# Fourier series based direct plug-in bandwidth selectors for kernel density estimation*

Carlos Tenreiro†

July 23, 2010

**Abstract**

A class of Fourier series based direct plug-in bandwidth selectors for kernel density estimation is considered in this paper. The proposed bandwidth estimators have a relative convergence rate $n^{-1/2}$ whenever the underlying density is smooth enough and the simulation results testify that they present a very good finite sample performance against the most recommended bandwidth selection methods in the literature.

KEYWORDS: Bandwidth selection, kernel density estimation, direct plug-in method, Fourier series based bandwidth selectors.

AMS 2010 SUBJECT CLASSIFICATIONS: 62G07, 62G20.

*This is an electronic version of an article published in Journal of Nonparametric Statistics (Vol. 23, 2011, 533–545), and available on line at `http://dx.doi.org/10.1080/10485252.2010.537337`

†CMUC, Department of Mathematics, University of Coimbra, Apartado 3008, 3001–454 Coimbra, Portugal. E-mail: `tenreiro@mat.uc.pt`. URL: `http://www.mat.uc.pt/~tenreiro/`.

# 1 Introduction

Let $X_1, \ldots, X_n$ be independent real-valued random variables with common density $f$ defined over a compact set of $\mathbb{R}$. Assuming that $f$ is smooth enough on $\mathbb{R}$, we propose in this paper Fourier series based direct plug-in bandwidth selectors for the kernel density estimator defined, for $x \in \mathbb{R}$, by

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

(Rosenblatt 1956, Parzen 1962) where $K_h(\cdot) = K(\cdot/h)/h$, for $h > 0$, $K$ is a bounded and symmetric density function on $\mathbb{R}$ (the kernel) and the smoothing parameter $h = h_n$ is a sequence of strictly positive real numbers converging to zero as $n$ tends to infinity (the bandwidth). See Devroye and Györfi (1985) and Bosq and Lecoutre (1987) for some of the most important asymptotic properties of $f_n$ as an estimator of $f$.

In the definition of $f_n$, the kernel $K$ and the bandwidth $h$ enter as unspecified parameters. For a fixed kernel the bandwidth is usually chosen on the basis of the data and this choice is crucial to the performance of the estimator. Too small an $h$ leads to an estimator with large variability which produces noisy estimates showing some features not shared by $f$ whereas too large an $h$ leads to a highly biased estimator producing flat estimates that do not reveal some interesting characteristics of $f$. Because of its relevancy, the selection of the bandwidth is one of the mostly studied topics in kernel density estimation and several approaches have been proposed for choosing $h$. A very good overview of the variety of methods that appeared in the literature since the late seventies can be found in Wand and Jones (1995; chap. 3), Jones et al. (1996) and Chiu (1996). For some important comments on bandwidth selection see also Loader (1999).

The direct plug-in method whose idea 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 for the bandwidth $h_0$ that minimizes the mean integrated square error $\mathrm{MISE}(f; n, h) = \mathrm{E}(\mathrm{ISE}(f; n, h)) = \mathrm{E} \| f_n - f \|_2^2$, where $\| \cdot \|_2$ denotes the $L_2$ distance:

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

See Chacón et al. (2007) for the existence and asymptotic behaviour of $h_0$. Under some moment and regularity conditions on $K$ and $f$, respectively, two asymptotic approximations to the optimal bandwidth $h_0$ are given by

$$h_1 = c_{1,K}\, \theta_2^{-1/5} n^{-1/5} \quad \text{and} \quad h_2 = c_{1,K}\, \theta_2^{-1/5} n^{-1/5} + c_{2,K}\, \theta_2^{-8/5}\, \theta_3\, n^{-3/5}, \tag{2}$$

where $\theta_r$ denotes the quadratic functional

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

with $r = 0, 1, \ldots,$ $f^{(r)}$ the $r$th derivative of $f$, and

$$c_{1,K} = k_0^{1/5} k_2^{-2/5}, \quad c_{2,K} = \frac{1}{20} k_0^{3/5} k_2^{-11/5} k_4, \tag{3}$$

with $k_0 = \int K(u)^2 du$, $k_j = \int u^j K(u) du$ for $j = 1, 2, \ldots$ (cf. Hall and Marron 1987, 1991). These asymptotic approximations to $h_0$ reduce the problem of estimating the optimal bandwidth to that of estimating the quadratic functionals $\theta_2$ and $\theta_3$. This is the idea of the direct plug-in approach to bandwidth selection.

The estimation of the quadratic functionals $\theta_r$, for $r = 0, 1, \ldots,$ has been studied by authors like Dmitriev and Tarasenko (1973), Levit (1978), Hall and Marron (1987, 1991), Bickel and Ritov (1988), Donoho and Nussbaum (1990), Jones and Sheather (1991), Chiu (1991), Efromovich and Low (1996) and Laurent (1996, 1997). The class of kernel estimators proposed by Hall and Marron (1987) and Jones and Sheather (1991) is widely used in practice leading to some well-known bandwidth selection methods such as those introduced by Sheather and Jones (1991) and Hall et al. (1991). The sample characteristic function based estimators proposed by Chiu (1991) leads to direct and adjusted plug-in bandwidths selectors with particularly good asymptotic properties. To our knowledge, much less attention has been paid to the practical performance of the plug-in bandwidth selectors based on the Fourier series estimators of $\theta_r$ studied by Laurent (1997) when the support of the underlying density function $f$ is known to be contained within a finite interval $[a, b]$. These estimators achieve the $n^{-1/2}$ rate whenever $f$ is smooth enough and they are efficient. Moreover, when the $n^{-1/2}$ rate is not achievable they achieve the optimal rate of convergence. Therefore, from a theoretical point of view they are natural candidates for the estimation of $\theta_2$ and $\theta_3$ in (2). The main aim of this paper is to describe the asymptotic and finite sample behaviour of the direct plug-in bandwidths based on (2) with $\theta_2$ and $\theta_3$ replaced by $\hat{\theta}_{2,\hat{m}}$ and $\hat{\theta}_{3,\hat{m}}$, respectively, where $\hat{\theta}_{r,m}$ denotes the Fourier series estimator of $\theta_r$, $m$ is the number of Fourier terms included in the estimator, and $\hat{m} = \hat{m}(X_1, \ldots, X_n)$ is a random sequence of positive integers.

The rest of the paper is organised as follows. In Section 2 we recall the definition of the Fourier series estimator $\hat{\theta}_{r,m}$ of $\theta_r$, and we establish the consistency, orders of convergence and asymptotic normality of the proposed direct plug-in bandwidth selectors in terms of the asymptotic behaviour of $\hat{m}$. In Section 3 we present two data-driven methods for choosing the number $m$ of Fourier terms which are based on the methods proposed by ) (1985) and Chiu (1992) to address related problems. In Section 4 a simulation study is carried out to compare the finite sample performance of the proposed Fourier series based direct plug-in bandwidths to those of the most recommended data-based bandwidths such as the direct plug-in and cross-validation methods proposed by Chiu (1991, 1992), the solve-the-equation method proposed by Sheather and Jones (1991) and the least squares cross-validation method introduced by Rudemo (1982) and Bowman (1984). The simulation results indicate that the Fourier series based methods perform very well in comparison to all the above-mentioned methods. Section 5 contains the proofs. The simulations and plots in this paper were performed using the R software (R Development Core Team 2009).

## 2 Fourier series based plug-in bandwidth selectors

If $\{p_\ell, \ell \geq 0\}$ is the orthonormal Fourier basis of $L_2([a,b])$ given by

$$\begin{aligned} p_0(x) &= \frac{1}{\sqrt{b-a}}, \\ p_{2\ell-1}(x) &= \sqrt{\frac{2}{b-a}} \sin\left(\xi_\ell\,(x-a)\right), \\ p_{2\ell}(x) &= \sqrt{\frac{2}{b-a}} \cos\left(\xi_\ell\,(x-a)\right), \end{aligned}$$

for $\ell > 0$, and $\xi_\ell = 2\pi\ell/(b-a)$, the projection estimator of $\theta_r$, for $r > 0$, studied in Laurent (1997) cames from the representation of $\theta_r$ as $\theta_r = \sum_{\ell=1}^{\infty} \xi_\ell^{2r} c_\ell$, where $c_\ell = a_{2\ell-1}^2 + a_{2\ell}^2$ with $a_\ell = \int_a^b f(x) p_\ell(x) dx$ the Fourier coefficients of $f$, and is defined by

$$\hat{\theta}_{r,m} = \sum_{\ell=1}^{m} \xi_\ell^{2r}\, \hat{c}_\ell,$$

where $\hat{c}_\ell$ is the unbiased estimator of $c_\ell$ given by

$$\hat{c}_\ell = \frac{2}{n(n-1)} \sum_{1 \leq j < k \leq n} \{p_{2\ell-1}(X_j)p_{2\ell-1}(X_k) + p_{2\ell}(X_j)p_{2\ell}(X_k)\}$$

and $m = m(n)$ is a sequence on integers converging to infinity. The number $m$ of Fourier terms plays the role of smoothing parameter and makes the trade-off between the variance and the bias of the estimator. A large value of $m$ implies a small bias but a large variance, whereas a small $m$ implies a large bias but a small variance. An interesting representation for the projection estimator $\hat{\theta}_{r,m}$ is given by

$$\begin{aligned} \hat{\theta}_{r,m} &= \frac{n}{n-1}\frac{2}{b-a} \sum_{\ell=1}^{m} \xi_\ell^{2r} \{|\tilde{\phi}(\xi_\ell)|^2 - 1/n\} \\ &= \frac{n}{n-1}\frac{1}{2\pi} \sum_{\ell=-m}^{m} \xi_\ell^{2r} \{|\tilde{\phi}(\xi_\ell)|^2 - 1/n\}(\xi_\ell - \xi_{\ell-1}), \end{aligned}$$

where $\tilde{\phi}(\lambda) = n^{-1} \sum_{j=1}^{n} \exp(i\,\lambda X_j)$ is the sample characteristic function. This shows that $\hat{\theta}_{r,m}$ is a sample characteristic function based estimator of $\theta_r$ and also links $\hat{\theta}_{r,m}$ with the plug-in estimator introduced in Chiu (1991) given by $\tilde{\theta}_{r,\Lambda} = (2\pi)^{-1} \int_{-\Lambda}^{\Lambda} \lambda^{2r} \{|\tilde{\phi}(t)|^2 - 1/n\} d\lambda$, where $\Lambda$ converges to infinity, which is based on the representation of $\theta_r$ as $\theta_r = (2\pi)^{-1} \int \lambda^{2r} |\phi(\lambda)|^2 d\lambda$, with $\phi(\lambda) = \int \exp(i\,\lambda x) f(x) dx$ the characteristic function of $f$.

As in practical situations the choice of $m$ should be based on the observations, this is, $m = \hat{m} = \hat{m}(X_1, \ldots, X_n)$, we consider the automatic estimators $\hat{\theta}_{r,\hat{m}}$ of $\theta_r$, and we use them to define direct plug-in bandwidths based on (2) given by

$$\hat{h}_{1,\hat{m}} = c_{1,K}\, \hat{\theta}_{2,\hat{m}}^{-1/5} n^{-1/5} \quad \text{and} \quad \hat{h}_{2,\hat{m}} = c_{1,K}\, \hat{\theta}_{2,\hat{m}}^{-1/5} n^{-1/5} + c_{2,K}\, \hat{\theta}_{2,\hat{m}}^{-8/5}\, \hat{\theta}_{3,\hat{m}}\, n^{-3/5}, \tag{4}$$

with $c_{1,K}$ and $c_{2,K}$ given by (3). The asymptotic behaviour of the relative errors $\hat{h}_{i,\hat{m}}/h_0 - 1$, which relies on the asymptotic behaviour of $\hat{\theta}_{2,\hat{m}}$ and $\hat{\theta}_{3,\hat{m}}$, is stated in the next result. Its proof is deferred to Section 5.

**Theorem 1.** *Let $K$ be a bounded and symmetric density function such that $\int |u|^5\, K(u)du < \infty$ and assume that for $s = p + \alpha > 4$, with $p \in \mathbb{N}$ and $\alpha \in ]0,1]$, $f$ is a density with support on $[a,b]$ which is $p$-times differentiable in $\mathbb{R}$ and $f^{(p)}$ satisfies the Lipschitz condition*

$$|f^{(p)}(x) - f^{(p)}(y)| \le C|x-y|^{\alpha}, \qquad x, y \in [a,b], \tag{5}$$

*where $C$ is a positive number. Finally, let $\hat{m}$ be such that $\hat{m} \xrightarrow{p} +\infty$ and $n^{-1}\hat{m}^5 \xrightarrow{p} 0$.*

*a)* Asymptotic behaviour of $\hat{h}_{1,\hat{m}}$. *We have*

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

*and if $\hat{m}$ is such that*

$$\mathrm{P}\big(C_1\, n^{\xi_1} \le \hat{m} \le C_2\, n^{\xi_2}\big) \to 1, \tag{6}$$

*where $C_1, C_2, \xi_1, \xi_2$ are strictly positive constants with*

$$\frac{1}{5(s-2)} < \xi_1 \le \xi_2 < \frac{3}{25}$$

*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 $\hat{h}_{2,\hat{m}}$. *We have*

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

*and if $\hat{m}$ satisfies (6) with*

$$\frac{1}{20(s-3)} < \xi_1 \leq \xi_2 < \frac{9}{70}$$

*then*

$$\frac{\hat{h}_{2,\hat{m}}}{h_0} - 1 = O_p\left(n^{-\min\{1/2, 1-5\xi_2, 2\xi_1(s-2)\}}\right).$$

*Moreover, if $s > 4 + 1/2$ and*

$$\frac{1}{4(s-2)} < \xi_1 \leq \xi_2 < \frac{1}{10}$$

*then*

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

*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* 1. When $4 < s < 4 + 1/2$ the best rate of convergence given for the relative error $\hat{h}_{2,\hat{m}}/h_0 - 1$ is $n^{-2(s-2)/(2s+1)}$ and is achieved whenever $\hat{m}$ satisfies (6) with $\xi_1 = \xi_2 = \frac{1}{2s+1}$. When $s \geq 4 + 1/2$ and (6) is fulfilled with $\xi_1 = \xi_2 = \frac{1}{10}$, we have $\hat{h}_{2,\hat{m}}/h_0 - 1 = O_p(n^{-1/2})$ and this is, in a minimax sense, the best possible rate of convergence as shown by Hall and Marron (1991).

*Remark* 2. When a lower bound $\underline{s} > 4 + 1/2$ for the degree of smoothness of $f$ is assumed to be known, we conclude that the asymptotic normality of the relative error of $\hat{h}_{2,\hat{m}}$ will take place whenever $\hat{m}$ satisfies (6) with $\xi_1 = \xi_2 = \frac{1}{2\underline{s}+1}$. Moreover, the variance $\sigma^2(f)$ is the same as the best possible constant coefficient derived by Fan and Marron (1992).

# 3 The automatic choice of $m$

The methods considered in this section for the automatic choice of $m$ are inspired by the equality $\hat{\theta}_{r,m} = \| \hat{f}_m^{(r)} \|_2^2 \, (1 + O(n^{-1})) + O(n^{-1}m^{2r+1})$ that connects $\hat{\theta}_{r,m}$ with the $L_2$-norm of the $r$-derivative of the Fourier estimator of $f$ defined by $\hat{f}_m(x) = \sum_{\ell=0}^{2m} \hat{a}_\ell p_\ell(x)$, where $\hat{a}_\ell = n^{-1}\sum_{j=1}^{n} p_\ell(X_j)$ are unbiased estimators of the Fourier coefficients $a_\ell$ (Tarter and Kronmal 1968). For a squared integrable density function $f$ with support contained within a finite interval $[a, b]$, Hart (1985) proves that the mean integrated square error of $\hat{f}_m$ can be expressed as

$$
\begin{aligned}
\mathrm{E}\| \hat{f}_m - f \|_2^2 &= \frac{2m}{n(b-a)} - \frac{n+1}{n} \sum_{\ell=1}^{2m} a_\ell^2 + \sum_{\ell=1}^{\infty} a_\ell^2 \\
&= \frac{2}{b-a} \left( \frac{m}{n} - \frac{n+1}{n} \sum_{\ell=1}^{m} |\phi(\xi_\ell)|^2 + \sum_{\ell=1}^{\infty} |\phi(\xi_\ell)|^2 \right) \qquad (7) \\
&= \frac{2}{b-a} \left( H(m) + \sum_{\ell=1}^{\infty} |\phi(\xi_\ell)|^2 \right),
\end{aligned}
$$

where $\xi_\ell = 2\pi\ell/(b-a)$, for $\ell \in \mathbb{N}$. Since the last term of these expressions does not depend on $m$, we follow Hart's idea and propose to take for $m$ the first integer $\hat{m}_{\mathrm{H}}$ satisfying

$$
\hat{m}_{\mathrm{H}} = \operatorname*{argmin}_{m \in \mathcal{M}_n} \hat{H}(m),
$$

where $\mathcal{M}_n = \{\mathrm{L}_n, \mathrm{L}_n + 1, \ldots, \mathrm{U}_n\}$, $\mathrm{L}_n < \mathrm{U}_n$ are deterministic sequences of positive integers, and $\hat{H}(m)$ is the unbiased estimator of $H(m)$ given by

$$
\hat{H}(m) = \frac{m}{n} - \frac{n+1}{n-1} \sum_{\ell=1}^{m} \left\{ |\tilde{\phi}(\xi_\ell)|^2 - \frac{1}{n} \right\}.
$$

Note that a similar idea was followed by Chiu (1991, 1992) for selecting the cut-off frequency $\Lambda$ appearing in the kernel density bandwidth selectors introduced in Chiu (1991). The proposed methods have natural counterparts in the present context. The method proposed in Chiu (1991; p. 1888) corresponds to take for $m$ the first local minimizer of $\hat{H}$ over $\mathbb{N}$. It is easy to see that this local minimizer is no more than the first positive integer $m$ satisfying $|\tilde{\phi}(\xi_{m+1})|^2 \leq 2/(n+1)$. The method proposed in Chiu (1992; p. 774) corresponds to take for $m$ the global minimizer of $\hat{H}$ over $\mathbb{N}$.

The value $\hat{m}_{\mathrm{H}}$ depends on $\mathcal{M}_n$ through the sequences $\mathrm{L}_n$ and $\mathrm{U}_n$ that need to be chosen by the user. If they are taken equal to $\mathrm{L}_n = \lfloor C_1 n^{\xi_1} \rfloor + 1$ and $\mathrm{U}_n = \lfloor C_2 n^{\xi_2} \rfloor$, where $\lfloor x \rfloor$ is the integral part of $x$ and $C_1, C_2, \xi_1, \xi_2$ are strictly positive constants satisfying the conditions of Theorem 1, we know that the data-dependent bandwidths $\hat{h}_{1,\hat{m}_{\mathrm{H}}}$ and $\hat{h}_{2,\hat{m}_{\mathrm{H}}}$ given by (4) will possess good asymptotic properties. Assuming for ease of explanation that $s \geq 5$ in Theorem 1, we deduce that the best orders of convergence for the relative errors of each one of the data-dependent bandwidths $\hat{h}_{1,\hat{m}_{\mathrm{H}}}$ and $\hat{h}_{2,\hat{m}_{\mathrm{H}}}$ will take place by choosing $\xi_1 = \xi_2 = 1/11$. In this case, since the power $n^{1/11}$ remains small for very large sample sizes, the sequences $\mathrm{L}_n$ and $\mathrm{U}_n$ are dominated by the size of the constants $C_1$ and $C_2$.

In order to gain some understanding about the practical choice of these constants, for two of the beta mixture distributions over the interval $[0, 1]$ considered in Section 4, we

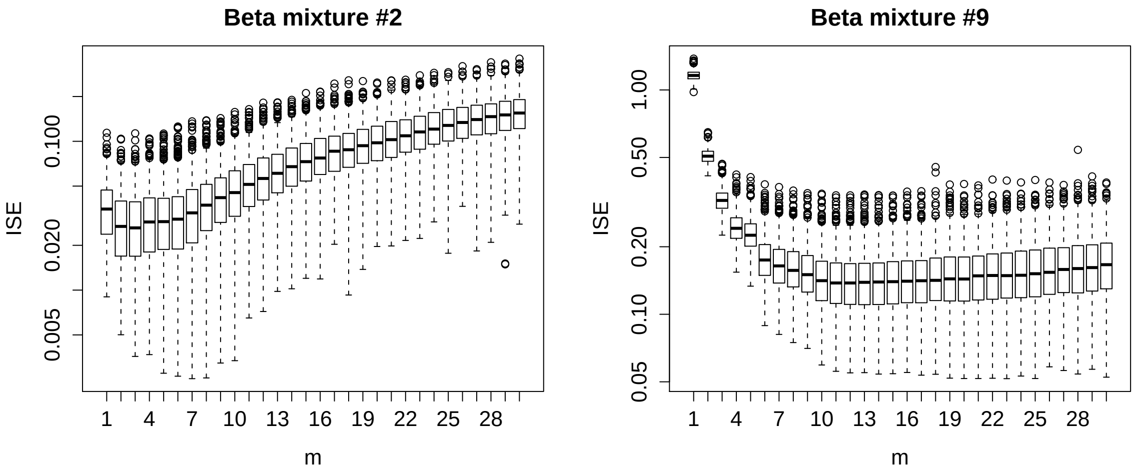


Figure 1: *Empirical distribution of* $\mathrm{ISE}(f; n, \hat{h}_{1,m})$ *as a function of* $m$ *for* $n = 200$*. The number of replications is* 500*.*

present in Figure 1 the empirical distribution of the integrated squared error associated to the kernel density estimator (1) with bandwidth $h = \hat{h}_{1,m}$ as a function of the number of terms $m$. We took for $K$ the standard Gaussian kernel, $[a, b] = [0, 1]$ and the integrated squared error $\mathrm{ISE}(f; n, h)$ was computed using the composite Simpson's rule. In order to assure the positivity of the estimates $\hat{\theta}_{r,m}$ in (4), whenever $\hat{\theta}_{r,m}$ is negative we will use $\| \hat{f}_m^{(r)} \|$ instead of $\hat{\theta}_{r,m}$. From expression (7) it is reasonable to expect that small values for $m$ are more appropriate when the Fourier coefficients of $f$ converge quickly to zero, which is the case of distribution #2, whereas large values for $m$ are the best choice for densities whose Fourier coefficients converge slowly to zero, which is the case of distribution #9. This is confirmed by Figure 1. Similar pictures can be observed for $h = \hat{h}_{2,m}$. As no significant finite sample differences were found, in particular for $n \geq 200$, between the bandwidth selectors based on $\hat{h}_{1,m}$ and $\hat{h}_{2,m}$, in the following we will restrict our attention to the plug-in bandwidth selector $\hat{h}_{1,\hat{m}}$.

From the previous considerations, we conclude that if we want to deal with a wide set of distributional characteristics, the sequences $\mathrm{L}_n$ and $\mathrm{U}_n$ should be chosen such that the set $\mathcal{M}_n$ contains very small and moderately large values of $m$. In the following we take $C_1 = 0.25$ and $C_2 = 25$ which leads to $\mathrm{L}_n = 1$ and $30 \leq \mathrm{U}_n \leq 87$ for $10 \leq n \leq 10^6$. As we will see later, the methods we introduce are quite robust against the choice of $C_2$ and their performance is not affected if larger values for $C_2$ are taken.

However, the previous choice of $C_2$ does not prevent us from getting excessively large values for $m$ which might lead to an overestimation of the quadratic functional $\theta_2$, 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 penalized much more by excessively small than by excessively large bandwidths. A similar situation was also pointed out by Hart (1985) and Chiu (1992) and two proposals to overcome this problem were suggested.

Hart (1985) suggested considering a weighted version of the criterion function $\hat{H}$ given by

$$\hat{H}_\gamma(m) = \frac{m}{n} - \gamma \frac{n+1}{n-1} \sum_{\ell=1}^{m} \left\{ |\tilde{\phi}(\xi_\ell)|^2 - \frac{1}{n} \right\},$$

for some $0 < \gamma < 1$. If we denote by $\hat{m}_{\mathrm{H}_\gamma}$ the first global minimizer of $\hat{H}_\gamma$ over $\mathcal{M}_n$, it is clear that $\hat{m}_{\mathrm{H}_{\gamma_1}} \leq \hat{m}_{\mathrm{H}_{\gamma_2}}$ whenever $\gamma_1 < \gamma_2$. Small values of $\gamma$ generally improve Hart's method for distributions whose Fourier coefficients converge quickly to zero, and large values of $\gamma$ are more appropriate for distributions with Fourier coefficients converging slowly to zero. In order to find a compromise between these to extreme situations we decide to take $\gamma = 0.5$. We shall denote the associated kernel density bandwidth $\hat{h}_{1,\hat{m}_{\mathrm{H}_\gamma}}$ by $\hat{h}_{\mathrm{H}}$.

An alternative modification of $\hat{H}$ is proposed by Chiu (1992). The basic idea is to take for $m$ the first local minimizer of $\hat{H}$ unless there exists a larger significant local minimizer. Using standard U-statistic techniques we note that for $q > p$, $\hat{H}(q) - \hat{H}(p)$ is asymptotically normal with mean $H(q) - H(p)$ and variance given by

$$\sigma^2_{p,q} = \left(\frac{n+1}{n}\right)^2 \sum_{\ell,\ell'=p+1}^{q} \left\{ \frac{2(n-2)}{n(n-1)} A_{\ell,\ell'} + \frac{1}{n(n-1)} B_{\ell,\ell'} - \frac{4n-6}{n(n-1)} C_{\ell,\ell'} \right\}, \tag{8}$$

where $A_{\ell,\ell'} = R\left(\phi(\xi_\ell + \xi_{\ell'})\overline{\phi(\xi_\ell)\phi(\xi_{\ell'})}) + \phi(\xi_\ell - \xi_{\ell'})\overline{\phi(\xi_\ell)\phi(-\xi_{\ell'})}\right)$, $B_{\ell,\ell'} = |\phi(\xi_\ell + \xi_{\ell'})|^2 + |\phi(\xi_\ell - \xi_{\ell'})|^2$ and $C_{\ell,\ell'} = |\phi(\xi_\ell)|^2 + |\phi(\xi_{\ell'})|^2$, where $\overline{z}$ and $R(z)$ denote, respectively, the conjugate and the real part of the complex number $z$. Therefore, denoting by $\hat{m}_l$ the first local minimizer of $\hat{H}$ over $\mathcal{M}_n$, Chiu's proposal leads to take for $m$ the global minimizer $\hat{m}_{\mathrm{H}_*}$ of $\hat{H}_*$ over $\mathcal{M}_n$ with

$$\hat{H}_*(m) = \hat{H}(m) + 1.645\, I_{\{m > \hat{m}_l\}}\, \hat{\sigma}_{\hat{m}_l,m},$$

where $I_A$, the indicator function of the set $A$, is defined to be identically one on $A$ and is zero elsewhere, $\hat{\sigma}^2_{p,q}$ is the estimator of $\sigma^2_{p,q}$ obtained from (8) by replacing $\phi$ by $\tilde{\phi}$, and we use the constant 1.645 because it is the 95th percentile of the standard normal distribution. The corresponding kernel density estimator bandwidth $\hat{h}_{1,\hat{m}_{\mathrm{H}_*}}$ will be denoted by $\hat{h}_{\mathrm{H}_*}$.

In order to implement the previous procedures we have to assume that the support of underlying density function $f$ is contained within some finite reference interval $[a, b]$. For a large set of beta mixture distributions and for different sample sizes we verified that $\hat{h}_{\mathrm{H}}$ and $\hat{h}_{\mathrm{H}_*}$ are quite robust against the choice of the reference interval. This is illustrated in Figure 2 for bandwidth $\hat{h}_{\mathrm{H}}$ where we present the empirical distribution of $\mathrm{ISE}(f; n, h)$ for

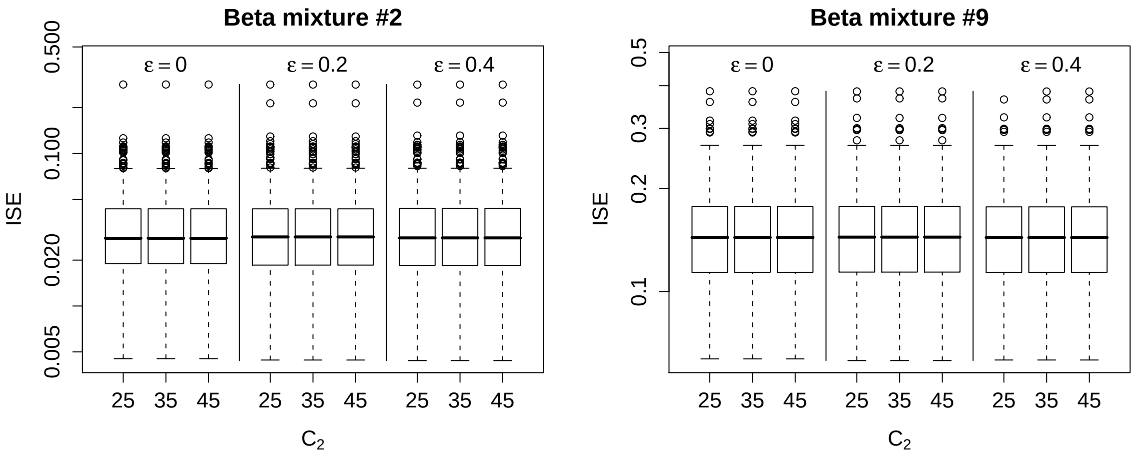


Figure 2: ISE$(f; n, h)$ *distribution for* $n = 200$ *and* $h = \hat{h}_{\mathrm{H}}$ *as a function of* $C_2$ *and* $\epsilon$*, where* $a = 0 - \epsilon$ *and* $b = 1 + \epsilon$*. The number of replications is* 500.

the above considered beta mixture distributions over the interval $[0, 1]$ and $n = 200$, where $\hat{h}_{\mathrm{H}}$ is evaluated by taking $a = 0 - \epsilon$ and $b = 1 + \epsilon$ for different values of $\epsilon > 0$. A similar behaviour was observed for $\hat{h}_{\mathrm{H}_*}$. The robustness of the proposed plug-in bandwidths against the choice of the reference interval is a desirable property since in practice the exact support of the underlying distribution may not be known. Finally note that the two procedures are strongly robust against the choice of the constant $C_2$ that we have fixed equal to 25. From Figure 2 we clearly see that the performance of $\hat{h}_{\mathrm{H}}$ is not affected if larger values for $C_2$ are taken.

# 4 Finite sample comparative study

We carried out a simulation study to compare the performance of the new bandwidths $\hat{h}_{\mathrm{H}}$ and $\hat{h}_{\mathrm{H}_*}$ (henceforth H and $\mathrm{H}_*$, respectively) with some of the most important bandwidth selection methods in the literature, namely the direct plug-in and cross-validation methods introduced by Chiu (1991) (henceforth CHcv and CHdpi, respectively), where the cut-off frequency selection is performed according to Chiu (1992), the solve-the-equation plug-in method proposed by Sheather and Jones (1991) (henceforth SJ) and the least squares cross-validation method (henceforth CV; some attractive asymptotic properties are described in Hall, 1983, Stone, 1984, and Hall and Marron, 1987). In the implementation of SJ we followed Wand and Jones (1995; p. 72). The normal density was used as the reference distribution and we have taken the estimate proposed by Silverman (1986; p. 47), as the estimate of scale. For CHcv and CV the minima were searched inside the interval $[0.001, 2]$, and for H and $\mathrm{H}_*$ we took $[a, b] = [-0.2, 1.2]$ as the reference interval. For all the cases $K$

| Mixture number | — Beta mixture distribution $\sum_i w_i B(p_i, q_i)$ — Weights $w$ | 1st shape parameters $p$ | 2nd shape parameters $q$ |
|---|---|---|---|
| #1 | $1$ | $4$ | $4$ |
| #2 | $(1/7, \ldots, 1/7)$ | $(4, \ldots, 4)$ | $(4, 6, 8, 10, 12, 14, 16)$ |
| #3 | $(1/7, \ldots, 1/7)$ | $(4, \ldots, 4)$ | $(10, 20, 30, 40)$ |
| #4 | $(1/2, 1/2)$ | $(7, 13)$ | $(13, 7)$ |
| #5 | $(1/2, 1/2)$ | $(4, 20)$ | $(20, 4)$ |
| #6 | $(1/2, 1/2)$ | $(6, 100)$ | $(10, 60)$ |
| #7 | $(1/4, 1/2, 1/4)$ | $(4, 8, 40)$ | $(40, 8, 4)$ |
| #8 | $(1/4, 1/2, 1/4)$ | $(10, 4, 200)$ | $(30, 4, 60)$ |
| #9 | $(5/11, 3/11, 2/11, 1/11)$ | $(25, 160, 320, 800)$ | $(60, 100, 80, 90)$ |

Table 1: *Beta mixture test distributions.*

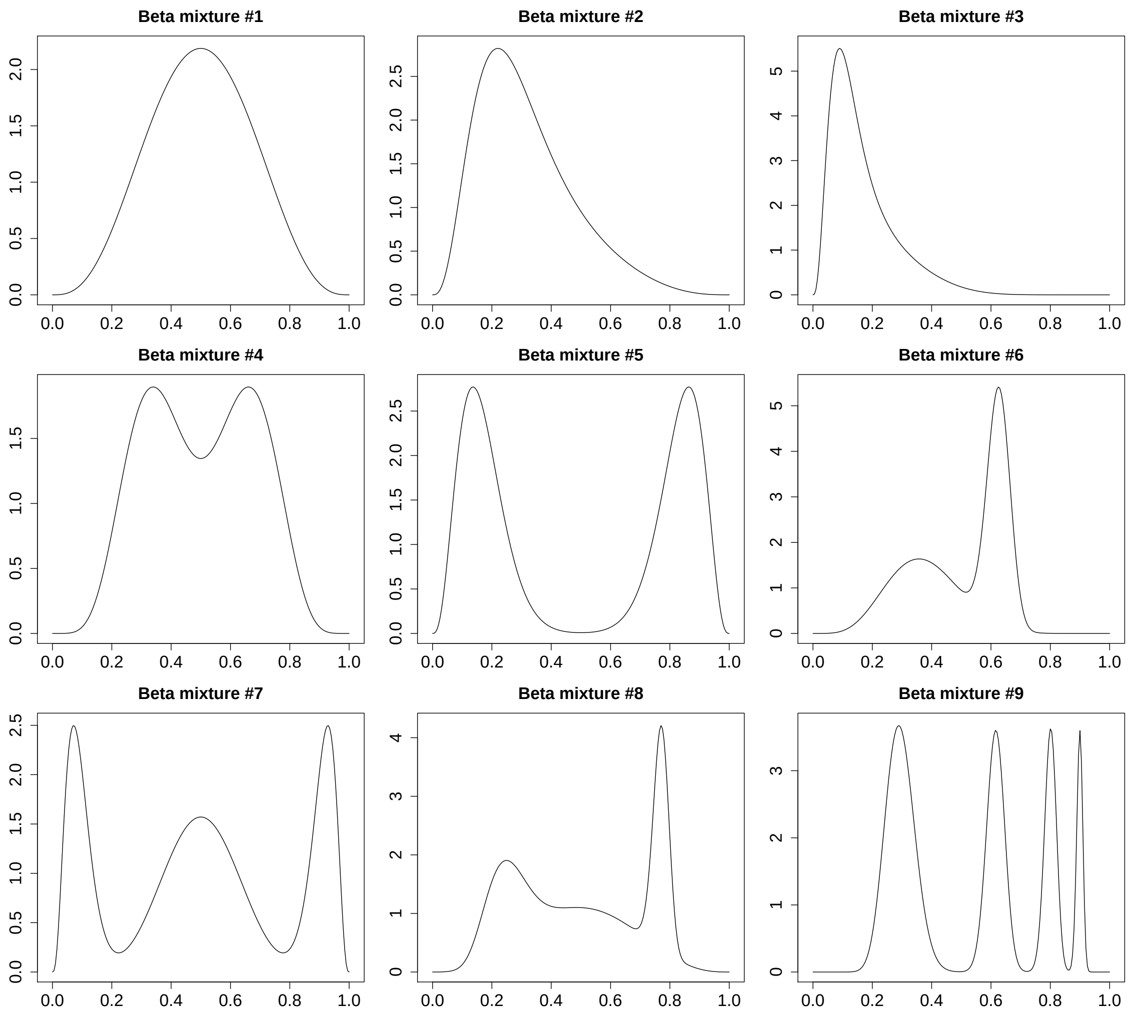


Figure 3: *Beta mixture test densities.*

| $n$ | H | CHdpi | SJ | CV |
|---|---|---|---|---|
| | Beta mixture #1 | | | |
| 100 | 3.56e-02 (1.54e-03) | 3.53e-02 (1.47e-03) | 3.73e-02 (1.35e-03) | 4.41e-02 (1.72e-03) |
| 200 | 1.98e-02 (7.11e-04) | 2.00e-02 (9.00e-04) | 2.04e-02 (6.41e-04) | 2.53e-02 (9.60e-04) |
| 400 | 1.21e-02 (4.35e-04) | 1.23e-02 (4.64e-04) | 1.22e-02 (3.8e-04) | 1.49e-02 (5.63e-04) |
| | Beta mixture #2 | | | |
| 100 | 5.41e-02 (1.60e-03) | 5.25e-02 (1.44e-03) | 5.20e-02 (1.40e-03) | 6.53e-02 (2.55e-03) |
| 200 | 3.36e-02 (8.82e-04) | 3.30e-02 (8.58e-04) | 3.24e-02 (8.19e-04) | 4.09e-02 (1.31e-03) |
| 400 | 1.98e-02 (5.39e-04) | 1.97e-02 (5.35e-04) | 1.89e-02 (4.45e-04) | 2.30e-02 (7.78e-04) |
| | Beta mixture #3 | | | |
| 100 | 1.36e-01 (3.77e-03) | 1.35e-01 (3.72e-03) | 1.27e-01 (3.25e-03) | 1.56e-01 (5.02e-03) |
| 200 | 7.63e-02 (1.84e-03) | 7.60e-02 (1.81e-03) | 7.34e-02 (1.65e-03) | 8.28e-02 (2.14e-03) |
| 400 | 4.42e-02 (1.02e-03) | 4.40e-02 (1.01e-03) | 4.30e-02 (9.09e-04) | 4.97e-02 (1.32e-03) |
| | Beta mixture #4 | | | |
| 100 | 5.85e-02 (1.59e-03) | 6.16e-02 (1.60e-03) | 5.07e-02 (1.32e-03) | 6.38e-02 (2.32e-03) |
| 200 | 3.13e-02 (8.07e-04) | 3.33e-02 (1.14e-03) | 3.09e-02 (7.50e-04) | 3.74e-02 (1.30e-03) |
| 400 | 1.85e-02 (4.67e-04) | 1.90e-02 (5.01e-04) | 1.85e-02 (4.53e-04) | 2.12e-02 (5.87e-04) |
| | Beta mixture #5 | | | |
| 100 | 9.46e-02 (2.38e-03) | 9.82e-02 (2.67e-03) | 9.40e-02 (1.97e-03) | 9.95e-02 (2.20e-03) |
| 200 | 5.29e-02 (1.16e-03) | 5.47e-02 (1.35e-03) | 5.45e-02 (1.09e-03) | 5.66e-02 (1.17e-03) |
| 400 | 3.27e-02 (6.96e-04) | 3.34e-02 (7.51e-04) | 3.29e-02 (6.22e-04) | 3.45e-02 (6.74e-04) |
| | Beta mixture #6 | | | |
| 100 | 1.40e-01 (3.62e-03) | 1.40e-01 (3.84e-03) | 1.37e-01 (3.33e-03) | 1.59e-01 (4.32e-03) |
| 200 | 8.31e-02 (1.96e-03) | 8.24e-02 (1.95e-03) | 8.24e-02 (1.97e-03) | 9.15e-02 (2.36e-03) |
| 400 | 4.79e-02 (1.09e-03) | 4.81e-02 (1.18e-03) | 4.77e-02 (1.06e-03) | 5.34e-02 (1.40e-03) |
| | Beta mixture #7 | | | |
| 100 | 1.33e-01 (2.11e-03) | 1.34e-01 (2.24e-03) | 1.50e-01 (2.04e-03) | 1.36e-01 (2.19e-03) |
| 200 | 7.70e-02 (1.21e-03) | 7.75e-02 (1.23e-03) | 8.98e-02 (1.26e-03) | 7.93e-02 (1.24e-03) |
| 400 | 4.65e-02 (7.20e-04) | 4.67e-02 (7.26e-04) | 5.25e-02 (7.69e-04) | 4.71e-02 (7.24e-04) |
| | Beta mixture #8 | | | |
| 100 | 1.74e-01 (3.46e-03) | 1.72e-01 (3.42e-03) | 1.89e-01 (3.03e-03) | 1.68e-01 (3.03e-03) |
| 200 | 9.30e-02 (1.56e-03) | 9.35e-02 (1.60e-03) | 1.10e-01 (1.79e-03) | 9.40e-02 (1.57e-03) |
| 400 | 5.53e-02 (8.87e-04) | 5.56e-02 (9.2e-04) | 6.41e-02 (1.04e-03) | 5.66e-02 (9.07e-04) |
| | Beta mixture #9 | | | |
| 100 | 2.45e-01 (3.58e-03) | 2.45e-01 (4.14e-03) | 3.35e-01 (3.08e-03) | 2.37e-01 (3.43e-03) |
| 200 | 1.47e-01 (2.25e-03) | 1.50e-01 (2.48e-03) | 2.09e-01 (2.09e-03) | 1.46e-01 (2.01e-03) |
| 400 | 8.94e-02 (1.20e-03) | 9.14e-02 (1.24e-03) | 1.26e-01 (1.28e-03) | 8.95e-02 (1.12e-03) |

Table 2: *Estimates for* $\mathrm{E}(\mathrm{ISE}(f;n,\hat{h}))$ *based on* 500 *replications for each case. The values inside the brackets are the estimated standard errors of the sample averages.*

is the standard Gaussian density.

A set of beta mixture distributions over the interval $[0,1]$, which represents a variety of different shapes, was taken as test distributions. Here we present the results of 9 typical cases defined in Table 1 and whose density functions are plotted in Figure 3. For each test distribution, three sample sizes $n = 100, 200$ and 400 were considered and 500 replications were used to estimate $\mathrm{E}(\mathrm{ISE}(f; n, h))$. As before the integrated squared error $\mathrm{ISE}(f; n, h)$ has been computed using the composite Simpson's rule.

The simulation results testify that H and $\mathrm{H}_*$ perform very well in comparison with all the considered methods. Similar results were obtained by H and $\mathrm{H}_*$ and by CHdpi and CHcv, but H and CHdpi are easy to implement and less time-consuming than $\mathrm{H}_*$ and CHcv, respectively. Therefore, the simulation results summarised in Table 2 exclusively concern the bandwidth selectors H, CHdpi, SJ and CV.

From Table 2 we see that H and CHdpi perform quite similarly, specially for moderate and large sample sizes, for all the considered test distributions. This can be explained by the strong connection between the estimators $\hat{\theta}_{2,m}$ and $\tilde{\theta}_{2,\Lambda}$ as explained in Section 2. Additionally, H and CHdpi behave very well for all the considered densities. This nice feature is not shared by SJ or CV. SJ is specially good for unimodal and bimodal densities but performs poorly for multimodal densities when many minor peaks are present, and CV shows a superior performance when the underlying density is multimodal but is less effective for some of the unimodal and bimodal densities. These conclusions are enlightened in Figure 4 where the empirical distribution of $\mathrm{ISE}(f; n, h)$ is shown for densities #2, #5 and #8 and $n = 200$.

Taking into account the previous simulation results, if we were to recommend one single method for general purposes, we would take the new H method which is less time-consuming than CHdpi specially for large sample sizes. The bandwidth $\hat{h}_{\mathrm{H}}$ is easy to obtain, presents an overall good performance against all the above considered bandwidths, and the simulation results indicate that it should be a reliable bandwidth for most practical situations.

# 5 Proof of Theorem 1

Let $\hat{m} = \hat{m}(X_1, \ldots, X_n)$ be a random sequence of positive integers, and consider the sequences $h_i$ and $\hat{h}_{i,\hat{m}}$, for $i = 1, 2$, defined by (2) and (4), respectively. Taking into account the decomposition

$$\frac{\hat{h}_{i,\hat{m}}}{h_0} - 1 = \frac{h_i}{h_0}\left(\frac{\hat{h}_{i,\hat{m}}}{h_i} - 1\right) + \frac{h_i}{h_0} - 1,$$

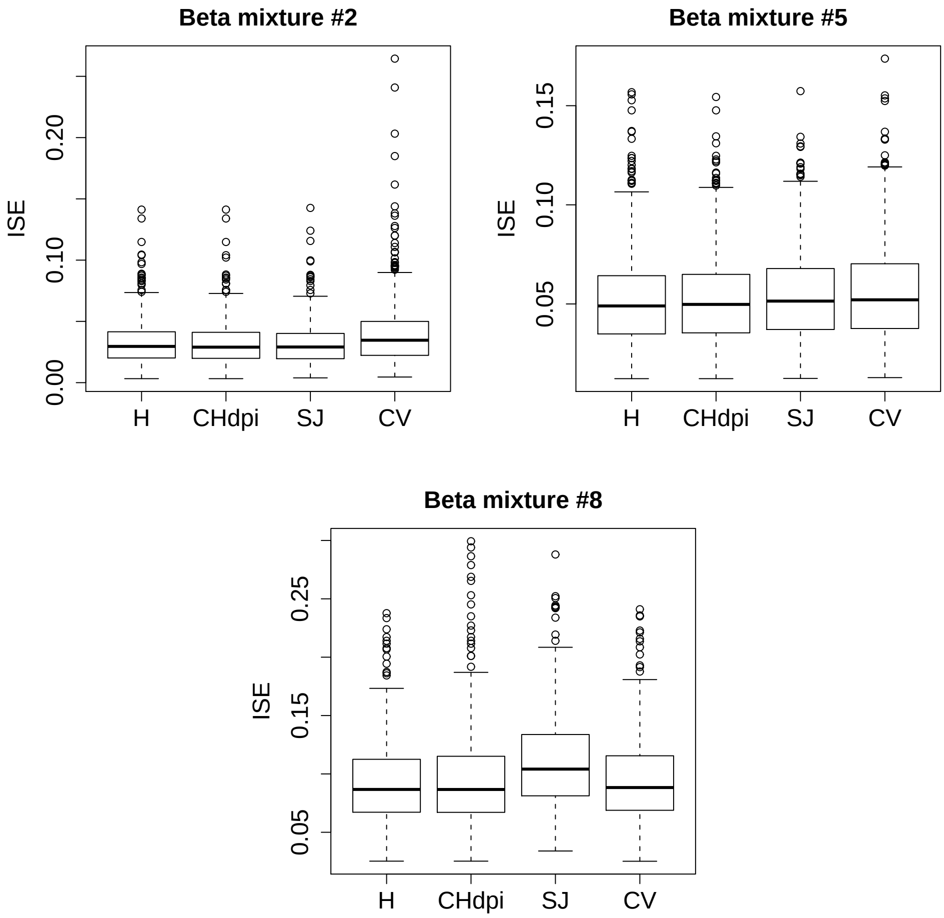


Figure 4: $\mathrm{ISE}(f;n,h)$ *distribution for* $n = 200$ *and* $h = \hat{h}_{\mathrm{H}}$, $h = \hat{h}_{\mathrm{CHdpi}}$, $h = \hat{h}_{\mathrm{SJ}}$ *and* $h = \hat{h}_{\mathrm{CV}}$. *The number of replications is* 500.

and the expansion

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

which is valid for $K$ a bounded and symmetric density function such that $\int |u|^5 K(u)du < \infty$ and $f$ a density defined over a compact set of $\mathbb{R}$ which is 4-times continuously differentiable in $\mathbb{R}$ (see Hall et al. 1991; section 2), the proof of Theorem 1 relies on the asymptotic behaviour of $\hat{\theta}_{r,\hat{m}}$ that we established in the following result.

**Lemma 1.** *For* $r > 0$ *and* $s = p + \alpha$, *with* $p \in \mathbb{N}$ *and* $\alpha \in ]0,1]$, *assume that* $f$ *is a density with support on* $[a,b]$ *which is* $p$*-times differentiable in* $[a,b]$ *with* $f^{(\ell)}(a) = f^{(\ell)}(b)$ *for* $\ell = 0,1,\ldots,p-1$, *and* $f^{(p)}$ *satisfies the Lipschitz condition (5).*
*a)* Consistency. *If* $\hat{m}$ *is such that* $\hat{m} \xrightarrow{p} +\infty$ *and* $n^{-1}\hat{m}^{2r+1} \xrightarrow{p} 0$ *then*

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

*b)* Rates of convergence. *If $\hat{m}$ satisfies (6) with*

$$0 < \xi_1 \leq \xi_2 < \frac{1}{2r+1},$$

*then*

$$\hat{\theta}_{r,\hat{m}} - \theta_r = O_p\left(n^{-\min\{1/2,1-\xi_2(2r+1),2\xi_1(s-r)\}}\right).$$

*c)* Asymptotic normality. *If $s > 2r + 1/2$ and $\hat{m}$ satisfies (6) with*

$$\frac{1}{4(s-r)} < \xi_1 \leq \xi_2 < \frac{1}{2(2r+1)},$$

*then*

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

The proof of Lemma 1 will be based on the double inequality

$$\hat{\theta}_{r,m_1} - c\,n^{-1}m_2^{2r+1} \leq \hat{\theta}_{r,\hat{m}} \leq \hat{\theta}_{r,m_2} + c\,n^{-1}m_2^{2r+1}, \tag{9}$$

where $c = 2\pi^{-1}(2\pi/(b-a))^{2r+1}$ and $m_1 = m_1(n)$ and $m_2 = m_2(n)$ are sequences of nonnegative integers with $m_1 \leq \hat{m} \leq m_2$, and on the following result that describes the asymptotic behaviour of $\hat{\theta}_{r,m}$ when $m = m(n)$ is a deterministic sequence of positive integers.

**Proposition 1** (Laurent, 1997, p. 190–204)**.** *For $r, m \in \mathbb{N}$, if $f$ satisfies the conditions of Lemma 1 with $s > r$ we have*

$$\mathrm{E}(\hat{\theta}_{r,m} - \theta_r)^2 \leq D_1\,n^{-1}m^{-\min\{0,2(s-2r-1/4)\}} + D_2\,n^{-2}m^{4r+1} + D_3\,m^{-4(s-r)},$$

*where $D_1, D_2$ and $D_3$ are positive constants independent of $n$ and $m$. Moreover, if $s > 2r + 1/4$ and $n^{-1/2}m^{2r+1/2} + n^{1/2}m^{-2(s-r)} \to 0$, as $n \to \infty$, we have*

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

**Proof of Lemma 1:** If $\hat{m}$ is such that $\hat{m} \xrightarrow{p} +\infty$ and $n^{-1}\hat{m}^{2r+1} \xrightarrow{p} 0$, we deduce that $\mathrm{P}\big(m_1 \leq \hat{m} \leq m_2\big) \to 1$, for all $N \in \mathbb{N}$ and $\xi > 0$, where $m_1 = N$ and $m_2 = \lfloor(\xi n)^{1/(2r+1)}\rfloor$. From the double inequality (9) and Proposition 1 we easily conclude that for all $\epsilon > 0$ and $\delta > 0$ there exist $N \in \mathbb{N}$, $\xi > 0$ and $n_0 \in \mathbb{N}$ such that $\mathrm{P}(|\,\hat{\theta}_{r,\hat{m}} - \theta_r| > \epsilon) < \delta$ for all $n \geq n_0$. This concludes the proof of part a).

From Proposition 1, we deduce that for $m = m_1 = \lfloor C_1\,n^{\xi_1}\rfloor$ or $m = m_2 = \lfloor C_2\,n^{\xi_2}\rfloor + 1$, with $C_1, C_2, \xi_1, \xi_2$ strictly positive constants, we have

$$\hat{\theta}_{r,m} - \theta_r = O_p\left(n^{-\min\{1/2,1-\xi_2(2r+1/2),2\xi_1(s-r)\}}\right),$$

if $0 < \xi_1 \leq \xi_2 < 1/(2r+1)$, and

$$\sqrt{n}\left(\hat{\theta}_{r,m} - \theta_r\right) \xrightarrow{d} N\left(0, 4\mathrm{Var}(f^{(2r)}(X_1))\right),$$

whenever $1/(s-r) < \xi_1 \leq \xi_2 < 1/(4r+1)$. These convergence results, together with (6) and (9), enable us to conclude the proof of parts b) and c).

□

**Acknowledgments.** The author expresses his thanks to the referee for the comments and suggestions. This research has been partially supported by the CMUC (Centre for Mathematics, University of Coimbra)/FCT.

# 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', *Statistica Sinica*, 17, 289–300.

Chiu, S.-T. (1991), 'Bandwidth selection for kernel density estimation', *The Annals of Statistics*, 19, 1883–1905.

Chiu, S.-T. (1992), 'An automatic bandwidth selector for kernel density estimation', *Biometrika*, 4, 771–782.

Chiu, S.-T. (1996), 'A comparative review of bandwidth selection for kernel density estimation', *Statistica Sinica*, 6, 129–145.

Deheuvels, P., Hominal, P. (1980), 'Estimation automatique de la densité', *Revue de Statistique Appliquée*, 28, 25–55.

Dmitriev, Y.G., Tarasenko, F.P. (1973), 'On the estimation of functionals of the probability density and its derivatives', *Theory of Probability and Its Applications*, 18, 628–633.

Devroye, L., Györfi, L. (1985), *Nonparametric Density Estimation: The $L_1$ View*, New York: Wiley.

Donoho, D., Nussbaum, M. (1990), 'Minimax quadratic estimation of a quadratic functional', *Journal of Complexity*, 6, 290–323.

Efromovich, S., Low, M. (1996), 'On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative', *The Annals of Statistics*, 24, 682–686.

Fan, J., Marron, J.S. (1992), 'Best possible constant for bandwidth selection', *The Annals of Statistics*, 20, 2057–2070.

Hall, P. (1983), 'Large sample optimality of least squares cross-validation in density estimation', *The Annals of Statistics*, 11, 1156–1174.

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

Hall, P., Marron, J.S. (1991), 'Lower bounds for bandwidth selection in density estimation', *Probability Theory and 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', *Journal of Statistical Computation and Simulation*, 21, 95–116.

Jones, M.C., Marron, J.S., Sheather, S.J. (1996), 'A brief survey of bandwidth selection for density estimation', *Journal of the American Statistical Association*, 91, 401–407.

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

Laurent, B. (1996), 'Efficient estimation of integral functionals of a density', *The Annals of Statistics*, 24, 659–681.

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

Levit, B.Ya. (1978), 'Asymptotically efficient estimation of nonlinear functionals', *Problems of Information Transmission*, 14, 64–72.

Loader, C.R. (1999), 'Bandwidth selection: classical or plug-in?', *The Annals of Statistics*, 27, 415–438.

Nadaraya, E.A. (1974), 'On the integral mean square error of some nonparametric estimates for the density function', *Theory of Probability and Its Applications*, 19, 133–141.

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

R Development Core Team (2009), *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', *The Annals of Mathematical Statistics*, 27, 832–837.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators', *Scandinavian Journal of Statistics*, 9, 65–78.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall

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

Stone, C.J. (1984), 'An asymptotically optimal window selection rule for kernel density estimates', *The Annals of Statistics*, 12, 1285–1297.

Tarter, M., Kronmal, R. (1968), 'The estimation of probability densities and cumulatives by Fourier series methods', *Journal of the American Statistical Association*, 63, 925–952.

Wand, M.P., Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman & Hall

Woodroofe, M. (1970), 'On choosing a delta-sequence', *The Annals of Mathematical Statistics*, 41, 1665–1671.