INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/06

Bandwidth Selection for Kernel Density Estimation with Doubly Truncated Data

MOREIRA, C. and I. VAN KEILEGOM

Bandwidth Selection for Kernel Density Estimation with Doubly Truncated Data

Carla Moreira [*] Ingrid Van Keilegom [§]

February 23, 2012

Abstract

In this work we introduce and compare several bandwidth selection procedures for kernel density estimation of a random variable that is sampled under random double truncation. The work is motivated by the fact that this type of incomplete data is often encountered in studies in astronomy and medicine. The bandwidth selection procedures we study are appropriate modifications of the normal reference rule, the least squares cross-validation procedure, two types of plug-in procedures, and a bootstrap based method. The methods are first shown to work from a theoretical point of view. A simulation study is then carried out to assess the finite sample behavior of these five bandwidth selectors. We also illustrate the use of the various practical bandwidth selectors by means of data regarding the luminosity of quasars in astronomy.

Key Words: Bandwidth selection; bootstrap; cross-validation; double truncation; kernel density estimation; normal reference rule; plug-in.

[*]University of Vigo, Lagoas - Marcosende, 36 310 Vigo, Spain, E-mail address: carlamgmm@gmail.com. The research has been carried out during a visit at the Université catholique de Louvain, Belgium.

[§]Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. E-mail address: ingrid.vankeilegom@uclouvain.be. Research supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy), by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

1 Introduction

This paper is concerned with the problem of how to select the bandwidth needed in kernel density estimation of a random variable that is sampled under random double truncation.

Randomly truncated data appear in a variety of fields, including astronomy, medicine, epidemiology and economics. Under random truncation, only values falling in a random set which varies across individuals are observed. For the recorded values, the truncation set is also observed. However, when the value of interest falls outside of the corresponding random set, nothing is observed. This issue typically introduces an observational bias, and hence proper corrections in statistical data analysis and inference are needed.

Nonparametric methods for one-sided (left or right) truncated data were introduced in the seminal paper Lynden-Bell (1971), see also Stute (1993) and Woodroofe (1985). Some authors have pointed out that the available information on the truncation time allows to construct more efficient estimators. See for example Wang (1989), Asgharian et al. (2002) and de Uña-Álvarez (2004).

When the data are subject to double truncation, the literature on nonparametric methods 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 both the theoretical developments and the practical implementation difficult. The first paper on nonparametric maximum likelihood estimation (NPMLE) of the distribution function under double truncation appeared in 1999 (Efron and Petrosian, 1999), and was motivated by a data set on quasars in astronomy. Shen (2010a) formally established the uniform strong consistency and the asymptotic normality of the NPMLE, while 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 literature also contains semiparametric approaches to estimate the distribution function under double truncation. Moreira and de Uña-Álvarez (2010b) investigated this problem when the distribution of the truncation times is assumed to belong to a given parametric family, see also Shen (2010b).

The estimation of a density function in the presence of random double truncation was introduced by Moreira and de Uña-Álvarez (2011). The authors proposed and studied both a nonparametric and a semiparametric estimator, and they also explored the asymptotic properties of the proposed estimators. The estimators are obtained as a convolution between a kernel function and one of the estimators of the cumulative distribution function mentioned above (namely Efron and Petrosian, 1999 for the nonparametric approach and Moreira and de Uña-Álvarez, 2010b for the semiparametric one).

Our aim in this paper is to propose and compare several automatic bandwidth selec-

tion procedures for the kernel density estimators introduced by Moreira and de Uña-Álvarez (2011). The procedures we study are appropriate modifications of the normal reference rule, the least squares cross-validation procedure, two types of plug-in procedures, and a bootstrap based method. The methods are first shown to work from a theoretical point of view. A simulation study is then carried out to assess the finite sample behavior of these bandwidth selectors. Although these selection procedures are well studied in the literature for completely observed data, their theoretical and practical behavior is quite different for doubly truncated data, and to the best of our knowledge, they have never been previously studied in the literature. Note that, from a historic point of view, the literature on bandwidth selection for kernel density estimation has, after a period of a clear preponderance of cross-validation methods, expanded in several directions, which ranged from a revival of the classical plug-in methods (see, e.g. Sheather and Jones, 1991) to the development of bootstrap-motivated techniques, see e.g. Cao et al. (1994) for completely observed data and Sánchez-Sellero et al. (1999), who investigated the bootstrap methodology for left-truncated and right-censored (LTRC) data.

The paper is organized as follows. In the next section we revisit the kernel density estimators proposed by Moreira and de Uña-Álvarez (2011). In Section 3 we present our five bandwidth selection procedures, and we give a theoretical justification for their definition. The finite sample performance of these methods is studied in Section 4 via a simulation study. In Section 5 we illustrate the use of the bandwidth selection methods by means of data on quasars in astronomy. The main conclusions of our study are summarized in Section 6.

2 Kernel density estimation with doubly truncated data

Let X^* be the random variable of interest with distribution function F, and assume that it is doubly truncated by the random pair (U^*, V^*) with joint distribution H, 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 $X^* > V^*$. We assume that X^* is independent of (U^*, V^*). Let (U_i, X_i, V_i), $i = 1, ..., n$, denote the sampling information, these are i.i.d. data with the same distribution as (U^*, X^*, V^*) given $U^* \leq X^* \leq V^*$. Introduce $\alpha = P(U^* \leq X^* \leq V^*)$, the probability of no-truncation. For any distribution 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 $H_1(u) = H(u, \infty)$ and $H_2(v) = H(-\infty, v)$ be the marginal distribution functions of U^* and V^*, respectively. When $a_{H_1} \leq a_F \leq a_{H_2}$ and $b_{H_1} \leq b_F \leq b_{H_2}$, F and H are both identifiable

(see Woodroofe, 1985).

To define the nonparametric kernel density estimator, proposed by Moreira and de Uña-Álvarez (2011), we first need to introduce the NPMLE of the distribution function of X^* (Efron and Petrosian, 1999). Let (we use the convention $0/0 = 0$)

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

where $\alpha_n = (\int_{-\infty}^{\infty} {G_n}^{-1}(t) F_n^*(dt))^{-1}$ is an estimator of α, $F_n^*(x) = n^{-1} \sum_{i=1}^{n} I_{[X_i \leq x]}$ is the ordinary empirical distribution function of the X_i's, and

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

is a nonparametric estimator of $G(t) = P(U^* \leq t \leq V^*)$, which is the probability of sampling a lifetime $X^* = t$. Here, $H_n(u, v) = \sum_{i=1}^{n} \widehat{\psi}_i I_{[U_i \leq u, V_i \leq v]}$ is the NPMLE of the joint distribution H of the truncation times, where the vector $\widehat{\psi} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_n)$ maximizes the likelihood

$$\mathcal{L}(\psi) = \prod_{i=1}^{n} \frac{\psi_i}{\Psi_i}$$

with respect to ψ, with $\Psi_i = \sum_{j=1}^{n} \psi_j I_{[U_j \leq X_i \leq V_j]}$ (see Moreira and de Uña-Álvarez, 2011 for more details).

Now, define

$$f_h(x) = \int K_h(x - t) F_n(dt) = \alpha_n \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) G_n^{-1}(X_i), \tag{2.1}$$

where K is a kernel density function, $K_h(\cdot) = K(\cdot/h)/h$, and $h = h_n$ is a bandwidth sequence tending to zero as n tends to infinity.

Suppose now that H belongs to a parametric family of distribution functions $\{H_\theta : \theta \in \Theta\}$, where Θ is a compact subset of $\mathbb{R}^k$. We estimate θ by maximizing the weighted likelihood of the (U_i, V_i)'s:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{h_\theta(U_i, V_i)}{G_\theta(X_i)},$$

where $h_\theta(u, v) = \frac{\partial^2}{\partial u \partial v} P(U^* \leq u, V^* \leq v) = H_\theta(du, dv)$ and

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

This leads to a semiparametric estimator of the distribution F:

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

where $\alpha_{\widehat{\theta}} = (\int_{-\infty}^{\infty} G_{\widehat{\theta}}^{-1}(t) F_n^*(dt))^{-1}$, and also to a semiparametric estimator of the density f:

$$f_{\widehat{\theta},h}(x) = \int K_h(x-t) F_{\widehat{\theta}}(dt) = \alpha_{\widehat{\theta}} \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) G_{\widehat{\theta}}^{-1}(X_i). \tag{2.2}$$

Moreira and de Uña-Álvarez (2010b) established the asymptotic normality of both $\widehat{\theta}$ and $F_{\widehat{\theta}}$. They also derived the asymptotic properties of $f_h(t)$ and $f_{\widehat{\theta},h}(t)$ through the analysis of their asymptotically equivalent (but unfeasible) version:

$$\overline{f}_h(x) = \int K_h(x-t) \overline{F}_n(dt) = \alpha \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) G^{-1}(X_i), \tag{2.3}$$

where

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

Consider now the following 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 $f(x)$ and $G^{-1}(x)f(x)$ are twice continuously differentiable in x.

Moreira and de Uña-Álvarez (2011) showed that under these assumptions the asymptotic mean and variance of $\overline{f}_h(x)$ are given by

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

and

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

As usual in kernel density estimation, the choice of the bandwidth h strongly influences the shape of the estimator $\overline{f}_h$. In order to select an optimal bandwidth, we need to choose a way to measure the discrepancy between the estimator $\overline{f}_h$ and its target density. The global error of $\overline{f}_h$ can be measured through the integrated MSE, namely

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

where $MSE(\overline{f}_h(x)) = \left[E\overline{f}_h(x) - f(x)\right]^2 + \text{Var}\left(\overline{f}_h(x)\right)$. Under assumptions (A1)-(A3), we have from the previous results:

$$MISE(\overline{f}_h) = AMISE(\overline{f}_h) + o((nh)^{-1}) + o(h^4), \tag{2.4}$$

where

$$AMISE(\overline{f}_h) = \frac{1}{4}h^4 R\left(f''\right)\mu_2^2(K) + (nh)^{-1}\alpha R(K)\int G^{-1}f \tag{2.5}$$

and $R\left(f''\right) = \int \left(f''\right)^2$. Minimization of $AMISE(\overline{f}_h)$ with respect to h leads to the asymptotically optimal bandwidth

$$h_{AMISE} = \left[\frac{\alpha R(K)\int G^{-1}f}{R\left(f''\right)\mu_2^2(K)}\right]^{1/5} n^{-1/5}. \tag{2.6}$$

Of course, this expression depends on unknown quantities involving f and G, which must be estimated in practice. In the next section we propose five automatic bandwidth selection methods in the context of doubly truncated data, which are all based on the minimization (with respect to h) of an appropriate estimator of $MISE(\overline{f}_h)$ or $AMISE(\overline{f}_h)$.

3 Automatic bandwidth selection

In this section we propose five bandwidth selection methods for the nonparametric estimator f_h and the semiparametric estimator $f_{\widehat{\theta},h}$. The five methods are appropriate adaptations to double truncation of the normal reference rule, the one- and two-stage plug-in procedure, the cross-validation procedure, and a bootstrap method. In order to simplify the presentation, we restrict attention in this section to the semiparametric estimator $f_{\widehat{\theta},h}$, but the proposed methods can be readably adapted to the nonparametric case.

3.1 Normal reference bandwidth selection

The estimation of the optimal bandwidth given in (2.6) involves the estimation of $R(f'')$ (apart from the estimation of α and $\int G^{-1}f$, which can be done using the estimators from Section 2). One simple and straightforward way to estimate $R(f'')$ is to assume that X^* follows a normal density $N(\mu, \sigma^2)$, in which case it can be easily shown that $R(f'') = \frac{3}{8}\pi^{-1/2}\sigma^{-5}$. If a Gaussian kernel is used, the window width obtained from (2.6) then equals

$$h_{AMISE} = \left(\frac{4}{3}\alpha\int G^{-1}f\right)^{1/5}\sigma n^{-1/5}. \tag{3.7}$$

A quick way of choosing the smoothing parameter, would be to estimate σ, α and $\int G^{-1}f$ from the data and then to substitute the estimates into (3.7). However, when f is not a normal density, the estimator of $R(f'')$ is in general not consistent. It may lead to oversmoothing if the population is multimodal, as a result of $\int (f'')^2$ being large relative to the standard deviation (see Silverman, 1986 and Wand and Jones, 1995). Better results can be obtained by using a robust measure of spread. If we write formula (3.7) in terms of the interquartile range IQR of the underlying normal distribution, we get that $h_{AMISE} = 0.79$ IQR $n^{-1/5}\left(\alpha \int G^{-1}f\right)^{1/5}$. The IQR is more robust against outliers if f has heavy tails (see Wand and Jones, 1995). Silverman (1986) recommended the use of the smaller between σ and IQR to reduce the chances of oversmoothing. This leads to the following bandwidth parameter:

$$\widehat{h}_{NR} = \left(\frac{4}{3}\alpha_{\widehat{\theta}} \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)\right)^{1/5} \min\left(\widehat{\sigma}, 0.79IQR\right) n^{-1/5}, \tag{3.8}$$

where, under double truncation, σ^2 can be estimated by

$$\widehat{\sigma}^2 = \alpha_{\widehat{\theta}} \int \left(t - m_{\widehat{\theta}}\right)^2 G_{\widehat{\theta}}^{-1}(t) F_n^*(dt), \tag{3.9}$$

with $m_{\widehat{\theta}} = \alpha_{\widehat{\theta}} \int t G_{\widehat{\theta}}^{-1}(t) F_n^*(dt)$ and $IQR = F_{\widehat{\theta}}^{-1}(0.75) - F_{\widehat{\theta}}^{-1}(0.25)$. Normal reference bandwidth selectors provide a quick 'first guess bandwidth' and can be expected to give reasonable answers when the data are close to normal. However, we need a more elaborated procedure to estimate $R(f'')$ for cases where the density is far away from a normal density. An appropriate nonparametric estimator for $R(f'')$ is discussed in the next subsection.

3.2 Plug-in bandwidth selection

Before explaining how to estimate $R(f'')$ in a more accurate way, we first need to consider the problem of estimating integrals of the form $R(f^{(s)}) = \int (f^{(s)}(x))^2 dx$ for positive integers s.

3.2.1 Estimation of the integrated squared density derivatives

Using integration by parts, we can write $R(f^{(s)}) = (-1)^s \int f^{(2s)}(x) f(x) dx$ under certain smoothness assumptions on f. It is therefore sufficient to study the estimation of integrals of the form $\psi_r = \int f^{(r)}(x) f(x) dx$ for r even (see Wand and Jones, 1995 for details). Note that $\psi_r = E\{f^{(r)}(X)\}$, which leads to the following semiparametric estimator under double truncation:

$$\widehat{\psi}_r(g) = \alpha_{\widehat{\theta}}\, n^{-1} \sum_{i=1}^{n} f_{\widehat{\theta},g}^{(r)}(X_i) G_{\widehat{\theta}}^{-1}(X_i) = \alpha_{\widehat{\theta}}^2\, n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} L_g^{(r)}(X_i - X_j) G_{\widehat{\theta}}^{-1}(X_i) G_{\widehat{\theta}}^{-1}(X_j)$$

(Jones and Sheather, 1991 and Wand and Jones, 1995), where the bandwidth g and the kernel L are possibly different from h and K. The asymptotic properties of $\widehat{\psi}_r(g)$ are important for the plug-in bandwidth selector described below. They will be derived under the following assumptions:

(i) The kernel L is a symmetric kernel of order k, $k = 2, 4, \ldots$, possessing r derivatives, such that $(-1)^{(r+k)/2+1} L^{(r)}(0)\mu_k(L) > 0$.

(ii) The functions f and $G^{-1}f$ have, respectively, $r + k$ and 2 continuous derivatives that are each ultimately monotone.

(iii) The bandwidth $g = g_n$ satisfies $g_n = o(1)$ and $ng_n^{2r+1} \to \infty$.

Note that we can write

$$\widehat{\psi}_r(g) = \alpha_{\widehat{\theta}}^2\, n^{-2} L_g^{(r)}(0) \sum_{i=1}^{n} G_{\widehat{\theta}}^{-2}(X_i) + \alpha_{\widehat{\theta}}^2\, n^{-2} \sum_i \sum_{j \neq i} L_g^{(r)}(X_i - X_j) G_{\widehat{\theta}}^{-1}(X_i) G_{\widehat{\theta}}^{-1}(X_j),$$

and define

$$\widetilde{\psi}_r(g) = \alpha^2\, n^{-2} L_g^{(r)}(0) \sum_{i=1}^{n} G^{-2}(X_i) + \alpha^2\, n^{-2} \sum_i \sum_{j \neq i} L_g^{(r)}(X_i - X_j) G^{-1}(X_i) G^{-1}(X_j),$$

which is the unfeasible estimator based on the true α and G. It can be easily seen using the asymptotic properties of $\widehat{\theta}$ given in Moreira and de Uña-Álvarez (2011), that $\widehat{\psi}_r(g)$ and $\widetilde{\psi}_r(g)$ are asymptotically equivalent. Moreover,

$$\begin{aligned} &E[\widetilde{\psi}_r(g)] \\ &= \alpha^2 n^{-1} L_g^{(r)}(0) E\left[G^{-2}(X) \middle| U \leq X \leq V\right] \\ &\quad + \alpha^2 (1 - n^{-1}) E\left[L_g^{(r)}(X_1 - X_2) G^{-1}(X_1) G^{-1}(X_2) \middle| U_1 \leq X_1 \leq V_1, U_2 \leq X_2 \leq V_2\right]. \end{aligned}$$

Using a Taylor expansion, the smoothness assumptions on f and $G^{-1}f$, and the fact that $L_g^{(r)}(0) = g^{-r-1} L^{(r)}(0)$, the bias can be written as

$$E[\widetilde{\psi}_r(g) - \psi_r] = \alpha\, n^{-1} g^{-r-1} L^{(r)}(0) \int G^{-1} f + (k!)^{-1} g^k \psi_{r+k} \mu_k(L) + O(g^{k+2}).$$

Note that the two main terms in this asymptotic bias cancel each other if we choose g equal to

$$g_{AMSE} = \left[-\frac{\alpha k! L^{(r)}(0) \int G^{-1} f}{\psi_{r+k} \mu_k(L)} \right]^{1/(r+k+1)} n^{-1/(r+k+1)}, \tag{3.10}$$

which is possible thanks to the assumption on the sign of $L^{(r)}(0)\mu_k(L)\psi_{r+k}$.

By similar calculations as done in Wand and Jones (1995) for the case of completely observed data, we can also obtain the formula of the asymptotic variance of $\widetilde{\psi}_r(g)$. We refer to Wand and Jones (1995) (page 69-70) for a discussion on the order of this variance in the general case and in the case where $g = g_{AMSE}$, and on how this variance depends on the larger value between k and r.

3.2.2 Plug-in bandwidth selection

We are now ready to present the plug-in bandwidth selection method, which is based on the idea of 'plugging-in' appropriate estimators of the unknown quantities that appear in the formula of the AMISE-optimal bandwidth (see (2.6)). First, rewrite this formula using the definition of ψ_4 as

$$h_{AMISE} = \left[\frac{\alpha R(K) \int G^{-1} f}{\psi_4 \mu_2^2(K)}\right]^{1/5} n^{-1/5}. \tag{3.11}$$

By replacing ψ_4 and $\alpha \int G^{-1} f$ by the estimators $\widehat{\psi}_4(g)$ and $\alpha_{\widehat{\theta}} \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)$ respectively, we obtain the direct plug-in (DPI) rule:

$$\widehat{h}_{DPI} = \left[\frac{\alpha_{\widehat{\theta}} R(K) \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)}{\widehat{\psi}_4(g) \mu_2^2(K)}\right]^{1/5} n^{-1/5}.$$

However, this formula still depends on the pilot bandwidth g, and is therefore not readably usable in practice. However, we can select g by making use of formula (3.10) with $r = 4$. If we take $L = K$, which can be any kernel of second order (so $k = 2$), we have:

$$g_{AMSE} = \left[-\frac{2\alpha K^{(4)}(0) \int G^{-1} f}{\psi_6 \mu_2(K)}\right]^{1/7} n^{-1/7}.$$

Again the same problem appears, in the sense that the estimation of this bandwidth formula necessitates an estimator of ψ_6, which requires again the selection of an appropriate bandwidth. It is clear that this process will never stop, since the formula of the optimal bandwidth for estimating ψ_r depends on ψ_{r+2} for any r. It is therefore necessary to estimate the functional ψ_r for a certain r by another 'simple' formula, which does not depend on any pilot bandwidth. We will use the normal reference rule described in Subsection 3.1 for this purpose, adapted to the estimation of derivatives of f (instead of f itself). Following Wand and Jones (1995), if f is a normal density with variance σ^2 then, for r even,

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1} (r/2)! \pi^{1/2}}, \tag{3.12}$$

which leads to the estimator

$$\widehat{\psi}_r^{NR} = \frac{(-1)^{r/2} r!}{(2\widehat{\sigma})^{r+1} (r/2)! \pi^{1/2}}, \tag{3.13}$$

where $\widehat{\sigma}$ is the estimator of σ defined in (3.9).

We now have a family of direct plug-in bandwidth selectors that depend on the number of stages of functional estimation before the normal reference rule is used. The plug-in procedure involving ℓ successive kernel functional estimations, is called an ℓ-stage plug-in bandwidth selector and is denoted by $\widehat{h}_{DPI,\ell}$. Note that the normal reference rule can be thought of as being a zero-stage direct plug-in bandwidth selector.

In practice, one has to decide how many steps will be included in this iterative procedure. Typically the more steps are included the smaller will be the bias but the more variable will be the procedure. In the simulation study carried out in the next section, we will work with $\ell = 0, 1$ and 2. To summarize, for $\ell = 1$ the procedure consists of the following steps:

(1) Calculate $\widehat{\psi}_6^{NR}$.

(2) Calculate $\widehat{\psi}_4(g_1)$, where $g_1 = \left[-\frac{2\alpha_{\widehat{\theta}} K^{(4)}(0) \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)}{\widehat{\psi}_6^{NR} \mu_2(K)} \right]^{1/7} n^{-1/7}$.

(3) The selected bandwidth is $\widehat{h}_{DPI,1} = \left[\frac{\alpha_{\widehat{\theta}} R(K) \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)}{\widehat{\psi}_4(g_1) \mu_2^2(K)} \right]^{1/5} n^{-1/5}$.

Similarly, the two-stage plug-in bandwidth selector can be described as follows:

(1) Calculate $\widehat{\psi}_8^{NR}$.

(2) Calculate $\widehat{\psi}_6(g_1)$, where $g_1 = \left[-\frac{2\alpha_{\widehat{\theta}} K^{(6)}(0) \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)}{\widehat{\psi}_8^{NR} \mu_2(K)} \right]^{1/9} n^{-1/9}$.

(3) Calculate $\widehat{\psi}_4(g_2)$, where $g_2 = \left[-\frac{2\alpha_{\widehat{\theta}} K^{(4)}(0) \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)}{\widehat{\psi}_6(g_1) \mu_2(K)} \right]^{1/7} n^{-1/7}$.

(4) The selected bandwidth is $\widehat{h}_{DPI,2} = \left[\frac{\alpha_{\widehat{\theta}} R(K) \int G_{\widehat{\theta}}^{-1}(t) F_{\widehat{\theta}}(dt)}{\widehat{\psi}_4(g_2) \mu_2^2(K)} \right]^{1/5} n^{-1/5}$.

3.3 Least-squares cross-validation bandwidth selection

The bandwidth selection procedures introduced in Subsections 3.1 and 3.2 are based on the asymptotic expression of the mean integrated squared error ($AMISE$). Another approach

is to directly try to estimate the $MISE$ and to minimize this estimator with respect to h. Given the estimator $f_{\widehat{\theta},h}$ of a density f, the mean integrated squared error can be written as

$$\begin{aligned} &MISE(f_{\widehat{\theta},h}) = E[ISE(f_{\widehat{\theta},h})] \\ &= E\Big[\int (f_{\widehat{\theta},h}(x) - f(x))^2 dx\Big] = E\Big[\int f^2_{\widehat{\theta},h}(x)\,dx - 2\int f_{\widehat{\theta},h}(x) f(x) dx + \int f^2(x) dx\Big]. \end{aligned}$$

The term $\int f^2(x)dx$ does not depend on h, and so minimizing $MISE(f_{\widehat{\theta},h})$ is equivalent to minimizing

$$S(f_{\widehat{\theta},h}) = MISE(f_{\widehat{\theta},h}) - \int f^2(x)dx = E\left[\int f^2_{\widehat{\theta},h}(x)\,dx - 2\alpha \int f_{\widehat{\theta},h}(x) G^{-1}(x) F^*(dx)\right].$$

In order to construct an estimator of $\mathrm{S}(f_{\widehat{\theta},h})$, let $f_{\widehat{\theta},h;-i}$ be the density estimator constructed from all data points except X_i, i.e.

$$f_{\widehat{\theta},h;-i}(x) = \int K_h(x-t)\widehat{F}_{\widehat{\theta};-i}(dt) = \alpha_{\widehat{\theta};-i}\frac{1}{n-1}\sum_{j\neq i} K_h(x - X_j) G^{-1}_{\widehat{\theta};-i}(X_j), \tag{3.14}$$

where $G_{\widehat{\theta};-i}(\cdot)$ is the estimator of $G(\cdot)$ defined in Section 2, except that the i-th data point is not used for estimating θ. Now define

$$LSCV(h) = \int f^2_{\widehat{\theta},h}(x)\,dx - 2n^{-1}\sum_{i=1}^{n} \alpha_{\widehat{\theta};-i} f_{\widehat{\theta},h;-i}(X_i) G^{-1}_{\widehat{\theta};-i}(X_i),$$

and estimate the optimal h by minimizing $LSCV(h)$ over h:

$$\widehat{h}_{LSCV} = \operatorname*{argmin}_h LSCV(h).$$

Note that

$$\begin{aligned} &E\Big[n^{-1}\sum_{i=1}^{n} \alpha_{\widehat{\theta};-i} f_{\widehat{\theta},h;-i}(X_i) G^{-1}_{\widehat{\theta};-i}(X_i)\Big] \\ &= E\Big[\alpha_{\widehat{\theta};-1} f_{\widehat{\theta},h;-1}(X_1) G^{-1}_{\widehat{\theta};-1}(X_1)\Big| U_1 \le X_1 \le V_1\Big] \\ &= E\Big[\alpha_{\widehat{\theta};-1}\int f_{\widehat{\theta},h;-1}(x) G^{-1}_{\widehat{\theta};-1}(x) F^*(dx)\Big], \end{aligned}$$

and this is asymptotically equivalent to $E[\alpha\int f_{\theta,h;-1}(x)G^{-1}(x)F^*(dx)] = E[\int f_{\theta,h;-1}(x)F(dx)]$ $= E[\int f_{\theta,h}(x)f(x)dx]$, where the latter equality follows from the fact that $E\{f_{\theta,h}(x)\}$ depends only on the kernel and the bandwidth, and not on the sample size. Hence, $E[LSCV(h)]$ is asymptotically equivalent to $S(f_{\widehat{\theta},h})$, which suggests that we can expect $\widehat{h}_{LSCV}$ to be close to the minimizer of $S(f_{\widehat{\theta},h})$ or $MISE(f_{\widehat{\theta},h})$.

3.4 Smoothed bootstrap bandwidth selection

The final bandwidth selector is based on the estimation of the $MISE$ of $f_{\widehat{\theta},h}(\cdot)$ by means of a smoothed bootstrap procedure. The reason we use a smoothed bootstrap procedure (as opposed to a non-smoothed one) is the same as in Silverman and Young (1987), namely without smoothing the bootstrap would be inconsistent. The bootstrap procedure can be described as follows. For fixed B and for $b = 1, \ldots, B$:

1. Let $X_{b,i}^{boot}$, $i = 1, \ldots, n$, be an i.i.d. sample from $f_{\widehat{\theta},g}$, where g is chosen to be $\widehat{g}_{DPI,2}$ (other choices for g are possible as well), and let $(U_{b,i}^{boot}, V_{b,i}^{boot})$, $i = 1, \ldots, n$, be an i.i.d. sample from H_n. Next, for each $i = 1, \ldots, n$, we keep the triplet $(U_{b,i}^{boot}, X_{b,i}^{boot}, V_{b,i}^{boot})$ in the resample only if the condition $U_{b,i}^{boot} \leq X_{b,i}^{boot} \leq V_{b,i}^{boot}$ is fulfilled. If not, the same resampling procedure is repeated until a triplet is found for which the inequality holds true.

2. Let $\widehat{\theta}_b^{boot}$ and $f_{\widehat{\theta}_b^{boot},b,h}^{boot}(\cdot)$ be the estimator of θ and of the density f respectively, obtained from the bootstrap sample $(U_{b,i}^{boot}, X_{b,i}^{boot}, V_{b,i}^{boot})$, $i = 1, \ldots, n$.

Note that in the first step above, no smoothing is required for the sampling distribution of the truncation times, as is the case for single truncation times (see Sánchez-Sellero et al. (1999), page 57-58). Next, let

$$BMISE(h) = B^{-1} \sum_{b=1}^{B} \int \left(f_{\widehat{\theta}_b^{boot},b,h}^{boot}(t) - f_{\widehat{\theta},g}(t) \right)^2 dt,$$

which for B large will approximate well the bootstrap $MISE$ given by

$$MISE^{boot}(h) = E^{boot}\Big[\int \left(f_{\widehat{\theta}^{boot},h}^{boot}(t) - f_{\widehat{\theta},g}(t) \right)^2 dt \Big],$$

where E^{boot} denotes the expected value conditionally on the original sample. Now, define

$$\widehat{h}_{boot} = \operatorname*{argmin}_h BMISE(h).$$

As is the case for the cross-validation procedure (see Subsection 3.3), we obtain here an estimator of the optimal bandwidth by minimizing an estimator of the (non-asymptotic) $MISE$ of $f_{\widehat{\theta},h}$, whereas the normal reference rule (see Subsection 3.1) and the plug-in bandwidth selection procedures (see Subsection 3.2) are based on the minimization of an appropriate estimator of the (asymptotic) $AMISE$. In the next section, we will examine the behavior of each of these bandwidth selectors for small samples via a thorough simulation study.

4 Simulations

In this section we illustrate through a simulation study the finite sample behavior of the five bandwidth selection methods: the normal reference bandwidth selector (NR), the one-stage plug-in bandwidth selector (DPI_1), the two-stage plug-in bandwidth selector (DPI_2), the least squares cross-validation selector (LSCV) and the bootstrap bandwidth selector (Boot), for both the nonparametric and the semiparametric kernel density estimator.

We consider two different situations of double truncation. In Case 1, U^*, V^* and X^* are mutually independent. In Case 2, we simulate U^* and let $V^* = U^* + \tau$ for some fixed constant $\tau > 0$. The density of (U^*, V^*) does not exist for Case 2, and hence the role of the joint density of (U^*, V^*) must be played by one of the marginal densities (see Moreira and de Uña-Álvarez, 2010a for more details).

Two different models are simulated for each of the two cases. 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 $\tau = 0.25$, $U^* \sim U(0,1)$, $X^* \sim U(0.25,1)$ (Model 2.1) and $\tau = 0.25$, $U^* \sim U(0,1)$, $X^* \sim 0.75N(0.5, 0.15) + 0.25$ (Model 2.2). Note that when we move from Model 1.1 (respectively 2.1) to Model 1.2 (respectively 2.2) we are changing the lifetime distribution while fixing the distribution of the truncation variables. On the other hand, when we move from Model 1.1 (respectively 1.2) to Model 2.1 (respectively 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 like to point out that, due to the random truncation, relatively small and moderate values of the lifetime are more likely to be observed under Models 1.1 and 1.2, while there is no observational bias under Models 2.1 and 2.2. Indeed, it can be easily seen that the function G is constant under Models 2.1 and 2.2, since $V^* = U^* + \tau$ and $U^* \sim U(0,1)$. In other words, the truncated distribution F^* coincides with the distribution of interest F, i.e. the truncation mechanism does not change the sampling probabilities.

For the computation of the semiparametric density estimator, as parametric information on (U^*, V^*) we always consider a $Beta(\theta_1, 1)$ for U^* and a $Beta(1, \theta_2)$ for V^* in Case 1. For each model, we simulate $M = 500$ samples of (final) sample size $n = 50$, 100, 250 or 500. For each generated sample, we estimate the optimal bandwidth by means of the five bandwidth selection methods (NR, DPI_1, DPI_2, LSCV and Boot) for both the nonparametric and the semiparametric kernel density estimator. In Tables 1-4 we report the median and interquartile range of the 500 estimated bandwidths for each method and each model, and we also present for each case the theoretical optimal bandwidth (h_{MISE}) for comparison purposes. In Figures 1-4 we represent the graphs of the densities of $\log_{10}(\widehat{h}) - \log_{10}(h_{MISE})$

with $\widehat{h}$ obtained by one of the five bandwidth selection methods. These graphs allow to compare the bias and variance in a visual way, and they give an idea of the shape of these densities, whereas this information cannot be distracted from the tables.

From the tables and figures it can be seen that in general (only 5 exceptions) the IQR's for the semiparametric estimator are smaller than the corresponding IQR's for the nonparametric estimator. Another interesting feature is that the IQR's are usually highest for LSCV, followed by Boot, and then followed by NR, DPI_1 and DPI_2, which usually behave quite similarly. In addition, even for samples of size $n = 500$, the LSCV remains quite variable. From Tables 1 and 3 (Models 1.1 and 2.1) we can conclude that in general all methods except NR and Boot (at least for $n = 500$) perform quite well, however the DPI_1 and DPI_2 perform almost as well. This can be explained by the fact that the target is an uniform density, which provokes strong boundary effects. The plug-in bandwidths are in all cases smaller than the bootstrap bandwidth. This confirms the findings in Cao et al. (1994). From Figures 1-4, we can conclude that the semiparametric estimator has less variance than the nonparametric one, and that the LSCV method has higher variance than all the other methods. All methods seem to converge to the optimum when the sample size increases.

	$n = 50$					$n = 100$			
	NP		SP			NP		SP	
Method	$h_{MISE} = 0.1620$		$h_{MISE} = 0.1490$		Method	$h_{MISE} = 0.1180$		$h_{MISE} = 0.1050$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.1006	0.0209	0.0996	0.0187	NR	0.0906	0.0142	0.0903	0.0114
DPI_1	0.0985	0.0229	0.0995	0.0217	DPI_1	0.0854	0.0152	0.0853	0.0143
DPI_2	0.0915	0.0257	0.0921	0.0233	DPI_2	0.0782	0.0185	0.0780	0.0167
LSCV	0.1240	0.1033	0.1310	0.0983	LSCV	0.0890	0.0823	0.0920	0.0780
Boot	0.1510	0.0633	0.1390	0.0502	Boot	0.1240	0.0483	0.1140	0.0370
	$n = 250$					$n = 500$			
	NP		SP			NP		SP	
Method	$h_{MISE} = 0.0750$		$h_{MISE} = 0.0670$		Method	$h_{MISE} = 0.0550$		$h_{MISE} = 0.0490$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0786	0.0097	0.0784	0.0092	NR	0.0700	0.0091	0.0699	0.0084
DPI_1	0.0709	0.0092	0.0707	0.0094	DPI_1	0.0560	0.0091	0.0603	0.0083
DPI_2	0.0636	0.0102	0.0642	0.0105	DPI_2	0.0528	0.0091	0.0529	0.0087
LSCV	0.0670	0.0490	0.0690	0.0470	LSCV	0.0460	0.0313	0.0450	0.0330
Boot	0.0930	0.0253	0.0880	0.0220	Boot	0.0720	0.0170	0.0690	0.0140

Table 1: Median and interquartile range of the 500 estimated bandwidths obtained from the five bandwidth selection methods: NR, DPI_1, DPI_2, LSCV and Boot, for both the nonparametric (NP) and the semiparametric (SP) kernel density estimators under Model 1.1. The exact value of h_{MISE} is also reported for comparison purposes.

Method	$n = 50$				Method	$n = 100$			
	NP		SP			NP		SP	
	$h_{MISE} = 0.0610$		$h_{MISE} = 0.0590$			$h_{MISE} = 0.0520$		$h_{MISE} = 0.0510$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0488	0.0091	0.0488	0.0092	NR	0.0440	0.0058	0.0441	0.0057
DPI_1	0.0524	0.0115	0.0527	0.0105	DPI_1	0.0467	0.0074	0.0468	0.0065
DPI_2	0.0512	0.0132	0.0519	0.0123	DPI_2	0.0458	0.0088	0.0459	0.0078
LSCV	0.0630	0.0260	0.0630	0.0250	LSCV	0.0520	0.0200	0.0530	0.0190
Boot	0.0690	0.0200	0.0670	0.0170	Boot	0.0600	0.0120	0.0590	0.0100
Method	$n = 250$				Method	$n = 500$			
	NP		SP			NP		SP	
	$h_{MISE} = 0.0430$		$h_{MISE} = 0.0420$			$h_{MISE} = 0.0370$		$h_{MISE} = 0.0360$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0369	0.0036	0.0369	0.0035	NR	0.0325	0.0022	0.0325	0.0022
DPI_1	0.0392	0.0039	0.0393	0.0037	DPI_1	0.0342	0.0027	0.0344	0.0024
DPI_2	0.0388	0.0049	0.0389	0.0047	DPI_2	0.0340	0.0034	0.0342	0.0032
LSCV	0.0430	0.0120	0.0430	0.0120	LSCV	0.0360	0.0090	0.0370	0.0090
Boot	0.0500	0.0070	0.0480	0.0050	Boot	0.0430	0.0050	0.0420	0.0040

Table 2: Median and interquartile range of the 500 estimated bandwidths obtained from the five bandwidth selection methods: NR, DPI_1, DPI_2, LSCV and Boot, for both the nonparametric (NP) and the semiparametric (SP) kernel density estimators under Model 1.2. The exact value of h_{MISE} is also reported for comparison purposes.

Method	$n = 50$				Method	$n = 100$			
	NP		SP			NP		SP	
	$h_{MISE} = 0.2300$		$h_{MISE} = 0.1490$			$h_{MISE} = 0.1800$		$h_{MISE} = 0.1080$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0994	0.0190	0.1024	0.0111	NR	0.0893	0.0103	0.0904	0.0064
DPI_1	0.0892	0.0180	0.0958	0.0116	DPI_1	0.0785	0.0086	0.0817	0.0063
DPI_2	0.0799	0.0187	0.0880	0.0172	DPI_2	0.0687	0.0117	0.0733	0.0101
LSCV	0.0800	0.0703	0.1140	0.0870	LSCV	0.0610	0.0523	0.0770	0.0643
Boot	0.1960	0.1043	0.1455	0.0540	Boot	0.1570	0.0663	0.1170	0.0663
Method	$n = 250$				Method	$n = 500$			
	NP		SP			NP		SP	
	$h_{MISE} = 0.1080$		$h_{MISE} = 0.0630$			$h_{MISE} = 0.0520$		$h_{MISE} = 0.0400$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0753	0.0056	0.0759	0.0036	NR	0.0661	0.0036	0.0662	0.0022
DPI_1	0.0638	0.0055	0.0646	0.0037	DPI_1	0.0539	0.0029	0.0542	0.0025
DPI_2	0.0556	0.0074	0.0569	0.0060	DPI_2	0.0461	0.0048	0.0468	0.0040
LSCV	0.0470	0.0320	0.0500	0.0310	LSCV	0.0350	0.0193	0.0365	0.0200
Boot	0.1090	0.0415	0.0820	0.0190	Boot	0.0760	0.0230	0.0630	0.0103

Table 3: Median and interquartile range of the 500 estimated bandwidths obtained from the five bandwidth selection methods: NR, DPI_1, DPI_2, LSCV and Boot, for both the nonparametric (NP) and the semiparametric (SP) kernel density estimators under Model 2.1. The exact value of h_{MISE} is also reported for comparison purposes.

Method	$n = 50$				Method	$n = 100$			
	NP		SP			NP		SP	
	$h_{MISE} = 0.0670$		$h_{MISE} = 0.0600$			$h_{MISE} = 0.0570$		$h_{MISE} = 0.0510$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0503	0.0096	0.0509	0.0085	NR	0.0461	0.0052	0.0458	0.0050
DPI_1	0.0519	0.0125	0.0536	0.0092	DPI_1	0.0466	0.0069	0.0469	0.0060
DPI_2	0.0503	0.0146	0.0524	0.0110	DPI_2	0.0456	0.0090	0.0460	0.0081
LSCV	0.0610	0.0330	0.0630	0.0270	LSCV	0.0570	0.0210	0.0540	0.0190
Boot	0.0790	0.0340	0.0690	0.0160	Boot	0.0650	0.0190	0.0590	0.0103
Method	$n = 250$				Method	$n = 500$			
	NP		SP			NP		SP	
	$h_{MISE} = 0.0420$		$h_{MISE} = 0.0420$			$h_{MISE} = 0.0370$		$h_{MISE} = 0.0360$	
	med	IQR	med	IQR		med	IQR	med	IQR
NR	0.0389	0.0031	0.0388	0.0031	NR	0.0334	0.0018	0.0339	0.0019
DPI_1	0.0393	0.0037	0.0396	0.0036	DPI_1	0.0334	0.0026	0.0344	0.0024
DPI_2	0.0389	0.0048	0.0391	0.0045	DPI_2	0.0339	0.0034	0.0342	0.0030
LSCV	0.0450	0.0140	0.0430	0.0120	LSCV	0.0370	0.0120	0.0360	0.0090
Boot	0.0525	0.0100	0.0490	0.0060	Boot	0.0430	0.0070	0.0425	0.0040

Table 4: Median and interquartile range of the 500 estimated bandwidths obtained from the five bandwidth selection methods: NR, DPI_1, DPI_2, LSCV and Boot, for both the nonparametric (NP) and the semiparametric (SP) kernel density estimators under Model 2.2. The exact value of h_{MISE} is also reported for comparison purposes.

5 Data analysis

Let us now apply the developed bandwidth selection methods to data on the luminosity of quasars in astronomy. One of the main aims of astronomers interested in quasars is to understand the evolution of the luminosity of quasars (see Efron and Petrosian, 1999, Shen, 2010a and Moreira et al., 2010). The motivating example presented in the paper by 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, that is bounded at both ends, and which is determined by detection limits.

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 apparent 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 dataset contains

at the end information about $n = 210$ quasars. The distribution function of the adjusted log-luminosity has been estimated by Moreira et al. (2010), using the NPMLE for doubly truncated data computed by the `R` package `DTDA`.

We first estimate the function $G(t)$, i.e. the probability that an observation is not truncated given that it equals t (or in other words the probability that the detection interval contains an adjusted log-luminosity of magnitude t). In the left panel of Figure 5 we observe that this function is varying a lot, which suggests that there is substantial observational bias, since a constant curve indicates that no observational bias is present. In particular we see that adjusted log luminosities below zero are observed with a particularly small probability (see also Moreira, 2010 and Moreira et al., 2010).

Next, we apply the nonparametric estimator and our five bandwidth selection methods to these data. The resulting kernel density estimators are depicted in the right panel of Figure 5. For comparison purposes, the naive kernel density estimator which does not correct for the presence of double truncation is also given. It can be clearly seen that by ignoring the presence of truncation we obtain a dramatically different kernel density estimator. We also observe that the plug-in methods (DPI_1 and DPI_2), the normal reference method (NR) and the smoothed bootstrap procedure (Boot) give very similar results, whereas the LSCV method yields a somewhat smoother curve. A possible explanation for this can be found in the fact that the data are rather scarce near the left endpoint of their support (there exist only four observations of adjusted log-luminosity below -1.7) and, in the case of the least squares cross-validation selector, it is clear that by leaving out one observation from the dataset, it may happen that no data points are left in a given window when the bandwidth is small. This feature forces the procedure to select a larger bandwidth, and this might explain why the cross-validation method yields a somewhat larger bandwidth estimator than the other methods.

6 Conclusions

In this paper we have considered the estimation of a density function by means of kernel smoothing, when the available data are subject to double truncation. In particular, we have proposed five bandwidth selection methods and have studied their finite sample behavior via a thorough simulation study. The proposed methods are: the normal reference rule, a one-stage plug-in bandwidth selector, a two-stage plug-in bandwidth selector, a least squares cross-validation selector and a smoothed bootstrap bandwidth selector. Both a nonparametric and a semiparametric kernel density estimator are studied. The simulations show that in general all methods perform well, with some exceptions in Models 1.1 and 2.1, in which the

target is an uniform density, provoking strong boundary effects. We can also conclude that the LSCV method, even though it performs rather well (at least for large sample sizes), has larger variance than all the other methods. As can be expected, the semiparametric estimator has smaller variance than the nonparametric one. The proposed methods are also applied to data on the luminosity of quasars in astronomy.

An interesting topic for future research is the study of automatic bandwidth selectors for the estimation of the hazard rate function, which is another important curve in survival analysis in the presence of double truncation.

References

Asgharian, M., C. MLan, and D. Wolfson (2002). Length-biased sampling with right-censoring: an unconditional approach. *Journal of the American Statistical Association 97*, 201–209.

Cao, R., A. Cuevas, and W. González-Manteiga (1994). A comparative study of several smoothing methods in density estimation. *Computational statistics & Data Analysis 17*, 153–176.

de Uña-Álvarez, J. (2004). Nonparametric estimation under length-biased sampling and type i censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics 56*, 667–681.

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

Jones, M. C. and S. J. Sheather (1991). Using non-stochastic terms to advantage in kernal-based estimation of integrated squared density derivatives. *Statistics and Probability Letters 11*, 511–514.

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. and J. de Uña-Álvarez (2011). Kernel density estimation with doubly truncated data. "Under revision".

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.

Sheather, S. and M. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of Royal Statistical Society 53*, 683–690.

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.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Volume 26 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

Silverman, B. and G. Young (1987). The bootstrap: To smooth or not to smooth? *Biometrika 74*, 469–479.

Sánchez-Sellero, C., W. González-Manteiga, and R. Cao (1999). Bandwidth selection in density estimation with truncated and censored data. *Annals of the Institute of Statistical Mathematics 51*, 51–70.

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

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. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association 84*, 742–748.

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

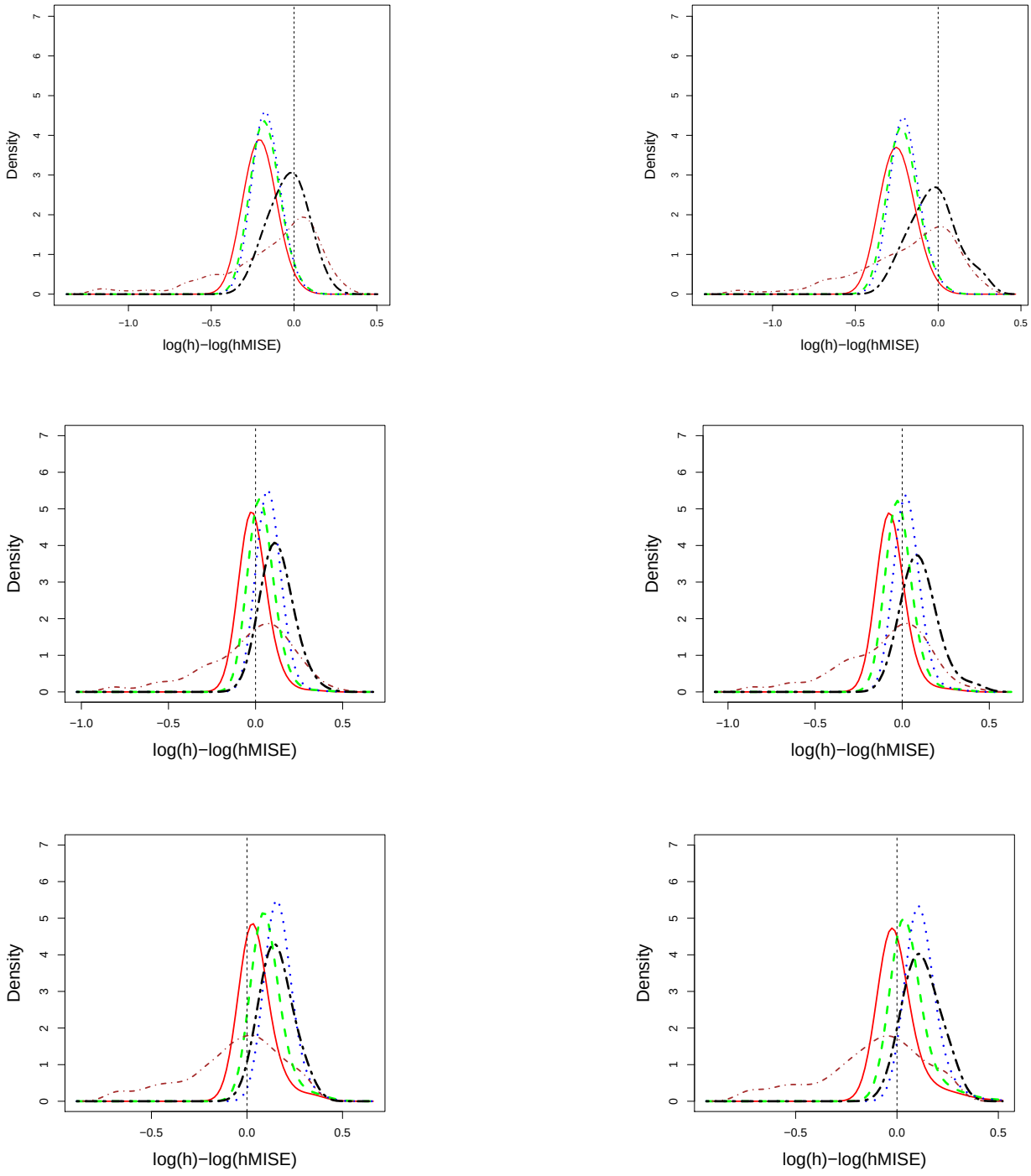

Figure 1: Density estimates of $\log_{10}(\widehat{h}) - \log_{10}(h_{MISE})$ for the semiparametric estimator (left panel) and the nonparametric estimator (right panel) with $\widehat{h}$ obtained by the five bandwidth selection methods: NR (dotted line); DPI_1 (dashed line); DPI_2 (solid line); LSCV (dot-dashed line) and Boot (two-dashed line). Selected bandwidths are based on 500 simulated samples for Model 1.1 with sample sizes $n = 50$ (top), 250 (middle) and 500 (bottom).

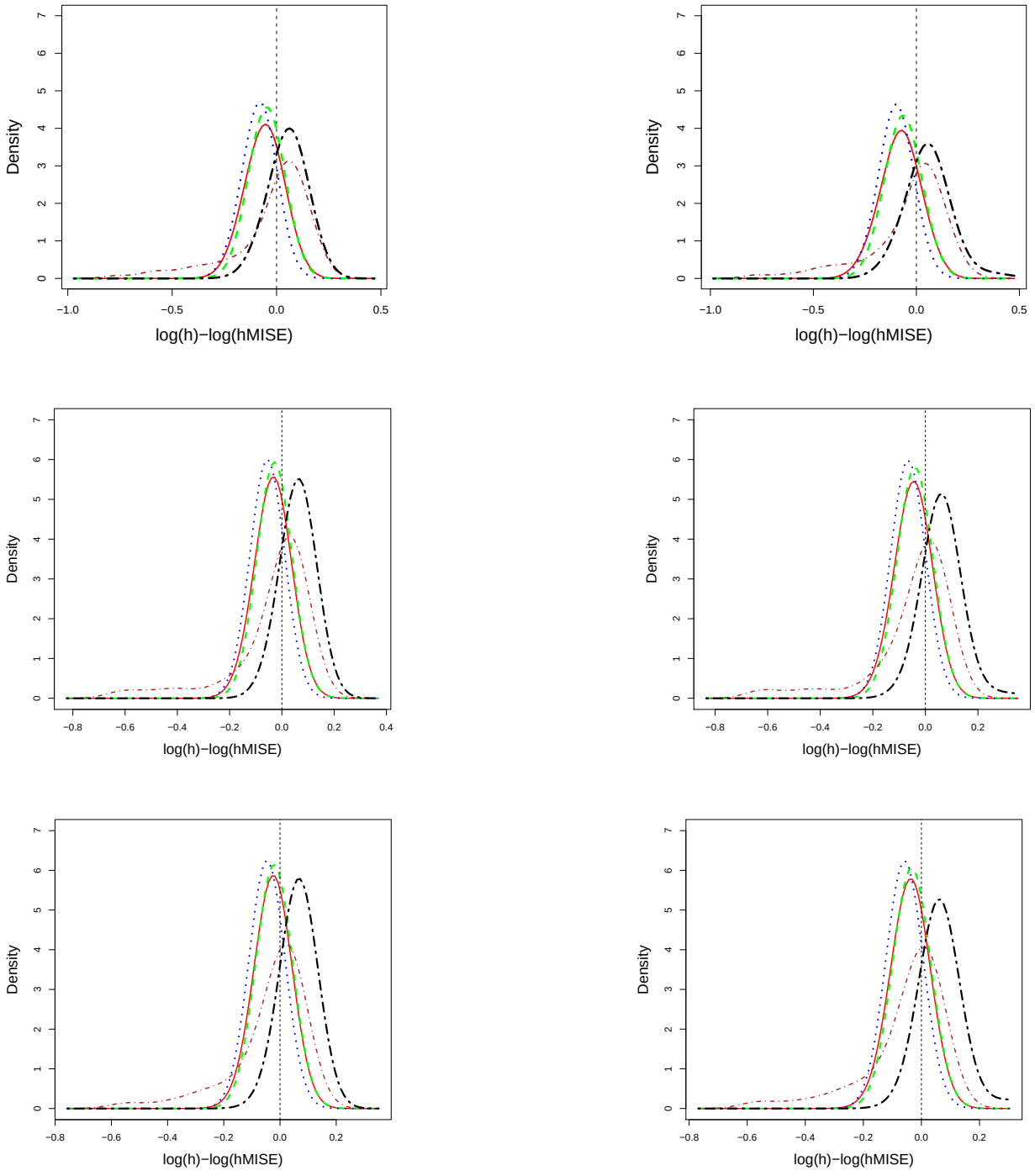

Figure 2: Density estimates of $\log_{10}(\widehat{h}) - \log_{10}(h_{MISE})$ for the semiparametric estimator (left panel) and the nonparametric estimator (right panel) with $\widehat{h}$ obtained by the five bandwidth selection methods: NR (dotted line); DPI_1 (dashed line); DPI_2 (solid line); LSCV (dot-dashed line) and Boot (two-dashed line). Selected bandwidths are based on 500 simulated samples for Model 1.2 with sample sizes $n = 50$ (top), 250 (middle) and 500 (bottom).

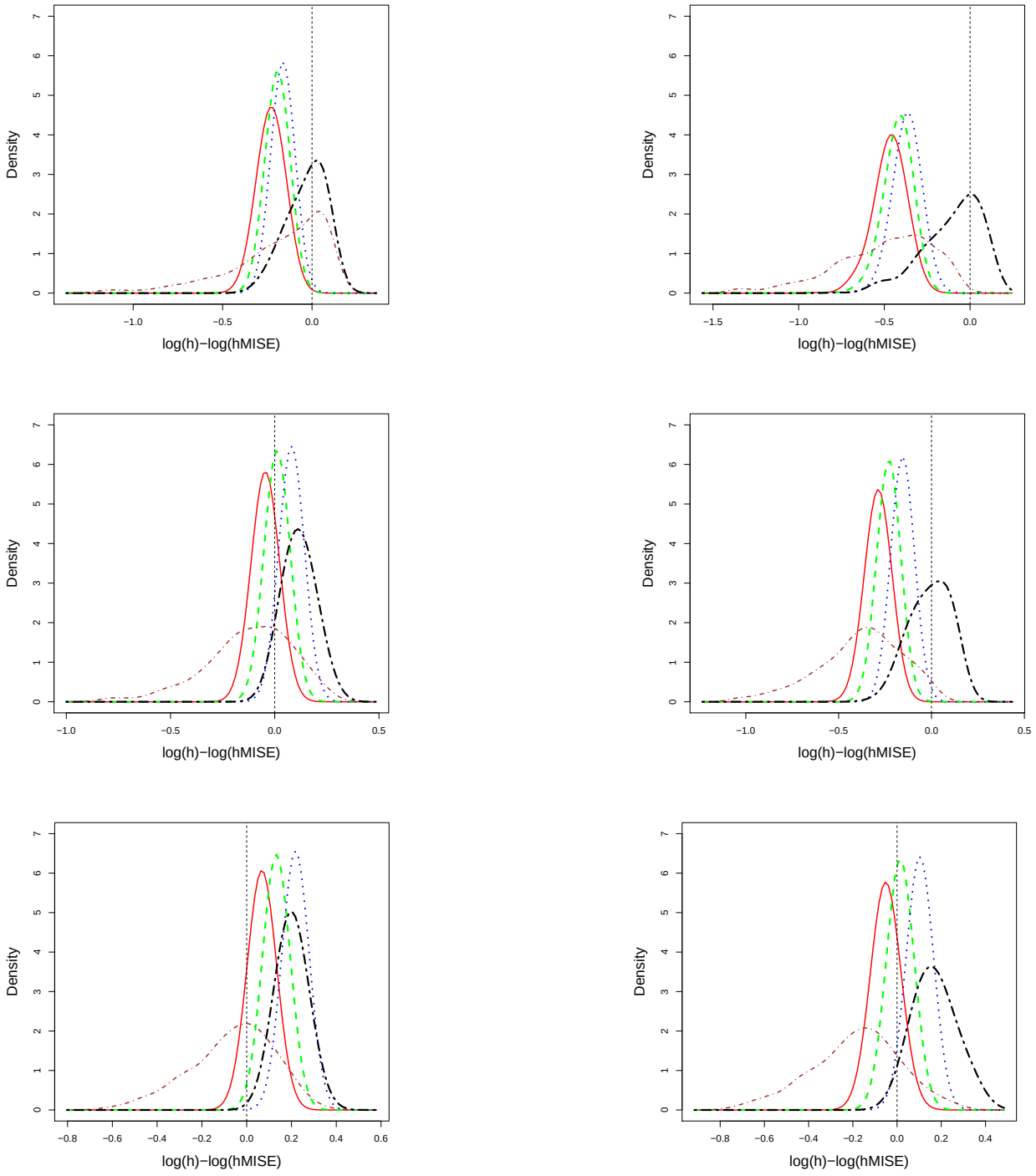

Figure 3: Density estimates of $\log_{10}(\widehat{h}) - \log_{10}(h_{MISE})$ for the semiparametric estimator (left panel) and the nonparametric estimator (right panel) with $\widehat{h}$ obtained by the five bandwidth selection methods: NR (dotted line); DPI_1 (dashed line); DPI_2 (solid line); LSCV (dot-dashed line) and Boot (two-dashed line). Selected bandwidths are based on 500 simulated samples for Model 2.1 with sample sizes $n = 50$ (top), 250 (middle) and 500 (bottom).

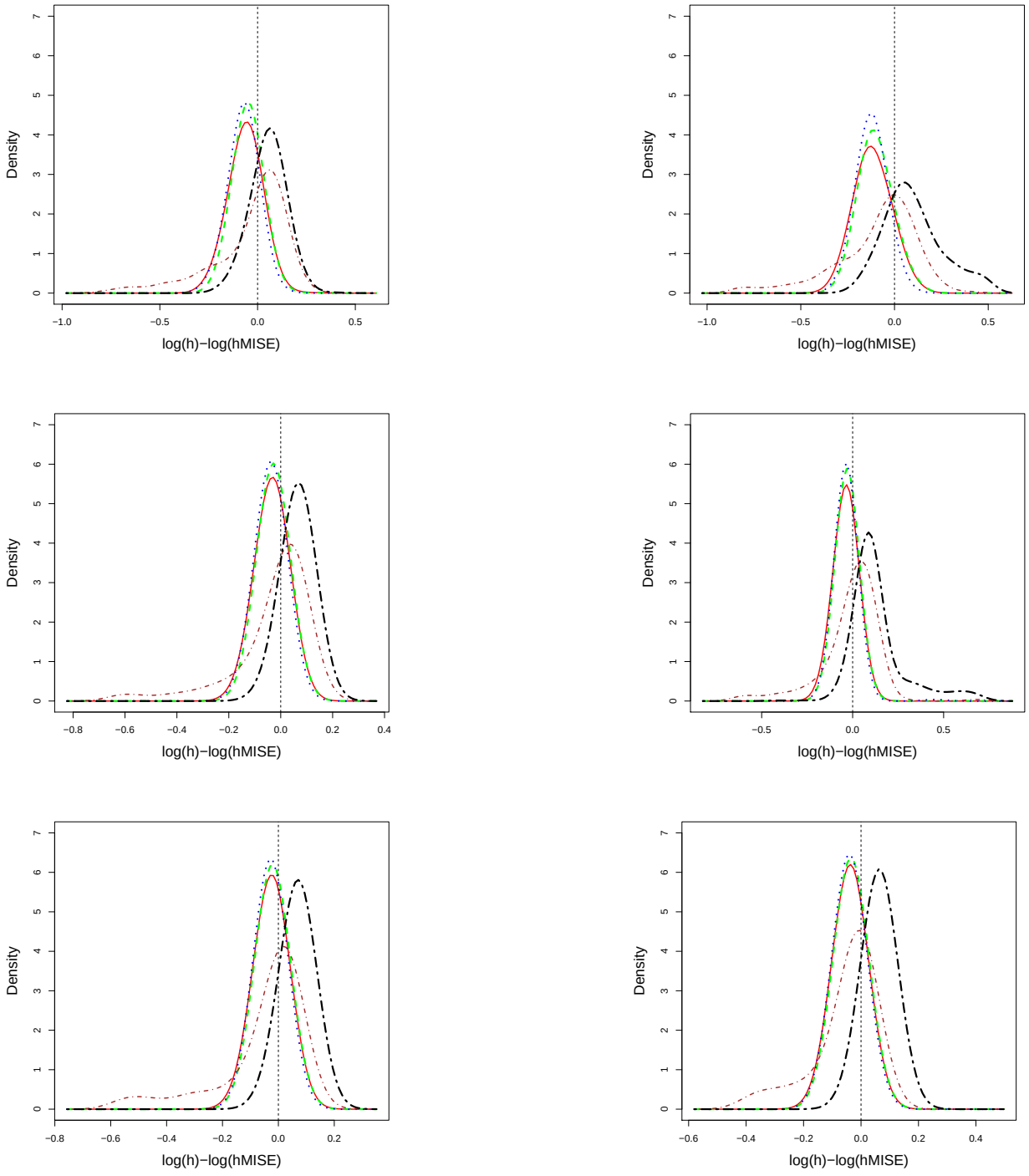

Figure 4: Density estimates of $\log_{10}(\widehat{h}) - \log_{10}(h_{MISE})$ for the semiparametric estimator (left panel) and the nonparametric estimator (right panel) with $\widehat{h}$ obtained by the five bandwidth selection methods: NR (dotted line); DPI_1 (dashed line); DPI_2 (solid line); LSCV (dot-dashed line) and Boot (two-dashed line). Selected bandwidths are based on 500 simulated samples for Model 2.2 with sample sizes $n = 50$ (top), 250 (middle) and 500 (bottom).

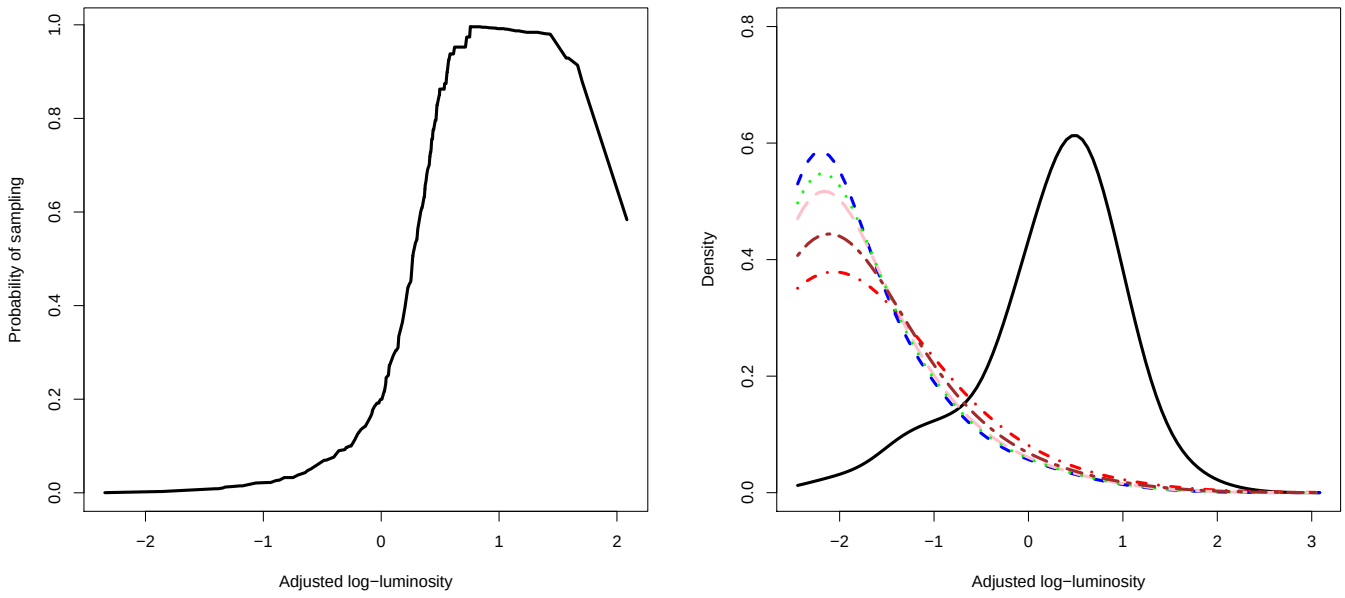

Figure 5: Left panel: Bias function for the quasars data. Right panel: Kernel density estimators for the log-luminosity of the quasars data with different bandwidth selectors. Naive (biased) estimator (solid line); NR (dashed line); DPI_1 (long-dashed line); DPI_2 (dotted line); LSCV (dot-dashed line) and Boot (two-dashed line).