|X_1=x]}{f^*(x)}, \\ &Var(\widehat{\alpha}^*(x)) \sim (nh)^{-1}R(K)\tfrac{Var[G(Y_1)^{-1}|X_1=x]}{f^*(x)}, \\ &Cov\left(\widehat{m}^*(x), \widehat{\alpha}^*(x)\right) \sim (nh)^{-1}R(K)\tfrac{E[Y_1 G(Y_1)^{-2}|X_1=x] - m^*(x)\alpha^*(x)}{f^*(x)} \end{aligned}$$

(for the variances and covariance), where $f^*(x)$ stands for the density of the observed covariate (X_1). Therefore, we obtain as $n \to \infty$

$$E[\widehat{\phi}(x)] \sim \tfrac{1}{2}\mu_2(K)h^2 B_0(x)$$

where

$$B_0(x) = \frac{1}{\alpha^*(x)} \frac{1}{f^*(x)} \left[m^{*\prime\prime}(x)f^*(x) + 2m^{*\prime}(x)f^{*\prime}(x)\right] - \frac{m^*(x)}{\alpha^*(x)^2} \frac{1}{f^*(x)} \left[\alpha^{*\prime\prime}(x)f^*(x) + 2\alpha^{*\prime}(x)f^{*\prime}(x)\right]$$

and

$$
\begin{aligned}
Var\ \widehat{\phi}(x) &= \frac{Var\ (\widehat{m}^*(x))}{\alpha^*(x)^2} + \frac{m^*(x)^2}{\alpha^*(x)^4} Var\ (\widehat{\alpha}^*(x)) - 2\frac{m^*(x)}{\alpha^*(x)^3} Cov\,(\widehat{m}^*(x), \widehat{\alpha}^*(x)) \\
&\sim (nh)^{-1} R(K) \frac{1}{f^*(x)} \left\{ \frac{Var(Y_1 G(Y_1)^{-1} | X_1 = x)}{\alpha^*(x)^2} + \frac{Var(G(Y_1)^{-1} | X_1 = x)}{\alpha^*(x)^4} m^*(x)^2 \right. \\
&- \left. 2\frac{m^*(x)}{\alpha^*(x)^3} \left[E(Y_1 G(Y_1)^{-2} | X_1 = x) - m^*(x)\alpha^*(x) \right] \right\} \\
&= (nh)^{-1} R(K) \frac{1}{f^*(x)} \left\{ E\left[\frac{Y_1^2 G(Y_1)^{-2}}{\alpha^*(x)^2} | X_1 = x \right] + E\left[\frac{G(Y_1)^{-2}}{\alpha^*(x)^4} m^*(x)^2 | X_1 = x \right] \right. \\
&- \left. 2E\left[\frac{Y_1 G(Y_1)^{-2}}{\alpha^*(x)^3} m^*(x) | X_1 = x \right] \right\} \\
&= (nh)^{-1} R(K) \frac{1}{f^*(x)\alpha^*(x)^2} E\left[(Y_1 - m(X_1))^2 G(Y_1)^{-2} | X_1 = x \right].
\end{aligned}
$$

Now, it is easily seen that the densities of X^* ($f(x)$) and X_1 ($f^*(x)$) are linked through $\alpha^{-1} f(x) = \alpha^*(x) f^*(x)$. Compute the second derivative of both sides of the equation to get

$$\alpha^{*\prime\prime}(x) f^*(x) + 2\alpha^{*\prime}(x) f^{*\prime}(x) = \alpha^{-1} f''(x) - \alpha^*(x) f^{*\prime\prime}(x).$$

Similarly, compute the second derivative of equation $m^*(x) f^*(x) = \alpha^{-1} m(x) f(x)$ to get

$$m^{*\prime\prime}(x) f^*(x) + 2m^{*\prime}(x) f^{*\prime}(x) = \alpha^{-1}(m''(x) f(x) + 2m'(x) f'(x)) + \alpha^{-1} m(x) f''(x) - m^*(x) f^{*\prime\prime}(x).$$

From these relationships we get $B_0(x) = [m''(x) f(x) + 2m'(x) f'(x)]/f(x)$ and the result on the asymptotic bias of the NW estimator is obtained. For the variance, just note
$\alpha\sigma^2(x)/f(x) = E\left[(Y_1 - m(X_1))^2 G(Y_1)^{-2} | X_1 = x \right] / f^*(x)\alpha^*(x)^2$ to conclude.

DPI bandwidth

Histograms are constructed by first partitioning the design interval $[a, b]$ into interval blocks $\mathcal{B}_j, j = 1, \ldots, N$. In this paper we have always use $N = 3$, as suggested by Härdle and Marron (1995). Let $\mathcal{B}$ denote a generic block $\mathcal{B}_j$, and let r and l denote the right and left boundaries of this block. The proportion of X_i's falling in each interval reflects the height of the density near the center of the block. Let $c = \frac{r+l}{2}$ denote the blockcenter and $r_b = \frac{r-l}{2}$ denote the block radius. The histogram density estimate (adapted to double truncation) is given by

$$\widehat{f}(c) = \frac{1}{2\sum_{i=1}^{n} G_n(Y_i)^{-1} r_b} \sum_{i=1}^{n} I(|c - X_i| \leq r_b) G_n(Y_i)^{-1}.$$

To estimate the derivative of f on $\mathcal{B}$ we use a simple differencing method. This requires two estimates of

f so we split the block into left and right halves. Estimate the frequencies on each half of the block by

$$n_l = n \sum_{i=1}^{n} I(l \leq X_i \leq c) \frac{G_n(Y_i)^{-1}}{\sum_{j=1}^{n} G_n(Y_j)^{-1}}$$

and

$$n_r = n \sum_{i=1}^{n} I(c \leq X_i \leq r) \frac{G_n(Y_i)^{-1}}{\sum_{j=1}^{n} G_n(Y_j)^{-1}}.$$

Forming a difference quotient based on histograms at the center of the two subblocks gives the derivative estimate

$$\widehat{f'}(c) = \frac{(n_r - n_l)/(nr_b)}{r_b}.$$

These two estimates are combined into the score function $\widehat{\frac{f'}{f}}(c)$ and, together with estimates of m' and m'', they are used to construct an estimate of $\int B^2(x)dx$. A natural and straight forward method for estimating m, m' and m'', is least-squares polynomial regression. The simplest version of this is a parabola, and to avoid lost of flexibility (see Härdle and Marron, 1995 for more details), we fit blockwise parabolas. To deal with the double truncation issue, the squares in the least-squares criterion are weighted by the $1/G_n(Y_i)$'s. Explicitly, by assuming $m(x) = \beta_1 + \beta_2 x + \beta_3 x^2$ locally, one sets $\widehat{B}(c_j) = 2\widehat{\beta}_{3j} + 2\left[2\widehat{\beta}_{3j}(c_j) + \hat{\beta}_{2j}\right] \widehat{\left(\frac{f'}{f}\right)}(c_j)$, where c_j is the center of the block $\mathcal{B}_j$. The final estimate of $B_2 = \int B^2(x)dx$ is given by $\widehat{B}_2 = \sum_{j=1}^{n} 2r_b \widehat{B}^2(c_j)$.

The estimation of $V(x)$ in (2.4) requires an estimator for $\left(\frac{\sigma^2(x)}{f(x)}\right)$. An estimate of this value at the center of a generic block $\mathcal{B}$ is given by

$$\widehat{\left(\frac{\sigma^2}{f}\right)}(c) = \frac{\widehat{\sigma}^2(c)}{\widehat{f}(c)},$$

where $\widehat{\sigma}^2(c) = \frac{1}{\sum_{X_i \in \mathcal{B}} G_n(Y_i)^{-1}} \sum_{X_i \in \mathcal{B}} (Y_i - \widehat{m}(X_i))^2 G_n(Y_i)^{-2}$, which leads to $\widehat{V}(c) = \widehat{\alpha} \widehat{\left(\frac{\sigma 2}{f}\right)}(c)(2\pi^{1/2})^{-1}$ when K is the Gaussian kernel and $\widehat{\alpha} = \left(\sum_{i=1}^{n} G_n(Y_i)^{-1}\right)^{-1}$. The blockwise approach (based on local parabolic pilot fits) leads to the following estimate of $V = \int V(x)dx$:

$$\widehat{V} = \sum_{j=1}^{n} 2r_b \widehat{V}(c_j).$$

Finally, the DPI bandwidth is computed as $h_{DPI} = (\widehat{V}/(4\widehat{B}_2 n))^{(1/5)}$.

References

Akritas, M. G. and M. P. LaValley (2005). A generalized product-limit estimator for truncated data. *Journal of Nonparametric Statistics 17*, 643–663.

Cristóbal, J. and T. Alcalá (2000). Nonparametric regression estimators for length biased data. *Journal of Statistical Planning and Inference 89*, 145–168.

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

Gross, S. T. and T. L. Lai (1996). Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association 91*, 1166–1180.

Härdle, W. and J. Marron (1995). Fast and simple scatterplot smoothing. *Computational Statistics & Data Analysis 20*, 1–17.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004). *Nonparametric and Semiparametric Models*. Berlin: Springer.

Iglesias-Pérez, M. and W. González-Manteiga (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics 10*, 213–244.

Kalbfleisch, J. D. and J. F. Lawless (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related aids. *American Statistical Association 84*, 360–372.

Klein, J. P. and M. L. Moeshberger (2003). *Survival Analysis. Techniques for Censored and Truncated Data*. New York: Springer.

Liang, H., J. de Uña-Álvarez, and M. Iglesias-Pérez (2011). Local polynomial estimation of a conditional mean function with dependent truncated data. *Test 20*, 653–677.

Martin, E. and R. A. Betensky (2005). Testing quasi-independence of failure and truncation times via conditional kendall's tau. *Journal of the American Statistical Association 100*, 484–492.

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. and J. de Uña-Álvarez (2012). Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics 6*, 501–521.

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.

Ould-Saïd, E. and M. Lemdani (2006). Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. *Annals of the Institute of Statistical Mathematics 58*, 357–378.

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

Sköld, M. (1999). Kernel regression in the presence of size-bias. *Journal of Nonparametric Statistics 12*, 41–51.

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.

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

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