Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 20–02

BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION: A HERMITE SERIES-BASED DIRECT PLUG-IN APPROACH

CARLOS TENREIRO

Abstract: In this paper we propose a new class of Hermite series-based direct plug-in bandwidth selectors for kernel density estimation and we describe their asymptotic and finite sample behaviours. Unlike the direct plug-in bandwidth selectors considered in the literature, the proposed methodology does not involve multistage strategies and reference distributions are no longer needed. The new bandwidth selectors show a good finite sample performance when the underlying probability density function presents not only "easy-to-estimate" but also "hard-to-estimate" distribution features. This quality, that is not shared by other widely used bandwidth selectors as the classical plug-in or the least-square cross-validation methods, is the most significant aspect of the Hermite series-based direct plug-in approach to bandwidth selection.

Keywords: Bandwidth selection; kernel density estimation; direct plug-in bandwidth selection; quadratic functionals; projection methods; Hermite series.
AMS Subject Classification (2010): 62G07; 62G20.

1. Introduction

If $X_1, \ldots, X_n$ are independent real-valued absolutely continuous random variables with common and unknown probability density function f, the Parzen-Rosenblatt estimator of f (Rosenblatt, 1956, Parzen, 1962) based on the observed sample is defined, for $x \in \mathbb{R}$, by

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

where $K_h(\cdot) = K(\cdot/h)/h$, for $h > 0$, with K a kernel in $\mathbb{R}$, that is, K is a bounded and integrable function such that $\int K(u)du = 1$, and the bandwidth $h = h_n$ is a sequence of strictly positive real numbers converging to zero as n tends to infinity (see Devroye and Györfi, 1985, Silverman, 1986, Bosq and Lecoutre, 1987, Wand and Jones, 1995, Simonoff, 1996, and Tsybakov, 2009, for general reviews on density estimation). Contrary to the selection of the

Received January 23, 2020.

kernel, the choice of the bandwidth is crucial to the performance of the estimator, this being one of the most studied topics in kernel density estimation, and several data-based approaches have been proposed for selecting h (see Wand and Jones, 1995, pp. 58–89, and also Tenreiro, 2017, p. 3440, where more recent bandwidth selection methods are mentioned).

The direct plug-in method, which dates back to Woodroofe (1970), Nadaraya (1974) and Deheuvels and Hominal (1980), is a very simple data-dependent method for choosing the bandwidth. It is based on asymptotic approximations of the bandwidth h_0 that minimizes the mean integrated square error $\mathrm{MISE}(f;n,h) = \mathrm{E}(\mathrm{ISE}(f;n,h)) = \mathrm{E}||f_{n,h} - f||_2^2$, where $||\cdot||_2$ denotes the L_2 distance:

$$h_0 = \underset{h>0}{\operatorname{argmin}}\, \mathrm{MISE}(f;n,h).$$

For a square integrable density f, the existence of this exact optimal bandwidth can be established whenever the kernel K is continuous at zero with $k_0 < 2K(0)$, where $k_0 = ||K||_2^2$ (see Chacón et al., 2007). Under some moment and regularity conditions on K and f, respectively (see Section 6.2), two asymptotic approximations of the optimal bandwidth h_0 are given by

$$h_1 = c_{1,K}\, \theta_2^{-1/5} n^{-1/5},$$

and

$$h_2 = c_{1,K}\, \theta_2^{-1/5} n^{-1/5} + c_{2,K}\, \theta_2^{-8/5}\, \theta_3\, n^{-3/5},$$

where θ_r, $r = 0, 1, \dots$, denotes the quadratic functional

$$\theta_r = \int f^{(r)}(x)^2 dx = ||f^{(r)}||_2^2,$$

with $f^{(r)} \in L_2$ the rth derivative of f, and the constants $c_{1,K}$ and $c_{2,K}$ depending on K and given by

$$c_{1,K} = k_0^{1/5} k_2^{-2/5} \quad \text{and} \quad c_{2,K} = \frac{1}{60}\, k_0^{3/5}\, k_2^{-16/5}(3k_2k_4 - 2k_3^2), \tag{1}$$

with $k_j = \int u^j K(u) du$ for $j = 1, 2, \dots$ (see Hall and Marron, 1987, 1991). These asymptotic approximations of h_0 reduce the problem of estimating the optimal bandwidth to that of estimating the quadratic functionals θ_2 and θ_3, this being the idea of the direct plug-in approach to bandwidth selection.

Although several methods for estimating the functionals θ_r, for $r = 0, 1, \dots$, have been studied in the literature (see the references given in Tenreiro, 2011, p. 534, and Chacón and Tenreiro, 2012, p. 524), the class of kernel estimators of θ_r proposed by Hall and Marron (1987) and Jones and Sheather (1991) is widely used in a bandwidth selection context. However, for these kernel estimators the asymptotically optimal bandwidth for estimating θ_r depends on θ_{r+2} (whenever a nonnegative and symmetric kernel is used). This makes the selection of the bandwidth into a somehow cyclic process. Although a multistage strategy could be used to overcome this problem (see Chacón and Tenreiro, 2013, for a detailed description of such a multistage procedure), the standard approach is to use a two-stage procedure with normal reference distribution leading to the popular two-stage direct plug-in bandwidth selector described in Wand and Jones (1995, pp. 71–72) and implemented by the function dpik of the R-package 'KernSmooth' (Wand, 2014, pp. 7–8).

When the support of the underlying density function f is known to be contained within a finite interval $[a, b]$, an alternative approach was followed by Tenreiro (2011) who proposed direct plug-in bandwidth selectors for the kernel density estimator based on the Fourier series estimators of θ_r studied by Laurent (1997). Prompted by the good practical performance of the proposed bandwidth selectors, the main purposes of this paper are: 1) to use estimators of θ_r based on the orthogonal projection of $f^{(r)}$ on the Hermite basis to extend the previous results to the case where the support of f is the whole real line; 2) to examine, from an asymptotic and finite sample point of view, the quality of the proposed Hermite series-based direct plug-in bandwidth selectors. Unlike the standard direct plug-in approach, the new implementation of the plug-in method does not involve multistage strategies and reference distributions are no longer needed.

The rest of this article is organised as follows. In Section 2 we consider Hermite series-based estimators of the quadratic functional θ_r, where the number of Hermite terms included in the estimators may depend on the observed sample, and we establish their consistency, probability orders of convergence and asymptotic normality. In Section 3 these results are used to describe the asymptotic behaviour of direct plug-in bandwidth selectors based on each one of the asymptotic approximations h_1 and h_2 of the exact optimal bandwidth h_0. In Section 4 we propose two data-driven methods for selecting the number of terms to be included in the Hermite series based estimators of θ_r, and in Section 5 we undertake a simulation study to analyse the finite sample behaviour of the proposed direct plug-in bandwidth selectors. The very good finite sample performance presented by the proposed bandwidth selectors when the underlying probability density function presents not only "easy-to-estimate" but also "hard-to-estimate" distribution features, is the most significant aspect with potential practical interest of the proposed methodology. This is a relevant attribute of the Hermite series-based bandwidth selectors proposed in this paper which is not shared by the generality of the existing bandwidth selector methods, which are usually high performing for "easy-to-estimate" densities, but, at the same time, they may be quite inefficient for densities presenting hard distribution features as strong asymmetry or multimodality. All the proofs and some auxiliar results are deferred to Section 6.

The simulations and plots in this paper were carried out using the R software (R Development Core Team, 2014).

2. Hermite series estimators of θ_r

Let $\{h_k, k=0,1,\dots\}$ be the Hermite orthonormal basis of L_2 defined by

$$h_k(x) = (2^k k! \pi^{1/2})^{-1/2} H_k(x) e^{-x^2/2},$$

with $x \in \mathbb{R}$, where H_k is the kth Hermite polynomial given by

$$H_k(x) = (-1)^k e^{x^2} (d^k/dx^k) e^{-x^2}.$$

For $r \in \{0,1,\dots\}$, if we assume that $f^{(r)}$ is square integrable, it is known that $f^{(r)}$ has the L_2 representation $f^{(r)} = \sum_{k=0}^{\infty} a_{r,k} h_k$, where $a_{r,k} = \int f^{(r)}(x) h_k(x) dx$ is the kth Hermite coefficient of $f^{(r)}$, and the quadratic functional of interest $\theta_r = ||f^{(r)}||_2^2$ can be written in terms of the Hermite coefficients of $f^{(r)}$ as

$$\theta_r = \sum_{k=0}^{\infty} a_{r,k}^2.$$

Using the fact that the kth Hermite coefficient of $f^{(r)}$ can be rewritten as

$$a_{r,k} = (-1)^r \int h_k^{(r)}(x) f(x) dx = (-1)^r \mathrm{E}\big(h_k^{(r)}(X_1)\big),$$

whenever f has bounded derivatives up to order r, it can be estimated without bias as in Greblicki and Pawlak (1984) by

$$\hat{a}_{r,k} = \frac{1}{n} \sum_{i=1}^{n} h_k^{(r)}(X_i),$$

which leads to the estimator of θ_r given by

$$\hat{\hat{\theta}}_{r,m} = \sum_{k=0}^{m} \hat{a}_{r,k}^2, \tag{2}$$

where $m = m(n)$ is a sequence of integers converging to infinity with n. A closely related alternative estimator of θ_r (see Section 6, Proposition 1), can be obtained by taking

$$\tilde{\hat{\theta}}_{r,m} = \sum_{k=0}^{m} \widehat{a^2}_{r,k}, \tag{3}$$

where $\widehat{a^2}_{r,k}$ is the unbiased estimator of $a_{r,k}^2$ given by

$$\widehat{a^2}_{r,k} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h_k^{(r)}(X_i) h_k^{(r)}(X_j).$$

As the choice of m should be based on the observed sample, we assume that $m = \hat{m}(X_1, \ldots, X_n)$ which leads to the automatic estimators $\hat{\hat{\theta}}_{r,\hat{m}}$ and $\tilde{\hat{\theta}}_{r,\hat{m}}$ of θ_r. Next we describe the asymptotic behaviour of these estimators that we simply denote by $\hat{\theta}_{r,\hat{m}}$. For $r, p = 0, 1, \ldots$ let $\mathscr{D}_{r,p}$ be the set of all densities f with bounded derivatives up to order $r+p$, where the functions $x \mapsto x^{r+p-i} f^{(i)}(x)$ are assumed to be square integrable, for $i = r, \ldots, r+p$. We denote by $s = r + p$ the order of smoothness of $\mathscr{D}_{r,p}$.

Theorem 1. *For $r = 0, 1, \ldots,$ assume that $f \in \mathscr{D}_{r,p}$, for some $p \in \{0, 1, \ldots\}$.*
(a) Consistency. *If $\hat{m}$ is such that $\hat{m} \xrightarrow{p} +\infty$ and $n^{-1}\hat{m}^{\max\{1, r+5/6\}} \xrightarrow{p} 0$, then*

$$\hat{\theta}_{r,\hat{m}} \xrightarrow{p} \theta_r.$$

(b) Rates of convergence. *Let $\hat{m}$ be such that*

$$P\big(Cn^\xi \leq \hat{m} \leq Dn^\xi\big) \to 1, \tag{4}$$

with $C, D > 0$ and $\xi > 0$. If $s > r$ and

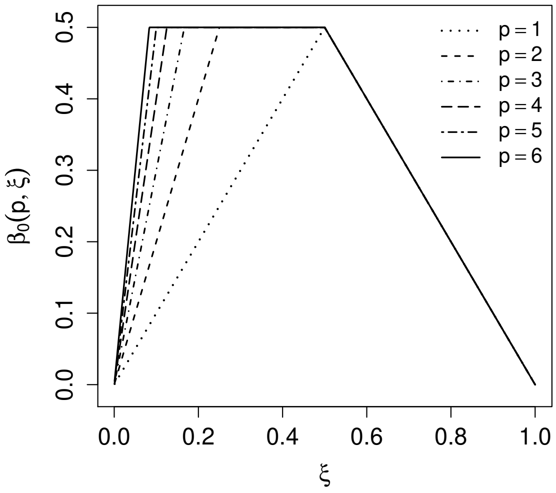
$$0 < \xi < \frac{1}{\max\{1, r + 5/6\}},$$

then

$$\hat{\theta}_{r,\hat{m}} - \theta_r = O_p\Big(n^{-\beta_r(p,\xi)}\Big),$$

where

$$\beta_r(p, \xi) = \min\big\{(1 - \xi\eta(r - p + 5/6))/2\,,\, 1 - \xi\eta(r + 5/6)\,,\, p\xi\big\}.$$



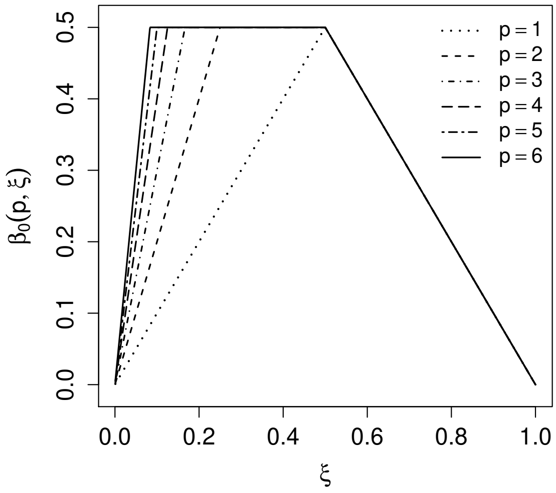


FIGURE 1. Rates of convergence of $\hat{\theta}_{r,\hat{m}}$ to θ_r for $r = 0$ (left) and $r = 2$ (right), as a function of $\xi \in]0, 1/\max\{1, r+5/6\}[$ and $p \in \{1, 2, \ldots, 6\}$.

and $\eta(t) = \max\{1, t\} I(t \geq 0)$.
(c) Asymptotic normality. *Additionally, if* $s \geq 2r + 1$ *and*

$$\frac{1}{2p} \leq \xi < \frac{1}{2\max\{1, r + 5/6\}},$$

then

$$n^{1/2}\big(\hat{\theta}_{r,\hat{m}} - \theta_r\big) \xrightarrow{d} N\big(0, 4\mathrm{Var}(f^{(2r)}(X_1)\big).$$

Remark 1. From part (b) of Theorem 1 we also conclude that: 1) if $s \geq 2r+1$ and $1/(2p) \leq \xi \leq 1/(2\max\{1, r+5/6\})$ the rate of convergence of $\hat{\theta}_{r,\hat{m}}$ to θ_r has the semi-parametric order $n^{-1/2}$. Moreover, the variance $4\mathrm{Var}(f^{(2r)}(X_1))$ is the same as the information bound for the nonparametric estimation of θ_r derived by Bickel and Ritov (1988). 2) if $r < s \leq 2r$, the best rate of convergence of $\hat{\theta}_{r,\hat{m}}$ to θ_r is obtained for $\xi = 1/(s + 5/6)$ and has the order $n^{-(s-r)/(s+5/6)}$. In the former case the same order of convergence can be obtained by the improved kernel-based estimator $\hat{S}_{\mathrm{D},r}$ of θ_r introduced in Jones and Sheather (1991) by employing a kernel of order $2r$. However, in the latter case the rate of convergence of $\hat{\theta}_{r,\hat{m}}$ to θ_r compares favourably with that achieved for $\hat{S}_{\mathrm{D},r}$ which is of order $n^{-(s-r)/(4r+1)}$.

3. Hermite series-based plug-in bandwidth selectors

In this section we describe the asymptotic behaviour of the relative errors associated to each one of the plug-in bandwidth selectors defined by

$$\hat{h}_{1,\hat{m}} = c_{1,K}\,\hat{\theta}_{2,\hat{m}}^{-1/5} n^{-1/5} \tag{5}$$

and

$$\hat{h}_{2,\hat{m}} = c_{1,K}\,\hat{\theta}_{2,\hat{m}}^{-1/5} n^{-1/5} + c_{2,K}\,\hat{\theta}_{2,\hat{m}}^{-8/5}\,\hat{\theta}_{3,\hat{m}}\, n^{-3/5}, \tag{6}$$

where $\hat{\theta}_{r,m}$ denotes either $\hat{\hat{\theta}}_{r,m}$ or $\hat{\hat{\hat{\theta}}}_{r,m}$ defined by (2) and (3), respectively, $c_{1,K}$ and $c_{2,K}$ are given by (1), and $\hat{m} = \hat{m}(X_1, \dots, X_n)$ is a random sequence of nonnegative integers. We will always assume that the kernel K is a kernel of order 2, that is, $\int u^2|K(u)|du < \infty$, with $k_1 = 0$ and $k_2 \neq 0$. We also assume that K is continuous at zero with $k_0 < 2K(0)$. As mentioned earlier, under these assumptions the existence of an exact optimal bandwidth h_0, in the sense of the minimisation of the mean integrated square error, can be established whenever f is square integrable (see Chacón et al., 2007, Theorem 1).

Theorem 2. *Let K be a kernel satisfying the previously stated conditions with $\int |u|^5|K(u)|du < \infty$. Assume that $f \in \mathscr{D}_{2,p}$, for some $p \in \{0, 1, \dots\}$, with bounded, integrable and continuous derivatives up to order 4. Finally, let $\hat{m}$ be such that $\hat{m} \xrightarrow{p} +\infty$ and $n^{-1}\hat{m}^{2+5/6} \xrightarrow{p} 0$.*
(a) Asymptotic behaviour of $\hat{h}_{1,\hat{m}}$. *We have*

$$\frac{\hat{h}_{1,\hat{m}}}{h_0} \xrightarrow{p} 1;$$

if $p \geq 1$ and $\hat{m}$ satisfies (4) with

$$0 < \xi < \frac{1}{3} \cdot \frac{18}{17}, \tag{7}$$

then

$$\frac{\hat{h}_{1,\hat{m}}}{h_0} - 1 = O_p\Big(n^{-\min\{\beta_2(p,\xi)\,,\,2/5\}}\Big),$$

where

$$\beta_2(p, \xi) = \min\{(1 - \xi\eta(17/6 - p))/2\,,\, 1 - 17\xi/6\,,\, p\xi\}.$$

Moreover, if $p \geq 3$ *and*

$$\frac{1}{5} \cdot \frac{2}{p} < \xi < \frac{1}{5} \cdot \frac{18}{17}, \tag{8}$$

then

$$n^{2/5}\left(\frac{\hat{h}_{1,\hat{m}}}{h_0} - 1\right) \xrightarrow{p} -c_{1,K}^{-1}\, c_{2,K}\, \theta_2^{-7/5}\theta_3.$$

(b) Asymptotic behaviour of $\hat{h}_{2,\hat{m}}$. *If* $p \geq 1$ *we have*

$$\frac{\hat{h}_{2,\hat{m}}}{h_0} \xrightarrow{p} 1;$$

if $\hat{m}$ *and* ξ *satisfy (4) and (7), respectively, we have*

$$\frac{\hat{h}_{2,\hat{m}}}{h_0} - 1 = O_p\Big(n^{-\beta_2(p,\xi)}\Big).$$

Moreover, if $p \geq 3$ *and*

$$\frac{1}{2p} \leq \xi < \frac{1}{6} \cdot \frac{18}{17},$$

then

$$n^{1/2}\left(\frac{\hat{h}_{2,\hat{m}}}{h_0} - 1\right) \xrightarrow{d} N\big(0, \sigma^2(f)\big),$$

with

$$\sigma^2(f) = \frac{4}{25}\left(\frac{\mathrm{E}(f^{(4)}(X_1)^2)}{\mathrm{E}^2(f^{(4)}(X_1))} - 1\right).$$

Remark 2. The order $n^{-1/2}$ obtained for the rate of convergence of the relative error $\hat{h}_{2,\hat{m}}/h_0 - 1$ by taking $\xi = 1/6$ when $p \geq 3$, is, in a minimax sense, the best possible rate of convergence as shown by Hall and Marron (1991). Moreover, the variance $\sigma^2(f)$ is the same as the best possible constant coefficient for bandwidth selection derived by Fan and Marron (1992).

4. The automatic selection of m

We are interested in estimating the unknown probability density function f by using the kernel estimator $f_{n,h}$, where the bandwidth h is one of the data-dependent bandwidths $\hat{h}_{1,\hat{m}}$ and $\hat{h}_{2,\hat{m}}$, defined by (5) and (6), respectively. As the estimator $\hat{\hat{\theta}}_{r,m}$ of θ_r defined by (3) may occasionally produce poor, sometimes negative, estimates of θ_r when the size of the sample is small, and it perform similarly to $\hat{\hat{\theta}}_{r,m}$ when the sample size is moderate or large, the data-dependent bandwidths based on the estimators $\hat{\hat{\theta}}_{r,m}$ are not considered hereafter. From now on we take for K the gaussian kernel, i.e., $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, $x \in \mathbb{R}$.

The bandwidths $\hat{h}_{1,\hat{m}}$ and $\hat{h}_{2,\hat{m}}$ depend on the random variable $\hat{m} = \hat{m}(X_1, \ldots, X_n)$, where $\hat{m}+1$ is the number of Hermite terms included in the estimators of θ_2 and θ_3 that appear in their definitions. In order to explore the distribution of $\mathrm{ISE}(f; n, \hat{h}_{i,m})$, for $i = 1, 2$, in each graph of Figure 2 we show 40 boxplots describing the empirical distribution of $\mathrm{ISE}(f; n, \hat{h}_{1,m})$ based on 500 simulated samples from densities #2, #3, and #13 of the Marron and Wand (1992) set of mixture of normal densities, for $m \in \{0, 1, ..., 10, 20, ..., 300\}$. Similar behaviours can also be observed for the bandwidth selector $\hat{h}_{2,m}$, but the corresponding graphs are not included here to save space. Also, we include a polygonal line going through the sample mean values of these distributions, thus giving an approximation of $\mathrm{EISE}(m) := \mathrm{E}(\mathrm{ISE}(f; n, \hat{h}_{1,m}))$. The solid red circle is used to point out the optimal value of m in the sense of minimising the approximation of the EISE function. Similar graphs were generated for all Marron and Wand (1992) densities and sample sizes $n = 25 \cdot 2^k$, $k = 0, 1, \ldots, 8$.

Densities #2 and #3, whose empirical distributions of $\mathrm{ISE}(f; n, \hat{h}_{1,m})$ are shown at the top of Figure 2, are representative members of two groups of densities we can identify among our 15 test densities. The pattern displayed by distribution #2 is shared by other densities having easy-to-identify features such as densities #1, #6, #7, #8 and #9, for which a small value of m seems to be the best choice. The same occurs for other hard-to-estimate densities only when the sample size is small or moderate. This is the case of densities #10, #13 and #15 for $n \leq 50$, #13 for $n \leq 400$, and #11 for $n \leq 1600$. As pointed out by Chacón and Tenreiro (2013, p. 2204) in a similar context, the reason for the good performance of a low value of m for such combinations of densities and sample sizes is that they present distribution features that are not revealed until the sample size is above some threshold. This situation is illustrated by the graphs at the bottom of Figure 2 where the empirical distribution of $\mathrm{ISE}(f; n, \hat{h}_{1,m})$ for density #13 is shown for sample sizes $n = 100$ and $n = 800$. The pattern displayed by distribution #3 is shared by other test densities for which using a large value of m seems to be highly advisable. Other than density #3, this second group of densities includes densities $\#4, \#5$ and $\#14$, and depending on the sample size also densities $\#10, \#12, \#13$ and $\#15$ for moderate and large sample sizes, and $\#11$ for very large values of n.

Based on these considerations, we conclude that if we want to deal with a wide set of distribution characteristics, any reasonable data-based selector $\hat{m}$ of m should take values on a set including small as well as large values of m. Two methods for selecting m will be considered. In both cases the value $\hat{m} = \hat{m}(X_1, \ldots, X_n)$ is obtained by minimising a certain criterion function over a set of integers

$$\mathscr{M}_n = \{L_n, L_n + 1, \ldots, U_n\},$$

where $L_n < U_n$ are deterministic sequences of nonnegative integers whose asymptotic behaviour determines that of the bandwidth selectors $\hat{h}_{1,\hat{m}}$ and $\hat{h}_{2,\hat{m}}$. Assuming that the underlying density f satisfies the conditions of Theorem 2 for some $p \geq 3$, we will take $L_n = \lfloor Cn^{\xi} \rfloor$ and $U_n = \lfloor Dn^{\xi} \rfloor$, with $C = 0.2$, $D = 80$ and $\xi = 1/6$. This leads to $L_n = 0$ and $117 \leq U_n \leq 330$ for $10 \leq n \leq 5 \cdot 10^3$.

Taking into account that choosing m among the set $\mathscr{M}_n$ is equivalent to selecting one of the bandwidths $\hat{h}_{i,m}$, for $m \in \mathscr{M}_n$, where $i = 1, 2$, and that for a squared integrable density function f the mean integrated square error of $f_{n,h}$ is given by $\mathrm{E}||f_{n,h} - f||_2^2 = W(h) + ||f||_2^2$, with $W(h) = \frac{k_0}{nh} + \int L_h(x - y) f(x) f(y) dx dy$, where $L = (1 - n^{-1}) K * \bar{K} - 2K$, with $\bar{K}(u) = K(-u)$ and $*$ denotes the convolution product, we can adapt the strategy followed in Chacón and Tenreiro (2013) in order to propose a first data-dependent method for selecting m. For $i = 1, 2$, it is defined by the first integer $\hat{m}_{i,W_\gamma}$ satisfying

$$\hat{m}_{i,W_\gamma} = \arg \min_{m \in \mathscr{M}_n} \hat{W}_\gamma\big(\hat{h}_{i,m}\big),$$

where $\hat{W}_\gamma(h)$ is the weighted cross-validation function defined, for $h > 0$, by

$$\hat{W}_\gamma(h) = \frac{k_0}{nh} + \frac{\gamma}{n(n-1)} \sum_{1 \leq i \ neqj \leq n} L_h(X_i - X_j),$$

where $0 < \gamma \leq 1$ needs to be chosen by the user. We refer the reader to Tenreiro (2017) for the weighted least-squares cross-validation bandwidth selector for kernel density estimation. For $\gamma = 1$, $\hat{W}_\gamma(h)$ is the standard least-squares cross-validation function proposed by Rudemo (1982) and Bowman (1984). Hearafter the bandwidths $\hat{h}_{i,\hat{m}_{i,W_\gamma}}$ will be simply denoted by $\hat{h}_{i,\hat{m}_{W_\gamma}}$.

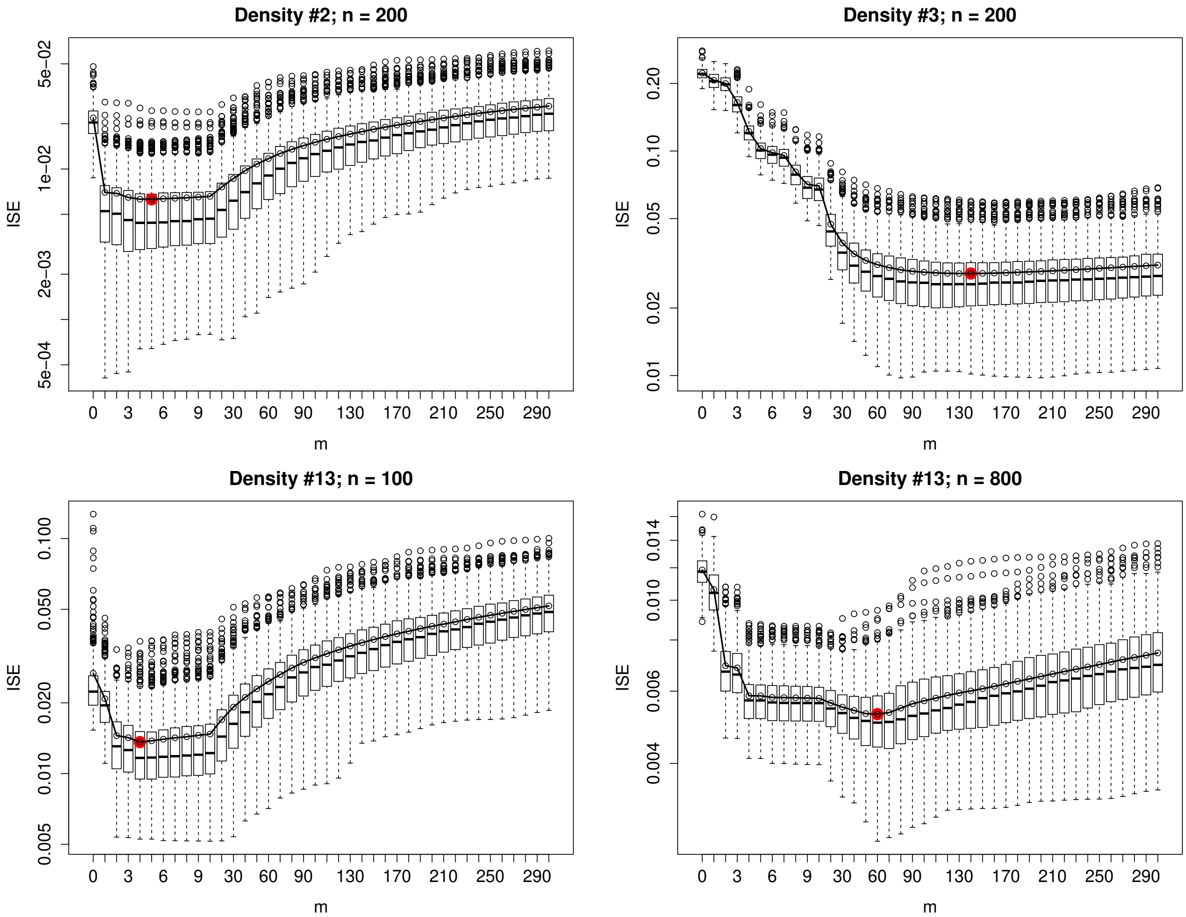


FIGURE 2. Empirical distribution of ISE$(f; n, \hat{h}_{1,m})$ depending on m for densities # 2 ($n = 200$), # 3 ($n = 200$) and # 13 ($n = 100, 800$) from the Marron and Wand (1992) set of normal mixture densities. The number of replications is 500.

The second method we consider for selecting m was used in the context of Fourier series-based plug-in bandwidth selectors by Tenreiro (2011). In this case the selection of m does not depend on the considered bandwidth selector. The idea is to take m in such a way that f can be well approximated, in the sense of the mean integrated squared error, by the Hermite series-based estimator of f defined by $\hat{f}_{n,m} = \sum_{k=0}^{m} \hat{a}_{0,k} h_k$. For a squared integrable density function f, Walter (1977) proves that the mean integrated square error of $\hat{f}_{n,m}$ is given by $\mathrm{E}||\hat{f}_{n,m} - f||_2^2 = H(m) + ||f||_2^2$, where $H(m) = \frac{1}{n}\sum_{k=0}^{m} \int h_k(x)^2 f(x)dx - \left(1 + \frac{1}{n}\right)\sum_{k=0}^{m} a_{0,k}^2$. Therefore, the second data-dependent method for selecting m we consider is defined by the first integer $\hat{m}_{H_\gamma}$ satisfying

$$\hat{m}_{H_\gamma} = \arg \min_{m \in \mathscr{M}_n} \hat{H}_\gamma(m),$$

where

$$\hat{H}_\gamma(m) = \frac{1}{n} \sum_{k=0}^{m} \frac{1}{n} \sum_{i=1}^{n} h_k(X_i)^2 - \gamma\Big(1 + \frac{1}{n}\Big) \sum_{k=0}^{m} \widehat{a^2}_{0,k},$$

for some $0 < \gamma \leq 1$. Although the motivation for this second method for selecting m can be considered less convincing than the previous one, because it is not related with the kernel density estimator of f we are interested in nor with the Hermite series-based estimators of θ_2 and θ_3 we are using, we will see that it performs quite well in practice, being less time consuming than the method based on $\hat{W}_\gamma$ especially for large sample sizes.

The inclusion of the correction parameter γ in the previous criterion functions is crucial for the good performance of both methods. To the best of our knowledge, a similar idea was for the first time suggested by Hart (1985) for selecting the number of terms to be used in a Fourier series-based density estimator. As the considered set $\mathscr{M}_n$ of possible values of m includes large values of m, some simulation experiments performed for all normal mixture densities of Marron and Wand (1992) reveal that taking $\gamma = 1$, in which case $\hat{W}_\gamma(h)$ and $\hat{H}_\gamma(m)$ are unbiased estimators of $\mathrm{E}||f_{n,h} - f||_2^2 - ||f||_2^2$ and $\mathrm{E}||\hat{f}_{n,m} - f||_2^2 - ||f||_2^2$, respectively, does not prevent the user from getting excessively large values of m, which leads to very poor results especially for densities with easy-to-estimate distribution features. In fact, excessively large values of m might lead to an overestimation of the quadratic functional θ_2, and therefore to an underestimation of the optimal bandwidth h_0. This is an undesirable situation since, as is well known, the kernel density estimator is penalised much more by excessively small than excessively large bandwidths. Taking into account that the functions $\gamma \mapsto \hat{h}_{i,\hat{m}_{W_\gamma}}$ are nonincreasing ($i = 1, 2$), and the function $\gamma \mapsto \hat{m}_{H_\gamma}$ is nondecreasing with probability one, we may expect to soften the above mentioned problems by including a correction parameter strictly less than one in the considered criterion functions. As suggested by these properties, the simulation results support the idea that small values of γ are more appropriate for easy-to-estimate densities, whereas large values of γ are more adequate for hard-to-estimate densities. In order to find a compromise between these two extreme situations, we decide to follow Tenreiro (2011) suggestion of taking $\gamma = 0.5$.

5. Simulation study

We present in this section the results of a simulation study carried out to analyse the finite sample behaviour of the Hermite series-based direct plug-in bandwidth selectors introduced in the previous sections, namely $\hat{h}_{1,\hat{m}_{W_\gamma}}$, $\hat{h}_{1,\hat{m}_{H_\gamma}}$, $\hat{h}_{2,\hat{m}_{W_\gamma}}$, and $\hat{h}_{2,\hat{m}_{H_\gamma}}$, with $\gamma = 0.5$. We use as test densities the 15 normal mixture densities of Marron and Wand (1992) that we referred to in Section 4. Two other bandwidth selectors are included in the study: the two-stage direct plug-in bandwidth selector (PI), implemented by the function dpik of the R-package 'KernSmooth', and the standard least-square cross-validation bandwidth selector (CV). It is well known that the PI method performs quite well for "easy-to-estimate" densities (e.g. $\#1, \#2, \#6, \#8, \#9$), whereas the CV method performs exceptionally well for "hard-to-estimate" densities (e.g. $\#3, \#4, \#5, \#14, \#15$), these being the main reasons for including these bandwidth selectors in our study.

For different sample sizes and for each one of the 15 test distributions the quality of each one of the considered bandwidths is analysed through the measure of stochastic performance defined by

$$L_2\text{–norm of } \mathrm{ISE}(f; n, \hat{h}) = \sqrt{\mathrm{Var}(\mathrm{ISE}(f; n, \hat{h})) + \mathrm{E}^2(\mathrm{ISE}(f; n, \hat{h}))}.$$

This performance measure takes into account not only the mean of the $\mathrm{ISE}(f; n, \hat{h})$ distribution, but also its variability. As the behaviour shown by the bandwidths $\hat{h}_{1,\hat{m}_{W_\gamma}}$ and $\hat{h}_{1,\hat{m}_{H_\gamma}}$ is close to that one of the bandwidths $\hat{h}_{2,\hat{m}_{W_\gamma}}$ and $\hat{h}_{2,\hat{m}_{H_\gamma}}$, respectively, only the behaviour of these two last bandwidths is reported in Figures 3, 4 and 5. In these figures the empirical L_2–norm of $\mathrm{ISE}(f; n, \hat{h})$, based on 500 replications, is shown for the bandwidth selectors $\hat{h}_{2,\hat{m}_{W_\gamma}}$, $\hat{h}_{2,\hat{m}_{H_\gamma}}$, $\hat{h}_{\mathrm{PI}}$ and $\hat{h}_{\mathrm{CV}}$ and sample sizes $n = 25 \cdot 2^k$, $k = 0, 1, \ldots, 7$.

As we can see from the graphics, the two Hermite series-based direct plug-in bandwidths perform similarly for all the test distributions, presenting a very good performance against the PI and CV methods. For some of the considered test densities, the new bandwidth selectors seem to mimic the behaviour of the best of these two classic bandwidths, retaining the good performance of the PI bandwidth for "easy-to-estimate" densities and sharing the superior performance of the CV bandwidth for "hard-to-estimate" densities. It is particularly interesting the cases of densities $\#10$ and $\#12$, where the new bandwidth selectors behave similarly to the PI selector for

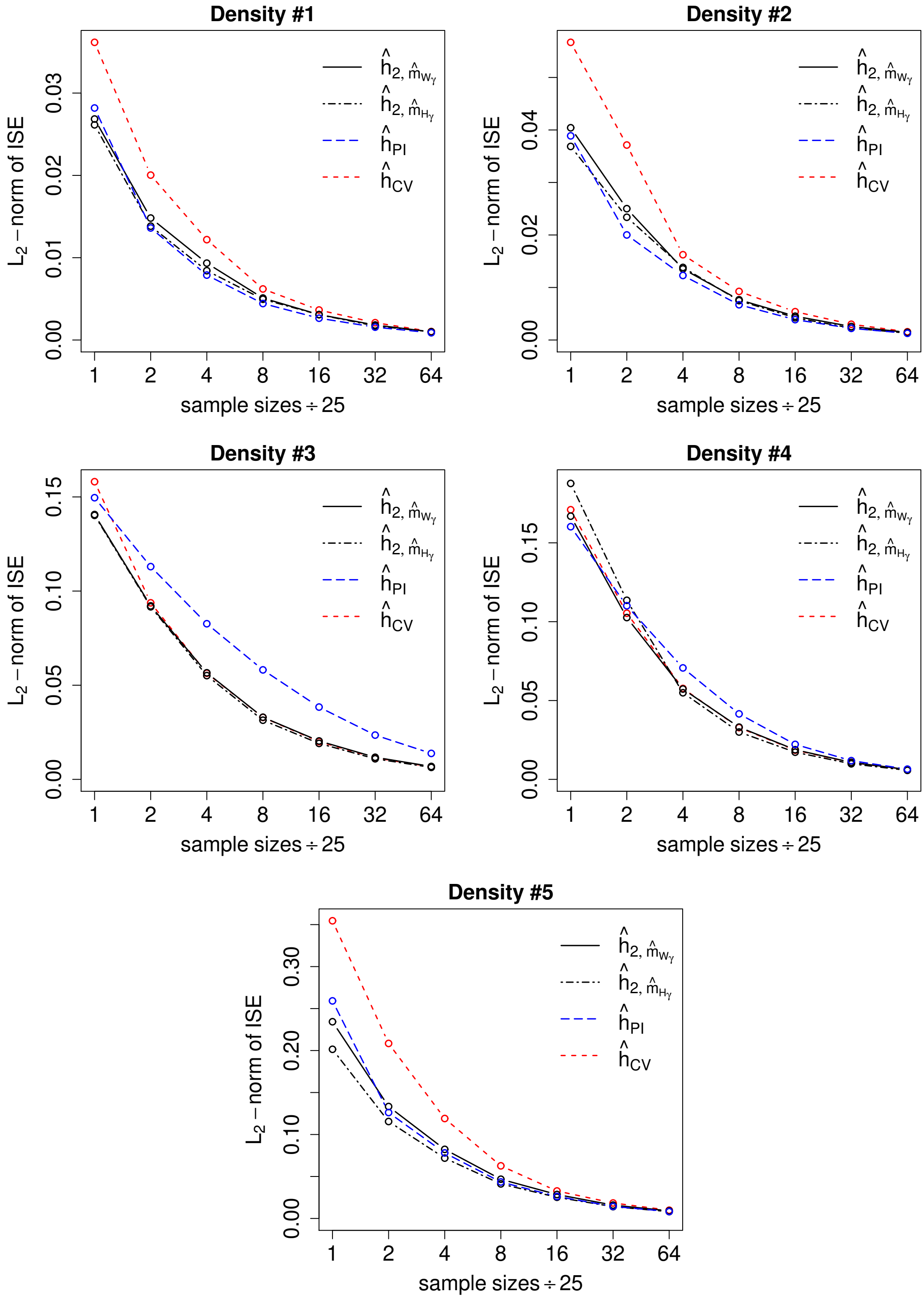


FIGURE 3. *Empirical L_2–norm of* $\mathrm{ISE}(f;n,\hat{h})$ *associated to the bandwidths* $\hat{h}_{2,\hat{m}_{W_\gamma}}$, $\hat{h}_{2,\hat{m}_{H_\gamma}}$ *(*$\gamma = 0.5$*),* $\hat{h}_{\mathrm{PI}}$ *and* $\hat{h}_{\mathrm{CV}}$*, for test densities #1 to #5. The number of replications is* 500.

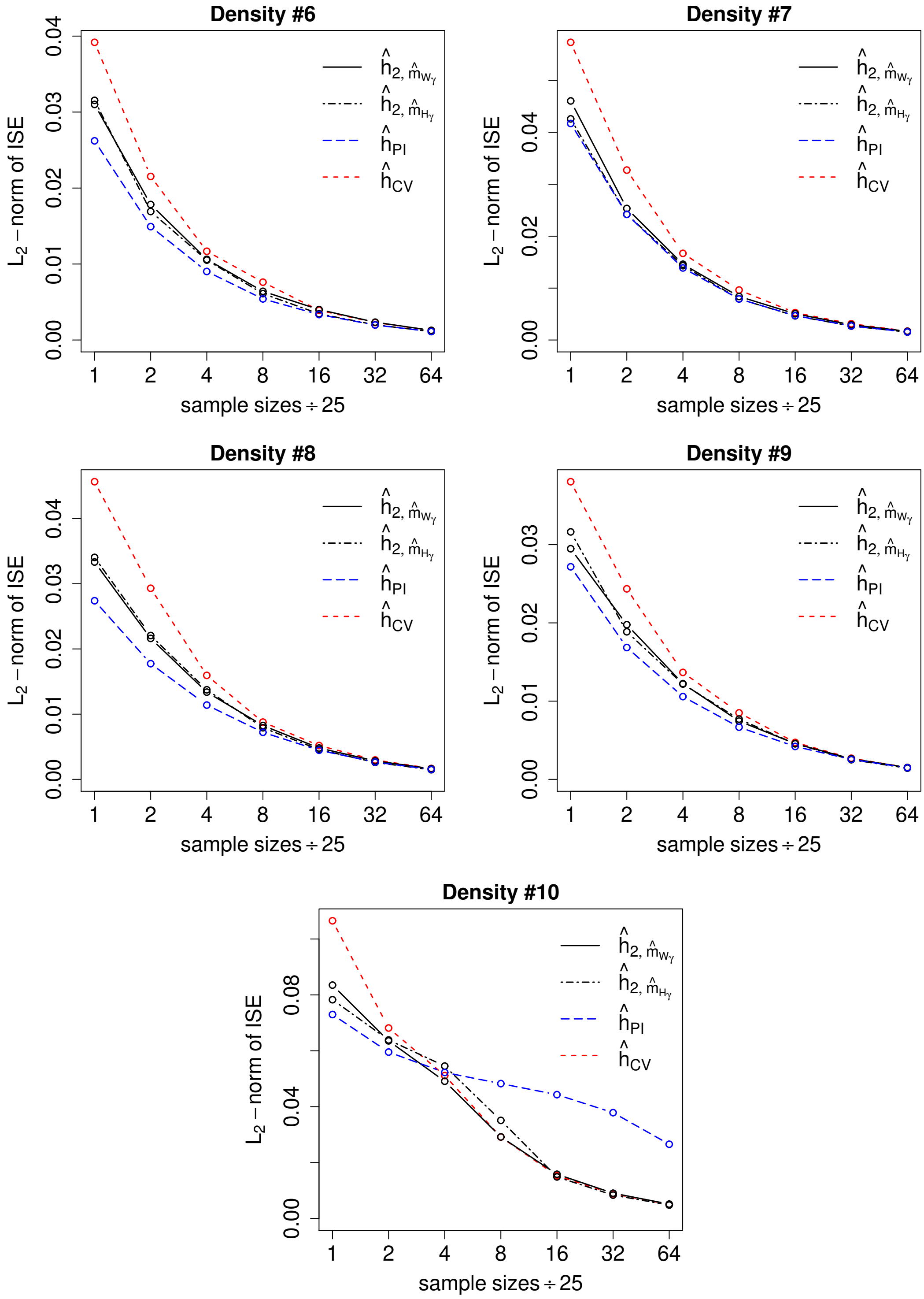


FIGURE 4. *Empirical L_2–norm of* $\mathrm{ISE}(f; n, \hat{h})$ *associated to the bandwidths* $\hat{h}_{2,\hat{m}_{W_\gamma}}$, $\hat{h}_{2,\hat{m}_{H_\gamma}}$ $(\gamma = 0.5)$, $\hat{h}_{\mathrm{PI}}$ *and* $\hat{h}_{\mathrm{CV}}$, *for test densities #6 to #10. The number of replications is* 500.

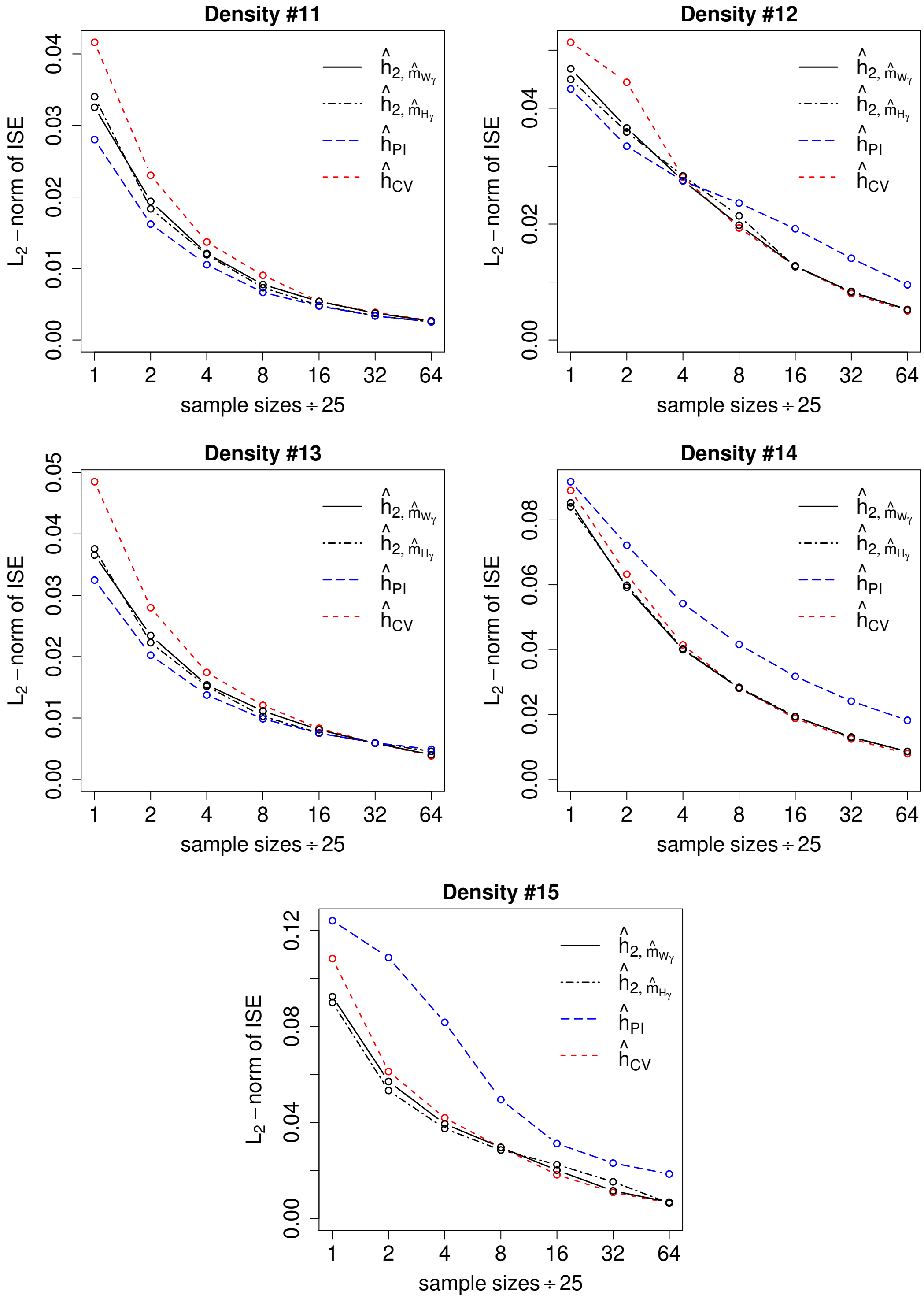


FIGURE 5. *Empirical L_2–norm of* $\mathrm{ISE}(f;n,\hat{h})$ *associated to the bandwidths* $\hat{h}_{2,\hat{m}_{W_\gamma}}$, $\hat{h}_{2,\hat{m}_{H_\gamma}}$ ($\gamma = 0.5$), $\hat{h}_{\mathrm{PI}}$ *and* $\hat{h}_{\mathrm{CV}}$, *for test densities #11 to #15. The number of replications is* 500.

small sample sizes, and similarly to the CV selector for moderate and large sample sizes.

Based on this evidence, we expect that the new direct plug-in bandwidths proposed in this paper might present a good overall performance for a wide range of density features, which is a distinctive quality in particular when no information about the underlying density shape is available or when a complex data structure is suspected.

6. Proofs

6.1. Proof of Theorem 1. We recall that $\hat{\bar{\theta}}_{r,\hat{m}}$ and $\hat{\tilde{\theta}}_{r,\hat{m}}$ are defined by (2) and (3), respectively, where $\hat{m} = \hat{m}(X_1, \ldots, X_n)$ is a random sequence of nonnegative integers. We will first set three preliminar propositions that will prove usefull.

Proposition 1. *For $r = 0, 1, \ldots$ and $n \geq 2$ we have*

$$\hat{\tilde{\theta}}_{r,\hat{m}} = \frac{n}{n-1}\,\hat{\bar{\theta}}_{r,\hat{m}} - R_{r,\hat{m}}, \tag{10}$$

where

$$0 \leq R_{r,m} \leq B_r n^{-1} m^{\max\{1, r+5/6\}},$$

and B_r is a constant independent of m. Moreover, if $m_1 = m_1(n)$ and $m_2 = m_2(n)$ are sequences of nonnegative integers such that $m_1 \leq \hat{m} \leq m_2$, then

$$\hat{\tilde{\theta}}_{r,m_1} - R_{r,m_2} \leq \hat{\tilde{\theta}}_{r,\hat{m}} \leq \hat{\tilde{\theta}}_{r,m_2} + R_{r,m_2}. \tag{11}$$

Proof: From the definitions of $\hat{\bar{\theta}}_{r,m}$ and $\hat{\tilde{\theta}}_{r,m}$ we easily see that equality (10) holds with

$$R_{r,m} = \frac{1}{n(n-1)} \sum_{k=0}^{m} \sum_{i=1}^{n} h_k^{(r)}(X_i)^2.$$

Taking into account that there exist constants $C_r > 0$, independent of k, such that

$$\sup_{x \in \mathbb{R}} |h_k^{(r)}(x)| \leq C_r (k+1)^{r/2 - 1/12}, \tag{12}$$

for $k = 0, 1, \ldots$ and $r = 0, 1, 2, \ldots$ (see Walter, 1977, pp. 1259–1260), we conclude that

$$\begin{aligned} 0 \leq R_{r,m} &\leq \frac{1}{n(n-1)} \sum_{k=0}^{m} \sum_{i=1}^{n} \big(C_r (k+1)^{r/2-1/12}\big)^2 \\ &\leq 2C_r^2 n^{-1} \sum_{k=0}^{m} (k+1)^{r-1/6} \leq B_r n^{-1} m^{\max\{1, r+5/6\}}, \end{aligned}$$

for some constant $B_r > 0$ independent of m. Finally, the double inequality (11) follows straightforward from (10) and the fact that $R_{r,m}$ is a nondecreasing function of m. ■

Proposition 2. *For $r = 0, 1, \ldots,$ assume that $f \in \mathscr{D}_{r,p}$, for some $p \in \{0, 1, \ldots\}$. Then for all $n, m \in \mathbb{N}$ we have*

$$\mathrm{E}\big(\hat{\hat{\theta}}_{r,m} - \theta_r\big)^2 \leq D_1 n^{-1} m^{\eta(r-p+5/6)} + D_2\, n^{-2} m^{2\eta(r+5/6)} + D_3\, m^{-2p} \nu_m,$$

where $D_1, D_2, D_3 > 0$ are constants independent of n and m, $\nu_m \geq 0$ is such that $\nu_m \to 0$, as $m \to \infty$, and $\eta(t) = \max\{1, t\} I(t \geq 0)$.

Proof: In order to establish the stated result, we use the classical decomposition

$$\mathrm{E}\big(\hat{\hat{\theta}}_{r,m} - \theta_r\big)^2 = \mathrm{Var}(\hat{\hat{\theta}}_{r,m}) + \big(\mathrm{E}(\hat{\hat{\theta}}_{r,m}) - \theta_r\big)^2.$$

We first examine the bias term. For $f \in \mathscr{D}_{r,p}$ we observe that the real-valued function $x \mapsto (x - d/dx)^p f^{(r)}(x)$ is square integrable and its $(k+p)$th Hermite coefficient, we denote by $b_{r,p,k+p}$, is related to the kth Hermite coefficient of $f^{(r)}$ by the expression

$$b_{r,p,k+p} = (2(k+p))^{1/2} (2(k+p-1))^{1/2} \ldots (2(k+1))^{1/2} a_{r,k},$$

for $k = 1, 2, \ldots$ (see Walter, 1977, pp. 1261). Thus we have

$$|a_{r,k}| \leq \big(2(k+1)\big)^{-p/2} |b_{r,p,k+p}|, \tag{13}$$

for $k = 0, 1, 2, \ldots,$ which leads to

$$\big(\mathrm{E}(\hat{\hat{\theta}}_{r,m}) - \theta_r\big)^2 = \Bigg(\sum_{k=m+1}^{\infty} a_{r,k}^2\Bigg)^2 \leq \sum_{k=m+1}^{\infty} \big(2(k+1)\big)^{-p} |b_{r,p,k+p}|^2 = O\big(m^{-2p} \nu_m\big), \tag{14}$$

where $\nu_m = \big(\sum_{k=m+1}^{\infty} |b_{r,p,k+p}|^2\big)^2$ converges to zero as m tends to infinity.

Turning now to the variance term, we notice that $\hat{\hat{\theta}}_{r,m}$ is a U-statistics as it can be written in the form

$$\hat{\hat{\theta}}_{r,m} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} H_{r,m}(X_i, X_j),$$

where $H_{r,m}$ is the symmetric function

$$H_{r,m}(x,y) = \sum_{k=0}^{m} h_k^{(r)}(x) h_k^{(r)}(y). \tag{15}$$

From Hoeffding's formula for the variance of a U-statistics (see Lee, 1990, Theorem 3, p. 12), we have

$$\mathrm{Var}\big(\hat{\hat{\theta}}_{r,m}\big) = \frac{2}{n(n-1)}\big(2(n-2)\sigma_{1,r,m}^2 + \sigma_{2,r,m}^2\big), \tag{16}$$

where $\sigma_{1,r,m}^2 = \mathrm{Var}(G_{r,m}(X_1))$ and $\sigma_{2,r,m}^2 = \mathrm{Var}(H_{r,m}(X_1, X_2))$, with

$$G_{r,m}(y) = \mathrm{E}(H_{r,m}(X_1, y)) = \sum_{k=0}^{m} \mathrm{E}\big(h_k^{(r)}(X_1)\big) h_k^{(r)}(y) = (-1)^r \sum_{k=0}^{m} a_{r,k} h_k^{(r)}(y). \tag{17}$$

From (12), (13) and the triangular inequality, we have

$$\begin{aligned}\sigma_{1,r,m}^2 &\le \mathrm{E}\big(G_{r,m}(X_1)^2\big) \le \left(\sum_{k=0}^{m} |a_{r,k}| \Big(\mathrm{E}\big(h_k^{(r)}(X_1)^2\big)\Big)^{1/2}\right)^2 \\ &\le C_r^2 2^{-p} \sum_{k=0}^{\infty} b_{r,p,k}^2 \sum_{k=0}^{m} (k+1)^{r-p-1/6} = O\big(m^{\eta(r-p+5/6)}\big). \end{aligned} \tag{18}$$

Regarding $\sigma_{2,r,m}^2$, from (12) and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sigma_{2,r,m}^2 &\le \mathrm{E}\big(H_{r,m}(X_1, X_2)^2\big) = \sum_{k,l=0}^{m} \big(\mathrm{E}\big(h_k^{(r)}(X_1) h_l^{(r)}(X_1)\big)\big)^2 \\ &\le \left(\sum_{k=0}^{m} \mathrm{E}\big(h_k^{(r)}(X_1)^2\big)\right)^2 = O\big(m^{2\eta(r+5/6)}\big). \end{aligned} \tag{19}$$

Therefore, from (16) we get

$$\mathrm{Var}\big(\hat{\hat{\theta}}_{r,m}\big) = O\big(n^{-1} m^{\eta(r-p+5/6)} + n^{-2} m^{2\eta(r+5/6)}\big),$$

which concludes the proof. ■

Proposition 3. *For $r = 0, 1, \ldots,$ assume that $f \in \mathscr{D}_{r,p}$, for some $p \in \{r+1, r+2, \ldots\}$. If $m = m(n)$ is a deterministic sequence of nonnegative integers such that $n^{-1/2} m^{\max\{1, r+5/6\}} \to 0$ and $n^{1/2} m^{-p} = O(1)$, then*

$$n^{1/2}\big(\hat{\tilde{\theta}}_{r,m} - \theta_r\big) \xrightarrow{d} N\big(0, 4\mathrm{Var}(f^{(2r)}(X_1)\big).$$

Proof: From the Hoeffding's decomposition (see Lee, 1990, Theorem 1, p. 26), we have

$$\hat{\tilde{\theta}}_{r,m} - \mathrm{E}(\hat{\tilde{\theta}}_{r,m}) = \frac{2}{n} \sum_{i=1}^{n} \{G_{r,m}(X_i) - \mathrm{E}(G_{r,m}(X_i))\} + U_{r,n},$$

where the degenerated U-statistics $U_{r,n}$ is defined by

$$U_{r,n} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \bar{H}_{r,m}(X_i, X_j),$$

with $\bar{H}_{r,m}(x, y) = H_{r,m}(x, y) - G_{r,m}(x) - G_{r,m}(y) + \mathrm{E}(H_{r,m}(X_1, X_2))$, and $H_{r,m}$ and $G_{r,m}$ are given by (15) and (17), respectively. From (18) and (19) we get

$$\begin{aligned}\mathrm{Var}(U_{r,n}) &= O\Big(n^{-2}\Big(\mathrm{E}\big(G_{r,m}(X_1)^2\big) + \mathrm{E}\big(H_{r,m}(X_1, X_2)^2\big)\Big)\Big) \\ &= O\big(n^{-2} m^{\eta(r-p+5/6)} + n^{-2} m^{2\eta(r+5/6)}\big).\end{aligned}$$

Using (14) and the assumptions on the sequence $m = m(n)$, we conclude that

$$n^{1/2}\big(\hat{\tilde{\theta}}_{r,m} - \theta_{r,m}\big) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \{G_{r,m}(X_i) - \mathrm{E}(G_{r,m}(X_i))\} + o_p(1).$$

The stated asymptotic normality follows now from the central limit theorem, whenever we prove that $\sup_{m \in \mathbb{N}} \sup_{x \in \mathbb{R}} |G_{r,m}(x)| < \infty$, and $G_{r,m}(x) \to (-1)^r f^{(2r)}(x)$, as $m \to \infty$, for all $x \in \mathbb{R}$, where $G_{r,m}$ is given by (17).

The first property follows from (12), (13) and the fact that $p \ge r + 1$. In fact, we have

$$\begin{aligned}\sup_{m \in \mathbb{N}} \sup_{x \in \mathbb{R}} |G_{r,m}(x)| &\le \sum_{k=0}^{\infty} |a_{r,k}| \sup_{x \in \mathbb{R}} |h_k^{(r)}(x)| \\ &\le 2^{-p/2} C_r \bigg(\sum_{k=0}^{\infty} (k+1)^{r-p-1/6}\bigg)^{1/2} \bigg(\sum_{k=0}^{\infty} b_{r,p,k+p}^2\bigg)^{1/2}.\end{aligned}$$

The pointwise convergence of $G_{r,m}$ to $(-1)^r f^{(2r)}$ follows from the differentiation theorem under the integral sign and the fact that the rth derivative of f can be expressed as $f^{(r)}(x) = \sum_{k=0}^{\infty} a_{r,k} h_k(x)$, for all $x \in \mathbb{R}$ (see Greblicki and Pawlak, 1985, Lemma 1). ■

Using the results established before, we may now prove Theorem 1. From the first part of Proposition 1, it is enough to consider the estimator $\hat{\hat{\theta}}_{r,\hat{m}}$.

Proof of part (a) of Theorem 1: It follows from the assumptions on $\hat{m}$ that $\mathrm{P}\big(A_n(M,N)\big) \to 1$, as $n \to \infty$, for all $M \in \mathbb{N}$ and $N > 0$, where $A_n(M,N) = \big\{M \le \hat{m} \le \lfloor (Nn)^{1/\max\{1,r+5/6\}} \rfloor\big\}$, with $\lfloor x \rfloor$ the integer part of x. Using Proposition 1 with $m_1(n) = M$ and $m_2(n) = \lfloor (Nn)^{1/\max\{1,r+5/6\}} \rfloor$, for $\epsilon > 0$ we have

$$\begin{aligned}\mathrm{P}\big(|\hat{\hat{\theta}}_{r,\hat{m}} - \theta_r| \ge \epsilon\big) \le \mathrm{P}\big(|\hat{\hat{\theta}}_{r,m_1} - \theta_r| \ge \epsilon/2\big) + \mathrm{P}\big(|\hat{\hat{\theta}}_{r,m_2} - \theta_r| \ge \epsilon/2\big)\\ + 2\mathrm{P}\big(R_{r,m_2} \ge \epsilon/2\big) + \mathrm{P}(A_n(M,N)^c),\end{aligned}$$

where

$$0 \le R_{r,m_2} \le B_r n^{-1} m_2^{\max\{1,r+5/6\}} \le B_r N,$$

and B_r is a constant independent of n. Moreover, from Proposition 2 and Markov's inequality we get

$$\begin{aligned}&\mathrm{P}\big(|\hat{\hat{\theta}}_{r,m_1} - \theta_r| \ge \epsilon/2\big)\\ &\quad \le \frac{4}{\epsilon^2}\Big(D_1 n^{-1} M^{\eta(r-p+5/6)} + D_2 n^{-2} M^{2\max\{1,r+5/6\}} + D_3 M^{-2p} \nu_M\Big)\end{aligned}$$

and

$$\mathrm{P}\big(|\hat{\hat{\theta}}_{r,m_2} - \theta_r| \ge \epsilon/2\big) \le \frac{4}{\epsilon^2}\Big(D_1 N + D_2 N^2 + D_3 n^{-2p/\max\{1,r+5/6\}} \nu_{m_2(n)}\Big).$$

Therefore, as $\nu_m \to 0$ when m tends to infinity, we easily conclude that for all $\epsilon > 0$ and $\delta > 0$ there exist $M \in \mathbb{N}$ large enough, $N > 0$ small enough, and $n_0 \in \mathbb{N}$ such that $\mathrm{P}(|\hat{\hat{\theta}}_{r,\hat{m}} - \theta_r| \ge \epsilon) < \delta$, for all $n \ge n_0$. ■

Proof of parts (b) and (c) of Theorem 1: Let $m_1 = m_1(n)$ and $m_2 = m_2(n)$ be two sequences of nonnegative integers such that $Cn^{\xi} - 1 \le m_1 < Cn^{\xi}$ and $Dn^{\xi} < m_2 \le Dn^{\xi} + 1$, for n large enough. As $\mathrm{P}\big(m_1 \le \hat{m} \le m_2\big) \to 1$, from Proposition 1 we also have

$$\mathrm{P}\big(\hat{\hat{\theta}}_{r,m_1} - \theta_r - R_{r,m_2} \le \hat{\hat{\theta}}_{r,\hat{m}} - \theta_r \le \hat{\hat{\theta}}_{r,m_2} - \theta_r + R_{r,m_2}\big) \to 1,$$

where $R_{r,m_2} = O_p\big(n^{-1}m_2^{\max\{1,r+5/6\}}\big) = O_p\big(n^{-(1-\xi\eta(r+5/6))}\big)$. Thus, part (b) of Theorem 1 follows from Proposition 2 as

$$\hat{\hat{\theta}}_{r,m_j} - \theta_r = O_p\big(n^{-\min\{(1-\xi\eta(r-p+5/6))/2\,,\,1-\xi\eta(r+5/6)\,,\,p\xi\}}\big),$$

for $j = 1, 2$, and part (c) of Theorem 1 follows from Proposition 3 as $R_{r,m_2} = o_p(n^{-1/2})$ and $n^{1/2}\big(\hat{\hat{\theta}}_{r,m_j} - \theta_r\big) \xrightarrow{d} N\big(0, 4\mathrm{Var}(f^{(2r)}(X_1)\big)$, for $j = 1, 2$. ■

6.2. Proof of Theorem 2. The asymptotic behaviour of the relative errors $\hat{h}_{i,\hat{m}}/h_0 - 1$, for $i = 1, 2$, where the plug-in bandwidth selectors $\hat{h}_{i,\hat{m}}$ are defined by (5) and (6), relies on Theorem 1 and on the following expansion of the exact optimal bandwidth

$$h_0 = c_{1,K}\,\theta_2^{-1/5}n^{-1/5} + c_{2,K}\,\theta_2^{-8/5}\,\theta_3\,n^{-3/5} + O(n^{-4/5}),$$

which holds when K is a kernel of order 2 with $\int |u|^5|K(u)|du < \infty$, and f has bounded, integrable and continuous derivatives up to order 4 (see Hall et al., 1991, sec. 2).

Proof of part (a) of Theorem 2: As

$$\frac{\hat{h}_{1,\hat{m}}}{h_0} - 1 = \frac{c_{1,K}\big(\hat{\theta}_{2,\hat{m}}^{-1/5} - \theta_2^{-1/5}\big)}{n^{1/5}h_0} - \frac{c_{2,K}\theta_2^{-8/5}\theta_3\,n^{-2/5}}{n^{1/5}h_0} + O(n^{-3/5}),$$

where $n^{1/5}h_0 \to c_{1,K}\,\theta_2^{-1/5}$, $n \to +\infty$, the stated convergence and order of convergence for the relative error $\hat{h}_{1,\hat{m}}/h_0 - 1$ follow from parts (a) and (b) of Theorem 1 with $r = 2$ and the fact that $\beta_2(p,\xi) > 2/5$ iff $p \geq 3$ and ξ satisfies (8). ■

Proof of part (b) of Theorem 2: We have

$$\frac{\hat{h}_{2,\hat{m}}}{h_0} - 1 = \frac{c_{1,K}\big(\hat{\theta}_{2,\hat{m}}^{-1/5} - \theta_2^{-1/5}\big)}{n^{1/5}h_0} + \frac{c_{2,K}\big(\hat{\theta}_{2,\hat{m}}^{-8/5}\hat{\theta}_{3,\hat{m}} - \theta_2^{-8/5}\theta_3\big)}{n^{1/5}h_0}n^{-2/5} + O(n^{-3/5}),$$

where

$$\begin{aligned}&\hat{\theta}_{2,\hat{m}}^{-8/5}\hat{\theta}_{3,\hat{m}} - \theta_2^{-8/5}\theta_3 \\ &\quad= \big(\hat{\theta}_{2,\hat{m}}^{-8/5} - \theta_2^{-8/5}\big)(\hat{\theta}_{3,\hat{m}} - \theta_3) + \big(\hat{\theta}_{2,\hat{m}}^{-8/5} - \theta_2^{-8/5}\big)\theta_3 + \theta_2^{-8/5}(\hat{\theta}_{3,\hat{m}} - \theta_3).\end{aligned}$$

From the part (a) of Theorem 1 with $r = 2$ we know that $\hat{\theta}_{2,\hat{m}} - \theta_2 = o_p(1)$. The convergence to zero of the relative error $\hat{h}_{2,\hat{m}}/h_0 - 1$ follows now from the convergence $n^{-2/5}(\hat{\theta}_{3,\hat{m}} - \theta_3) = o_p(1)$, which can be established by reasoning as in the proof of part (a) of Theorem 1 using the fact that $f \in \mathscr{D}_{3,p-1}$ with $p-1 \geq 0$.

If $\hat{m}$ satisfies (4), from Theorem 1 we have $\hat{\theta}_{2,\hat{m}} - \theta_2 = O_p\big(n^{-\beta_2(p,\xi)}\big)$, and $\hat{\theta}_{3,\hat{m}} - \theta_3 = O_p\big(n^{-\beta_3(p-1,\xi)}\big)$. Therefore, the stated order of convergence for the relative error $\hat{h}_{2,\hat{m}}/h_0 - 1$ follows from the fact that $\beta_2(p,\xi) < \beta_3(p-1,\xi) + 2/5$, for $p \geq 1$ and $0 < \xi < 6/17$. Finally, from part (c) of Theorem 1 we have $n^{1/2}\big(\tilde{\theta}_{2,\hat{m}}^{-1/5} - \theta_2^{-1/5}\big) \xrightarrow{d} N\big(0, \theta_2^{-2/5}\sigma^2(f)\big)$, whenever $p \geq 3$ and $1/(2p) \leq \xi < 3/17$, from which we deduce the stated asymptotic normality of the relative error of $\hat{h}_{2,\hat{m}}$. ■

Funding

Research partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324/2019, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

Bickel, P.J., Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhya Ser. A* 50, 381–393.

Bosq, D., Lecoutre, J.-P. (1987). *Théorie de l'estimation fonctionnelle*. Paris: Economica.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71, 353–360.

Chacón, J.E., Montanero, J., Nogales, A.G., Pérez, P. (2007). On the existence and limit behavior of the optimal bandwidth for kernel density estimation. *Statist. Sinica* 17, 289–300.

Chacón, J.E., Tenreiro, C. (2012). Exact and asymptotically optimal bandwidths for kernel estimation of density functionals. *Methodol. Comput. Appl. Probab.* 14, 523–548.

Chacón, J.E., Tenreiro, C. (2013). Data-based choice of the number of pilot stages for plug-in bandwidth selection. *Comm. Statist. Theory Methods* 42, 2200–2214.

Deheuvels, P., Hominal, P. (1980). Estimation automatique de la densité. *Rev. Statist. Appl.* 28, 25–55.

Devroye, L., Györfi, L. (1985). *Nonparametric density estimation: the L_1 view.* New York: Wiley.

Fan, J., Marron, J.S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* 20, 2057–2070.

Greblicki, W., Pawlak, M. (1984). Hermite series estimates of a probability density and its derivatives. *J. Multivariate Anal.* 15, 174–182.

Greblicki, W., Pawlak, M. (1985). Pointwise consistency of the Hermite series density estimate. *Statis. Probab. Lett.* 3, 65–69.

Hall, P., Marron, J.S. (1987). Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* 74, 567–581.

Hall, P., Marron, J.S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields* 90, 149–173.

Hall, P., Sheather, S.J., Jones, M.C., Marron, J.S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* 78, 263–269.

Hart, J.D. (1985). On the choice of a truncation point in Fourier series density estimation. *J. Stat. Comput. Simul.* 21, 95–116.

Jones, M.C., Sheather, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statis. Probab. Lett.* 11, 511–514.

Laurent, B. (1997). Estimation of integral functionals of a density and its derivatives. *Bernoulli* 3, 181–211.

Lee, A.J. (1990). *U-statistics, theory and practice.* New York: Marcel Dekker.

Marron, J.S., Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* 20, 712–736.

Nadaraya, E.A. (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theory Probab. Appl.* 19, 133–141.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* 33, 1065–1076.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* 27, 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* 9, 65–78.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* London: Chapman and Hall.

Simonoff, J.S. (1996). *Smoothing methods in statistics.* New York: Springer.

Tenreiro, C. (2011). Fourier series based direct plug-in bandwidth selectors for kernel density estimation. *J. Nonparametr. Stat.* 23, 533–545.

Tenreiro, C. (2017). A weighted least-squares cross-validation bandwidth selector for kernel density estimation. *Comm. Statist. Theory Methods* 46, 3438–3458.

Tsybakov, A.B. (2009). *Introduction to nonparametric estimation.* London: Springer.

Walter, G. (1977). Properties of Hermite series estimation of probability density. *Ann. Statist.* 5, 1258–1264.

Wand, M.P., Jones, M.C. (1995). *Kernel smoothing.* New York: Chapman & Hall.

Wand, M.P. (2014). KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2.23-13. http://CRAN.R-project.org/package=KernSmooth

Woodroofe, M. (1970). On choosing a delta-sequence. *Ann. Math. Statist.* 41, 1665–1671.

Carlos Tenreiro
CMUC, Department of Mathematics, University of Coimbra, 3001–501 Coimbra, Portugal
E-mail address: tenreiro@mat.uc.pt