Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 15–26

A WEIGHTED LEAST-SQUARES CROSS-VALIDATION BANDWIDTH SELECTOR FOR KERNEL DENSITY ESTIMATION

CARLOS TENREIRO

Abstract: Since the late eighties several methods have been considered in the literature to reduce the sample variability of the least-squares cross-validation bandwidth selector for kernel density estimation. In this paper a weighted version of this classical method is proposed and its asymptotic and finite sample behaviour is studied. The simulation results attest that the weighted cross-validation bandwidth performs quite well presenting a better finite sample performance than the standard cross-validation method for "easy-to-estimate" densities, and retaining the good finite sample performance of the standard cross-validation method for "hard-to-estimate" ones.

Keywords: Kernel density estimation; bandwidth selection; cross-validation.
AMS Subject Classification (2010): 62G07, 62G20.

1. Introduction

Let $X_1, \ldots, X_n$ be independent random variables from an absolutely continuous probability distribution with unknown density f on $\mathbb{R}$, and let

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

be the Parzen-Rosenblatt estimator of f (Rosenblatt, 1956, Parzen, 1962) based on kernel K and bandwidth h, where K is an integrable function with $\int K(u)du = 1$, $h = h_n$ is a sequence of strictly positive real numbers converging to zero as n tends to infinity, and $\alpha_h(\cdot) = \alpha(\cdot/h)/h$, for an arbitrary real function α (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).

The bandwidth controls the smoothness of the resulting curve estimate and its choice is a crucial step in estimating f. Due to its relevancy, this is one of the mostly studied topics in kernel density estimation and several approaches have been proposed for selecting h. One of such approaches is

Received June 2, 2015.

the now classical least-squares cross-validation method proposed by Rudemo (1982) and Bowman (1984). The cross-validation bandwidth is defined as

$$\hat{h}_{\mathrm{CV}} = \arg\min_{h>0} \mathrm{CV}(h),$$

where the least-squares cross-validation criterion function is given by

$$\mathrm{CV}(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} L_h(X_i - X_j),$$

with $L = (1-n^{-1})K * \bar{K} - 2K$, where $*$ denotes the convolution product, $\bar{\alpha}(u) = \alpha(-u)$ and $R(\alpha) = \int \alpha(x)^2 dx$, whenever α is square integrable. For each $h > 0$, $\mathrm{CV}(h)$ is a minimum variance unbiased estimator of $\mathrm{MISE}(f; n, h) - R(f)$ (see Serfling, 1980, p. 176) where

$$\mathrm{MISE}(f; n, h) = \mathrm{E}(\mathrm{ISE}(f; n, h)) = \mathrm{E}\left(\int \{f_n(x) - f(x)\}^2 dx\right)$$

is the mean integrated square error of the kernel density estimator f_n. Denoting by h_{MISE} the exact optimal bandwidth in the sense that

$$h_{\mathrm{MISE}} = \arg\min_{h>0} \mathrm{MISE}(f; n, h)$$

(see Chacón et al., 2007, for the existence and asymptotic behaviour of h_{MISE}), it is well known that under some regularity and moment conditions on K and f we have

$$\frac{\hat{h}_{\mathrm{CV}}}{h_{\mathrm{MISE}}} \xrightarrow{a.s.} 1,$$

and also

$$n^{1/10}\left(\frac{\hat{h}_{\mathrm{CV}}}{h_{\mathrm{MISE}}} - 1\right) \xrightarrow{d} N\big(0, \sigma^2_{\mathrm{CV}}(f; K)\big),$$

where

$$\sigma^2_{\mathrm{CV}}(f; K) = \frac{2R(\rho_K)\,\theta(f)}{25[R(K)^9 \mu_2(K)^2]^{1/5}}, \tag{1}$$

with $\rho_K(x) = x(K * \bar{K})'(x) - 2x(K^s)'(x)$, where α^s is the symmetrisation of α defined by $\alpha^s = (\alpha + \bar{\alpha})/2$, $\theta(f) = R(f)R(f'')^{-1/5}$ and $\mu_k(\alpha) = \int x^k \alpha(x) dx$ denotes the k-th moment of α whenever $|\mu_k|(\alpha) = \int |x^k \alpha(x)| dx$ is finite (see Hall, 1983, Stone, 1984, Nolan and Pollard, 1987, Hall and Marron, 1987, and Park and Marron, 1990; all these authors take for K a symmetric kernel in which case $K^s = K$).

Despite the inferior asymptotic performance presented by the cross-validation bandwidth in comparison with some other bandwidth selectors $\hat{h} = \hat{h}(X_1, \ldots, X_n)$, that achieve both optimal root n order of convergence and optimal asymptotic variance for the relative error $\hat{h}/h_{\mathrm{MISE}} - 1$ (see Hall and Marron, 1991, Hall et al., 1991, and Fan and Marron, 1992), it is well known that the cross-validation bandwidth presents a very good finite sample behaviour for "hard-to-estimate" densities, that is, densities with distributional characteristics such as strong asymmetry or multimodality usually leading to large values of the density functional $R(f'')$ and therefore small values of the density functional $\theta(f)$ (see the examples discussed in Loader, 1999). However, the cross-validation bandwidth also presents a large sample variability for "easy-to-estimate" densities. In this latter case the cross-validation criterion quite often selects too small a bandwidths leading to undersmoothing. This is an unattractive feature of $\hat{h}_{\mathrm{CV}}$ because the kernel density estimator is penalised much more by excessively small rather than by excessively large bandwidths (see Simonoff, 1996, p. 76, and references therein).

Several alternative methods have been considered in the literature in order to reduce the sample variability of the cross-validation bandwidth. These include the biased cross-validation method proposed by Scott and Terrel (1987), the smoothed cross-validation method of Hall et al. (1992) and its variants, including bootstrap bandwidth selection, and the direct plug-in method whose idea dates back to the works of Woodroofe (1970) and Nadaraya (1974). Reviews of all these methods can be found in Cao et al. (1994), Chiu (1996) and Chacón et al. (2008). In all these attempts the classical least-squares cross-validation function is replaced by new criterion functions that are better estimators of the entire MISE function, or an asymptotic approximation of it, than the least-squares cross-validation function is as an estimator of $\mathrm{MISE}(h) - R(f)$. More recently, new least-squares cross-validation based procedures have appeared in the literature such as those proposed by Hall and Robinson (2009) (bootstrap aggregation), Martínez-Miranda et al. (2009) (one-sided cross-validation) and Savchuk et al. (2010) (indirect cross-validation). An alternative approach that can be seen as a hybrid between cross-validation and direct plug-in bandwidths is considered in Chacón and Tenreiro (2013). For a recent simulation study that includes some of the previous bandwidth selectors see Heidenreich et al. (2013).

Although the generality of these methods perform quite well for "easy-to-estimate" densities improving the finite-sample performance of the cross-validation method, they occasionally also present a poor behaviour for "hard-to-estimate" densities being clearly outperformed by the classical cross-validation method. This is an undesirable feature in particular when no information about the underlying density is available.

A simple alternative to the standard cross-validation approach, that we call weighted least-squares cross-validation, is considered in this paper. We propose to replace the cross-validation function CV by a weighted version defined by

$$\mathrm{CV}_\gamma(h) = \frac{R(K)}{nh} + \frac{\gamma}{n(n-1)} \sum_{1 \leq i \neq j \leq n} L_h(X_i - X_j), \tag{2}$$

where the weight $\gamma = \gamma_n$, with $0 < \gamma \leq 1$, is at this point a deterministic value that needs to be chosen by the user, and we consider the weighted cross-validation bandwidth given by

$$\hat{h}_\gamma = \arg\min_{h>0} \mathrm{CV}_\gamma(h). \tag{3}$$

To the best of our knowledge a similar idea was for the first time mentioned by Hart (1985) for selecting the number of terms to be used in a Fourier series density estimator (see also Tenreiro, 2011). The major motivation for considering the previous weighted cross-validation function, is the fact that the function $\gamma \mapsto \hat{h}_\gamma - \hat{h}_{\mathrm{CV}}$ is nonnegative and nonincreasing for $0 < \gamma \leq 1$. Therefore, by choosing an appropriate weight value γ we introduce a positive bias with respect to the standard cross-validation bandwidth through which we expect to control undersmoothing. Besides, at least from an asymptotic point of view, $\hat{h}_\gamma$ presents a significant reduction in variability with respect to $\hat{h}_{\mathrm{CV}}$ as described below. These two features, with an apparent predominance of the former, are on the basis of the finite-sample properties of the weighted cross-validation bandwidth selector that we describe in this work.

The rest of the paper is organised as follows. In Section 2, we begin by stating the asymptotic equivalence, with probability one, between $\hat{h}_\gamma$ and the bandwidth h_γ that minimizes the function

$$h \mapsto \mathrm{ECV}_\gamma(f; n, h) := \mathrm{E}(\mathrm{CV}_\gamma(h)), \tag{4}$$

that is,

$$h_\gamma = \arg\min_{h>0} \mathrm{ECV}_\gamma(f; n, h), \tag{5}$$

and we conclude that the weighted cross-validation bandwidth is asymptotically equivalent to the optimal bandwidth h_{MISE} whenever γ converges to one as n tends to infinity. Moreover, we present an asymptotic expansion in probability for the relative error $\hat{h}_\gamma/h_{\mathrm{MISE}} - 1$ which is the main result of this paper. It enables us to quantify the above mentioned bias and variance effects and gives us a quite complete understanding of the role played by the weighted least-squares cross-validation function. In Section 3 we address the automatic choice of γ, which leads us to define the automatic weighted cross-validation bandwidth $\hat{h}_{\mathrm{WCV}}$. In Section 4 we undertake a simulation study to analyse its finite sample behaviour. The simulation results confirm that the weighted cross-validation bandwidth performs quite well presenting a better finite sample performance than the cross-validation method for "easy-to-estimate" densities, and retaining the good finite sample performance of the standard cross-validation method for "hard-to-estimate" ones. This is an important property that is not shared by the generality of the bandwidth selector methods proposed in the literature. All the proofs and some auxiliar results are deferred to Section 5. The simulations and plots in this paper were carried out using the R software (R Development Core Team, 2014).

2. Asymptotic behaviour of $\hat{h}_\gamma$

Next we describe the asymptotic behaviour of the weighted least-squares cross-validation bandwidth $\hat{h}_\gamma$ given by (3). Taking into account that $\hat{h}_\gamma = \hat{h}_{\mathrm{CV}}$ whenever $\gamma = 1$, the results presented in this section include specifically those obtained by authors such as Hall (1983), Stone (1984), Hall and Marron (1987), Nolan and Pollard (1987) and Park and Marron (1990).

Consider the following set of assumptions on f and K:

(D.1) f has continuous, bounded and integrable derivatives up to order 2.

(K.1) K is a continuous function on $\mathbb{R}$ such that $K(u) \to 0$, as $|u| \to \infty$, and

$$R(K) < 2K(0).$$

(K.2) K is a second order kernel, that is,

$$\mu_0(K) = 1, \quad \mu_1(K) = 0, \quad \text{and} \quad \mu_2(K) \neq 0.$$

(K.3) K is of bounded variation on $\mathbb{R}$.

The previous assumptions on K are satisfied by the kernels usually considered in the literature. In particular, the inequality $R(K) < 2K(0)$ is fulfilled for nonnegative kernels with $K(0) = \sup_{u\in\mathbb{R}} K(u)$. This set of assumptions is slightly different from the set of assumptions of Hall and Marron (1987). In particular, no symmetry or compact support conditions are imposed to K. This can be largely explained by the proof technique we have employed which relies on the uniform almost sure limit theorems of Pollard (1986) and Nolan and Pollard (1987). Finally, note that, by using the arguments of Stone (1984) and Chacón et al. (2007), we can conclude that the minima of the functions $h \mapsto \mathrm{CV}_\gamma(h)$ and $h \mapsto \mathrm{ECV}_\gamma(f; n, h)$ given by (2) and (4), respectively, are actually taken on at some $h > 0$ (the former with probability one) for

$$n > \frac{1-\gamma}{\gamma}\frac{R(K)}{2K(0) - R(K)}.$$

Therefore, under the previous general conditions the sequences $\hat{h}_\gamma$ and h_γ given by (3) and (5) are well defined for n large enough whenever the sequence of weights (γ_n) is such that $\liminf \gamma > 0$.

We start by stating the asymptotic equivalence with probability one between $\hat{h}_\gamma$ and h_{MISE} whenever the sequence of weights converges to one as n tends to infinity.

Theorem 1. *Under assumptions (D.1), (K.1), (K.2) and (K.3), if γ is such that $\liminf \gamma > 0$, then*

$$\frac{\hat{h}_\gamma}{h_\gamma} \xrightarrow{a.s.} 1.$$

Moreover, $\gamma \to 1$, as $n \to \infty$, if and only if

$$\frac{\hat{h}_\gamma}{h_{\mathrm{MISE}}} \xrightarrow{a.s.} 1.$$

Under some additional assumptions on f and K, we can establish the asymptotic normality of the weighted cross-validation bandwidth:

(D.2) f has continuous, bounded and integrable derivatives up to order 3.

(K.4) K has derivatives up to order 2 such that the functions $x \mapsto x^i K^{(j)}(x)$ are of bounded variation on $\mathbb{R}$ and $|\mu_6|(K^{(j)}) < \infty$, for $0 \le i \le j$ and $j = 0, 1, 2$.

Our main result is as follows:

Theorem 2. *Under assumptions (D.2), (K.1), (K.2) and (K.4), if* γ *is such that* $\liminf \gamma > 0$*, then*

$$\frac{\hat{h}_\gamma}{h_{\mathrm{MISE}}} - 1 = (\gamma^{-1/5} - 1) + \gamma^{7/10} n^{-1/10} \sigma_{\mathrm{CV}}(f; K) Z + o_p(n^{-1/10}),$$

where Z *is asymptotically normal* $N(0,1)$ *and* $\sigma^2_{\mathrm{CV}}(f;K)$ *is given by (1). Moreover, if* $\gamma = 1 + o(n^{-1/10})$ *we have*

$$n^{1/10}\left(\frac{\hat{h}_\gamma}{h_{\mathrm{MISE}}} - 1\right) \xrightarrow{d} N\big(0, \sigma^2_{\mathrm{CV}}(f;K)\big).$$

Based on the previous asymptotic expansion we conclude that by taking a large, but not too large weight γ, the weighted cross-validation bandwidth $\hat{h}_\gamma$ introduces a small positive bias in the exact optimal bandwidth estimation process but, at the same time, it reduces the variability associated to the standard cross-validation bandwidth. As n increases γ should vary in such a way that each of the components of the asymptotic mean square error of $\hat{h}_\gamma / h_{\mathrm{MISE}} - 1$ becomes smaller. Therefore, we should take $\gamma = \tilde{\gamma}$ where $\tilde{\gamma}$ is defined as the minimiser of the function $\gamma \mapsto (\gamma^{-1/5} - 1)^2 + \gamma^{7/5} n^{-1/5} \sigma^2_{\mathrm{CV}}(f;K)$. It is easy to see that $\tilde{\gamma} = \tilde{\eta}^5$, where $\tilde{\eta}$ is the unique root of the equation

$$\frac{7}{2}\, n^{-1/5} \sigma^2_{\mathrm{CV}}(f;K) \eta^9 + \eta - 1 = 0. \tag{6}$$

Moreover, $\tilde{\gamma}$ is a decreasing function of $\sigma^2_{\mathrm{CV}}(f;K)$ that converges to one as n tends to infinity with

$$n^{1/5}(\tilde{\gamma} - 1) \to -\frac{35}{2}\, \sigma^2_{\mathrm{CV}}(f;K).$$

Under the conditions of Theorem 2 we conclude that the ideal weighted cross-validation bandwidth defined by

$$\hat{h}_{\mathrm{IWCV}} := \hat{h}_{\tilde{\gamma}} = \mathrm{argmin}_{h>0} \mathrm{CV}_{\tilde{\gamma}}(h), \tag{7}$$

is such that

$$n^{1/10}\left(\frac{\hat{h}_{\mathrm{IWCV}}}{h_{\mathrm{MISE}}} - 1\right) \xrightarrow{d} N\big(0, \sigma^2_{\mathrm{CV}}(f;K)\big).$$

Thus, from an asymptotic point of view, no first-order differences exist between weighted and standard least-squares cross-validation bandwidths as estimators of h_{MISE}. However, as we will see later, from a finite sample

	Density number					
n	#1	#2	#3	#8	#12	#15
25	0.624	0.642	0.818	0.720	0.897	0.894
50	0.646	0.664	0.835	0.741	0.908	0.905
100	0.669	0.686	0.850	0.761	0.918	0.915
200	0.690	0.708	0.865	0.780	0.927	0.925
400	0.712	0.729	0.878	0.799	0.935	0.933
$\sigma_{\mathrm{CV}}(f;K)$	0.339	0.320	0.175	0.250	0.119	0.122

TABLE 1. *Values of $\tilde{\gamma}$ for some of the Marron and Wand's normal mixture densities where K is the standard Gaussian density.*

point of view considerable improvements can be obtained by the weighted cross-validation bandwidth for "easy-to-estimate" densities. Table 1 gives us a first indication in this direction. In this table we present the theoretical weights $\tilde{\gamma}$ for some of the densities of the well known Marron and Wand (1992) set of normal mixture densities, where we take for K the standard Gaussian density. For densities with small values of $\sigma_{\mathrm{CV}}(f;K)$, we see that, even for small sample sizes, $\tilde{\gamma}$ is close to one, which implies that marked differences between weighted and standard cross-validation bandwidths are not expected for such densities. This is in fact a desirable property because, as mentioned before, the standard cross-validation method performs quite well for this class of densities. For densities with large standard deviations $\sigma_{\mathrm{CV}}(f;K)$, the weights $\tilde{\gamma}$ given in Table 1 reveal that an improvement of the weighted cross-validation bandwidth over the standard one may be expected in this case, especially for small and moderated sample sizes. We shall return to this point later.

3. The automatic weighted cross-validation bandwidth

The exact evaluation of the optimal weight $\tilde{\gamma}$ defined in the previous section, and therefore, also the evaluation of the ideal weighted cross-validation bandwidth given by (7), is not possible in practice because the asymptotic variance $\sigma^2_{\mathrm{CV}}(f;K)$ given by (1) depends on the unknown density function f throughout the density functional $\theta(f) = R(f)R(f'')^{-1/5}$. Denoting by $\hat{\eta}$ the unique root of the equation obtained from (6) by replacing in $\sigma^2_{\mathrm{CV}}(f;K)$ the unknown parameter $\theta(f)$ by a strongly consistent estimator $\hat{\theta}_n$ (that is, $\hat{\theta}_n \to \theta(f)$ *a.s.*), and taking $\hat{\gamma} = \hat{\eta}^5$, we shall consider the automatic weighted

cross-validation bandwidth defined by

$$\hat{h}_{\mathrm{WCV}} := \hat{h}_{\hat{\gamma}} = \arg\min_{h>0} \mathrm{CV}_{\hat{\gamma}}(h),$$

that we will use as a surrogate for the ideal weighted cross-validation bandwidth $\hat{h}_{\mathrm{IWCV}}$.

Next we describe the asymptotic behaviour of $\hat{h}_{\mathrm{WCV}}$.

Theorem 3. *Assume that the density function f is such that $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is a strongly consistent estimator of $\theta(f)$. Under the assumptions of Theorem 1 we have*

$$\frac{\hat{h}_{\mathrm{WCV}}}{h_{\mathrm{MISE}}} \xrightarrow{a.s.} 1.$$

Moreover, under the assumptions of Theorem 2 we have

$$n^{1/10}\left(\frac{\hat{h}_{\mathrm{WCV}}}{h_{\mathrm{MISE}}} - 1\right) \xrightarrow{d} N\big(0, \sigma^2_{\mathrm{CV}}(f; K)\big).$$

Although the asymptotic behaviour of the automatic weighted cross-validation bandwidth $\hat{h}_{\hat{\gamma}}$ does not depend on the considered strongly consistent estimator $\hat{\theta}_n$ of $\theta(f) = \psi_0(f)\psi_4(f)^{-1/5}$, where for the convenience of notation we write $\psi_r(f) = (-1)^{r/2} R(f^{(r/2)})$ for an even positive integer r, it is natural to expect that its finite-sample performance could depend on $\hat{\theta}_n$. In this paper we consider the estimator of $\theta(f)$ defined by

$$\hat{\theta}_n = \tilde{\psi}_0 \tilde{\psi}_4^{-1/5}, \tag{8}$$

where $\tilde{\psi}_r$ denotes the two-stage direct plug-in kernel estimator of ψ_r given by

$$\tilde{\psi}_r = \hat{\psi}_r\left(\varphi_r\left(\left|\hat{\psi}_{r+2}\left(\varphi_{r+2}\left(|\hat{\psi}^{\mathrm{NR}}_{r+4}|\right)\right)\right|\right)\right), \tag{9}$$

where:

- $\hat{\psi}_r(g)$ is the kernel estimator of ψ_r introduced by Jones and Sheather (1991) (see also Hall and Marron, 1987a, and Wand and Jones, 1995, p. 67–70) defined by

$$\hat{\psi}_r(g) = \frac{1}{n^2}\sum_{i,j=1}^{n} \phi_g^{(r)}(X_i - X_j),$$

where ϕ denotes the standard Gaussian density, $g > 0$ is the bandwidth and $\phi_g^{(r)}$ represents the rth derivative of the function $\phi_g(x) = \phi(x/g)/g$, that is, $\phi_g^{(r)}(x) = \phi^{(r)}(x/g)/g^{r+1}$;

- φ_r is the real valued function defined for $t > 0$ by

$$\varphi_r(t) = \left(\frac{r!}{2^{(r-1)/2}(r/2)!\sqrt{\pi}\, n\, t} \right)^{1/(r+3)};$$

- $\hat{\psi}_r^{\mathrm{NR}}$ is a quick and simple estimator of the functional ψ_r based on the normal reference distribution given by

$$\hat{\psi}_r^{\mathrm{NR}} = \frac{(-1)^{r/2} r!}{(2\hat{\sigma})^{r+1}(r/2)!\sqrt{\pi}},$$

where $\hat{\sigma}$ is any scale estimator.

See Wand and Jones (1995, p. 71–74) for the motivation of this type of multistage kernel estimators and Tenreiro (2003) for their weak consistency and asymptotic normality. Other strong consistency results for this class of kernel estimators can be found in Liebscher (1998). For the sake of completeness we finish this section by presenting a set of sufficient conditions for the strong consistency of the two-stage direct plug-in kernel estimator $\tilde{\psi}_r$ defined by (9).

Theorem 4. *If the density function f has continuous, bounded and integrable derivatives up to order $r+2$, and there exists $\sigma_f > 0$ such that $\hat{\sigma} \to \sigma_f$ a.s., then $\tilde{\psi}_r \to \psi_r$ a.s.*

Taking for $\hat{\sigma}$ the scale estimator proposed by Silverman (1986, p. 47), that is, $\hat{\sigma} = \min\{s, IQR/1.34\}$, where s is the sample standard deviation and IQR the sample interquartile range, the previous assumption on $\hat{\sigma}$ is fulfilled whenever the density function f has a finite second moment and the corresponding distribution function is not flat in a right-neighbourhood of its first and third quartiles (see Serfling, 1980, p. 74–75). Therefore, under these general assumptions we conclude that $\hat{\theta}_n = \tilde{\psi}_0 \tilde{\psi}_4^{-1/5}$ is a strongly consistent estimator of $\theta(f)$ whenever f has continuous, bounded and integrable derivatives up to order 6.

4. Simulation study

We present in this section the results of a simulation study carried out to analyse the finite sample behaviour of the automatic weighted cross-validation bandwidth (WCV). Three other bandwidth selection methods are included in the study: the standard cross-validation method (CV), the bootstrap aggregation or bagging method (BAGG) proposed by Hall and Robinson (2009) and the one-sided cross-validation method (OSCV) considered in Martínez-Miranda et al. (2009) (see also Mammen et al., 2011). These latter methods are meant to reduce the stochastic variability of the cross-validation bandwidth, which explains their inclusion in our simulation study. In implementing all these bandwidths we take for K the standard Gaussian density. For the WCV bandwidth we take for $\hat{\sigma}$ the above mentioned scale estimator proposed by Silverman (1986, p. 47), and in the implementation of the BAGG method we follow the recommendations of Hall and Robinson (2009) by considering the approach based on half-sample bagging $\hat{h}_{\mathrm{CV}}$. As the considered kernel K is symmetric, note that the implemented OSCV bandwidth agrees with the right-sided cross-validation and do-validation bandwidths considered in Mammen et al. (2011). We use as test densities the 15 normal mixtures densities of Marron and Wand (1992) that we have refereed to in Section 2. All the simulations results are based on 500 samples of sizes $n = 25, 50, 100, 200$ and $n = 400$, from each test density in the study.

We start by presenting in Figure 1 the average and the standard deviation of the bandwidths produced by the ideal (IWCV) and automatic weighted cross-validation bandwidth selectors for densities #2 and #3. For comparative proposes we also present the average and standard deviation of the bandwidths selected by the standard cross-validation method. In view of this figure we see that $\hat{h}_{\mathrm{WCV}}$ is a better surrogate for the ideal bandwidth $\hat{h}_{\mathrm{IWCV}}$ for the "easy-to-estimate" density #2, than for the "hard-to-estimate" density #3. We know that the two-stage direct plug-in kernel estimators $\tilde{\psi}_r$ defined by (9) can perform poorly when the true underlying distribution deviates severely from normality because of their use of the normal reference density, which may explain the observed numerical results. In this particular case, the overestimation of $\theta(f)$ produced by the considered kernel estimator (8), leads to systematic smaller estimated weights $\hat{\gamma}$ than ideal weights $\tilde{\gamma}$, which explains the fact that the bandwidths produced by the automatic WCV method being larger that those generated by its ideal version. In

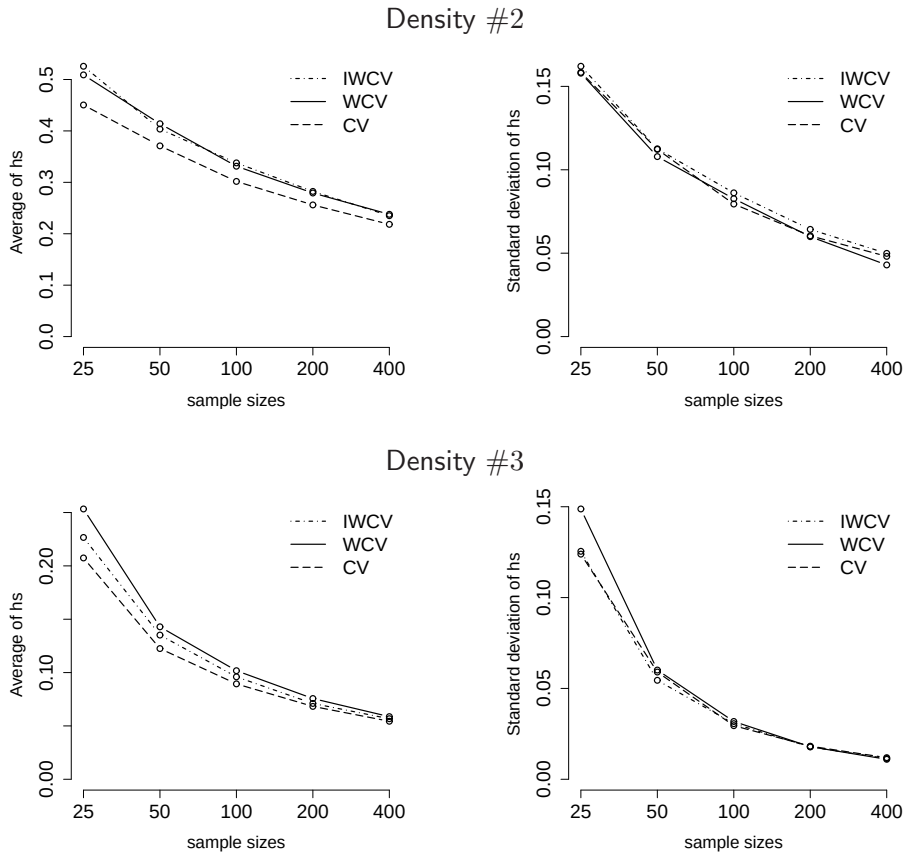

FIGURE 1. *Average and standard deviation of the bandwidths produced by the bandwidth selection methods* IWCV*,* WCV *and* CV*. The number of replications is* 500*.*

both cases the simulation results do not reveal any significant bandwidth variability reduction in comparison to the standard cross-validation bandwidth as we could expect from the asymptotic theory presented in Theorem 2. However, they clearly show the positive systematic bias introduced by the weighted cross-validation bandwidth with respect to the standard cross-validation bandwidth selector. As we will see below, this effect seems to dominate the finite sample behaviour of the weighted cross-validation bandwidth $\hat{h}_{\mathrm{WCV}}$.

For each one of the bandwidths WCV, CV, OSCV and BAGG, we describe in Figures 2 and 3 the behaviour of the following measure of the stochastic performance of the bandwidth selector $\hat{h}$:

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

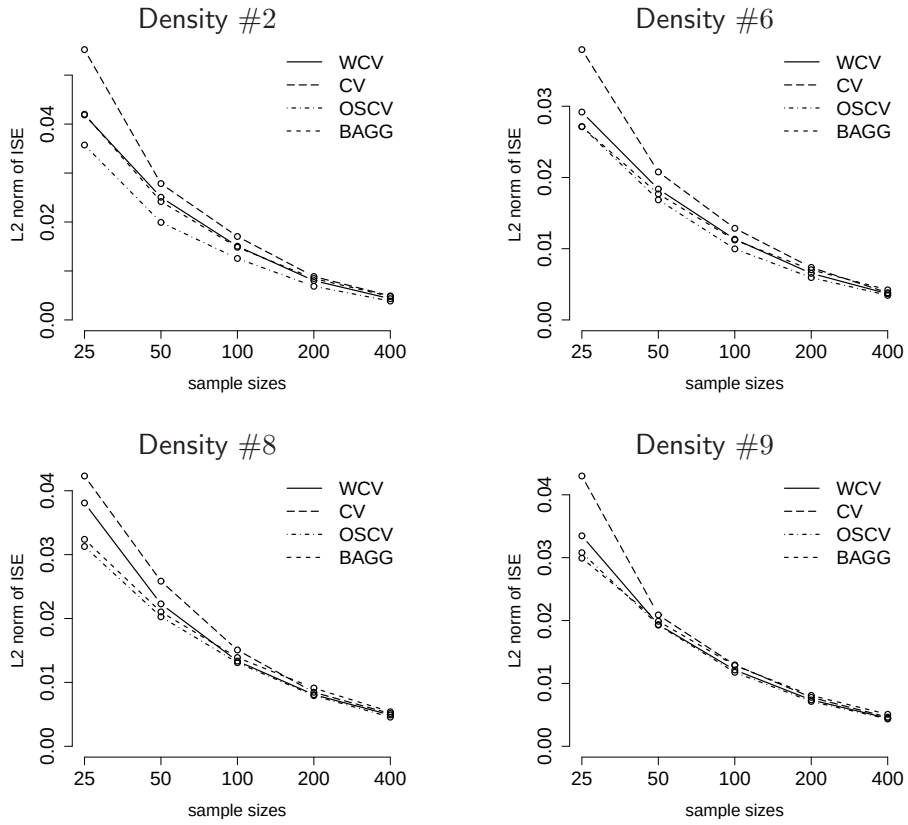

FIGURE 2. *Empirical L^2–norm of* $\mathrm{ISE}(f; n, \hat{h})$ *associated to the bandwidth selector methods* WCV*,* CV*,* OSCV *and* BAGG*, for some "easy-to-estimate" densities. The number of replications is* 500*.*

In Figure 2 we include some "easy-to-estimate" Marron and Wand's densities such as densities $\#2, \#6, \#8$ and $\#9$, whereas in Figure 3 some "hard-to-estimate" densities, such as densities $\#3, \#4, \#12$ and $\#15$, are considered.

As we can see from the graphics the OSCV and BAGG bandwidths present the best results for "easy-to-estimate" densities whereas the WCV and CV bandwidth selectors are the best methods for "hard-to-estimate" densities. With the exception of the multimodal densities #12 and #15 where the OSCV method shows a poor behaviour, the different bandwidths perform similarly when the sample size is large. Although inferior to OSCV and BAGG bandwidths for "easy-to-estimate" densities, the WCV bandwidth presents a better finite sample performance than the CV method for such densities, in particular for small sample sizes. Moreover, it retains the good finite sample performance of the CV method for "hard-to-estimate" densities.

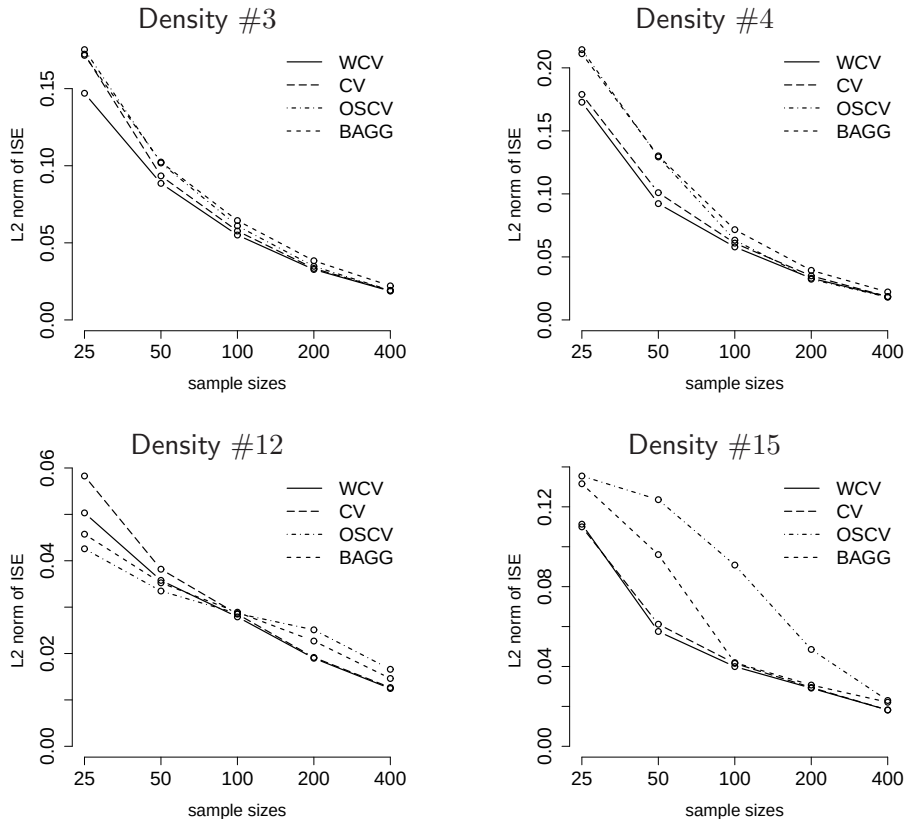

FIGURE 3. *Empirical* L^2*–norm of* $\mathrm{ISE}(f; n, \hat{h})$ *associated to the bandwidth selector methods* WCV*,* CV*,* OSCV *and* BAGG*, for some "hard-to-estimate" densities. The number of replications is* 500*.*

Based on this evidence, we expect that the new data-based bandwidth might present a good overall performance for a wide range of density features, which is an important property in particular when no information about the underlying density is available or when a complex data structure is suspected. Note also that this relevant attribute is not shared by the generality of the bandwidth selector methods proposed in the literature, which are usually high performing for "easy-to-estimate" densities, but, at the same time, they may be quite inefficient for densities presenting hard distributional features as strong asymmetry or multimodality.

5. Proofs

In this section we prove the asymptotic results stated in Sections 2 and 3. Before that several preliminar and auxiliar results are needed. We start

by studying the limit behaviour of the deterministic sequence h_γ, where we assume that the sequence of weights (γ_n) is such that $\liminf \gamma > 0$. This condition, that we assumed valid from now on, assures that the sequences $\hat{h}_\gamma$ and h_γ are well defined for n large enough whenever the kernel K is a continuous function that vanish at infinity with $R(K) < 2K(0)$, and f is a square integrable density (see the arguments of Stone, 1984, and Chacón et al., 2007). Although the results presented in this paper can be extended to continuous kernels on $\mathbb{R}\backslash\{0\}$ with finite one-sided limits $K(0^-)$ and $K(0^+)$, we always assume that K is a continuous function on $\mathbb{R}$.

Taking into account that $h_\gamma = h_{\mathrm{MISE}}$ for $\gamma = 1$, the next two results include the corresponding ones obtained by Chacón (2004) and Chacón et al. (2007). The technique used in their proofs was borrowed from these two references. For notational ease we write $\mathrm{ECV}_\gamma(n,h)$ or $\mathrm{ECV}_\gamma(h)$ instead of $\mathrm{ECV}_\gamma(f;n,h) := \mathrm{E}(\mathrm{CV}_\gamma(h))$. As the functions $\mathrm{CV}_\gamma(h)$ and $\mathrm{CV}(h)$ are related by

$$\mathrm{CV}_\gamma(h) = (1-\gamma)\frac{R(K)}{nh} + \gamma \mathrm{CV}(h), \tag{10}$$

the function $\mathrm{ECV}_\gamma(h)$ can be expressed in terms of $\mathrm{MISE}(h)$ by

$$\mathrm{ECV}_\gamma(h) = (1-\gamma)\frac{R(K)}{nh} + \gamma(\mathrm{MISE}(h) - R(f)). \tag{11}$$

Proposition 1. *Assume that f is square integrable and let K be a kernel satisfying conditions (K.1) and $\mu_2(K) \neq 0$. We have*

$$\lim h_\gamma = 0 \quad \textit{and} \quad \lim nh_\gamma = \infty.$$

Proof: We start by proving that $\gamma \mapsto h_\gamma$ is a non-increasing function of $\gamma \in]0,1]$. For that we use the fact that h_γ minimises $h \mapsto \mathrm{ECV}_\gamma(h)$ to first write that $\gamma_2 \mathrm{ECV}_{\gamma_1}(h_{\gamma_1}) + \gamma_1 \mathrm{ECV}_{\gamma_2}(h_{\gamma_2}) \leq \gamma_2 \mathrm{ECV}_{\gamma_1}(h_{\gamma_2}) + \gamma_1 \mathrm{ECV}_{\gamma_2}(h_{\gamma_1})$, for every $\gamma_1, \gamma_2 \in]0,1]$. As the previous inequality is equivalent to $(h_{\gamma_1} - h_{\gamma_2})(\gamma_2 - \gamma_1) \geq 0$, the non-increasing monotonicity of $\gamma \mapsto h_\gamma$ is established. From this property we deduce that $h_{\mathrm{MISE}} \leq h_\gamma$ and therefore $\liminf nh_{\mathrm{MISE}} \leq \liminf nh_\gamma$. This shows that $\lim nh_\gamma = \infty$ because $\lim nh_{\mathrm{MISE}} = \infty$ (see Chacón et al., 2007, Theorem 2, p. 291). Next, taking into account (11) and the fact that $\mathrm{ECV}_\gamma(h_\gamma) \leq \mathrm{ECV}_\gamma(h_{\mathrm{MISE}})$, we deduce that

$$0 \leq (1-\gamma)\frac{R(K)}{nh_\gamma} + \gamma \mathrm{MISE}(h_\gamma) \leq (1-\gamma)\frac{R(K)}{nh_{\mathrm{MISE}}} + \gamma \mathrm{MISE}(h_{\mathrm{MISE}}),$$

where $\lim \mathrm{MISE}(h_{\mathrm{MISE}}) = 0$ (see Chacón et al., 2007, proof of Theorem 2, p. 297). Therefore, we have $\lim \gamma\mathrm{MISE}(h_\gamma) = 0$, and also $\lim \mathrm{MISE}(h_\gamma) = 0$ as $\liminf \gamma > 0$. Finally, using the fact that K is a kernel whose Fourier transform is not identically equal to 1 in a neighbourhood of the origin, we conclude that $\lim h_\gamma = 0$ (see Chacón et al., 2007, proof of Theorem 3, p. 299). ■

Proposition 2. *Under assumptions (D.1), (K.1) and (K.2) we have:*

a) $0 < \liminf n^{1/5} h_\gamma \leq \limsup n^{1/5} h_\gamma < \infty$.

b) $h_\gamma / h_\gamma^* \to 1$, *where* $h_\gamma^* = \gamma^{-1/5} c_0(f) n^{-1/5}$

with $c_0(f) = R(K)^{1/5} \mu_2(K)^{-2/5} R(f'')^{-1/5}$.

Additionally, if f satisfies assumption (D.2) and $|\mu_4|(K) < \infty$, we have:

c) $h_\gamma / h_\gamma^* - 1 = O\left(n^{-2/5}\right)$.

Proof: a) For $h \to 0$ consider the asymptotic expansion

$$\mathrm{MISE}(h) = \frac{R(K)}{nh} + \frac{h^4}{4} \mu_2(K)^2 R(f'') + O(n^{-1}) + o(h^4) \tag{12}$$

(see Bosq and Lecoutre, 1987, p. 80–81). Using (11) we get, as $n \to \infty$,

$$\begin{aligned} & n^{4/5} \big(\mathrm{ECV}_\gamma(h_\gamma^*) + \gamma R(f) \big) \\ & \quad = (1-\gamma) \frac{R(K)}{n^{1/5} h_\gamma^*} + \gamma n^{4/5} \mathrm{MISE}(h_\gamma^*) \\ & \quad = \frac{R(K)}{n^{1/5} h_\gamma^*} + \gamma \frac{(n^{1/5} h_\gamma^*)^4}{4} \mu_2(K)^2 R(f'') + o(1) \\ & \quad = \frac{R(K)}{\gamma^{-1/5} c_0(f)} + \gamma \frac{(\gamma^{-1/5} c_0(f))^4}{4} \mu_2(K)^2 R(f'') + o(1). \end{aligned} \tag{13}$$

Therefore,

$$\limsup n^{4/5} \big(\mathrm{ECV}_\gamma(h_\gamma^*) + \gamma R(f) \big) \leq R(K)^{4/5} \mu_2(K)^{2/5} R(f'')^{1/5} < \infty. \tag{14}$$

On the other hand, from Proposition 1 we have

$$\begin{aligned} & n^{4/5} \big(\mathrm{ECV}_\gamma(h_\gamma) + \gamma R(f) \big) \\ & = \frac{R(K)}{n^{1/5} h_\gamma} + \gamma \frac{(n^{1/5} h_\gamma)^4}{4} \mu_2(K)^2 R(f'') + O(n^{-1/5}) + o((n^{1/5} h_\gamma)^4). \end{aligned} \tag{15}$$

Together with (14) this enables us to get the stated order of convergence for h_γ.

b) In order to prove that $\gamma^{1/5}n^{1/5}h_\gamma$ converge to $c_0(f)$, let $\gamma_{n_k}^{1/5}n_k^{1/5}h_\gamma(n_k) \to \lambda \in]0,\infty[$ be a convergent subsequence of $\gamma^{1/5}n^{1/5}h_\gamma$. From (13) and (15) we get

$$\frac{R(K)}{\lambda} + \frac{\lambda^4}{4}\mu_2(K)^2R(f'') \leq \frac{R(K)}{c_0(f)} + \frac{c_0(f)^4}{4}\mu_2(K)^2R(f''),$$

which implies that $\lambda = c_0(f)$. This concludes the proof of statement b).

c) Taking into account that the function $h \mapsto \mathrm{MISE}(h)$ is twice differentiable for $h > 0$ (see Hall and Marron, 1987, p. 569), the same is true for $h \mapsto \mathrm{ECV}_\gamma(h)$, which enables us to use the Taylor's formula to get the expansion

$$0 = \mathrm{ECV}'_\gamma(h_\gamma) = \mathrm{ECV}'_\gamma(h^*_\gamma) + \mathrm{ECV}''_\gamma(\tilde{h}_\gamma)(h_\gamma - h^*_\gamma),$$

with $\tilde{h}_\gamma$ between h_γ and h^*_γ. Therefore

$$\frac{h_\gamma}{h^*_\gamma} - 1 = -\left(\mathrm{ECV}''_\gamma(\tilde{h}_\gamma)h^*_\gamma\right)^{-1}\mathrm{ECV}'_\gamma(h^*_\gamma),$$

where

$$\mathrm{ECV}'_\gamma(h^*_\gamma) = -(1-\gamma)\frac{R(K)}{n{h^*_\gamma}^2} + \gamma\mathrm{MISE}'(h^*_\gamma)$$

and

$$n^{2/5}\mathrm{ECV}''_\gamma(\tilde{h}_\gamma) = (1-\gamma)\frac{2R(K)}{(n^{1/5}\tilde{h}_\gamma)^3} + \gamma n^{2/5}\mathrm{MISE}''(\tilde{h}_\gamma).$$

In order to conclude it suffices to use part b) and the following asymptotic expansions that can be derived by standard methods (see Hall and Marron, 1987, and Hall et al., 1991):

$$\mathrm{MISE}'(h) = -\frac{R(K)}{nh^2} + h^3\mu_2(K)^2R(f'') + O(n^{-1}h) + O(h^5), \qquad (16)$$

and

$$\mathrm{MISE}''(h) = \frac{2R(K)}{nh^3} + 3h^2\mu_2(K)^2R(f'') + O(n^{-1}) + O(h^4). \qquad \blacksquare$$

Proof of Theorem 1: Taking into account that f is a bounded density function with a square integrable second order derivative, and K is a second order kernel of bounded variation on $\mathbb{R}$, we first note that by using the

uniform almost sure limit theorems of Pollard (1986) and Nolan and Pollard (1987) we have

$$\sup_{h>0}\left|\frac{\mathrm{CV}(h)+R(f)-Z_n}{\mathrm{MISE}(h)}-1\right|\xrightarrow{a.s.}0, \tag{17}$$

where

$$Z_n=\frac{2}{n}\sum_{i=1}^{n}\big(f(X_i)-\mathrm{E}f(X_i)\big)$$

(see Pollard, 1986, p. 16–19, and Nolan and Pollard, 1987, p. 794–795). Hence, from (10), (11) and the fact that $\mathrm{ECV}_\gamma(h)+\gamma R(f)\geq\gamma\mathrm{MISE}(h)>0$ we get

$$\sup_{h>0}\left|\frac{\mathrm{CV}_\gamma(h)+\gamma R(f)-\gamma Z_n}{\mathrm{ECV}_\gamma(h)+\gamma R(f)}-1\right|\xrightarrow{a.s.}0.$$

Now by the arguments of Nolan and Pollard (1987, p. 794) we conclude that

$$\frac{\mathrm{ECV}_\gamma(\hat{h}_\gamma)+\gamma R(f)}{\mathrm{ECV}_\gamma(h_\gamma)+\gamma R(f)}\xrightarrow{a.s.}1,$$

which, together with (11) and (12), enables us to deduce the stated almost sure asymptotic equivalence between $\hat{h}_\gamma$ and h_γ. Finally, using Proposition 2.b) and the equality

$$\frac{\hat{h}_\gamma}{h_{\mathrm{MISE}}}=\frac{\hat{h}_\gamma}{h_\gamma}\frac{h_\gamma}{h_\gamma^*}\frac{h_\gamma^*}{h_{\mathrm{MISE}}}=\frac{\hat{h}_\gamma}{h_\gamma}\frac{h_\gamma}{h_\gamma^*}\frac{h_1^*}{h_1}\gamma^{-1/5},$$

we deduce that $\gamma\to1$ if and only if $\hat{h}_\gamma/h_{\mathrm{MISE}}\xrightarrow{a.s.}1$. ■

The following lemmas are crucial for the proof of Theorem 2 allowing us to establish the asymptotic normality of the relative error $\hat{h}_\gamma/h_\gamma-1$. Consider the U-statistic

$$U_\varphi(h)=\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\{\varphi_h(X_i-X_j)-\mathrm{E}(\varphi_h(X_i-X_j))\},$$

where $h=h_n$ is a non-random sequence of positive numbers converging to zero as n tends to infinity, and φ is a symmetric real-valued function.

The next lemma presents useful although standard expansions for the expectation $\mathrm{E}(\varphi_h(X_1-X_2))$, for a general, non necessarily symmetric function φ.

Lemma 1. *Assume that f has bounded and continuous derivatives up to order $s \in \mathbb{N}$ and let φ be a (non necessarily symmetric) real-valued function. The following expansions hold:*

a) If $|\mu_{2s-2}|(\varphi) < \infty$ we have

$$\mathrm{E}(\varphi_h(X_1 - X_2)) = \sum_{\ell=0}^{2s-3} \frac{h^\ell}{\ell!}\, \mu_\ell(\varphi)(\bar{f} * f)^{(\ell)}(0) + O(h^{2s-2}).$$

b) If $|\mu_{2s-1}|(\varphi) < \infty$ we have

$$\mathrm{E}(\varphi_h(X_1 - X_2)) = \sum_{\ell=0}^{2s-2} \frac{h^\ell}{\ell!}\, \mu_\ell(\varphi)(\bar{f} * f)^{(\ell)}(0) + O(h^{2s-1}).$$

c) If $|\mu_{2s}|(\varphi) < \infty$ we have

$$\mathrm{E}(\varphi_h(X_1 - X_2)) = \sum_{\ell=0}^{2s-1} \frac{h^\ell}{\ell!}\, \mu_\ell(\varphi)(\bar{f} * f)^{(\ell)}(0) + \frac{h^{2s}}{(2s)!}\mu_{2s}(\varphi)R(f^{(s)})(1 + o(1)).$$

Next we describe the asymptotic variance and the asymptotic distribution of $U_\varphi(h)$, when φ is a symmetric and square integrable function with vanishing moments up to order $k-1$ for some even integer k (see Lee, 1990, p. 12, Hall, 1984, and Tenreiro, 1997).

Lemma 2. *Assume that f has bounded and continuous derivatives up to order $s \in \mathbb{N}_0$, and let φ be a symmetric and square integrable function such that $|\mu_k|(\varphi) < \infty$ and $\mu_j(\varphi) = 0$, $j = 0, 1, \ldots, k-1$, for some even integer $k \in \{0, 2, 4, ...\}$. We have*

$$\mathrm{Var}(U_\varphi(h)) = 2n^{-2}h^{-1}R(\varphi)R(f) + O(n^{-1}h^{2(k\wedge s)}) + o(n^{-2}h^{-1}).$$

Moreover, if h is such that $nh^{2k+1} \to 0$, $\liminf n^\delta h > 0$, for some $0 < \delta < 2$, and $\limsup nh^{2s+1} < \infty$, we have

$$nh^{1/2}U_\varphi(h) \xrightarrow{d} N\big(0, 2R(\varphi)R(f)\big).$$

The following lemma gives an uniform in h almost sure upper bound for $U_\varphi(h)$ that relies on the uniform almost sure limit theorems of Pollard (1986) and Nolan and Pollard (1987).

Lemma 3. *Under the assumptions of Lemma 2, let φ be of bounded variation on $\mathbb{R}$. If β is a fixed natural number, then for all $0 < c_1 < c_2 < \infty$ we have*

$$\sup_{c_1 n^{-1/(2\beta+1)} \le h \le c_2 n^{-1/(2\beta+1)}} |U_\varphi(h)| = o\left(n^{-2(\beta \wedge s \wedge k)/(2\beta+1)}\right) \quad a.s.$$

Proof of Theorem 2: Consider the expansion

$$\frac{\hat{h}_\gamma}{h_{\mathrm{MISE}}} - \gamma^{-1/5} = \frac{h_\gamma}{\gamma^{-1/5} h_{\mathrm{MISE}}} \gamma^{-1/5} \left(\frac{\hat{h}_\gamma}{h_\gamma} - 1\right) + \gamma^{-1/5} \left(\frac{h_\gamma}{\gamma^{-1/5} h_{\mathrm{MISE}}} - 1\right),$$

where, taking into account Proposition 2.c), we have

$$\frac{h_\gamma}{\gamma^{-1/5} h_{\mathrm{MISE}}} - 1 = \frac{h_1^*}{h_1} \left(\left(\frac{h_\gamma}{h_\gamma^*} - 1\right) - \left(\frac{h_1}{h_1^*} - 1\right)\right) = O\big(n^{-2/5}\big).$$

Thus, Theorem 2 follows if we can prove that

$$\gamma^{-9/10} n^{1/10} \left(\frac{\hat{h}_\gamma}{h_\gamma} - 1\right) \xrightarrow{d} N\big(0, \sigma^2_{\mathrm{CV}}(f; K)\big),$$

with $\sigma^2_{\mathrm{CV}}(f; K)$ given by (1). Because K is twice differentiable on $\mathbb{R}$, we start by using Taylor's formula in order to write

$$-\mathrm{CV}'_\gamma(h_\gamma) = \mathrm{CV}'_\gamma(\hat{h}_\gamma) - \mathrm{CV}'_\gamma(h_\gamma) = \mathrm{CV}''_\gamma(\tilde{h}_\gamma)(\hat{h}_\gamma - h_\gamma),$$

and also

$$\frac{\hat{h}_\gamma}{h_\gamma} - 1 = -\left(\mathrm{CV}''_\gamma(\tilde{h}_\gamma) h_\gamma\right)^{-1} \mathrm{CV}'_\gamma(h_\gamma),$$

for some random variable $\tilde{h}_\gamma$ between $\hat{h}_\gamma$ and h_γ.

Now consider the expansion

$$\begin{aligned}\gamma^{-9/10} n^{1/10} \left(\frac{\hat{h}_\gamma}{h_\gamma} - 1\right) = &- \left(\gamma^{-2/5} n^{3/5} \mathrm{CV}''_\gamma(\tilde{h}_\gamma) h_\gamma\right)^{-1} \\ &\times \Big\{\gamma^{-3/10} n^{7/10} \big(\mathrm{CV}'(h_\gamma) - \mathrm{E}(\mathrm{CV}'(h_\gamma))\big) + n^{7/10} \mathrm{E}(\mathrm{CV}'_\gamma(h_\gamma))\Big\}.\end{aligned}$$

The asymptotic behaviour of each one of the right-hand side terms will be described in the following propositions. Together with the previous expansion, they enable us to conclude the proof of Theorem 2.

Proposition 3. *Under assumptions (D.1), (K.1) and (K.2), if K has a bounded derivative and $|\mu_5|(K^{(j)}) < \infty$, for $j = 0, 1$, we have*

$$\gamma^{-3/10} n^{7/10} \left(\mathrm{CV}'(h_\gamma) - \mathrm{E}(\mathrm{CV}'(h_\gamma))\right) \xrightarrow{d} N(0, \sigma^2),$$

where

$$\sigma^2 = 2R(K)^{-3/5} \mu_2(K)^{6/5} R(\rho_K) R(f) R(f'')^{3/5},$$

and $\rho_K(x) = x(K * \bar{K})'(x) - 2x(K^s)'(x)$.

Proof: We have

$$\begin{aligned} \mathrm{CV}'(h) = &-\frac{R(K)}{nh^2} - \frac{1}{nh(n-1)} \sum_{1 \le i \ne j \le n} \dot{M}_h(X_i - X_j) \\ &+ \frac{1}{n^2 h(n-1)} \sum_{1 \le i \ne j \le n} \dot{N}_h(X_i - X_j), \end{aligned} \tag{18}$$

where $M = K * \bar{K} - 2K$, $N = K * \bar{K}$, and $\dot{\alpha}(x) = \alpha(x) + x\alpha'(x)$ for a differentiable real valued function α. Thus

$$\begin{aligned} &\mathrm{CV}'(h_\gamma) - \mathrm{E}(\mathrm{CV}'(h_\gamma)) \\ &\quad = -\frac{2}{nh_\gamma(n-1)} \sum_{1 \le i < j \le n} \{\dot{M}_{h_\gamma}(X_i - X_j) - \mathrm{E}(\dot{M}_{h_\gamma}(X_i - X_j))\} \\ &\qquad + \frac{2}{n^2 h_\gamma(n-1)} \sum_{1 \le i < j \le n} \{\dot{N}_{h_\gamma}(X_i - X_j) - \mathrm{E}(\dot{N}_{h_\gamma}(X_i - X_j))\} \\ &\quad = -\frac{1}{h_\gamma} U_{\dot{M}^s}(h_\gamma) + \frac{1}{nh_\gamma} U_{\dot{N}}(h_\gamma), \end{aligned} \tag{19}$$

where $\dot{M}^s$ and $\dot{N}$ are symmetric and square integrable functions.

Taking into account that $\mu_j(\dot{N}) = 0$, $j = 0, 1$, and $\mu_2(\dot{N}) = -4\mu_2(K) \ne 0$ (notice that, by using the integration by parts formula, we have $\mu_j(\alpha') = -j\mu_{j-1}(\alpha)$, for every differentiable and integrable function α with bounded derivative such that $|\mu_j|(\alpha) < \infty$ and $|\mu_j|(\alpha') < \infty$, with $j \in \mathbb{N}$), by using Lemma 2 (with $k = s = 2$) and Proposition 2, we get

$$\frac{1}{nh_\gamma} U_{\dot{N}}(h_\gamma) = O_p\left(\frac{1}{nh_\gamma}\left(n^{-1} h_\gamma^{-1/2} + n^{-1/2} h_\gamma^2\right)\right) = O_p(n^{-17/10}). \tag{20}$$

On the other hand, we have $\mu_j(\dot{M}^s) = 0$, $j = 0, 1, 2, 3$, and $\mu_4(\dot{M}^s) = -24\mu_2(K)^2 \ne 0$. Therefore, by using Lemma 2 (with $k = 4$ and $s = 2$) and

Proposition 2, we get

$$nh_\gamma^{1/2} U_{\dot{M}^s}(h_\gamma) \xrightarrow{d} N\left(0, 2R(\dot{M}^s)R(f)\right). \tag{21}$$

Finally, taking into account (19), (20), (21) and Proposition 2 we get

$$\gamma^{-3/10} n^{7/10}(\mathrm{CV}'(h_\gamma) - \mathrm{E}(\mathrm{CV}'(h_\gamma))) = -\frac{1}{(\gamma^{1/5} n^{1/5} h_\gamma)^{3/2}} nh_\gamma^{1/2} U_{\dot{M}^s}(h_\gamma) + O_p(n^{-1}),$$

where the last term is negligible and the first one is asymptotically normal with zero mean and variance

$$\frac{2R(\dot{M}^s)R(f)}{c_0(f)^3} = 2R(K)^{-3/5}\mu_2(K)^{6/5} R(\dot{M}^s)R(f)R(f'')^{3/5},$$

with

$$\begin{aligned} R(\dot{M}^s) &= R(M^s) + R(x(M^s)') + \int 2x(M^s)'(x)M^s(x)dx \\ &= R(x(M^s)') = R(\rho_K). \end{aligned}$$

■

Proposition 4. *Under assumptions (D.1), (K.1) and (K.2), if K has a bounded derivative and $|\mu_6|(K^{(j)}) < \infty$, for $j = 0, 1$, we have*

$$\mathrm{E}(\mathrm{CV}'_\gamma(h_\gamma)) = O(n^{-4/5}).$$

Proof: Using (18) and Lemma 1.a) and b), we have

$$\begin{aligned} \mathrm{E}(\mathrm{CV}'_\gamma(h)) &= -(1-\gamma)\frac{R(K)}{nh^2} + \gamma \mathrm{E}(\mathrm{CV}'(h)) \\ &= -\frac{R(K)}{nh^2} - \frac{\gamma}{h}\mathrm{E}(\dot{M}_h(X_1 - X_2)) + \frac{\gamma}{nh}\mathrm{E}(\dot{N}_h(X_1 - X_2)), \end{aligned}$$

where

$$\mathrm{E}(\dot{M}_h(X_1 - X_2)) = \frac{h^4}{4!}\mu_4(\dot{M})R(f'') + O(h^5) = -h^4\mu_2(K)^2 R(f'') + O(h^5),$$

and

$$\mathrm{E}(\dot{N}_h(X_1 - X_2)) = O(h^2),$$

whenever h converges to zero. Thus from Proposition 2 we get

$$\mathrm{E}(\mathrm{CV}'_\gamma(h_\gamma)) = -\frac{R(K)}{nh_\gamma^2} + \gamma h_\gamma^3 \mu_2(K)^2 R(f'') + O(n^{-4/5}). \tag{22}$$

On the other hand, from (11) and (16) we have

$$\begin{aligned} 0 = \mathrm{MISE}'_\gamma(h_\gamma) &= -(1-\gamma)\frac{R(K)}{nh_\gamma^2} + \gamma \mathrm{MISE}'(h_\gamma) \\ &= -\frac{R(K)}{nh_\gamma^2} + \gamma h_\gamma^3 \mu_2(K)^2 R(f'') + O(n^{-4/5}). \end{aligned}$$

Together with (22) this enables us to conclude the proof. ■

Proposition 5. *Under assumptions (D.2), (K.1), (K.2) and (K.4), we have*

$$\gamma^{-2/5} n^{3/5} \mathrm{CV}''_\gamma(\tilde{h}_\gamma) h_\gamma = 5R(K)^{3/5} \mu_2(K)^{4/5} R(f'')^{2/5} + o(1) \quad a.s. \tag{23}$$

Proof: We have

$$\begin{aligned} \mathrm{CV}''(h) = \frac{2R(K)}{nh^3} &+ \frac{2}{nh^2(n-1)} \sum_{1 \le i < j \le n} \ddot{M}_h(X_i - X_j) \\ &- \frac{2}{n^2 h^2 (n-1)} \sum_{1 \le i < j \le n} \ddot{N}_h(X_i - X_j), \end{aligned}$$

where $\ddot{\alpha}(x) = 2\alpha(x) + 4x\alpha'(x) + x^2\alpha''(x)$, for a twice differentiable function α, and $\ddot{M}$ and $\ddot{N}$ are square integrable with $|\mu_4|(\ddot{M}) < \infty$ and $|\mu_2|(\ddot{N}) < \infty$, where $\mu_j(\ddot{M}) = 0$ for $j = 0, 1, 2, 3$, $\mu_4(\ddot{M}) = 72\mu_2(K)^2 \neq 0$, $\mu_j(\ddot{N}) = 0$ for $j = 0, 1$, and $\mu_2(\ddot{N}) = 4\mu_2(K) \neq 0$. Thus, by using Lemma 1.a) and c) we have

$$\begin{aligned} \mathrm{E}(\mathrm{CV}''_\gamma(h)) &= (1-\gamma)\frac{2R(K)}{nh^3} + \gamma \mathrm{E}(\mathrm{CV}''(h)) \\ &= \frac{2R(K)}{nh^3} + \frac{\gamma}{h^2} \mathrm{E}(\ddot{M}_h(X_1 - X_2)) - \frac{\gamma}{nh^2} \mathrm{E}(\ddot{N}_h(X_1 - X_2)) \\ &= \frac{2R(K)}{nh^3} + \frac{\gamma}{h^2} \frac{h^4}{4!} \mu_4(\ddot{M}) R(f'')(1 + o(1)) + \frac{\gamma}{nh^2} O(h^2) \\ &= \frac{2R(K)}{nh^3} + 3\gamma h^2 \mu_2(K)^2 R(f'') + O(n^{-1}) + o(h^2). \end{aligned}$$

Using Proposition 2 and taking $h = \tilde{h}_\gamma$ we get (recall that $\tilde{h}_\gamma / h_\gamma \to 1$ *a.s.*)

$$\begin{aligned}
&\gamma^{-2/5} n^{3/5} \mathrm{E}(\mathrm{CV}''_\gamma(h)) h_\gamma \\
&= \frac{2R(K)}{(\gamma^{1/5} n^{1/5} h_\gamma)^2} + 3(\gamma^{1/5} n^{1/5} h_\gamma)^3 \mu_2(K)^2 R(f'') + o(1) \quad a.s. \\
&= 5R(K)^{3/5} \mu_2(K)^{4/5} R(f'')^{2/5} + o(1) \quad a.s.
\end{aligned}$$

Thus, (23) follows if we can prove that

$$\sup_{c_1 n^{-1/5} \le h \le c_2 n^{-1/5}} |\mathrm{CV}''_\gamma(h) - \mathrm{E}(\mathrm{CV}''_\gamma(h))| = o(n^{-2/5}) \quad a.s. \tag{24}$$

for all $c_1, c_2 > 0$. For that, let us write

$$\begin{aligned}
&\mathrm{CV}''_\gamma(h) - \mathrm{E}(\mathrm{CV}''_\gamma(h)) \\
&= \frac{2}{nh^2(n-1)} \sum_{1 \le i < j \le n} \{\ddot{M}_h(X_i - X_j) - \mathrm{E}(\ddot{M}_h(X_i - X_j))\} \\
&\quad - \frac{2}{n^2 h^2 (n-1)} \sum_{1 \le i < j \le n} \{\ddot{N}_h(X_i - X_j) - \mathrm{E}(\ddot{N}_h(X_i - X_j))\} \\
&= \frac{1}{h^2} U_{\ddot{M}^s}(h) - \frac{1}{nh^2} U_{\ddot{N}}(h),
\end{aligned}$$

where $\ddot{M}^s$ and $\ddot{N}$ are symmetric and square integrable functions of bounded variation on $\mathbb{R}$ with $|\mu_4|(\ddot{M}^s) < \infty$ and $|\mu_2|(\ddot{N}) < \infty$, with $\mu_j(\ddot{M}^s) = 0$ for $j = 0, 1, 2, 3$, and $\mu_j(\ddot{N}) = 0$ for $j = 0, 1$. Hence, (24) follows from Lemma 3 by taking $\beta = 2$, $\varphi = \ddot{M}^s$ and $\varphi = \ddot{N}$. ■

Proof of Theorem 3: Taking into account the convergence $\hat{\sigma}^2_{\mathrm{CV}} \to \sigma^2_{\mathrm{CV}}(f; K)$ *a.s.*, we first notice that

$$n^{1/5}\big(\hat{\gamma} - 1\big) \xrightarrow{a.s.} -\frac{35}{2}\, \sigma^2_{\mathrm{CV}}(f; K). \tag{25}$$

As in the proof of Theorem 1, in view of (17) and (25) we can prove that

$$\sup_{h>0} \left| \frac{\mathrm{CV}_{\hat{\gamma}}(h) + \hat{\gamma} R(f) - \hat{\gamma} Z_n}{\hat{\gamma}\mathrm{MISE}(h)} - 1 \right| \xrightarrow{a.s.} 0.$$

This implies the convergence

$$\frac{\mathrm{MISE}(\hat{h}_{\hat{\gamma}})}{\mathrm{MISE}(h_{\mathrm{MISE}})} \xrightarrow{as} 1,$$

from which we deduce the almost sure asymptotic equivalence between $\hat{h}_{\mathrm{WCV}} = \hat{h}_{\hat{\gamma}}$ and h_{MISE}, which concludes the proof of the first part of Theorem 3.

Reasoning as in the proof of Theorem 2 we can write

$$\frac{\hat{h}_{\mathrm{WCV}}}{h_{\mathrm{MISE}}} - 1 = -\left(\mathrm{CV}''_{\hat{\gamma}}(\tilde{h})h_{\mathrm{MISE}}\right)^{-1} \mathrm{CV}'_{\hat{\gamma}}(h_{\mathrm{MISE}}), \tag{26}$$

for some random variable $\tilde{h}$ between $\hat{h}_{\mathrm{WCV}}$ and h_{MISE}. Taking into account the equalities

$$\mathrm{CV}'_{\hat{\gamma}}(h_{\mathrm{MISE}}) = -(1-\hat{\gamma})\frac{R(K)}{nh_{\mathrm{MISE}}^2} + \hat{\gamma}\mathrm{CV}'(h_{\mathrm{MISE}})$$

and

$$\mathrm{CV}''_{\hat{\gamma}}(\tilde{h}) = 2(1-\hat{\gamma})\frac{R(K)}{n\tilde{h}^3} + \hat{\gamma}\mathrm{CV}''(\tilde{h}),$$

the asymptotic normality of $\hat{h}_{\mathrm{WCV}}/h_{\mathrm{MISE}} - 1$ follows now from (26), the asymptotic normality

$$n^{7/10}\mathrm{CV}'_{\hat{\gamma}}(h_{\mathrm{MISE}}) \xrightarrow{d} N(0,\sigma^2),$$

that can be derived from Proposition 3, and the almost sure convergence

$$n^{3/5}\mathrm{CV}''_{\hat{\gamma}}(\tilde{h})h_{\mathrm{MISE}} \xrightarrow{a.s.} 5R(K)^{3/5}\mu_2(K)^{4/5}R(f'')^{2/5},$$

which is a consequence of Proposition 5. ■

Proof of Theorem 4: We will show that if f has continuous, bounded and integrable derivatives up to order s ($s \geq 0$ is even) then $\hat{\psi}_s(g_s) \to \psi_s$ *a.s.* whenever the bandwidth g_s takes the form $g_s = g_{s,n}n^{-1/(s+3)}$ with $g_{s,n} \to g_{s,\infty}$ *a.s.*, for some deterministic positive value $g_{s,\infty}$. Two applications of this result for $s = r+2$ and $s = r$, gives the convergence $\tilde{\psi}_r \to \psi_r$ *a.s.* for every density function f with continuous, bounded and integrable derivatives up to order $r+2$.

Taking into account that

$$\mathrm{E}(\hat{\psi}_s(g)) = n^{-1}g^{-s-1}\phi^{(s)}(0) + (1-n^{-1})\int \phi(u)f^{(s)} * \bar{f}(gu)du$$

(see Chacón and Tenreiro, 2012, p. 525, 539), from the continuity of $f^{(s)} * \bar{f}$ and the Lebesgue's dominated convergence theorem we conclude that

$$\mathrm{E}(\hat{\psi}_s(g)) \xrightarrow{as} f^{(s)} * \bar{f}(0) = \psi_r, \text{ for } g = g_s. \tag{27}$$

Moreover, we have

$$\hat{\psi}_s(g) - \mathrm{E}(\hat{\psi}_s(g)) = \frac{1}{g^s}\frac{2}{n^2}\sum_{1\le i<j\le n}\{\varphi_g(X_i - X_j) - \mathrm{E}(\varphi_g(X_i - X_j))\}$$
$$= \frac{1}{g^s}(1 - n^{-1})U_\varphi(g),$$

where $\varphi = \phi^{(s)}$ is a symmetric function of bounded variation on $\mathbb{R}$, with $|\mu_s|(\varphi) < \infty$ and $\mu_j(\varphi) = 0$, for $j = 0, 1, \ldots, s-1$. Thus, from Lemma 3 we get

$$\sup_{c_1 n^{-1/(s+3)} \le g \le c_2 n^{-1/(s+3)}} |\hat{\psi}_s(g) - \mathrm{E}(\hat{\psi}_s(g))| \xrightarrow{a.s.} 0, \text{ for all } c_1, c_2 > 0,$$

from which we deduce that

$$\hat{\psi}_s(g) - \mathrm{E}(\hat{\psi}_s(g)) \xrightarrow{a.s.} 0, \text{ for } g = g_s.$$

Together with (27) this implies that $\hat{\psi}_s(g_s) \to \psi_s$ *a.s.* ■

Acknowledgments

This work was partially supported by the Centre for Mathematics of the University of Coimbra – UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

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.

Cao, R., Cuevas, A., González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Comput. Statist. Data Anal.* 17, 153–176.

Chacón, J.E. (2004). *Estimación de densidades: algunos resultados exactos y asintóticos*. PhD Thesis. Universidad de Extremadura, Spain.

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

Chacón, J.E., Montanero, J., Nogales, A.G. (2008). Bootstrap bandwidth selection using an h-dependent pilot bandwidth. *Scand. J. Statist.* 35, 139–157.

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.

Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statist. Sinica* 6, 129–145.

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.

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

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* 14, 1–16.

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. (1987a). Estimation of integrated squared density derivatives. *Statis. Probab. Lett.* 6, 109–115.

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.

Hall, P., Marron, J.S., Park, B.U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* 92, 1–20.

Hall, P., Robinson, A.P. (2009). Reducing variability of crossvalidation for smoothing-parameter choice. *Biometrika* 96, 175–186.

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

Heidenreich, N.-B., Schindler, A., Sperlich, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Adv. Stat. Anal.* 97, 403–433.

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

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

Liebscher, E. (1998). On a class of plug-in methods of bandwidth selection for kernel density estimators. *Statist. Decisions* 16, 229–243.

Loader, C.R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.*, 27, 415–438.

Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S. (2011). Do-validation for kernel density estimation. *J. Amer. Statist. Assoc.* 106, 651–660.

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

Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S. (2009). One sided crossvalidation for density estimation with an application to operational risk. In *Operational Risk Towards Basel III: Best Practices and Issues in Modelling, Management, and Regulation*, ed. G.N. Gregoriou, 177–196. Wiley, Hoboken.

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

Nolan, D., Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.* 15, 780–799.

Park, B.U., Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* 85, 66–72.

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

Pollard, D. (1986). Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. Unpublished manuscript.

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.

Savchuk, O.Y., Hart, J.D., Sheather, S.J. (2010). Indirect cross-validation for density estimation. *J. Amer. Statist. Assoc.* 105, 415–423.

Scott, D.W., Terrel, G.R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* 82, 1131–1146.

Serfling, R.J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

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.

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* 12, 1285–1297.

Tenreiro, C. (1997). Loi asymptotique des erreurs quadratiques intégrées des estimateurs à noyau de la densité et de la régression sous des conditions de dépendance. *Portugal. Math.* 54, 187–213.

Tenreiro, C. (2003). On the asymptotic normality of multistage integrated density derivatives kernel estimators. *Statist. Probab. Lett.* 64, 311–322.

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

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

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

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