Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 09–18

FOURIER SERIES BASED BANDWIDTH SELECTORS FOR KERNEL DENSITY ESTIMATION

CARLOS TENREIRO

Abstract: A class of Fourier series based plug-in bandwidth selectors for kernel density estimation is considered in this paper. The proposed data-dependent bandwidths are simple to obtain, easy to interpret and consistent for a wide class of compact supported distributions. Some of them present good finite sample comparative performances against the classical two-stage direct plug-in method or the least squares cross-validation method, being good alternatives to these classical methods. Finally, we argue that the flexibility of the proposed class of bandwidths makes it suitable for the family approach to density estimation.

Keywords: Kernel density estimation, bandwidth selection, Fourier series based selectors, family approach.
AMS Subject Classification (2000): 62G05, 62G07.

1. Introduction

The kernel estimator for probability density functions introduced by Rosenblatt [29] and Parzen [26], known as Parzen-Rosenblatt estimator, is given, for $x \in \mathbb{R}$, by

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

where $X_1, \ldots, X_n$ are independent real-valued random variables with common density function f, $K_h(\cdot) = K(\cdot/h)/h$, for $h > 0$, K is a kernel on $\mathbb{R}$, i.e., an integrable function such that $\int_{\mathbb{R}} K(u)du = 1$, and $h = h_n$ is a sequence of strictly positive real numbers converging to zero as $n \to \infty$.

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 that present some features not shared by f whereas too large an h leads to a high biased estimator that does not revel some interesting characteristics of f (see Devroye and Györfi [7] and Bosq and Lecoutre [2] for some of the most important properties of f_n as an estimator of f). Due to its relevancy, the selection of the bandwidth

Received May 19, 2009.
This research has been partially supported by the CMUC (Centre for Mathematics, University of Coimbra)/FCT.

is one of the most studied topics in kernel density estimation and several approaches have been proposed for choosing h. A quite complete overview of the variety of methods appeared in the literature since the late seventies can be found in Wand and Jones [38] (Cap. 3), Jones *et al.* [17] and Chiu [5]. For some important comments on bandwidth selection see also Loader [21] and Sheather [33].

The reference distribution method and the direct plug-in method whose ideas go back to Deheuvels [6] and Woodroofe [39], respectively, are very simple data-dependent methods for choosing the bandwidth. They are based on asymptotic approximations for the bandwidth h_0 that minimizes the mean integrated square error $\mathrm{MISE}(f;n,h) = \mathrm{E}||f_n - f||_2^2$, where $||\cdot||_2$ denotes the L_2 distance:

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

(see Chacón *et al.* [4] for the existence and asymptotic behaviour of h_0). If we take for K a symmetric probability density function it is well known that, under some moment and regularity conditions on K and f, respectively, two asymptotic approximations to the optimal bandwidth h_0 are given by

$$h_1 = c_{1,K}||f''||_2^{-2/5} n^{-1/5} \tag{1}$$

and

$$h_2 = c_{1,K}||f''||_2^{-2/5} n^{-1/5} - c_{2,K}||f'''||_2^2\,||f''||_2^{-16/5} n^{-3/5},$$

where

$$c_{1,K} = \mu_2(K)^{-2/5}||K||_2^{2/5}, \quad c_{2,K} = \frac{1}{20}\,\mu_2(K)^{-11/5}\,\mu_4(K)||K||_2^{6/5}$$

and $\mu_j(K) = \int u^j K(u)du$ for $j = 1, 2, \ldots$ (cf. Hall and Marron [11, 12]). These asymptotic approximations to h_0 depend on the unknown functionals $||f''||_2$ and $||f'''||_2$ that need to be estimated from the data. The reference distribution and direct plug-in methods differ on the way such quantities are estimated. In the former case a parametric estimator based on some reference distribution family is used (usually the normal one). In the latter one these unknown quantities are replaced by consistent estimators.

The problem of estimating $||f^{(r)}||_2$, for some $r = 0, 1, \ldots$, has been studied by authors like Levit [20], Hall and Marron [11, 12], Bickel and Ritov [1], Jones and Sheather [18], Efromovich [8] and Laurent [19], and several estimators have been proposed by these authors. The class of kernel estimators proposed by Hall and Marron [11] and Jones and Sheather [18] is widely used in practice leading to some well-known bandwidth selection methods like those introduced by Sheather and Jones [34] (multistage solve-the-equation bandwidth selector), and Jones and Sheather [18] and Wand and Jones [38] (multistage direct plug-in bandwidth selector). To our knowledge, much less attention have been payed to the Fourier series based estimators studied by Efromovich [8] and Laurent [19].

It is the main purpose of this paper to introduce easy to use Fourier series based reference distribution and plug-in bandwidth selectors for kernel density estimation. Their finite sample behaviour mainly depends on the maximum number of terms used to model the underlying density through a truncated Fourier series which makes the considered class of bandwidth selectors quite flexible and of easy interpretation. Since the finite sample behaviour of the proposed methods for each one of the previous asymptotic approximations to h_0 was found very similar we restrict our attention to the methods based on the asymptotic approximation h_1. Moreover, only the kernel density setting is considered in this paper. However, the same approach might be used to the histogram estimator, the kernel distribution function estimator or the boundary kernel density or distribution function estimators (see Tenreiro [37]). For all these cases the asymptotic optimal bandwidth depends on some functionals of the form $||f^{(r)}||_2$, for some $r = 0, 1, \ldots$.

This paper is organized as follows: The Fourier series reference distribution method and the direct plug-in Fourier series based method are introduced in Sections 2 and 3, respectively. In Section 3 we also establish the consistency of the plug-in bandwidth for a wide class of compact supported distributions. In Section 4 the finite sample performance of the Fourier series based plug-in method is studied and compared with other well known data-based methods for choosing the bandwidth like the least squares cross-validation method introduced by Rudemo [28] and Bowman [3] (some attractive asymptotic properties are described in Hall [10], Stone [35] and Hall and Marron [11]) and the two-stage direct plug-in method proposed by Wand and Jones [38] (p. 72) (for large sample asymptotics see Hall *et al.* [13], Fan and Marron [9] and Tenreiro [36]). The bandwidths obtained by selecting a small or a large number of Fourier terms present good finite sample comparative performances against the two-stage direct plug-in method and the least squares cross-validation method, respectively, being good and simple alternatives to these classical methods. For an intermediate number of Fourier terms the proposed bandwidths present intermediate properties which makes this class of Fourier series based bandwidth selectors suitable for the family approach to density estimation recommended by several authors like Scott [31] p. 161, Marron and Chung [24] and Sheather [33], that advised looking at a family of density estimates based on a family of smoothing parameters. This procedure is illustrated in Section 5 for sets of simulated and real data. Finally, in Section 6 we provide some overall conclusions. All the proofs are postponed to Section 7. The simulations and plots in this paper were obtained using the R software [27].

2. The Fourier series reference distribution method

If the support of f is known to be on the finite interval $[a, b]$ and f is continuous in $[a, b]$ it is well known that the density f can be expanded in Fourier series (uniformly and in L_2 in $[a, b]$)

$$f(x) = \frac{1}{b-a} + \sum_{k=1}^{\infty} (a_k \phi_k(x) + b_k \psi_k(x)),$$

with

$$\phi_k(x) = \sqrt{\frac{2}{b-a}} \cos\left(\frac{k\pi}{b-a}(2x - a - b)\right),$$

$$\psi_k(x) = \sqrt{\frac{2}{b-a}} \sin\left(\frac{k\pi}{b-a}(2x - a - b)\right),$$

$$a_k = \int_a^b \phi_k(x) f(x) dx$$

and

$$b_k = \int_a^b \psi_k(x) f(x) dx$$

for $k \in \mathbb{N}$ (see Sansone [30]). Therefore, for a fixed nonnegative integer N it is natural to take for reference distribution family the family $\mathcal{F}_N$ of probability density functions f over the interval $[a, b]$ that take the form

$$f(x) = \frac{1}{b-a} + \sum_{k=1}^{N} (a_k \phi_k(x) + b_k \psi_k(x)),$$

for $x \in [a, b]$ with $a_N \neq 0$ or $b_N \neq 0$.

For $f \in \mathcal{F}_N$ we have

$$||f''||_2^2 = \frac{16\pi^4}{(b-a)^4} \sum_{k=1}^{N} k^4 c_k^2$$

with

$$c_k^2 = a_k^2 + b_k^2,$$

and, following Hart [15], an unbiased estimator of c_k^2 is given by

$$\widehat{c_k^2} := \widehat{a_k^2} + \widehat{b_k^2} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \{\phi_k(X_i)\phi_k(X_j) + \psi_k(X_i)\psi_k(X_j)\}\,. \quad (2)$$

Consequently, the reference distribution method based on the family $\mathcal{F}_N$ and on the asymptotic approximation h_1 to h_0 given by (1) leads to the data-driven bandwidth

$$\hat{h}_{1,N} = c_{1,K}\hat{\psi}_N^{-1/5} n^{-1/5},$$

where

$$\hat{\psi}_N := \frac{16\pi^4}{(b-a)^4} \sum_{k=1}^{N} k^4 \widehat{c_k^2}. \quad (3)$$

Taking into account that $\hat{\psi}_N$ is a nondegenerate U-statistic we easily conclude (cf. Hoeffding [16], p. 305) that the bandwidth relative error $\hat{h}_{1,N}/h_1 - 1$ is asymptotically normal and

$$\hat{h}_{1,N}/h_1 - 1 = O_p\left(n^{-1/2}\right),$$

whenever f belongs to $\mathcal{F}_M$ for some $M \leq N$. This rate of convergence is usually shared by a reference distribution bandwidth in relation to its reference distribution model. This is the case of the well known normal reference bandwidth method (cf. Deheuvels [6] and Silverman [32]). However, when the reference distribution family is $\mathcal{F}_N$ the practical interest of this property is stressed by the fact that a general continuous density f with support on $[a, b]$ can be approximated (uniformly and in L_2) by a density in $\mathcal{F}_N$ for some large value of N. This feature, which is not shared by the usually considered reference distribution models, is on the basis of the direct plug-in Fourier series based method introduced in the following.

3. A direct plug-in Fourier series based method

The finite sample performance of $\hat{\psi}_N$, and therefore the one of $\hat{h}_{1,N}$, strongly depends on the choice of N. If we do not have any relevante information about f that could guide the choice of the parametric model dimension N, it is natural to try to develop a data-based method for choosing N. We show in this section that if $\hat{N}$ is a possibly random sequence of positive integers, $\hat{N} = \hat{N}(X_1, \ldots, X_n)$, satisfying some general conditions, the associated data-based bandwidth given by

$$\hat{h} := \hat{h}_{1,\hat{N}} = c_{1,K}\hat{\psi}_{\hat{N}}^{-1/5} n^{-1/5},$$

is consistent for a wide class of probability distributions. Contrary to each one of the bandwidths $\hat{h}_{1,N}$, $\hat{h}$ is in fact a plug-in bandwidth since its consistency is not confined to the distributions of each one of the parametric models $\mathcal{F}_1, \ldots, \mathcal{F}_N$.

When $\hat{N} = N(n)$ is a deterministic sequence converging to infinity, $\hat{\psi}_{\hat{N}}$ is the estimator of $||f''||_2^2$ considered in Efromovich [8] and Laurent [19] which optimal rates of convergence are obtained for f belonging to some Lipschitz class of order $s + \alpha > 2$. However, since for a non-deterministic sequence $\hat{N}$ the statistic $\hat{\psi}_{\hat{N}}$ has no longer a U-statistic structure, the rates of convergence we present in the following are suboptimal (although close to the optimal ones).

We shall denote by $\mathcal{L}(s+\alpha)$, where $s \geq 0$ is an integer and $\alpha \in]0,1]$, the set of all densities f with support on $[a,b]$ which are s-times differentiable in $[a,b]$ with $f^{(\ell)}(a) = f^{(\ell)}(b)$ for $\ell = 0, 1, \ldots, s$, and $f^{(s)}$ satisfies the Lipschitz condition

$$|f^{(s)}(x) - f^{(s)}(y)| \leq C|x-y|^{\alpha}, \qquad x, y \in [a,b], \tag{4}$$

where C is a positive number.

Remark that if $f \in \mathcal{F}_M$ for some $M \in \mathbb{N}$, then $f \in \mathcal{L}(s+\alpha)$ for all $s \geq 0$ and $\alpha \in]0,1]$.

Theorem 1. *Assume that f belongs to $\mathcal{L}(\nu)$ for some $\nu = s + \alpha > 2 + 1/2$.*

(i) *If $\hat{N}$ is such that $\hat{N} \xrightarrow{p} +\infty$ and $n^{-1/5}\hat{N} = o_p(1)$ then*

$$\hat{h}/h_1 - 1 = o_p(1).$$

(ii) *If $\nu < 5$ and*

$$\mathrm{P}\left(C\, n^{(2\nu-5)/(4\nu(\nu-2))} \leq \hat{N} \leq D\, n^{1/(2\nu)}\right) \to 1,$$

for some positive constants C, D, we have

$$\hat{h}/h_1 - 1 = O_p\left(n^{-(2\nu-5)/(2\nu)}\right).$$

(iii) *If $\nu = 5$ and*

$$\mathrm{P}\left(C\, n^{1/12} \leq \hat{N} \leq D\, n^{1/10}\right) \to 1,$$

for some positive constants C, D, we have

$$\hat{h}/h_1 - 1 = O_p\left(n^{-1/2}\log n\right).$$

(iv) *If $\nu > 5$ and*

$$\mathrm{P}\left(C\, n^{1/(4\nu-2)} \leq \hat{N} \leq D\, n^{1/10}\right) \to 1,$$

for some positive constants C, D, we have

$$\hat{h}/h_1 - 1 = O_p\left(n^{-1/2}\right).$$

(v) *If $f \in \mathcal{F}_M$, for some $M \in \mathbb{N}$, and*

$$\mathrm{P}\left(M \leq \hat{N} \leq D\,n^{1/10}\right) \to 1,$$

for some positive constant D, we have

$$\hat{h}/h_1 - 1 = O_p\left(n^{-1/2}\right).$$

A simple data-based method for choosing N can be based on the natural connection between estimating the dimension of the parametric reference model and the estimation of the truncation point of a Fourier series density estimator. Hart [15] proposed to take for N the value $\hat{N}$ that minimizes $\widehat{H}(N)$ over the set $\{1, \dots, \mathrm{M}_n\}$ where $\mathrm{M}_n \in \mathbb{N}$ converges to infinity and

$$\widehat{H}(N) = \frac{2N}{n(b-a)} - \gamma_n \frac{n+1}{n} \sum_{k=1}^{N} \widehat{c_k^2}, \tag{5}$$

where γ_n, usually taken equal to 1, is a sequence of strictly positive real numbers converging to some strictly positive number $\gamma > 0$. If $c_M^2 > 0$ for some $M \in \mathbb{N}$, which is in particular true if $f \in \mathcal{F}_M$, we can prove that $\mathrm{P}(\hat{N} \geq M) \to 1$ as n tends to infinity (see Hart [15] p. 116). Therefore, if (M_n) is such that $\mathrm{M}_n \leq Dn^{1/10}$ for some positive constant D, the consistency (and rate of convergence) of the plug-in bandwidth associated to $\hat{N}$ for $f \in \mathcal{F}_M$ follows from part (v) of Theorem 1. Otherwise, if $f \in \mathcal{L}(\nu)$ for some $\nu = s + \alpha > 2 + 1/2$, and $f \notin \mathcal{F}_N$ for all $N \in \mathbb{N}$, we conclude that $\hat{N} \xrightarrow{p} +\infty$. Therefore, if (M_n) is such that $n^{-1/5}\mathrm{M}_n = o(1)$ the consistency of the plug-in bandwidth associated to $\hat{N}$ for such distributions follows from part (i) of Theorem 1. Additionally, if the minimization of $\widehat{H}(N)$ is restricted to a set of positive integers $\{\mathrm{L}_n, \mathrm{L}_n + 1, \dots, \mathrm{M}_n\}$, the conditions imposed to $\hat{N}$ in parts (ii)–(iv) of Theorem 1 are fulfilled for suitable choices of the sequences L_n and M_n which enables us to obtain rates of convergence for the relative error $\hat{h}/h_1 - 1$ where $\hat{h}$ is the plug-in bandwidth associated to $\hat{N}$.

4. A simulation study

A simulation study that includes some of the distributions considered by Marron and Wand [23] is undertaken in this section to analyze the finite sample behaviour of the Fourier series based plug-in bandwidth selector

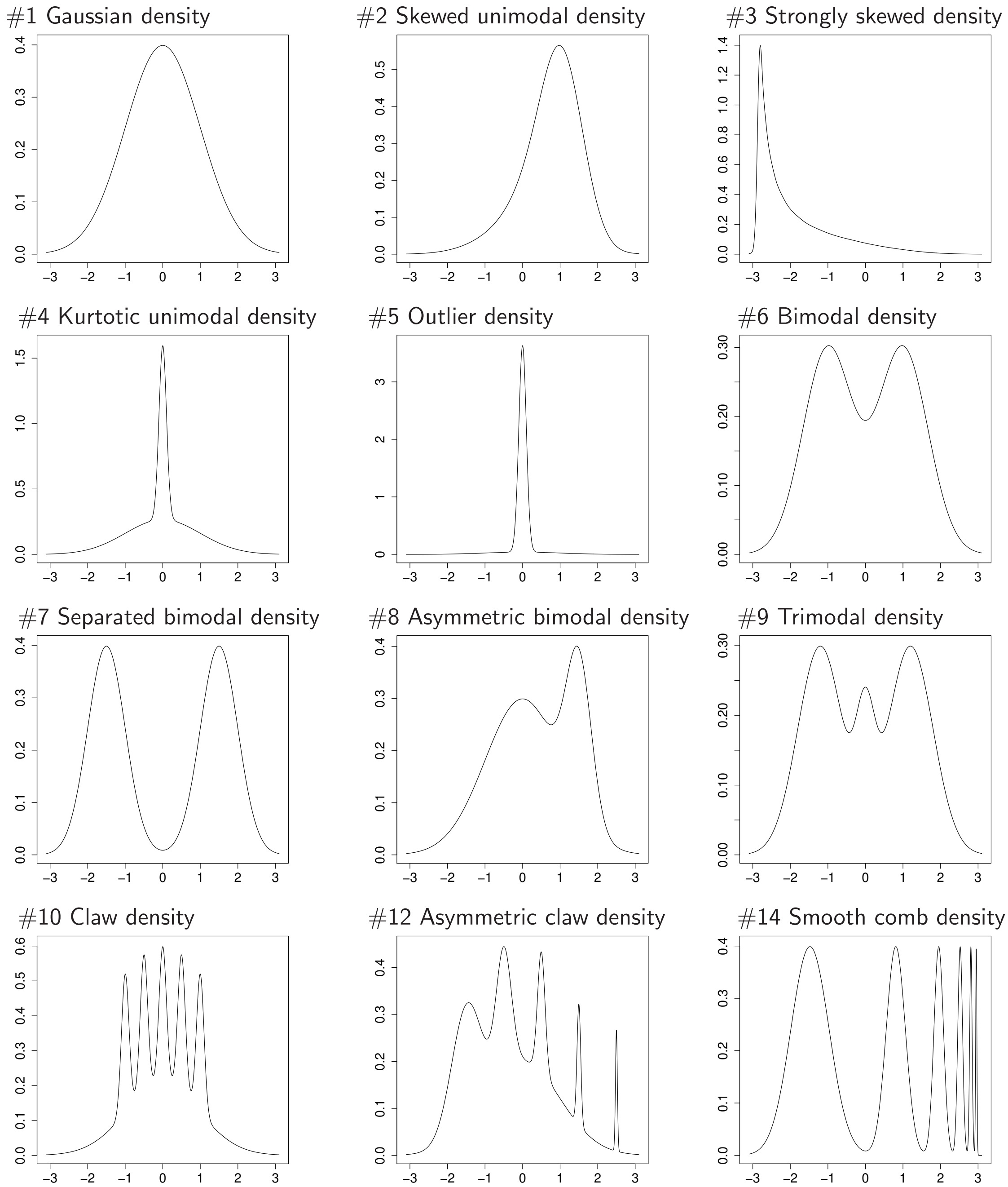

Figure 1. Densities used in the simulation study.

introduced in the previous section. The considered distributions are shown in Figure 1 where we keep the numeration used in Marron and Wand [23].

Taking into account that M_n is the maximum number of Fourier terms to be considered to model f through a truncated Fourier series, it is natural to expect that small values of M_n could be appropriate for densities #1, 2, 5, 6, 7, 8, 9 (easy-to-estimate densities) whereas large values of M_n could be adequate for densities #3, 4, 10, 12, 14 (hard-to-estimate densities: strongly multimodal, strongly asymmetric or strongly kurtotic).

Some preliminary simulation results confirm this expectable behaviour but also revealed that the choice $\gamma_n = 1$ in (5) does not prevent us against getting very large values for $\hat{N}$ whenever $f \in \mathcal{F}_K$ for some small value K. In practice this problem can be overcame by considering a smaller value for γ_n. If fact, in the limit case $\gamma_n \to 0$, we have $\mathrm{P}(\hat{N} \leq K) \to 1$ as $n \to \infty$, whenever $f \in \mathcal{F}_K$. In the rest of the paper we have considered $\gamma_n = 0.75$. Moreover, the low rates of convergence to infinity prescribed in Theorem 1 for the sequences L_n and M_n, lead us to consider the choices $\mathrm{L}_n = 1$ and $\mathrm{M}_n = \mathrm{M}$ with $\mathrm{M} = 2, 3, 5, 7, 10, 15$. The associated plug-in bandwidth $\hat{h}$ is therefore denoted by $\hat{h}_{\mathrm{M}}$.

In the rest of the paper two minor modifications of the estimators $\widehat{a_k^2}$ and $\widehat{b_k^2}$ given by (2) are considered. Firstly, in order to avoid negative estimates to the nonnegative quantities a_k^2 and b_k^2, they will be replaced by $\max(\widehat{a_k^2}, 0)$ and $\max(\widehat{b_k^2}, 0)$, respectively. Additionally, if the modified estimate of $||f''||_2^2$ is equal to zero, it will be replaced by $(\widehat{a}_1)^2 + (\widehat{b}_1)^2$, where $\widehat{a}_1$ and $\widehat{b}_1$ are the natural unbiased estimators of a_1 and b_1, respectively. This way, the considered estimator $\hat{\psi}_{\hat{N}}$ of $||f''||_2^2$ is strictly positive. Finally, since the previous distributions do not have compact support we have evaluated $\hat{h}_{\mathrm{M}}$ by considering $a = \max(\mathrm{Min}, \mathrm{Q}_1 - 1.5{\times}\mathrm{IQR})$ and $b = \min(\mathrm{Max}, \mathrm{Q}_3 + 1.5{\times}\mathrm{IQR})$, where Min, Max, Q_1, Q_3 and IQR are the sample minimum, maximum, first quartile, third quartile and interquartile range, respectively. The quartiles were evaluated using the R function quantile(·,type=7).

From each distribution we generated 100 samples of sizes $n = 100, 200, 400$ and a comparative analysis of the different methods behaviour is made by displaying the sample distribution of $\mathrm{ISE}(f; h)/\mathrm{ISE}(f; h_0)$ where the standard normal density was used as the kernel function and we have followed Marron and Wand [23] in the evaluation of $\mathrm{ISE}(f; h)$ and h_0. For comparative proposes we have taken the least squares cross-validation bandwidth $\hat{h}_{\mathrm{CV}}$ and the two-stages plug-in bandwidth $\hat{h}_{\mathrm{PI}_2}$. As remarked by Park and Marron [25], Jones *et al.* [17] and by other authors, the least squares cross-validation suffers from a great deal of sample variability and its performance has been often disappointing. However, as reported by Jones *et al.* [17], the distribution of $\hat{h}_{\mathrm{CV}}$ appears to have mean near h_0. For comparative proposes this is particularly important for hard-to-estimate densities where $\hat{h}_{\mathrm{PI}_2}$ fail completely the target. The observed results are shown in Figure 2.

For easy-to-estimate densities the best results are obtained by $\hat{h}_2$, $\hat{h}_3$ and $\hat{h}_{\mathrm{PI}_2}$, and these last two bandwidths show a similar performance for all

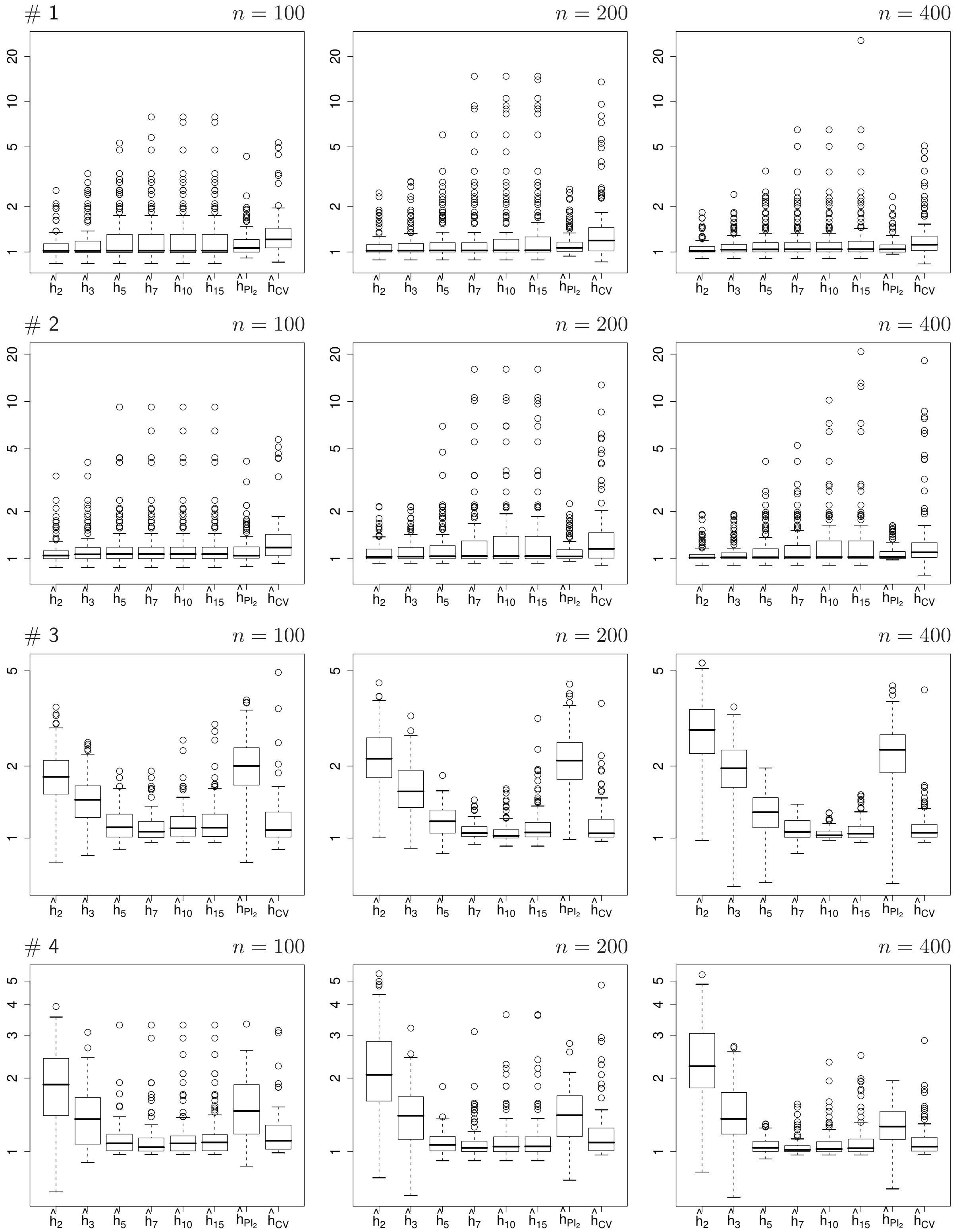

FIGURE 2. Empirical distribution of the relative errors $\mathrm{ISE}(f;h)/\mathrm{ISE}(f;h_0)$ for each h of $\hat{h}_{\mathrm{M}}$, $\mathrm{M} = 2,3,5,7,10,15$, $\hat{h}_{\mathrm{PI}_2}$ and $\hat{h}_{\mathrm{CV}}$, from 100 Monte Carlo samples of distributions #1, 2, 3 and 4.

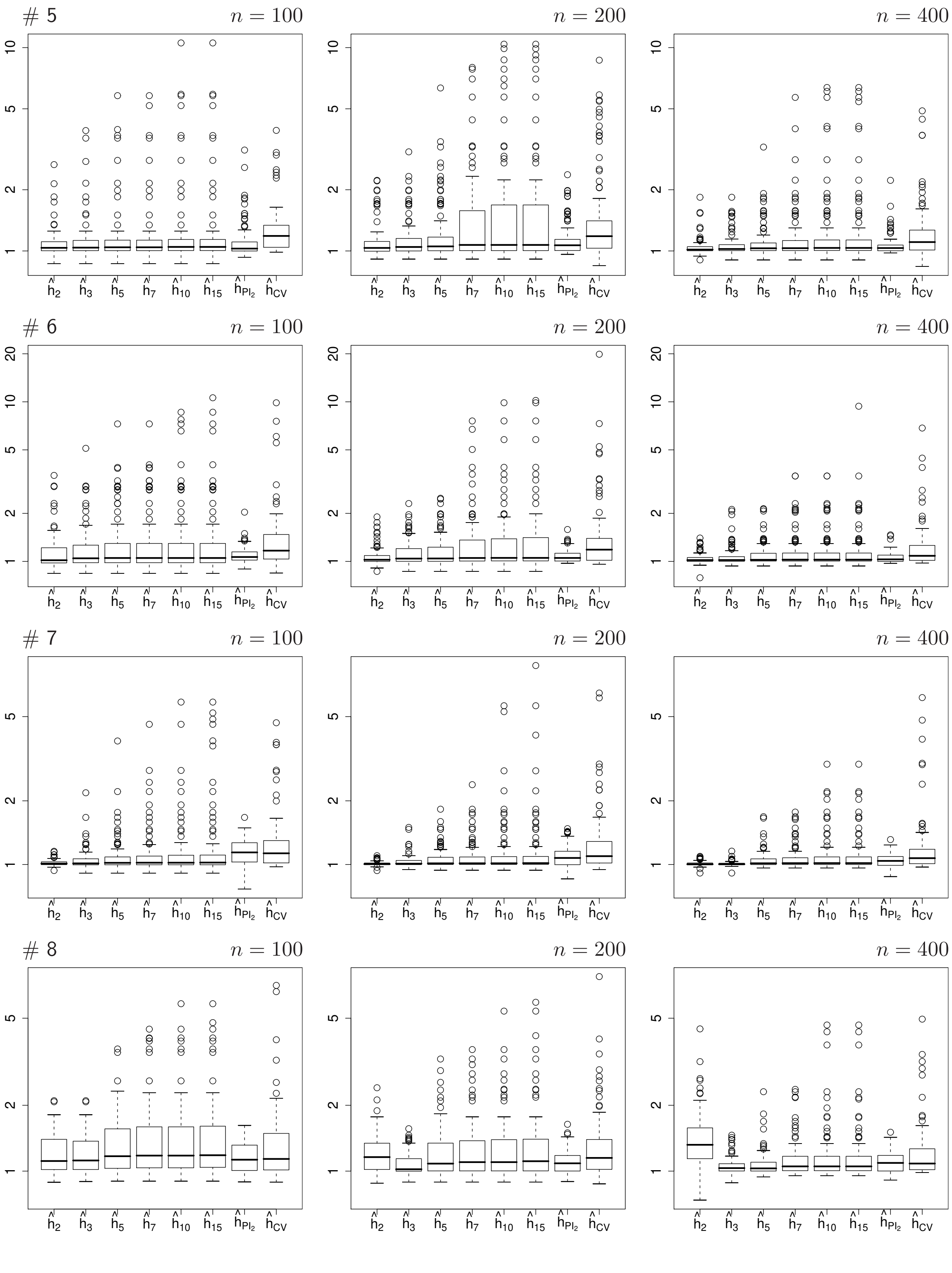

FIGURE 2. (cont.) Empirical distribution of the relative errors $\mathrm{ISE}(f;h)/\mathrm{ISE}(f;h_0)$ for each h of $\hat{h}_{\mathrm{M}}$, $\mathrm{M} = 2, 3, 5, 7, 10, 15$, $\hat{h}_{\mathrm{PI}_2}$ and $\hat{h}_{\mathrm{CV}}$, from 100 Monte Carlo samples of distributions #5, 6, 7 and 8.

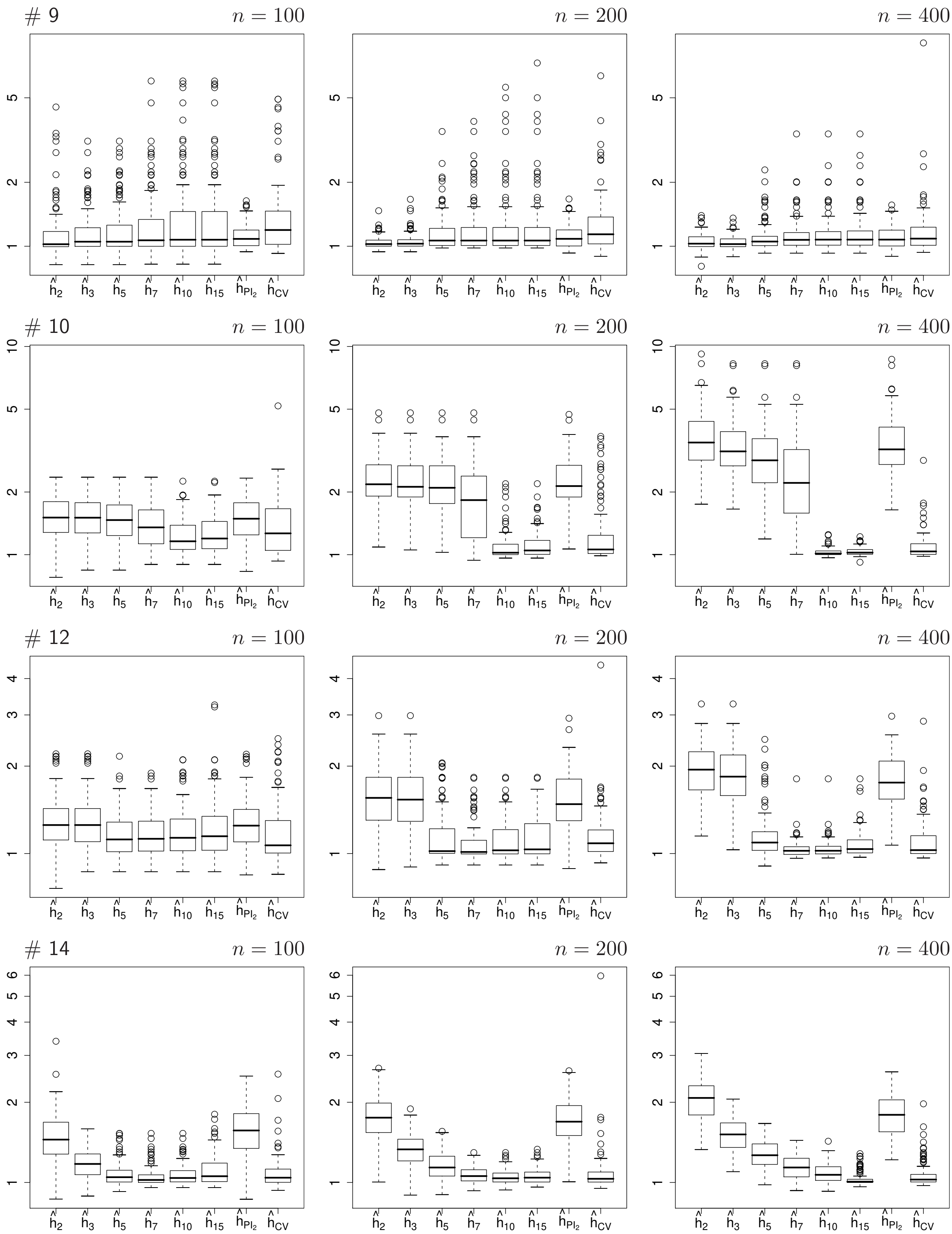

FIGURE 2. (cont.) Empirical distribution of the relative errors $\mathrm{ISE}(f;h)/\mathrm{ISE}(f;h_0)$ for each h of $\hat{h}_{\mathrm{M}}$, $\mathrm{M} = 2, 3, 5, 7, 10, 15$, $\hat{h}_{\mathrm{PI}_2}$ and $\hat{h}_{\mathrm{CV}}$, from 100 Monte Carlo samples of distributions #9, 10, 12 and 14.

the considered distributions. Therefore, $\hat{h}_3$ appears to be a good alternative to the commonly used two-stage direct plug-in bandwidth. Also remark the exceptional good comparative results obtained by $\hat{h}_2$ for some of these distributions and particularly for large sample sizes.

For easy-to-estimate densities the best results are obtained by $\hat{h}_2$, $\hat{h}_3$ and $\hat{h}_{\mathrm{PI}_2}$, and these last two bandwidths show a similar performance for all the considered distributions. Therefore, $\hat{h}_3$ appears to be a good alternative to the commonly used two-stage direct plug-in bandwidth. Also remark the exceptional good comparative results obtained by $\hat{h}_2$ for some of these distributions and particularly for large sample sizes. However, the performance of these bandwidths for the considered set of hard-to-estimate distributions is in general very poor. For these densities the best performance is achieved by $\hat{h}_{10}$ and $\hat{h}_{15}$ and slightly inferior results were obtained by $\hat{h}_{\mathrm{CV}}$. Taking into account the results observed for all the considered densities, $\hat{h}_{10}$ seems to be a better choice than the least squares cross-validation bandwidth. For intermediate values of M the bandwidth $\hat{h}_{\mathrm{M}}$ is in general better than $\hat{h}_3$ and worse than $\hat{h}_{10}$ for hard-to-estimate densities but worse than $\hat{h}_3$ and better than $\hat{h}_{10}$ for easy-to-estimate densities. Although the high variability shown by these bandwidths for the set of easy-to-estimate densities, if no information about the underlying density is available that enables us to classify it as an easy-to-estimate density, $\hat{h}_{10}$ appear to be the best of the considered data-based methods for choosing the bandwidth since its distribution appears to have median near h_0.

5. The family approach: some examples

In practice the choice of the maximum number of terms used to model the underlying density through a truncated Fourier series could be a hard task in particular if there is no available information about the underlying density that enables us to classify it as an easy-to-estimate or hard-to-estimate density. As stressed by Sheather [33] pp. 593–594, several authors like Scott [31] p. 161, and Marron and Chung [24] recommended the family approach to density estimation. They suggested that density estimates should be drawn with more than one value of the bandwidth. Taking into account the finite sample properties previously described, the family of bandwidth selectors introduced in this paper seems to be suitable to achieve this goal by considering the recommended bandwidth $\hat{h}_{10}$ but also the bandwidths $\hat{h}_{1,N}$ for some values of N between 1 and 15 (say). We expect that this set of estimates could give a more accurate picture about

	Datasets				
	1	2	3	4	5
$N = \mathrm{M}$	$\hat{h}_{1,N}$ / $\hat{N}$				
1	0.2796 / 1	0.2192 / 1	0.2561 / 1	0.0980 / 1	0.1527 / 1
2	0.2336 / 2	0.1417 / 2	0.1923 / 2	0.0695 / 2	0.1527 / 1
3	0.2336 / 2	0.1075 / 3	0.1332 / 3	0.0603 / 3	0.1527 / 1
4	0.2336 / 2	0.0906 / 4	0.1164 / 4	0.0540 / 3	0.1527 / 1
5	0.1670 / 2	0.0864 / 5	0.1060 / 5	0.0540 / 3	0.1527 / 1
6	0.1519 / 2	0.0762 / 6	0.0975 / 6	0.0367 / 3	0.1527 / 1
7	0.1519 / 2	0.0635 / 7	0.0862 / 7	0.0362 / 3	0.1527 / 1
8	0.1519 / 2	0.0580 / 8	0.0783 / 8	0.0339 / 3	0.0584 / 1
9	0.0914 / 2	0.0571 / 8	0.0747 / 8	0.0244 / 3	0.0584 / 1
10	0.0734 / 2	0.0549 / 10	0.0747 / 8	0.0223 / 3	0.0584 / 1
11	0.0620 / 2	0.0526 / 11	0.0747 / 8	0.0218 / 3	0.0584 / 1
12	0.0620 / 2	0.0475 / 12	0.0747 / 8	0.0218 / 3	0.0584 / 1
13	0.0620 / 2	0.0462 / 12	0.0747 / 8	0.0218 / 3	0.0584 / 1
14	0.0620 / 2	0.0453 / 12	0.0533 / 8	0.0208 / 3	0.0321 / 1
15	0.0620 / 2	0.0438 / 12	0.0523 / 8	0.0208 / 3	0.0296 / 1
	$\hat{h}_{\mathrm{PI}_2}$				
	0.2525	0.1201	0.1655	0.0759	0.1597
	$\hat{h}_{\mathrm{CV}}$				
	0.2867	0.0543	0.1026	0.0443	0.1906

TABLE 1. Bandwidths $\hat{h}_{1,N}$ and data dependent Fourier model dimension $\hat{N}$ for the considered datasets.

the underlying density function than the single density estimate associated to $\hat{h}_{10}$.

This procedure is illustrated in this section by considering simulated samples from distributions #2 and #3 of sizes 200 (dataset 1) and 500 (dataset 2), respectively, and a widely used dataset in density estimation literature that consists of eruption durations (in minutes) of the Old Faithful geyser (dataset 3). This set comes from Härdle [14] and is composed by 272 observations. Finally, the last two datasets we consider are obtained by splitting the eruption duration observations in two sets of short (dataset 4) and long (dataset 5) duration eruptions with sizes 97 and 195, respectively.

For each one of these datasets we give in Table 1 the bandwidths $\hat{h}_{1,N}$ for $N = 1, \ldots, 15$, and the values of $\hat{N}$, for $\mathrm{M} = 1, \ldots, 15$. Since $\hat{h}_{\mathrm{M}} = \hat{h}_{1,\hat{N}}$ and $\hat{N} \le \mathrm{M}$, from Table 1 we easily obtain $\hat{h}_{\mathrm{M}}$ for $\mathrm{M} = 1, \ldots, 15$. For comparative proposes we also give the bandwidths $\hat{h}_{\mathrm{PI}_2}$ and $\hat{h}_{\mathrm{CV}}$. For some values of N the corresponding kernel density estimates are displayed in Figures 3–7. The solid line always represent the density estimate based on

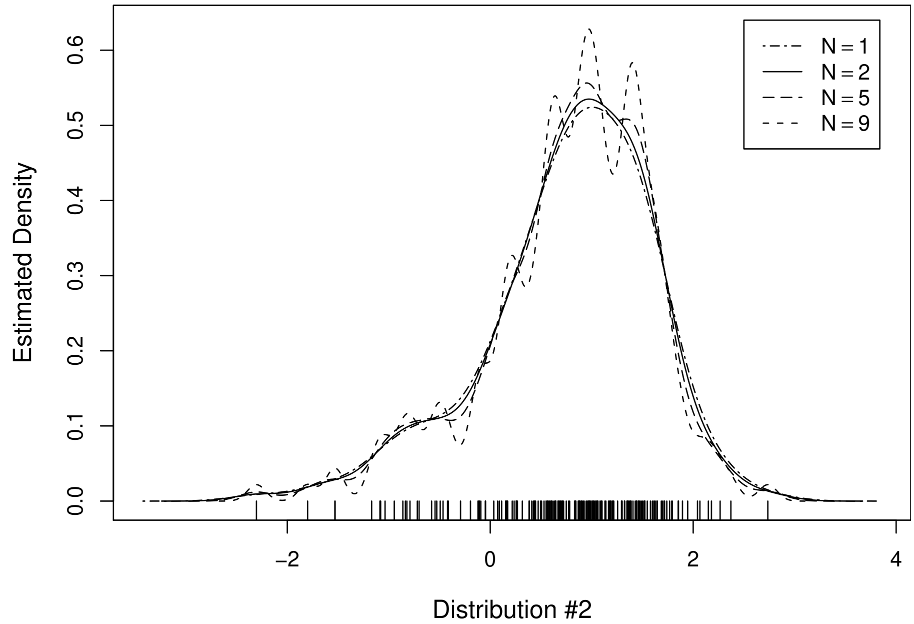

Figure 3. Kernel density estimates for a sample of size 200 from distribution #2 for the bandwidths $\hat{h}_{1,N}$ with $N = 2, 5, 7,$ 9. The solid line corresponds to the estimate based on $\hat{h}_{10}$.

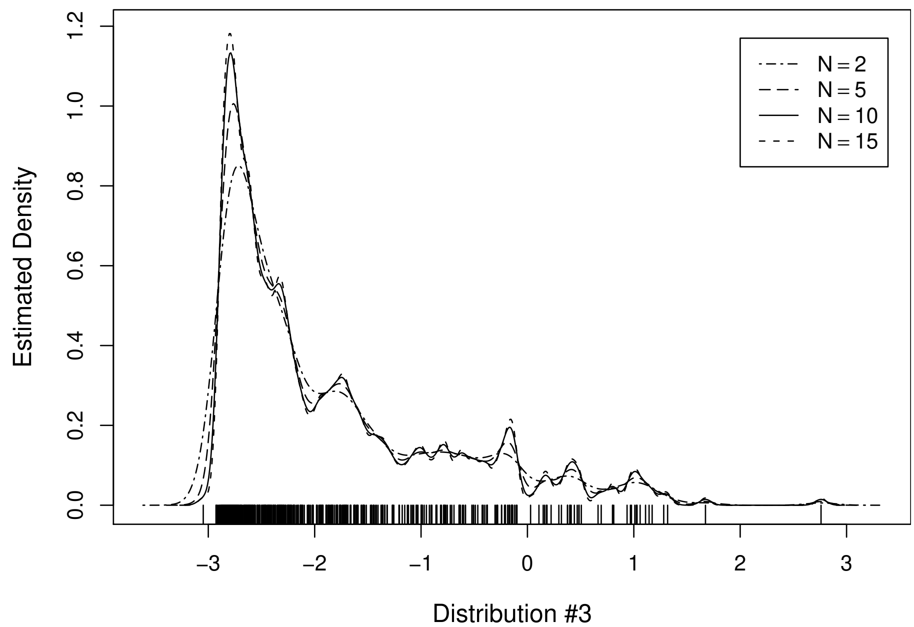

Figure 4. Kernel density estimates for a sample of size 500 from distribution #3 for the bandwidths $\hat{h}_{1,N}$ for $N = 2, 5, 10,$ 15. The solid line corresponds to the estimate based on $\hat{h}_{10}$.

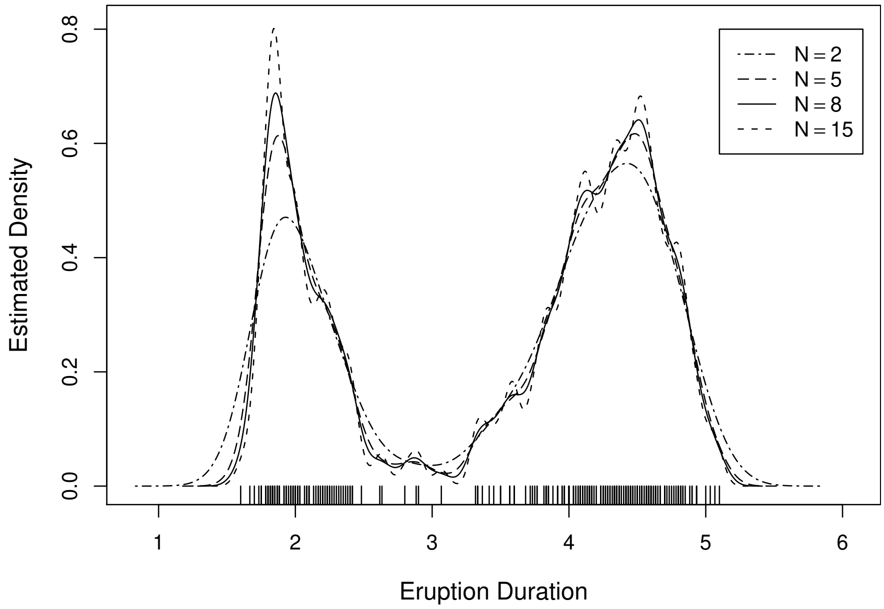

FIGURE 5. Kernel density estimates for the Old Faithful dataset for the bandwidths $\hat{h}_{1,N}$ for $N = 2, 5, 8, 15$. The solid line corresponds to the estimate based on $\hat{h}_{10}$.

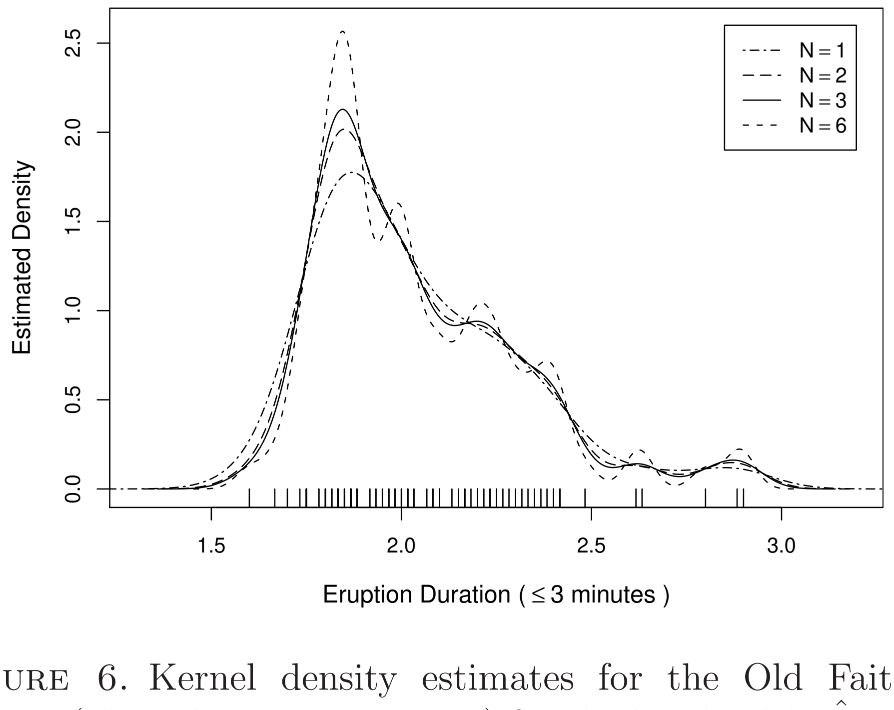

FIGURE 6. Kernel density estimates for the Old Faithful dataset (short duration eruptions) for the bandwidths $\hat{h}_{1,N}$ for $N = 1, 2, 3, 6$. The solid line corresponds to the estimate based on $\hat{h}_{10}$.

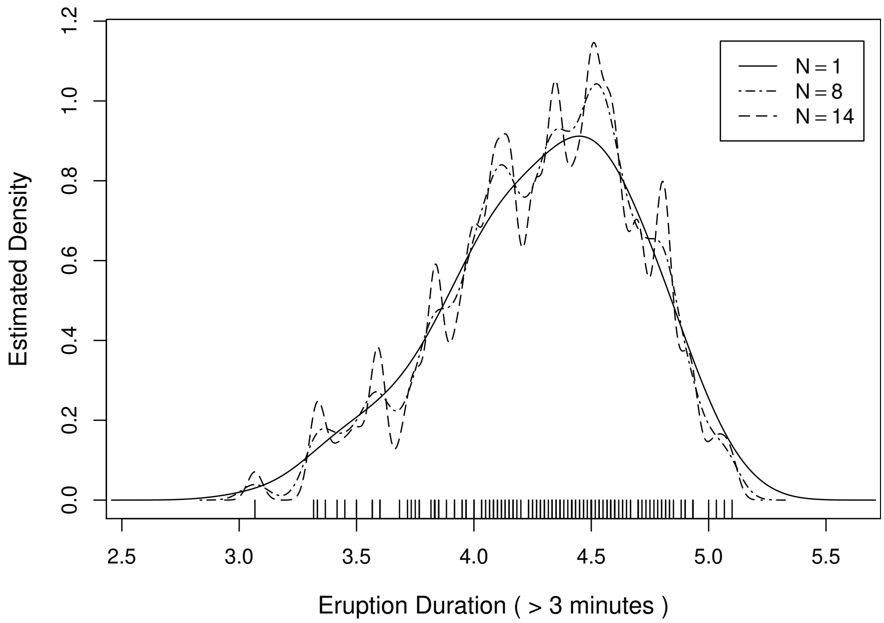

FIGURE 7. Kernel density estimates for the Old Faithful dataset (long duration eruptions) for the bandwidths $\hat{h}_{1,N}$ for $N = 1, 8, 14$. The solid line corresponds to the estimate based on $\hat{h}_{10}$.

$\hat{h}_{10}$. Taking into account these plots it appears that the density estimate based on $\hat{h}_{10}$ gives a good account of the underlying probability structure for all the considered datasets. Even for dataset 2 we cannot say that $\hat{h}_{10}$ is a less successful choice since it stresses an important feature of f and, as a byproduct, leads to a very local noisy estimate at the right support of the distribution. This is the price to pay for choosing a kernel estimator with a global bandwidth.

6. Conclusions

The class of Fourier series based plug-in bandwidth selectors for kernel density estimation proposed in this paper gives us a set of simple to use and large sample consistent data-dependent bandwidths. Their finite sample behaviour is easy to understand since it mainly depends on the maximum number of terms used to model the underlying density through a truncated Fourier series which makes the considered class of bandwidth selectors quite flexible and suitable to the family approach to density estimation. Additionally, the undertaken finite sample analysis gives us some additional inside about behaviour of the least squares cross-validation and the direct two-stage plug-in approaches to bandwidth selection. The two-stage plug-in bandwidth seems to behave like a reference distribution bandwidth based on a lower dimensional model. It behaves very well for densities close to the model but it cannot capture density features that are not taken into account by it. The least squares cross-validation bandwidth appears to behave like a true nonparametric procedure. In consequence, it presents large variability from sample to sample but it can manage with a wide set of distributional characteristics.

7. Proof of Theorem 1

Let f be such that $f \in \mathcal{L}(s+\alpha)$ for some integer $s \geq 2$ and $\alpha \in]0,1]$. From the continuity of f'' in $[a,b]$ and the fact that $f^{(\ell)}(a) = f^{(\ell)}(b)$, for $\ell = 0, 1$, we have

$$f''(x) = -\frac{4\pi^2}{(b-a)^2} \sum_{k=1}^{\infty} k^2 (a_k \phi_k(x) + b_k \psi_k(x))$$

in a L_2 sense in $[a,b]$ (see Sansone [30] p. 46) and

$$||f''||_2^2 = \frac{16\pi^4}{(b-a)^4} \sum_{k=1}^{\infty} k^4 (a_k^2 + b_k^2). \tag{6}$$

Therefore

$$\hat{\psi}_{\hat{N}} - ||f''||_2^2 = C_0 \sum_{k=1}^{\hat{N}} k^4 (\widehat{c_k^2} - c_k^2) - C_0 \sum_{k=\hat{N}+1}^{\infty} k^4 c_k^2,$$

where $C_0 = 16\pi^4/(b-a)^4$ and $\widehat{c_k^2}$ and c_k^2 have been defined in Section 2.

If L_n and M_n are sequences of natural numbers such that $\mathrm{L}_n \leq \hat{N} \leq \mathrm{M}_n$ we get

$$\left| \hat{\psi}_{\hat{N}} - ||f''||_2^2 \right| \leq C_0 \sum_{k=1}^{\mathrm{M}_n} k^4 |\widehat{c_k^2} - c_k^2| + C_0 \sum_{k=\mathrm{L}_n+1}^{\infty} k^4 c_k^2,$$

and consequently for

$$\Upsilon_n = \left(\hat{\psi}_{\hat{N}} - ||f''||_2^2 \right) \mathrm{I}(\mathrm{L}_n \leq \hat{N} \leq \mathrm{M}_n)$$

we have

$$\mathrm{E}\, |\Upsilon_n| = O\left(\sum_{k=1}^{\mathrm{M}_n} k^4 \mathrm{Var}(\widehat{c_k^2})^{1/2} + \sum_{k=\mathrm{L}_n+1}^{\infty} k^4 c_k^2 \right). \tag{7}$$

The following lemmas will be useful to control each one of the previous terms. The positive number C is the Lipschitz constant of f in (4).

Lemma 1. *For $f \in \mathcal{L}(s+\alpha)$ we have*

$$\operatorname{Var}(\widehat{c_k^2}) \leq \frac{C_1}{nk^{2s+2\alpha}} + \frac{C_2}{n^2},$$

with $C_1 = 32C^2\left((b-a)/(2\pi)\right)^{2s+2\alpha}$ *and* $C_2 = 32/(b-a)^2$.

Proof. This result follows from the standard technique for the calculation of the variance of the U-statistic $\widehat{c_k^2}$ and the following upper bound for the Fourier coefficients of f (see Sansone [30] p. 53)

$$\max(|a_k|, |b_k|) \leq 2C\pi^{1/2}\left(\frac{b-a}{2\pi}\right)^{s+\alpha+1/2}\frac{1}{k^{s+\alpha}}.$$

□

Lemma 2 (Lorentz's Inequality)**.** *For $f \in \mathcal{L}(s+\alpha)$ and $m \in \mathbb{N}$ we have*

$$\sum_{k=m+1}^{\infty} k^{2s}c_k^2 \leq \frac{C_3}{m^{2\alpha}}$$

where $C_3 = 2C^2(b-a)^{2s+2\alpha+1}/((2\pi)^{2s}4^\alpha(4^\alpha-1))$.

This lemma is proved in Devroye and Györfi [7] (pp. 304–308).

From (7) and the previous lemmas we get

$$\mathrm{E}\,|\Upsilon_n| = O\left(n^{-1/2}\sum_{k=1}^{\mathrm{M}_n} k^{-(s+\alpha-4)} + n^{-1}\mathrm{M}_n^5 + \mathrm{L}_n^{-2(s+\alpha-2)}\right),$$

and the announced probability rates of convergence for the relative error $\hat{h}/h_1 - 1$ in parts (i) to (iv) follow for suitable choices of the sequences L_n and M_n. Finally, if $f \in \mathcal{F}_M$ for some $M \in \mathbb{N}$, the part (v) of the theorem follows from the equality

$$\mathrm{E}\,|\Upsilon_n| = O\left(n^{-1/2} + n^{-1}\mathrm{M}_n^5\right),$$

which is a consequence of (7) by choosing $L_n = M$ and of Lemma 1 by taking $s+\alpha > 5$.

References

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

[2] Bosq, D., Lecoutre, J.-P. (1987). *Théorie de l'Estimation Fonctionnelle.* Economica, Paris.

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

[4] 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.

[5] Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statist. Sinica* 6, 129–145.
[6] Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl.* 25, 5–42.
[7] Devroye, L., Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.
[8] Efromovich, S. (1986). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* 30, 557–568.
[9] Fan, J., Marron, J.S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* 20, 2057–2070.
[10] Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* 11, 1156–1174.
[11] 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.
[12] Hall, P., Marron, J.S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields* 90, 149–173.
[13] 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.
[14] Härdle, W. (1991). *Smoothing techniques: with implementation in S.* Springer, New York.
[15] Hart, J.D. (1985). On the choice of a truncation point in Fourier series density estimation. *J. Stat. Comput. Simul.* 21, 95–116.
[16] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* 19, 293–325.
[17] Jones, M.C., Marron, J.S., Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* 91, 401–407.
[18] Jones, M.C., Sheather, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Letters*, 11, 511–514.
[19] Laurent, B. (1997). Estimation of integral functionals of a density and its derivatives. *Bernoulli* 3, 181–211.
[20] Levit, B.Ya. (1978). Asymptotically efficient estimation of nonlinear functionals. *Problems Inform. Transmission* 14, 64–72.
[21] Loader, C.R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.* 27, 415–438.
[22] Marron, J.S. (1989). Comments on a data based bandwidth selector. *Comput. Statist. Data Anal.* 8, 155–170.
[23] Marron, J.S., Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* 20, 712–736.
[24] Marron, J.S., Chung, S.S. (2001). Presentation of smoothers: the family approach. *Comput. Statist.* 16, 195-207.
[25] Park, B.U., Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* 85, 66–72.
[26] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* 33, 1065–1076.
[27] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.
[28] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* 9, 65–78.
[29] Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* 27, 832–837.
[30] Sansone, G. (1959). *Orthogonal functions.* Interscience Publ., New York.
[31] Scott, D.W. (1992). *Multivariate Density Estimation.* Wiley, New York.

[32] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.
[33] Sheather, S.J. (2004). Density Estimation. *Statist. Sci.* 19, 588–597.
[34] Sheather, S.J., Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Methodological* 53, 683–690.
[35] Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* 12, 1285–1297.
[36] Tenreiro, C. (2003). On the asymptotic normality of multistage integrated density derivatives kernel estimators. *Statis. Probab. Lett.* 64, 311–322.
[37] Tenreiro, C. (2008). Boundary kernels for distribution function estimation. *Pré-publicações do DMUC*, 08–25, University of Coimbra.
[38] Wand, M.P., Jones, M.C. (1995). *Kernel Smoothing*. Chapman & Hall, London.
[39] Woodroofe, M. (1970). On choosing a delta-sequence. *Ann. Math. Statist.* 41, 1665–1671.

Carlos Tenreiro
CMUC, Department of Mathematics, University of Coimbra, 3001–454 Coimbra, Portugal

E-mail address: tenreiro@mat.uc.pt
URL: http://www.mat.uc.pt/∼tenreiro/