

A nonstandard application of cross-validation to estimate density functionals

José E. Chacón^{*} and Carlos Tenreiro[†]

April 23, 2024

Abstract

Cross-validation is usually employed to evaluate the performance of a given statistical methodology. When such a methodology depends on a number of tuning parameters, cross-validation proves to be helpful to select the parameters that optimize the estimated performance. In this paper, however, a very different and nonstandard use of cross-validation is investigated. Instead of focusing on the cross-validated parameters, the main interest is switched to the estimated value of the error criterion at optimal performance. It is shown that this approach is able to provide consistent and efficient estimates of some density functionals, with the noteworthy feature that these estimates do not rely on the choice of any further tuning parameter, so that, in that sense, they can be considered to be purely empirical. Here, a base case of application of this new paradigm is developed in full detail, while many other possible extensions are hinted as well.

Keywords: cross-validation, density functional, empirical estimation, kernel smoothing, tuning parameter selection

^{*}IMUEx, Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain. E-mail: jechacon@unex.es

[†]CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal. Email: tenreiro@mat.uc.pt

1 Introduction

Cross-validation is a very general and widely used technique in statistics and machine learning (see Arlot and Celisse, 2010). Generally speaking, it is used to evaluate how a statistical procedure, constructed on the basis of some training data, performs on a set of independent test data. Usually, the statistical procedure of interest depends on a number of tuning parameters, and hence cross-validation is typically employed to choose those tuning parameters in order to optimize the estimated performance.

Here, we explore a completely different use of cross-validation, which apparently seems not to have been exploited before. In the following, this new approach is fully developed in the context of a specific problem in nonparametric estimation, though we note that its scope of applicability is much wider, and we also briefly examine other contexts where this principle yields new methodology.

The base case on which we focus is the estimation of the integral of a squared density, since it represents the simplest problem to which our novel methodology can be applied. This functional arises when studying the efficiency of rank-based statistics, and in some other contexts like projection pursuit and symmetry testing or, more recently, in relation to causal estimation (Kennedy, Balakrishnan and Wasserman, 2020). The problem has been studied by many authors, through different methodologies, and it is by now very well understood (see Chacón and Tenreiro, 2012, for a detailed account of contributions on the subject). Yet, it still continues to motivate new research on the topic; see, for instance, the exhaustive investigations of Goldenshluger and Lepski (2022a,b).

Among the existing estimators, a number of them are based on kernel smoothing; see Sheather, Hettmansperger and Donald (1994), and references therein, or Giné and Nickl (2008) and Wu *et al.* (2014), for more recent contributions. Other approaches make use of orthogonal series, as in Laurent (1996), Klemelä (2006) or Tenreiro (2020); or wavelets, as in Kerkycharian and Picard (1996) and Prakasa Rao (1999); or are motivated by model selection ideas, see Laurent and Massart (2000) and Laurent (2005). In addition, estimators based on spacings were studied in Hall (1984), Khasimov (1989), or van Es (1992).

All those previous approaches share one feature in common: they depend on some tuning parameter that needs to be chosen (a bandwidth, a cutoff, a spacing order, etc). In contrast, the main novelty in our proposal is that it is fully empirical, in the sense that the new estimator is computed solely from the data, with no tuning parameters involved. Despite this simplicity, the method is shown to be efficient, in the sense that it reaches the information bound and the fastest convergence rate for the problem, provided the density is smooth enough.

In Section 2 we develop the new methodology for the base problem, illustrate how the proposed estimator can be obtained via three different motivations, and accurately describe its large-sample behaviour. Section 3 contains a brief simulation study that shows the

exceptional performance of the new proposal in practice. Section 4 deals with some selected extensions to highlight the wide applicability of the introduced principle. More intricate developments of the same idea are shown in Section 5 to yield novel, fully empirical selectors of the smoothing parameters for histogram and kernel density estimation. Finally, Section 6 contains a discussion on the implications of this new paradigm, and suggests further possibilities for future applications.

2 Estimation of the integral of a squared density

Consider independent and identically distributed univariate random variables X_1, \dots, X_n , with a common continuous distribution function F with density f . The goal of this section is to present and study the properties of a new estimator of the statistical functional $\psi = \int_{\mathbb{R}} f(x)^2 dx = \int_{\mathbb{R}} f(x) dF(x)$. As noted in the Introduction, this problem can be considered as the base case on which we will develop our novel methodology in full detail, but we will extend it in many other directions afterwards.

Following the terminology in Jones and Sheather (1991), from the above integral expressions of ψ two different but related ‘diagonals-in’ kernel estimators can be proposed: first, $\tilde{\psi}_D^*(g) = \int_{\mathbb{R}} \hat{f}(x; g)^2 dx$, but also $\tilde{\psi}_D(g) = \int_{\mathbb{R}} \hat{f}(x; g) dF_n(x) = n^{-1} \sum_{i=1}^n \hat{f}(X_i; g)$. Here, F_n denotes the empirical distribution function and $\hat{f}(x; g) = n^{-1} \sum_{j=1}^n L_g(x - X_j)$ stands for the kernel density estimator, with kernel L (a symmetric integrable function with unit integral) and bandwidth $g > 0$, and $L_g(x) = L(x/g)/g$ is the scaled kernel. More explicitly,

$$\tilde{\psi}_D^*(g) = n^{-2} \sum_{i,j=1}^n (L * L)_g(X_i - X_j) \quad \text{and} \quad \tilde{\psi}_D(g) = n^{-2} \sum_{i,j=1}^n L_g(X_i - X_j),$$

where $*$ is the convolution product. This shows that both estimators are in fact of the same type, with the only difference that $\tilde{\psi}_D^*(g)$ employs the convolved kernel $L * L$ while $\tilde{\psi}_D(g)$ simply makes use of L . In addition, previously Hall and Marron (1987) noted that $\tilde{\psi}_D(g)$ contains the non-stochastic term $n^{-1}L_g(0)$ and suggested to remove it and consider instead the closely-related ‘no-diagonals’ estimator $\tilde{\psi}_{ND}(g) = \{n(n-1)\}^{-1} \sum_{i \neq j} L_g(X_i - X_j)$. Both estimators can be shown to be consistent if $g \equiv g_n$ satisfies $g \rightarrow 0$ and $ng \rightarrow \infty$ as $n \rightarrow \infty$ (Bhattacharya and Roussas, 1969; Giné and Nickl, 2008; Chacón and Tenreiro, 2012).

Nevertheless, the choice of g in practice can be problematic: it is known that the optimal bandwidth depends on the integral of squared density derivatives of higher order (Wand and Jones, 1995, Section 3.5), so this raises the problem of estimating these higher order functionals. The options here are to use yet another kernel estimation stage to estimate those functionals, with some pilot bandwidth (and face again some further bandwidth selection problem), or, more simply, to adjust the smoothing level in order to be optimal for some location-scale family based on a reference distribution. In fact, in order to avoid an infinitely cyclic reasoning, the reference distribution approach must be employed at some

initial step to start this process of multiple kernel stages. However, as noted in Jones, Marron and Sheather (1996), if a simple unimodal distribution is taken as a reference, then any bandwidth selector relying on such an initial stage will not pass a ‘bimodality test’, in the sense that its performance could be made arbitrarily bad by sampling from a bimodal distribution with far enough modes.

Here we present a new estimator of ψ which, despite being closely connected with the former kernel methods, does not require the specification of a bandwidth or any other tuning parameter. Thus, it allows us to overcome the bandwidth selection difficulties.

The rationale for our novel proposal is the following: recall, on the other hand, that to measure the performance of $\hat{f}(\cdot; g)$ as an estimator of f it is common to use the mean integrated squared error $\text{MISE}(g) = \mathbb{E}\{\text{ISE}(g)\}$, where $\text{ISE}(g) = \int_{\mathbb{R}} \{\hat{f}(x; g) - f(x)\}^2 dx$. An unbiased estimator of $\text{MISE}(g) - \psi$ is given the cross-validation criterion (Rudemo, 1982; Bowman, 1984), which can be explicitly written as

$$\text{CV}(g) = (ng)^{-1}R(L) + \{n(n-1)\}^{-1} \sum_{i \neq j} \{(1-n^{-1})(L * L)_g - 2L_g\}(X_i - X_j),$$

where $R(\alpha) = \int_{\mathbb{R}} \alpha(x)^2 dx$ for any square integrable function α . Since $\text{CV}(g)$ is unbiased for $\text{MISE}(g) - \psi$ for each $g > 0$ and $\min_{g>0} \text{MISE}(g) \rightarrow 0$ as $n \rightarrow \infty$ (Chacón *et al.*, 2007), it is reasonable to expect that $\min_{g>0} \text{CV}(g) \rightarrow -\psi$ in probability. This suggests considering the estimator of ψ defined as

$$\hat{\psi} = -\min_{g>0} \text{CV}(g).$$

Such a minimum is actually attained if $R(L) < 2L(0)$ (Stone, 1984), so all the kernels considered henceforth are assumed to satisfy this condition.

A few remarks about this new estimator are in order: first, note that $\hat{\psi}$ is fully empirical, in the sense that it does not require the specification of any further tuning parameter. Second, observe that $\hat{\psi}$ makes quite an unconventional use of cross-validation: whereas its classical application focuses on its minimizer $\hat{g}_{\text{CV}} = \text{argmin}_{g>0} \text{CV}(g)$, which is seen as a data-based bandwidth that seeks to mimic the behaviour of $g_{\text{MISE}} = \text{argmin}_{g>0} \text{MISE}(g)$, here the main interest lies on the minimum value that the cross-validation criterion attains, as an estimator of $-\psi$.

Moreover, an additional independent motivation for $\hat{\psi}$ can be given. Notice that the cross-validation criterion can be written in terms of the previously introduced kernel estimators of ψ , namely $\text{CV}(g) = \tilde{\psi}_{\text{D}}^*(g) - 2\tilde{\psi}_{\text{ND}}(g)$, so that

$$\hat{\psi} = -\min_{g>0} \{\tilde{\psi}_{\text{D}}^*(g) - 2\tilde{\psi}_{\text{ND}}(g)\} = \max_{g>0} \{2\tilde{\psi}_{\text{ND}}(g) - \tilde{\psi}_{\text{D}}^*(g)\}. \quad (1)$$

Since $\tilde{\psi}_{\text{D}}(g)$ and $\tilde{\psi}_{\text{ND}}^*(g)$ are consistent estimators of ψ as long as $g \rightarrow 0$ and $ng \rightarrow \infty$, then the same applies to the estimator $\tilde{\psi}(g) = 2\tilde{\psi}_{\text{ND}}(g) - \tilde{\psi}_{\text{D}}^*(g)$. But $\hat{g}_{\text{CV}}/g_{\text{MISE}} \rightarrow 1$ in probability (Hall, 1983), with $g_{\text{MISE}} \rightarrow 0$ and $ng_{\text{MISE}} \rightarrow 0$ under fairly general assumptions

(Chacón *et al.*, 2007), which means that we can expect $\hat{\psi} = -\text{CV}(\hat{g}_{\text{CV}})$ to be a consistent estimator of ψ .

A further third motivation for $\hat{\psi}$ stems from the fact that we can also express

$$\hat{\psi} = \max_{g>0} \{\tilde{\psi}_{\text{ND}}^{\bullet}(g) - W(g)\}, \quad (2)$$

where $\tilde{\psi}_{\text{ND}}^{\bullet}(g) = \{n(n-1)\}^{-1} \sum_{i \neq j} (2L - L * L)_g(X_i - X_j)$ is of the same type as $\tilde{\psi}_{\text{ND}}(g)$, but based on the “twicing” kernel $2L - L * L$. Here, $W(g) = \tilde{\psi}_{\text{ND}}^{\bullet}(g) - 2\tilde{\psi}_{\text{ND}}(g) + \tilde{\psi}_{\text{D}}^*(g) = (ng)^{-1}R(L) + O_P(n^{-1})$ is such that $0 \leq W(g) \leq (ng)^{-1}R(L)$ if the kernel L is positive (Devroye, 1989, Lemma 3), so it can be understood as a penalization term, and this unveils $\hat{\psi}$ as closely connected to the estimators based on model selection (see Laurent and Massart, 2000; Laurent, 2005).

Thus, $\hat{\psi}$ is indeed based on kernel smoothing, but its most remarkable and striking feature is that its bandwidth is implicitly chosen, no multistage process or reference distribution are required. It remains intriguing, however, why the maximum of $\tilde{\psi}(g)$ over the possible bandwidths should result in a good choice. The key to understand this feature lies in the behaviour of $\tilde{\psi}(g)$ as an estimator of ψ , which is analyzed in detail in Appendix A of the Supplementary Material. The crucial observation is that minimizing the mean squared error of $\tilde{\psi}(g)$ is asymptotically equivalent to minimizing its squared bias; this is an attribute that it shares with estimators of type $\tilde{\psi}_{\text{D}}(g)$ (Jones and Sheather, 1991; Chacón and Tenreiro, 2012). Moreover, it is clear that the bias of $\tilde{\psi}(g) = -\text{CV}(g)$ equals $-\text{MISE}(g) < 0$; hence the bandwidth g that minimizes the squared bias is precisely g_{MISE} , which is in turn estimated by the cross validation bandwidth \hat{g}_{CV} that is implicitly used in $\hat{\psi}$.

However, it is a common conclusion in comparative bandwidth selection studies that the cross-validation bandwidth is unacceptably unstable for its large variability (see, e.g., Cao, Cuevas and González-Manteiga, 1994), therefore, the fact that $\hat{\psi}$ implicitly depends on \hat{g}_{CV} may suggest that this could be the case for this estimator too. Notwithstanding, the next result shows yet another unforeseen and remarkable property of this estimator: it is consistent under minimal assumptions and, for a smooth enough density, it is asymptotically efficient since it is \sqrt{n} -consistent and attains the information bound for this problem.

The smoothness of f will be defined in terms of Sobolev spaces. Denote by $\varphi_f(t) = \int_{\mathbb{R}} e^{itx} f(x) dx$ the characteristic function associated to f and define the Sobolev class of densities of order $\beta > 0$ as $\mathcal{H}_{\beta} = \{\text{densities } f \text{ such that } \int_{\mathbb{R}} |t|^{2\beta} |\varphi_f(t)|^2 dt < \infty\}$. The asymptotic behaviour of the new estimator $\hat{\psi}$ is described in the next result.

Theorem 1. *If the density f is bounded and L is a symmetric density of bounded variation, continuous at zero and such that $R(L) < 2L(0)$, then*

$$\left(\hat{\psi} - \psi - 2n^{-1} \sum_{i=1}^n Y_i \right) / \text{MISE}(g_{\text{MISE}}) \rightarrow -1 \quad \text{almost surely,}$$

where $Y_i = f(X_i) - \psi$. As a consequence, $\hat{\psi}$ is a strongly consistent estimator of ψ .

In addition, assume that $\int_{\mathbb{R}} x^2 L(x) dx < \infty$ and $f \in \mathcal{H}_\beta$ for some $\beta > 0$. If $\beta > 1/2$, then $\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N(0, 4\tau^2)$, where $\tau^2 = \text{Var}\{f(X_1)\} = \int_{\mathbb{R}} f(x)^3 dx - \psi^2$. If $0 < \beta \leq 1/2$, then $\hat{\psi} - \psi = O_P(n^{-2\beta/(2\beta+1)})$.

Remark 1. The proof of the first part of Theorem 1 is given in Section 7 below. It follows that $\hat{\psi} - \psi = 2n^{-1} \sum_{i=1}^n Y_i + O_P\{\text{MISE}(g_{\text{MISE}})\}$ so, asymptotically, $\hat{\psi} - \psi$ behaves as (twice) an average of centred random variables plus a remainder. To prove the second statement of the theorem, note that the first term is easily described with a central limit theorem, so it suffices to study the rate of convergence of $\text{MISE}(g_{\text{MISE}})$. Proceeding as in the proof of Theorem 1.5 in Tsybakov (2009) it can be shown that $\text{MISE}(g_{\text{MISE}}) = O(n^{-2m/(2m+1)})$, where $m = \min\{\beta, 2\}$. Hence, the remainder is indeed negligible if $\beta > 1/2$, but otherwise becomes the dominant term, in which case it is of order $O(n^{-2\beta/(2\beta+1)})$.

Remark 2. Unfortunately, Theorem 1 shows that the estimator ψ is not rate-adaptive. Bickel and Ritov (1988) showed that ψ can be estimated at rate $n^{-1/2}$ if f is β -Hölder with $\beta > 1/4$ and at rate $n^{-4\beta/(4\beta+1)}$ if $\beta \leq 1/4$; however, in the related Sobolev framework, the estimator $\hat{\psi}$ needs $\beta > 1/2$ to be asymptotically efficient. In any case, this is not a too restrictive smoothness assumption since it holds if f is absolutely continuous and its derivative is square integrable. Furthermore, $\hat{\psi}$ is still \sqrt{n} -consistent (though with an asymptotic variance greater than $4\tau^2$) if $\beta = 1/2$, which covers the case of a density with a finite number of discontinuity points (van Eeden, 1985). Again, the main advantage of $\hat{\psi}$ is that it does not require any other further tuning parameter to be chosen.

Remark 3. As noted above, the motivation of $\hat{\psi}$ in terms of a penalized criterion reveals some interesting connections with the estimation of ψ based on model selection. In that framework, a judicious choice of the penalization allows for the construction of rate-adaptive estimators; therefore, this raises the open question of whether a suitable modification of the penalty term in (2) could yield a rate-adaptive estimator in this context as well.

3 A brief numerical study

To explore the performance of the new estimator $\hat{\psi}$ in practice, a brief simulation study was carried out. Due to their nice convolution properties, the test densities included in this study were the 15 normal mixture densities introduced in Marron and Wand (1992), plus the 10-modal normal mixture described in Loader (1999).

For the sake of brevity, only the behaviour of three estimators of ψ is reported here: the 2-stage solve-the-equation kernel estimator of Sheather, Hettmansperger and Donald (1994), the 2-stage direct-plug-in kernel estimator of Jones and Sheather (1991) and the new method introduced in Section 2. They are labeled as $\hat{\psi}_{\text{SH}}$, $\hat{\psi}_{\text{JS}}$ and $\hat{\psi}_{\text{CT}}$, respectively, in the sequel. In all cases, the standard normal density was used as the kernel. These

three estimators share the feature that the choice of all the tuning parameters required for their implementation (none, in the case of the new proposal) were clearly specified in their respective references.

Another estimator that fulfils that condition is the one proposed in Wu (1995); however, it showed an unsatisfactory performance in some preliminary experiments and was left out of the final comparison. Many other proposals in the literature, even if shown to enjoy nice theoretical properties, depended on some unspecified constants, with no explicit indication on how to choose them in practice, and for this reason they were not included in the study either.

Two sample sizes were explored, $n = 100$ and $n = 1000$. For each of these two scenarios, $B = 500$ samples of size n from each test density were drawn. The performance of a certain estimator of ψ , say $\hat{\psi}_{XY}$, with regard to some test density f , was measured by the relative root mean squared error, $\text{RRMSE} = \{B^{-1} \sum_{b=1}^B (\hat{\psi}_{XY,b} - \psi)^2\}^{1/2} / \psi$, where $\hat{\psi}_{XY,b}$ is the estimate obtained from the b -th synthetic sample.

To compare the performance of the estimators in the study, for each test density we computed the RRMSE of each estimator divided by the minimum RRMSE among the three of them for that density model. Such a ratio represents how bad an estimator is with respect to the best, which corresponds to a ratio value of 1.000. An extended description with more detailed results for each density model and sample size is included in Appendix B of the Supplementary Material. Here, in Table 3, to save space we only include three summary statistics along the 16 test densities: the mean, minimum and maximum RRMSE ratio for each estimator.

	$n = 100$			$n = 1000$		
	CT	SH	JS	CT	SH	JS
Mean	1.095	1.261	1.335	1.044	1.300	2.246
Median	1.080	1.113	1.007	1.019	1.029	1.032
Minimum	1.000	1.000	1.000	1.000	1.000	1.000
Maximum	1.262	3.778	3.828	1.220	5.219	16.672

Table 1: Summary statistics for the RRMSE ratios of the three compared methods, along the 16 test densities. A smaller ratio corresponds to a lower error.

The main conclusion from Table 3 is that the new method is always the best or very close to the best in terms of performance. On average, it is about 10% worst than the best for $n = 100$ and 5% worst than the best for $n = 1000$. The fact that the minimum along the 16 density models is 1.000 for all the methods indicates that each of them is the best for at least one of the considered test densities.

The relatively high maximum ratios observed for the estimators $\hat{\psi}_{SH}$ and $\hat{\psi}_{JS}$ both correspond to the case of Loader’s ten-modal density, for which $\hat{\psi}_{CT}$ performs best. While

$\hat{\psi}_{\text{CT}}$ obtains reasonably good results already for $n = 100$, the other two show a relatively bad performance even for $n = 1000$. It is worth remarking that this is a density model for which the ‘local variability’ in each of the ten components is very poorly modelled by a global, unimodal reference, and this may be the crucial factor hindering the performance of $\hat{\psi}_{\text{SH}}$ and $\hat{\psi}_{\text{JS}}$ (which eventually rely on a normal-reference initial stage).

4 Extensions

A great deal of extensions of the main new idea are possible, which shows its applicability in other contexts. In the following two sections we briefly explore some of these possibilities.

4.1 Directional data

The use of a reference distribution in an initial estimation step is especially problematic for the case of directional data. Attempting to fit circular data using a simple unimodal reference distribution, when they were truly drawn from a density with two antipodal modes, often leads to a fit that is close to the uniform distribution (see Oliveira, Crujeiras and Rodríguez-Casal, 2012) and, therefore, to severely oversmoothed kernel estimates if they are based on the reference distribution approach.

Next we show how our new methodology can be easily adapted to estimate the integral of the squared density of a circular distribution. Let $\Theta_1, \dots, \Theta_n$ be a random sample from a circular density $f: [0, 2\pi) \rightarrow \mathbb{R}$. The goal here is to estimate $\psi = \int_0^{2\pi} f(\theta)^2 d\theta$, and we will rely on the kernel density estimator in this context, given by $\hat{f}(\theta; \nu) = n^{-1} \sum_{i=1}^n K_\nu(\theta - \Theta_i)$, where now $\nu > 0$ is a concentration parameter, which acts as the bandwidth, and K_ν is a circular kernel function. Quite commonly, the von Mises kernel with concentration ν is employed, so that $K_\nu(\theta - \Theta_i) = \exp\{\nu(\theta - \Theta_i)\} / \{2\pi I_0(\nu)\}$, where I_0 denotes the 0-order modified Bessel function of the first kind (see Taylor, 2008).

The cross-validation criterion $\text{CV}(\nu)$ for density estimation with spherical data was introduced in Hall, Watson and Cabrera (1987). There, it was shown that its minimum expected value satisfies $\min_{\nu > 0} \mathbb{E}\{\text{CV}(\nu)\} \rightarrow -\psi$ as $n \rightarrow \infty$. Thus, arguing as in the linear case above, a reasonable estimator of ψ can be defined as $\hat{\psi} = -\min_{\nu > 0} \text{CV}(\nu)$.

The practical performance of this estimator was investigated in Chacón (2017), by comparing it with the usual kernel estimator proposed in Di Marzio, Panzera and Taylor (2011) on the basis of the 20 circular mixture models introduced in Oliveira, Crujeiras and Rodríguez-Casal (2012). The new estimator performed best or second best (and quite close to the best) for all the models. For those distributions with antipodal modes, the classical 2-stage kernel estimator with an initial von Mises reference showed a specially bad performance, as expected due to the aforementioned fitting issue; meanwhile, on the contrary, those models posed no problem at all for the new proposal, which does not rely on any reference distribution.

4.2 Measurement errors

The problem of estimating ψ has also been considered for the case in which the data present measurement errors (Delaigle and Gijbels, 2002). In that context, instead of observing the error-free variables X_1, \dots, X_n , we are given a sample Y_1, \dots, Y_n such that $Y_j = X_j + \varepsilon_j$, where the unobservable errors $\varepsilon_1, \dots, \varepsilon_n$ are independent of X_1, \dots, X_n and are assumed to have a fully-known density f_ε . Hence, the problem is that we are interested in estimating $\psi = \int_{\mathbb{R}} f_X(x)^2 dx$ but we observe a sample with density $f_Y = f_X * f_\varepsilon$, where $f \equiv f_X$ in this section.

The classical approach here is based on deconvolution kernel density estimators: noting that the characteristic functions of Y , X and ε are related through $\varphi_Y = \varphi_X \varphi_\varepsilon$ with fully known φ_ε , then by Fourier inversion f_X is estimated by $\hat{f}_X(x; g) = (ng)^{-1} \sum_{i=1}^n K_\varepsilon(\frac{x-Y_i}{g}; g)$, where $K_\varepsilon(x; g) = (2\pi)^{-1} \int_{\mathbb{R}} \exp(-itx) \varphi_K(t) / \varphi_\varepsilon(t/g) dt$, with $\varphi_K(t) = \int_{\mathbb{R}} \exp(itx) K(x) dx$ standing for the Fourier transform of a kernel function K . The estimator of ψ studied in Delaigle and Gijbels (2002) is then $\tilde{\psi}_D^*(g) = \int_{\mathbb{R}} \hat{f}_X(x; g)^2 dx$, by analogy with the error-free case. A multi-stage procedure with a parametric start is proposed by those authors to select the bandwidth g .

In contrast, Hesse (1999) derived a cross-validation criterion $CV(g)$ for the density estimator $\hat{f}_X(x; g)$ which, as in the error-free case, satisfies $\mathbb{E}[CV(g)] = \text{MISE}(g) - \psi$, where now $\text{MISE}(g) = \mathbb{E} \int_{\mathbb{R}} \{\hat{f}_X(x; g) - f_X(x)\}^2 dx$. Therefore, a bandwidth-free estimator can be analogously defined as $\hat{\psi} = -\min_{g>0} CV(g)$ in this context. In this case, both the theoretical and practical investigation of this estimator are open for future research.

4.3 Entropy estimation

Entropy estimation is a topic of great interest, both for statisticians (see Berrett, Samworth and Yuan, 2019; Han *et al.*, 2020) but also within the Information Theory community and other areas (see the survey of Verdú, 2019). A number of nonparametric entropy estimators and its applications were reviewed in Beirlant *et al.* (1997).

The entropy of a distribution with density f is defined as $H = -\int_{\mathbb{R}} f(x) \log f(x) dx$. This statistical functional presents some similarities with ψ . For instance, it can be alternatively written as $H = -\int_{\mathbb{R}} \log f(x) dF(x)$, and this suggests two possible estimators: first, $\tilde{H}_D^*(g) = -\int_{\mathbb{R}} \hat{f}(x; g) \log \hat{f}(x; g) dx$, but also $\tilde{H}_D(g) = -n^{-1} \sum_{i=1}^n \log \hat{f}(X_i; g)$. The latter is more commonly used in practice, since it avoids numerical integration. Both of them contain some non-stochastic terms, so a third possibility is to use $\tilde{H}_{ND}(g) = -n^{-1} \sum_{i=1}^n \log \hat{f}_i(X_i; g)$, where $\hat{f}_i(x; g) = \{(n-1)h\}^{-1} \sum_{j \neq i} K\{(x - X_j)/h\}$.

Bandwidth selection for these estimators is a difficult and relatively unexplored topic. However, the methodology hereby introduced also circumvents this problem as follows: the last of the previous three estimators satisfies $\tilde{H}_{ND}(g) = -\text{LCV}(g)$, where the latter acronym stands for likelihood cross-validation, a criterion related to measuring the density estima-

tion error through the Kullback-Leibler divergence $\text{KL}(g) = \int_{\mathbb{R}} f(x) \log\{f(x)/\hat{f}(x; g)\} dx$. In fact, Hall (1987) showed that $\mathbb{E}[\tilde{H}_{\text{ND}}(g)] = -\mathbb{E}[\text{LCV}(g)] = \mathbb{E}[\text{KL}(g)] + H$ so, since $0 \leq \min_{g>0} \mathbb{E}[\text{KL}(g)] \rightarrow 0$ under appropriate conditions, it follows that a reasonable, fully empirical entropy estimator is given by

$$\hat{H} = \min_{g>0} \tilde{H}_{\text{ND}}(g) = -n^{-1} \max_{g>0} \sum_{i=1}^n \log \hat{f}_i(X_i; g).$$

Incidentally, this coincides exactly with the bandwidth selection recommendation for $\tilde{H}_{\text{ND}}(g)$ given in Hall and Morton (1993, Section 3), albeit with a slightly different motivation. Thus, the theoretical properties and practical performance of \hat{H} can be consulted in that paper.

Recently, Devroye and Györfi (2022) posed the open problem of finding a fully data-driven entropy estimate that is consistent under the only condition that $H < \infty$. Given that cross-validation techniques are typically consistent under minimal conditions, we believe that \hat{H} is a firm candidate to satisfy the requirements of that open problem.

4.4 Estimation of integrated squared density derivatives

The problem of estimating ψ is a particular instance of the more general problem of estimating the integrated squared r -th density derivative $\theta_r = \int_{\mathbb{R}} f^{(r)}(x)^2 dx$. This functional is of great interest for the problem of automatic smoothing parameter selection for density estimation, especially the cases $r = 1$ for the histogram estimator, and $r = 2$ for the kernel estimator (see Hall and Marron, 1987; Jones and Sheather, 1991; Wand, 1997).

This more general problem is connected to the estimation of the r -th density derivative $f^{(r)}(x)$. The usual estimator is constructed by taking the r -th derivative of the kernel density estimator $\hat{f}^{(r)}(x; g) = (ng^{r+1})^{-1} \sum_{i=1}^n L^{(r)}\{(x - X_i)/g\}$, and its error is globally measured through $\text{MISE}_r(g) = \mathbb{E}\{\text{ISE}_r(g)\}$, where $\text{ISE}_r(g) = \int_{\mathbb{R}} \{\hat{f}^{(r)}(x; g) - f^{(r)}(x)\}^2 dx$. For kernel density derivative estimation, the cross-validation criterion $\text{CV}_r(g)$ was first introduced by Härdle, Marron and Wand (1990). It is shown in Chacón and Duong (2013) that $\mathbb{E}[\text{CV}_r(g)] = \text{MISE}_r(g) - \theta_r$ so, since $\min_{g>0} \text{MISE}_r(g) \rightarrow 0$ (Chacón, Duong and Wand, 2011), again this justifies defining the fully empirical estimator of θ_r as $\hat{\theta}_r = -\min_{g>0} \text{CV}_r(g)$.

The theoretical properties of $\hat{\theta}_r$ could be deduced by following the same steps as for the case $r = 0$ (i.e., the estimation of ψ) in Theorem 1 and Remark 1. The error $\hat{\theta}_r - \theta_r$ can be written as an average of centred random variables plus a remainder of the same order as $\min_{g>0} \text{MISE}_r(g)$. The latter can be shown to be $O(n^{-2m/(2m+2r+1)})$ when $f^{(r)} \in \mathcal{H}_\beta$ with $\beta > r$ and L has a finite second-order moment, where $m = \min\{\beta - r, 2\}$, thus leading to a \sqrt{n} -consistent estimator whenever $\beta \geq 2r + 1/2$, provided $r \in \{0, 1\}$. However, preliminary numerical work seems to suggest that the performance of $\hat{\theta}_r$ might not be as satisfactory for $r \geq 1$ as for $r = 0$, so this remains an avenue for further research.

5 Applications for density estimation

All the previous extensions deal with the problem of estimating a real-valued statistical functional. In a step further, in terms of complexity, this section investigates how to apply the new methodology for the estimation of the error function of density estimators.

5.1 Bandwidth selection for kernel density estimation

Here we show how an application of the same principles introduced in this paper could also be useful to suggest a new bandwidth selector for the classical kernel density estimator, closely related to the smoothed cross validation methodology, but with an automatic, implicit choice of the pilot bandwidth.

Consider now the kernel density estimator $\hat{f}(x; h) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$ with kernel K and bandwidth $h > 0$. As before, let us measure its performance in terms of $\text{MISE}(h) = \mathbb{E} \int_{\mathbb{R}} \{\hat{f}(x; h) - f(x)\}^2 dx$ and denote the optimal level of smoothing in this sense by $h_{\text{MISE}} = \arg\min_{h>0} \text{MISE}(h)$. The goal is to find an estimator of the MISE function, which would suggest an automatic bandwidth choice by minimizing such an error estimate.

First, notice that the MISE can be expressed as

$$\text{MISE}(h) = \{(nh)^{-1}R(K) - n^{-1}R(K_h * f)\} + \{R(K_h * f) - 2R_{K,h}(f) + R(f)\}, \quad (3)$$

where we are denoting $R(\alpha) = \int_{\mathbb{R}} \alpha(x)^2 dx$ for any square integrable function α and $R_{\alpha,h}(f) = \int_{\mathbb{R}} (\alpha_h * f)(x)f(x)dx$ for any integrable function α (Chacón and Duong, 2018). The first term in curly brackets in (3) corresponds to the integrated variance of $\hat{f}(\cdot; h)$, while the second one is the integrated squared bias.

The MISE is usually simplified by keeping only the dominant term of the integrated variance, namely $(nh)^{-1}R(K)$, since this can be shown to have no effect on bandwidth selection (Chacón and Duong, 2011, Theorem 1). This leads to the criterion

$$M(h) = (nh)^{-1}R(K) + \{R(K_h * f) - 2R_{K,h}(f) + R(f)\}, \quad (4)$$

so that the problem of estimating the MISE reduces to estimating the integrated squared bias. In addition, note that $R(K_h * f) = R_{K*K,h}(f)$ for symmetric K . Therefore, it suffices to investigate how to estimate functionals of the type $R_{\alpha,h}(f)$.

It is convenient to introduce here the notation K_0 for the Dirac delta function, as in Jones, Marron and Park (1991). It is not a proper function, but a generalized function (see Gel'fand and Shilov, 1964) which acts within an integral so that $\int_{\mathbb{R}} \alpha(x)K_0(x)dx = \alpha(0)$ for any integrable function α ; in particular, this implies that $K_0 * \alpha = \alpha$. The notation K_0 is well suited because $K_h * \alpha(x) \rightarrow \alpha(x) = K_0 * \alpha(x)$ as $h \rightarrow 0$ if x is a point of continuity of a bounded function α (Wheeden and Zygmund, 2015, Theorem 9.8). With this notation, the error criterion (4) can be simply written as

$$M(h) = (nh)^{-1}R(K) + R_{K*K-2K+K_0,h}(f), \quad (5)$$

so it suffices to estimate $R_{\alpha,h}(f)$ for this particular instance of $\alpha = K * K - 2K + K_0$.

Next, by writing $R_{\alpha,h}(f)$ as $\psi_{\alpha,h} = \int_{\mathbb{R}} (\alpha_h * f)(x) dF(x)$ it is immediate to recognize its similarity with $\psi = \int_{\mathbb{R}} f(x) dF(x)$. The latter is the same as the former, except for the convolution with α_h in the integrand and, besides, for the Dirac delta $\alpha_h = K_0$ we have $\psi_{\alpha,0} = \psi$. Therefore, the previously introduced methodology regarding ψ can be adapted to estimate $\psi_{\alpha,h}$ and, hence, the error criterion $M(h)$.

For instance, the smoothed cross validation criterion (Hall, Marron and Park, 1992) is obtained by estimating $R_{\alpha,h}(f) = \psi_{\alpha,h}$ in (5) by

$$\tilde{\psi}_{\alpha,h}^*(g) = \int_{\mathbb{R}} \alpha_h * \hat{f}(x; g) \hat{f}(x; g) dx = n^{-2} \sum_{i,j=1}^n \alpha_h * (L * L)_g(X_i - X_j), \quad (6)$$

thus leading to $\text{SCV}(h; g) = (nh)^{-1} R(K) + \tilde{\psi}_{K*K-2K+K_0,h}^*(g)$ as an estimator of $M(h)$. A very careful choice of the pilot bandwidth g , depending on h , can make the SCV selector $\hat{h}_{\text{SCV}} \equiv \hat{h}_{\text{SCV}}(g) = \text{argmin}_{h>0} \text{SCV}(h; g)$ attain the fastest possible relative convergence rate towards h_{MISE} (Jones, Marron and Park, 1991).

The choice of such a pilot bandwidth g , though, is a very delicate issue that involves a multi-stage kernel estimation process and, eventually, the use of an initial reference distribution, with all the aforementioned problems associated to this strategy. However, with the new methodology hereby introduced, it is possible to define an estimator of $\psi_{\alpha,h}$ which makes use of an implicitly defined pilot smoothing level, so that it does not need the specification of any further tuning parameters (although it is also based on kernel smoothing). Let us elaborate on this.

Any of the three motivations for the estimator $\hat{\psi}$ of ψ can be adapted for the estimation of $\psi_{\alpha,h}$, but perhaps the most straightforward technique is the one that combines the diagonals-in estimator introduced in (6) with the no-diagonals estimator. Explicitly, the new, bandwidth-free estimator of $\psi_{\alpha,h}$ is defined as

$$\hat{\psi}_{\alpha,h} = \max_{g>0} \left\{ \frac{2}{n(n-1)} \sum_{i \neq j} \alpha_h * L_g(X_i - X_j) - \frac{1}{n^2} \sum_{i,j=1}^n \alpha_h * (L * L)_g(X_i - X_j) \right\}.$$

By using this estimator we obtain a fully empirical estimator of the error criterion $M(h)$, namely $\widehat{M}(h) = (nh)^{-1} R(K) + \hat{\psi}_{K*K-2K+K_0,h}$. The corresponding novel bandwidth selector $\hat{h} = \text{argmin}_{h>0} \widehat{M}(h)$ can thus be seen as a variant of the SCV approach, but with the remarkable difference that no pilot bandwidth choice is needed for its implementation.

Not relying on any auxiliary tuning parameter, the new \hat{h} can be easily implemented (once the involved double optimization process is handled carefully), and its practical performance can be inspected in a simulation study. Indeed, some preliminary results were reported in Chacón (2015), showing that \hat{h} effectively reduces the instability of the classical cross-validation selector for simple models (as expected from its pilot pre-smoothing feature), while at the same time it does not exhibit a too large bias in more complex scenarios

(since it does not need an initial stage depending on a reference distribution). Its theoretical analysis, however, is quite more complicated than usual (partly because of the double optimization as well), so for the moment it stands as an open problem.

5.2 Histogram estimates

Another popular density estimator is the histogram. Even if it is by now well-understood that histograms are not as efficient as kernel density estimators, they still stand among the most commonly used density estimators, due to their simplicity. A detailed study of histograms as density estimators is contained in Chapter 3 of Scott (2015).

To construct a histogram the first step is to divide the real line into bins $\{B_k\}_{k \in \mathbb{Z}}$. To fix ideas we will simply consider the anchor point to be the origin, and all bins to have the same binwidth b , so that $B_k = [kb, (k+1)b)$. Then, the histogram density estimator is defined as $\tilde{f}(x; b) = (nb)^{-1} \sum_{k \in \mathbb{Z}} \nu_k I_{B_k}(x)$, where I_A denotes the indicator function of a set A and $\nu_k \equiv \nu_k(b)$ is the number of sample points falling within B_k .

The MISE of the histogram estimate can be exactly written as

$$\text{MISE}(b) = (nb)^{-1} - (nb)^{-1}(n+1) \sum_{k \in \mathbb{Z}} p_k^2 + R(f), \quad (7)$$

where $p_k \equiv p_k(b) = \int_{B_k} f(x)dx$. On the other hand, its shifted version $\text{MISE}(b) - R(f)$ is unbiasedly estimated by $\text{CV}(b) = 2/\{(n-1)b\} - (n+1)/\{n^2(n-1)b\} \sum_{k \in \mathbb{Z}} \nu_k^2$ (Scott, 2015). So again, as in Section 2, a sensible, fully empirical estimate of $R(f)$ can be proposed as $\check{\psi} = -\min_{b>0} \text{CV}(b)$.

To highlight the similarities of the histogram-type estimate $\check{\psi}$ and the kernel-based estimate $\hat{\psi}$, it is convenient to note that

$$\sum_{k \in \mathbb{Z}} \nu_k^2 = \sum_{i,j=1}^n \left\{ \sum_{k \in \mathbb{Z}} I_{B_k}(X_i) I_{B_k}(X_j) \right\} = \sum_{i \neq j} \left\{ \sum_{k \in \mathbb{Z}} I_{B_k}(X_i) I_{B_k}(X_j) \right\} + n.$$

The former shows how the V -statistic $V(b) = (n^2b)^{-1} \sum_{k \in \mathbb{Z}} \nu_k^2$ can be expressed as $U(b) + (nb)^{-1}$, where $U(b) = (n^2b)^{-1} \sum_{i \neq j} \left\{ \sum_{k \in \mathbb{Z}} I_{B_k}(X_i) I_{B_k}(X_j) \right\}$ is the corresponding U -statistic. Hence, modulo $n-1 \approx n \approx n+1$ we can write

$$\check{\psi} \approx \max_{b>0} \{V(b) - 2/(nb)\} = \max_{b>0} \{2U(b) - V(b)\},$$

which is the equivalent to (1) for the histogram estimate.

The cross-validation criterion here is derived by replacing the unknown term $b^{-1} \sum_{k \in \mathbb{Z}} p_k^2$ in the MISE expression with the empirical estimate $U(b)$. Following the analogy with the kernel estimator, a smooth cross-validation criterion is obtained when $b^{-1} \sum_{k \in \mathbb{Z}} p_k^2$ is estimated thorough $V_b(c) = b^{-1} \sum_{k \in \mathbb{Z}} \{\int_{B_k} \tilde{f}(x; c) dx\}^2$, where $\tilde{f}(x; c)$ is the histogram density estimator with pilot binwidth c . More explicitly, if $\tilde{f}(x; c)$ is based on the partition $\{C_\ell\}_{\ell \in \mathbb{Z}}$

with $C_\ell = [\ell c, (\ell + 1)c)$, then $\int_{B_k} \tilde{f}(x; c) dx = (nc)^{-1} \sum_{i=1}^n \sum_{\ell \in \mathbb{Z}} I_{C_\ell}(X_i) \lambda(B_k \cap C_\ell)$ where λ denotes the Lebesgue measure in \mathbb{R} ; therefore,

$$V_b(c) = b^{-1} (nc)^{-2} \sum_{i,j=1}^n \sum_{k \in \mathbb{Z}} \sum_{\ell, \ell' \in \mathbb{Z}} I_{C_\ell}(X_i) I_{C_{\ell'}}(X_j) \lambda(B_k \cap C_\ell) \lambda(B_k \cap C_{\ell'}). \quad (8)$$

The main difficulty here lies on the choice of the pilot binwidth c . However, constructing the corresponding U -statistic $U_b(c)$ as in (8), but with the first double sum restricted to $i \neq j$, and reasoning analogously as for the kernel estimator, a sensible estimate of the unknown $b^{-1} \sum_{k \in \mathbb{Z}} p_k^2$ is given by $\max_{c>0} \{2U_b(c) - V_b(c)\}$. Replacing this estimate in (7) gives a smooth cross validation criterion for the histogram, which automatically (internally) adjusts the pilot binwidth, and hence leads to a new, fully empirical, SCV-type binwidth selector for histogram density estimation.

6 Discussion

When a certain methodology depends on some tuning parameter, and cross-validation is employed to estimate its performance, its most common utility lies on using its minimiser as a data-driven choice for the tuning parameter. Here, a very different feature is explored, namely how the corresponding optimal cross-validated performance defines an estimator of a certain statistical functional.

This approach is fully developed for the base case of cross-validation for kernel density estimation, where a new estimator of the integrated squared density is found, with the remarkable property that it does not rely on the choice of any further tuning parameter. The whole analysis is conducted in the univariate setting, but it could be extended in a straightforward manner to the multivariate case, where the standard kernel estimate of this functional does depend on the choice of a smoothing parameter (see Chacón and Duong, 2010, Section 3.2).

In addition, here the focus is on independent data, but it must be noted that cross-validation techniques for density estimation with dependent data are also available (see Hart and Vieu, 1990), so an analogue procedure for the estimation of the integrated squared density with dependent data could be derived in a similar fashion.

Many other applications are outlined in Sections 4 and 5, which shows the potential of the novel approach. Those related to smoothing parameter selection for density estimation surely deserve further attention, from both theoretical and practical perspectives. Particularly, it appears that the bandwidth selector introduced in Section 5.1 might have some connection to the methodology developed by Goldenshluger and Lepski (2011), albeit with a completely different motivation. Studying its asymptotic and finite-sample properties certainly constitutes a future research challenge.

7 Proof of Theorem 1

The proof of Theorem 1 makes extensive use of the results in Nolan and Pollard (1987), which show that several random criteria $C(g)$, say, are almost surely uniformly equivalent to $\text{MISE}(g)$. Precisely, this means that $\sup_{g>0} |C(g)/\text{MISE}(g) - 1| \rightarrow 0$ almost surely.

To begin with, under the conditions of Theorem 1, Equation (13) in Nolan and Pollard (1987) assures that

$$\sup_{g>0} \left| \frac{\text{ISE}(g) - \text{CV}(g) - \psi - Z_n}{\text{MISE}(g)} \right| \rightarrow 0 \quad \text{almost surely,}$$

where $Z_n = 2n^{-1} \sum_{i=1}^n Y_i$. Moreover, Equation (12) in the same reference states that $\sup_{g>0} |\text{ISE}(g)/\text{MISE}(g) - 1| \rightarrow 0$ almost surely, so that it is possible to replace $\text{ISE}(g)$ with $\text{MISE}(g)$ in the previous display. Hence, the random variable

$$A_n(g) = 1 - \frac{\text{MISE}(g)}{\text{CV}(g) + \psi + Z_n}$$

satisfies $\sup_{g>0} |A_n(g)| \rightarrow 0$ almost surely. This further implies that with probability one $\text{CV}(g) + \psi + Z_n > 0$ for all $g > 0$ for large enough n , and since

$$\begin{aligned} \{1 - A_n(g_{\text{MISE}})\} \{\text{CV}(g_{\text{MISE}}) + \psi + Z_n\} &= \text{MISE}(g_{\text{MISE}}) \leq \text{MISE}(\hat{g}_{\text{CV}}) \\ &= \{1 - A_n(\hat{g}_{\text{CV}})\} \{\text{CV}(\hat{g}_{\text{CV}}) + \psi + Z_n\}, \end{aligned}$$

it follows that with probability one, for large enough n ,

$$\{1 - A_n(g_{\text{MISE}})\} \frac{\text{CV}(g_{\text{MISE}}) + \psi + Z_n}{\text{CV}(\hat{g}_{\text{CV}}) + \psi + Z_n} \leq 1 - A_n(\hat{g}_{\text{CV}}).$$

Therefore,

$$\frac{\text{CV}(g_{\text{MISE}}) + \psi + Z_n}{\text{CV}(\hat{g}_{\text{CV}}) + \psi + Z_n} \rightarrow 1 \quad \text{almost surely}$$

and, from the equivalence with the MISE, it also holds that

$$\frac{\text{MISE}(g_{\text{MISE}})}{\text{CV}(\hat{g}_{\text{CV}}) + \psi + Z_n} \rightarrow 1 \quad \text{almost surely,}$$

which finishes the proof of Theorem 1.

Funding. J. E. Chacón has been partially supported by Spanish Ministerio de Ciencia e Innovación grants PID2019-109387GB-I00 and PID2021-124051NB-I00, C. Tenreiro has been partially supported by the Centre for Mathematics of the University of Coimbra (funded by the Portuguese Government through FCT/MCTES, doi: 10.54499/UIDB/00324/2020).

Code. An R script implementing the estimators introduced in Sections 2 and 5.1 is available at the webpage of the first author (matematicas.unex.es/~jechacon).

References

- Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40–79.
- Beirlant, J., Dudewicz, E. J., Györfi, L. and van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, **6**, 17–39.
- Berrett, T. B., Samworth, R. J. and Yuan, M. (2019). Efficient multivariate entropy estimation via k -nearest neighbour distances. *Annals of Statistics*, **47**, 288–318.
- Bhattacharyya, G. K. and Roussas, G. G. (1969). Estimation of a certain functional of a probability density function. *Scandinavian Actuarial Journal*, **1969**, 201–206.
- Bickel, J. P. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā, Series A*, **50**, 381–393.
- Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.
- Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, **17**, 153–176.
- Chacón, J. E. (2015). Pilot pre-smoothing without reference distributions. In *Programme and Abstracts of the 8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics*, 118.
- Chacón, J. E. (2017). Estimation of the integrated squared density of a circular random variable. In *Programme and Book of Abstracts of the ADISTA17 International Directional Statistics Workshop*, 38.
- Chacón, J. E. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, 375–398.
- Chacón, J. E. and Duong, T. (2011). Unconstrained pilot selectors for smoothed cross-validation. *Australian and New Zealand Journal of Statistics*, **53**, 331–351.
- Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, **7**, 499–532.
- Chacón, J. E. and Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*, Chapman and Hall/CRC, Boca Raton.

- Chacón, J. E., Duong, T. and Wand, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, **21**, 807–840.
- Chacón, J. E., Montanero, J., Nogales, A. G. and Pérez, P. (2007). On the existence and limit behavior of the optimal bandwidth for kernel density estimation. *Statistica Sinica*, **17**, 289–300.
- Chacón, J. E. and Tenreiro, C. (2012). Exact and asymptotically optimal bandwidths for kernel estimation of density functionals. *Methodology and Computing in Applied Probability*, **14**, 523–548.
- Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society, Series B*, **64**, 869–886.
- Devroye, L. (1989). On the non-consistency of the L_2 -cross-validated kernel density estimate. *Statistics & Probability Letters*, **8**, 425–433.
- Devroye, L. and Györfi, L. (2022). On the consistency of the Kozachenko-Leonenko entropy estimate. *IEEE Transactions on Information Theory*, **68**, 1178–1185.
- Di Marzio, M., Panzera, A. and Taylor, C. C. (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, **141**, 2156–2173.
- Gel’fand, I. M. and Shilov, G. E. (1964). *Generalized Functions. Volume 1: Properties and Operations*. Academic Press, New York.
- Giné, E. and Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli*, **14**, 47–61.
- Goldenshluger, A. and Lepski, O. V. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Annals of Statistics*, **39**, 1608–1632.
- Goldenshluger, A. and Lepski, O. V. (2022a). Minimax estimation of norms of a probability density: I. Lower bounds. *Bernoulli*, **28**, 1120–1154.
- Goldenshluger, A. and Lepski, O. V. (2022b). Minimax estimation of norms of a probability density: II. Rate-optimal estimation procedures *Bernoulli*, **28**, 1155–1178.
- Kennedy, E. H., Balakrishnan, S. and Wasserman, L. (2020). Discussion of “On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by Machine Learning”. *Statistical Science*, **35**, 540–544.
- Khasimov, Sh. A. (1989). Asymptotic properties of functions of spacings. *Theory of Probability and its Applications*, **34**, 298–307.

- Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, **11**, 1156–1174.
- Hall, P. (1984). Limit theorems for sums of general functions of m -spacings. *Mathematical Proceedings of the Cambridge Philosophical Society*, **96**, 517–532.
- Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Annals of Statistics*, **15**, 1491–1519.
- Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, **6**, 109–115.
- Hall, P., Marron, J. S. and Park, B. U. (1992) Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1–20.
- Hall, P. and Morton, S. C. (1993). On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, **45**, 69–88.
- Hall, P., Watson, G. P. and Cabrera, J. (1987). Kernel density estimation for spherical data. *Biometrika*, **74**, 751–762.
- Han, Y., Jiao, J., Weissman, T. and Wu, Y. (2020). Optimal rates of entropy estimation over Lipschitz balls. *Annals of Statistics*, **48**, 3228–3250.
- Härdle, W., Marron, J. S. and Wand, M. P. (1990). Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society, Series B*, **52**, 223–232.
- Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, **18**, 873–890.
- Hesse, C. H. (1999). Data-driven deconvolution. *Journal of Nonparametric Statistics*, **10**, 343–373.
- Jones, M. C., Marron, J. S. and Park, B. U. (1991). A simple root n bandwidth selector. *Annals of Statistics*, **19**, 1919–1932.
- Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, **11**, 337–381.
- Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics & Probability Letters*, **11**, 511–514.
- Kerkycharian, G. and Picard, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *Annals of Statistics*, **24**, 485–507.

- Klemelä, J. (2006). Sharp adaptive estimation of quadratic functionals. *Probability Theory and Related Fields*, **134**, 539–564.
- Laurent, B. (1996). Efficient estimation of integral functionals of a density. *Annals of Statistics*, **24**, 659–681.
- Laurent, B. (2005). Adaptive estimation of a quadratic functional of a density by model selection. *ESAIM: Probability and Statistics*, **9**, 1–18.
- Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, **28**, 1302–1338.
- Loader, C. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics*, **27**, 415–438.
- Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics*, **20**, 712–736.
- Nolan, D. and Pollard, D. (1987). *U*-processes: rates of convergence. *Annals of Statistics*, **15**, 780–799.
- Oliveira, M., Crujeiras, R. M. and Rodríguez-Casal, A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis*, **56**, 3898–3908.
- Prakasa Rao, B. L. S. (1999). Estimation of the integrated squared density derivatives by wavelets. *Bulletin of Informatics and Cybernetics*, **31**, 47–65.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.
- Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice and Visualization*, 2nd Edition. Wiley, Hoboken.
- Sheather, S. J., Hettmansperger, T. P. and Donald, M. R. (1994). Data-based bandwidth selection for kernel estimators of the integral of $f^2(x)$. *Scandinavian Journal of Statistics*, **21**, 265–275.
- Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, 1285–1297.
- Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493–3500.

- Tenreiro, C. (2020). Bandwidth selection for kernel density estimation: a Hermite series-based direct plug-in approach. *Journal of Statistical Computation and Simulation*, **90**, 3433–3453.
- Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.
- van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Annals of the Institute of Statistical Mathematics*, **37**, 461–572.
- van Es, B. (1992). Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, **19**, 61–72.
- Verdú, S. (2019). Empirical estimation of information measures: A literature guide. *Entropy*, **21**, 720.
- Wand, M. P. (1997). Data-based choice of histogram bin width. *The American Statistician*, **51**, 59–64.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.
- Wheeden, R. L. and Zygmund, A. (2015). *Measure and Integral: An Introduction to Real Analysis*, 2nd Edition. Chapman & Hall, Boca Raton.
- Wu, T.-J. (1995). Adaptive root n estimates of integrated squared density derivatives. *Annals of Statistics*, **23**, 1474–1495.
- Wu, T.-J., Hsu, C.-Y., Chen, H.-Y. and Yu, H.-C. (2014). Root n estimates of vectors of integrated density partial derivative functionals. *Annals of the Institute of Statistical Mathematics*, **66**, 865–895.

Supplementary material to “A nonstandard application of cross-validation to estimate density functionals”

José E. Chacón¹ and Carlos Tenreiro²

Abstract

This document contains two appendices as supplementary material to the manuscript *A nonstandard application of cross-validation to estimate density functionals*. Appendix A provides a detailed theoretical study of one of the auxiliary estimators of the integrated squared density considered in the paper. Appendix B supplies an elaborated analysis of the results of the simulation study contained in the paper.

¹IMUEx, Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain. E-mail: jechacon@unex.es

²CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal. E-mail: tenreiro@mat.uc.pt

A Detailed theoretical analysis of $\check{\psi}(g)$

In this appendix we provide a detailed analysis of the estimator of $\psi = \int_{\mathbb{R}} f(x)^2 dx$ defined by $\check{\psi}(g) = -\text{CV}(g)$, where CV is the cross-validation function associated to the kernel density estimator $\hat{f}(\cdot; g)$ of f . More precisely, we prove that there exists a bandwidth g_{MSE} that minimizes the mean squared error $\text{MSE}(g)$ of $\check{\psi}(g)$ and that it is asymptotically equivalent to the bandwidth g_{MISE} that minimizes the mean integrated squared error of $\hat{f}(\cdot; g)$. These results support the idea of taking g_{MISE} as the target bandwidth for $\check{\psi}(g)$, giving us an additional motivation for estimating ψ by $\hat{\psi} = \check{\psi}(\hat{g}_{\text{CV}})$, where the cross-validation bandwidth $\hat{g}_{\text{CV}} = \text{argmin}_{g>0} \text{CV}(g)$ seeks to mimic the behaviour of g_{MISE} .

A.1 Introduction

Given X_1, \dots, X_n independent real-valued random variables with common probability density function f , we are interested in the estimation of the functional

$$\psi = \int_{\mathbb{R}} f(x)^2 dx.$$

For that we consider the estimator of ψ given by

$$\check{\psi}(g) = -\text{CV}(g),$$

where $g \equiv g_n > 0$ is a sequence of real numbers converging to zero as n tends to infinity, and $\text{CV}(g)$, given by

$$\text{CV}(g) = (ng)^{-1}R(L) + \{n(n-1)\}^{-1} \sum_{i \neq j} \{(1-n^{-1})(L * L)_g - 2L_g\}(X_i - X_j),$$

is the cross-validation criterion function associated to the kernel density estimator

$$\hat{f}(x; g) = n^{-1} \sum_{j=1}^n L_g(x - X_j), \quad (9)$$

where we denote $R(\alpha) = \int_{\mathbb{R}} \alpha(x)^2 dx$ for an arbitrary square integrable real function α , L is a kernel, that is, a bounded, symmetric and integrable function with unit integral, $*$ denotes the convolution product, and $L_g(x) = L(x/g)/g$ is the scaled kernel associated to L .

The rest of this appendix is organized as follows. In Section A.2 we provide mild conditions on the kernel and the density that ensure the existence of a bandwidth g_{MSE} , called exact optimal bandwidth, that minimizes the mean squared error $\text{MSE}(g)$ of $\check{\psi}(g)$. In Section A.3 we study the asymptotic properties of this bandwidth. In Section A.4 we prove that g_{MSE} is asymptotically equivalent to the bandwidth g_{MISE} that minimizes the mean integrated squared error of the kernel density estimator (9), and we obtain the relative rates of convergence of g_{MSE} to g_{MISE} . These results suggest that using a kernel L of order higher than 2 may not be convenient as the quality of the approximation between g_{MSE} and g_{MISE} decreases when the kernel order increases. All the proofs are deferred to Section A.5.

A.2 Existence of an exact optimal bandwidth

For $g > 0$, we have

$$\check{\psi}(g) = -(ng)^{-1}R(L) + U_{M_g} + n^{-1}U_{N_g},$$

where M_g and N_g are the scaled kernels associated to

$$M = 2L - L * L \quad \text{and} \quad N = L * L,$$

respectively, and for a given bounded, symmetric and integrable function φ , U_{φ_g} is the U -statistic

$$U_{\varphi_g} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \varphi_g(X_i - X_j).$$

Taking into account that $\mathbb{E}U_{\varphi_g} = R_{\varphi,g}(f)$, with

$$R_{\varphi,g}(f) = \int_{\mathbb{R}} \varphi_g * f(x) f(x) dx,$$

we deduce that the bias $B(g)$ of $\check{\psi}(g)$ can be written as

$$B(g) = \mathbb{E}\check{\psi}(g) - \psi = -(ng)^{-1}R(L) + R_{M,g}(f) + n^{-1}R_{N,g}(f) - \psi. \quad (10)$$

From this equality and Equation (10) in Chacón et al. (2007b, p. 296) we see that $B(g) = -\text{MISE}(g)$, where $\text{MISE}(g) = \mathbb{E} \int_{\mathbb{R}} \{\hat{f}(x; g) - f(x)\}^2 dx$ is the mean integrated squared error of the kernel density estimator (9).

On the other hand, from the covariance formula between two U -statistics (see Lee, 1990, Theorem 2, p. 17), we deduce that the variance $V(g)$ of $\check{\psi}(g)$ is given by

$$\begin{aligned} V(g) &= \text{Var}(U_{M_g}) + 2n^{-1} \text{Cov}(U_{M_g}, U_{N_g}) + n^{-2} \text{Var}(U_{N_g}) \\ &= 4(n-2)\{n(n-1)\}^{-1} \{T_{M,M,g}(f) + 2n^{-1}T_{M,N,g}(f) + n^{-2}T_{N,N,g}(f)\} \\ &\quad - (4n-6)\{n(n-1)\}^{-1} \{R_{M,g}(f)^2 + 2n^{-1}R_{M,g}(f)R_{N,g}(f) + n^{-2}R_{N,g}(f)^2\} \\ &\quad + 2g^{-1}\{n(n-1)\}^{-1} \{R_{M^2,g}(f) + 2n^{-1}R_{M \cdot N,g}(f) + n^{-2}R_{N^2,g}(f)\}, \end{aligned} \quad (11)$$

with

$$T_{\varphi,\phi,g}(f) = \int_{\mathbb{R}} \varphi_g * f(x) \phi_g * f(x) f(x) dx,$$

where φ and ψ are bounded, symmetric and integrable functions.

Combining equations (10) and (11) we obtain the following exact formula for the mean squared error $\text{MSE}(g)$ of the estimator $\check{\psi}(g)$:

$$\begin{aligned} \text{MSE}(g) &= \{(ng)^{-1}R(L) - R_{M,g}(f) - n^{-1}R_{N,g}(f) + \psi\}^2 \\ &\quad + 4(n-2)\{n(n-1)\}^{-1} \{T_{M,M,g}(f) + 2n^{-1}T_{M,N,g}(f) + n^{-2}T_{N,N,g}(f)\} \\ &\quad - (4n-6)\{n(n-1)\}^{-1} \{R_{M,g}(f)^2 + 2n^{-1}R_{M,g}(f)R_{N,g}(f) + n^{-2}R_{N,g}(f)^2\} \\ &\quad + 2g^{-1}\{n(n-1)\}^{-1} \{R_{M^2,g}(f) + 2n^{-1}R_{M \cdot N,g}(f) + n^{-2}R_{N^2,g}(f)\}. \end{aligned} \quad (12)$$

This exact error formula is the analogue of formula (5) in Chacón and Tenreiro (2012, p. 526) and will be useful to explore the existence and limit behavior of the optimal bandwidth.

In the following results we will make the next assumptions on the kernel L and the density f :

(L1) L is a bounded, symmetric and integrable function with unit integral, which is continuous at zero, with $R(L) < 2L(0)$.

(D1) f is bounded.

The next result shows that under mild conditions there is always an exact optimal bandwidth, that is, a bandwidth which minimizes the exact MSE of estimator $\tilde{\psi}(g)$. It is the analogue of Theorem 1 in Chacón and Tenreiro (2012, p. 526) for the ‘diagonals-in’ kernel estimator $\tilde{\psi}_D(g) = n^{-2} \sum_{i,j=1}^n L_g(X_i - X_j)$.

Theorem 2. *Under assumptions (L1) and (D1), there exists $g_{\text{MSE}} = g_{\text{MSE},n}(f)$ such that $\text{MSE}(g_{\text{MSE}}) \leq \text{MSE}(g)$, for all $g > 0$.*

A.3 Limit behavior of g_{MSE}

From formula (12) and Lemma 1 in Section A.5 below it follows that $\text{MSE}(g) \rightarrow 0$ for any bandwidth sequence $g = g_n$ such that $g \rightarrow 0$ and $ng \rightarrow \infty$, as $n \rightarrow \infty$. Therefore, conditions $g \rightarrow 0$ and $ng \rightarrow \infty$ are sufficient for $\tilde{\psi}(g)$ to be consistent. It is natural, then, to wonder if the bandwidth g_{MSE} also fulfill the previous consistency conditions. We will see that the second condition holds quite generally but the same is not necessarily true for the first one. This is similar to the situation with the exact optimal bandwidth for the ‘diagonals-in’ kernel estimator $\tilde{\psi}_D(g)$, as shown in Chacón and Tenreiro (2012).

Theorem 3. *Under assumptions (L1) and (D1), we have $ng_{\text{MSE}} \rightarrow \infty$, as $n \rightarrow \infty$.*

For the analysis of the limit behavior of the sequence g_{MSE} we use the notation $\varphi_K(t) = \int_{\mathbb{R}} e^{itx} K(x) dx$, $t \in \mathbb{R}$, for the characteristic function of an integrable function K , and for every density f and every kernel L , we denote

$$\begin{aligned} C_f &= \sup\{r \geq 0 : \varphi_f(t) \neq 0 \text{ a.e. for } t \in [0, r]\}, \\ D_f &= \sup\{t \geq 0 : \varphi_f(t) \neq 0\}, \\ S_L &= \inf\{t \geq 0 : \varphi_L(t) \neq 1\} \text{ and} \\ T_L &= \inf\{r \geq 0 : \varphi_L(t) \neq 1 \text{ a.e. for } t \geq r\}. \end{aligned}$$

A detailed discussion about these quantities is presented in Chacón et al. (2007b). In particular, we note that all these exist, with C_f, D_f possibly being infinite, $S_L, T_L \in [0, \infty)$, $C_f \leq D_f$ and $S_L \leq T_L$. By definition, $S_L > 0$ for superkernels and $S_L = 0$ if L is a kernel of finite order ν (even), that is, if $m_j(L) = 0$ for $j = 1, 2, \dots, \nu - 1$ and $m_\nu(L) \neq 0$ with

$|m_\nu|(L) < \infty$, where $m_j(L) = \int_{\mathbb{R}} u^j L(u) du$ and $|m_j|(L) = \int_{\mathbb{R}} |u^j L(u)| du$ (see Chacón et al., 2007a).

An additional assumption on the kernel is needed to show that g_{MSE} converges to zero:

(L2) L is such that $\varphi_L(t) \leq 1$, for all $t \in \mathbb{R}$.

In the following result we show that g_{MSE} converges to zero under very general conditions. In particular, if L is a kernel of finite order, the convergence to zero takes place with no additional conditions on f other than being bounded. The same property occurs in the superkernel case whenever the characteristic function of f has unbounded support.

Theorem 4. *Under assumptions (L1), (L2) and (D1), if $S_L = 0$ or $D_f = \infty$ then $g_{\text{MSE}} \rightarrow 0$, as $n \rightarrow \infty$.*

A.4 The bandwidths g_{MSE} and g_{MISE}

In order to study the order of convergence to zero of the exact optimal bandwidth g_{MSE} , we need some additional assumptions on the kernel L and the density f :

(L3) L is a kernel of finite order ν (even) such that $(-1)^{\nu/2} m_\nu(L) < 0$ and $|m|_{2\nu}(L) < \infty$.

(D2) f has bounded and integrable derivatives up to order ν .

Under conditions (L1)–(L3), (D1) and (D2), from Lemma 2 and the fact that $M = 2L - L * L$ and $N = L * L$ are kernels of orders 2ν and ν , respectively, with $m_{2\nu}(M) = -(2\nu)! m_\nu(L)^2 / (\nu!)^2$ and $m_\nu(N) = 2m_\nu(L)$, we conclude that the bias and variance of $\tilde{\psi}(g)$ given by equations (10) and (11), respectively, admit the asymptotic expansions, as $g \rightarrow 0$,

$$B(g) = -(ng)^{-1} R(L) - g^{2\nu} R(f^{(\nu)}) m_\nu(L)^2 / (\nu!)^2 + O(n^{-1}) + o(g^{2\nu}) \quad (13)$$

and

$$V(g) = 4n^{-1} \text{Var } f(X_1) + O(n^{-1} g^{2\nu} + n^{-2} g^{-1}),$$

from which we get the following asymptotic expansion for the mean square error of $\tilde{\psi}(g)$:

$$\begin{aligned} \text{MSE}(g) &= 4n^{-1} \text{Var } f(X_1) + \left\{ (ng)^{-1} R(L) + g^{2\nu} R(f^{(\nu)}) m_\nu(L)^2 / (\nu!)^2 \right\}^2 \\ &\quad + O(n^{-2} g^{-1}) + o(n^{-1} g^{2\nu-1} + g^{2\nu}). \end{aligned} \quad (14)$$

Based on these asymptotic expansions, in the following result we start by establishing that the exact optimal bandwidth g_{MSE} is of order $n^{-1/(2\nu+1)}$. Therefore, for g of order $n^{-1/(2\nu+1)}$ and the fact that $B(g) = -\text{MISE}(g)$, we get the equality $\text{MSE}(g) = 4n^{-1} \text{Var } f(X_1) + \text{MISE}(g)^2(1 + o(1))$, that suggests that the optimal bandwidth g_{MSE} may be asymptotically equivalent to the bandwidth g_{MISE} that minimizes the mean integrated squared error of the kernel density estimator $\hat{f}(\cdot; g)$ given by (9) (regarding the existence and asymptotic behaviour of g_{MISE} , see Chacón et al., 2007b). This fact, together with the order of convergence of the relative error $g_{\text{MSE}}/g_{\text{MISE}} - 1$, is established in the following result.

Theorem 5. *Under assumptions (L1)–(L3), (D1) and (D2), we have:*

(a) *The bandwidths g_{MSE} and g_{MISE} are of the same order, that is,*

$$0 < \liminf n^{1/(2\nu+1)} g_{\text{MSE}} \leq \limsup n^{1/(2\nu+1)} g_{\text{MSE}} < \infty.$$

(b) *The bandwidths g_{MSE} and g_{MISE} are asymptotically equivalent, that is,*

$$g_{\text{MSE}}/g_{\text{MISE}} \rightarrow 1.$$

(c) *There exists a constant C , depending on L and f , such that*

$$n^{1/(2\nu+1)}(g_{\text{MSE}}/g_{\text{MISE}} - 1) = C(1 + o(1)).$$

This result supports the idea of taking g_{MISE} as the target bandwidth for $\check{\psi}(g)$, giving us an additional motivation for estimating ψ by $\hat{\psi} = \check{\psi}(\hat{g}_{\text{CV}})$. Moreover, it also enables us to recommend the use of a kernel L of second order ($\nu = 2$) because the order of convergence to zero of the relative error $g_{\text{MSE}}/g_{\text{MISE}} - 1$ is a decreasing function of the kernel order.

A.5 Proofs

We start by establishing the continuity and limit behaviour of the functions $g \mapsto R_{\varphi,g}(f)$ and $g \mapsto T_{\varphi,\phi,g}(f)$ defined in Section A.2.

Lemma 1. *Under assumption (D1), assume that φ, ϕ are bounded and integrable functions.*

(a) *The function $g \mapsto R_{\varphi,g}(f)$ is continuous with*

$$\lim_{g \rightarrow 0} R_{\varphi,g}(f) = R(f) \int_{\mathbb{R}} \varphi(u) du.$$

In addition, if φ is continuous at zero, then $\lim_{g \rightarrow \infty} g R_{\varphi,g}(f) = \varphi(0)$.

(b) *The function $g \mapsto T_{\varphi,\phi,g}(f)$ is continuous with*

$$\lim_{g \rightarrow 0} T_{\varphi,\phi,g}(f) = R(f^{3/2}) \int_{\mathbb{R}} \varphi(u) du \int_{\mathbb{R}} \phi(u) du.$$

In addition, if φ and ϕ are continuous at zero, then $\lim_{g \rightarrow \infty} g^2 T_{\varphi,\phi,g}(f) = \varphi(0)\phi(0)$.

Proof: Taking into account that

$$R_{\varphi,g}(f) = \int_{\mathbb{R}} \varphi(u) (\bar{f} * f)(ug) du, \quad (15)$$

where $\bar{f}(u) = f(-u)$, part (a) follows from the dominated convergence theorem and the boundedness and the continuity of the convolution product of square integrable functions (see also Chacón et al., 2007b, Lemma, p. 296). Moreover, we have

$$T_{\varphi,\phi,g}(f) = \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(u) \phi(v) (\bar{f} \odot \bar{f} \odot f)(ug, vg) du dv, \quad (16)$$

where we are denoting

$$(\alpha \odot \beta \odot \gamma)(x, y) = \int_{\mathbb{R}} \alpha(x - z) \beta(y - z) \gamma(z) dz.$$

Therefore, part (b) follows from the dominated convergence theorem and the boundedness and the continuity of $\alpha \odot \beta \odot \gamma$ when α, β and γ are bounded and integrable functions (see Chacón and Tenreiro, 2012, proof of Lemma 1, p. 539). ■

Proof of Theorem 2: From the expression (12) for the MSE function, and the properties of the functions $g \mapsto R_{\varphi, g}(f)$ and $g \mapsto T_{\varphi, \phi, g}(f)$ shown in Lemma 1, we conclude that $g \mapsto \text{MSE}(g)$ is a continuous function such that $\lim_{g \rightarrow 0} \text{MSE}(g) = \infty$ and $\lim_{g \rightarrow \infty} \text{MSE}(g) = \psi^2$. Moreover, from the hypotheses on L we have

$$\lim_{g \rightarrow \infty} g^2 (\text{MSE}(g) - \psi^2) = -\infty,$$

which enables us to conclude that we can choose $g > 0$ big enough so that $\text{MSE}(g) < \psi^2$. This concludes the proof. ■

Proof of Theorem 3: Suppose that ng_{MSE} does not converge to infinity. Then ng_{MSE} has a subsequence which is upper bounded by some positive constant C . Therefore, along that subsequence we have $g_{\text{MSE}} \rightarrow 0$. From (10) this implies that

$$\limsup_{n \rightarrow \infty} \text{MSE}(g_{\text{MSE}}) \geq \limsup_{n \rightarrow \infty} B(g_{\text{MSE}})^2 = \limsup_{n \rightarrow \infty} \{n^{-1} g_{\text{MSE}}^{-1} R(L)\}^2 \geq \{R(L)/C\}^2 > 0,$$

which contradicts the fact that $\lim_{n \rightarrow \infty} \text{MSE}(g_{\text{MSE}}) = 0$, as we can deduce from $0 \leq \text{MSE}(g_{\text{MSE}}) \leq \text{MSE}(g) \rightarrow 0$, for any bandwidth sequence $g = g_n$ such that $g \rightarrow 0$ and $ng \rightarrow \infty$, as $n \rightarrow \infty$. ■

Proof of Theorem 4: Denote by $\Lambda_{f, L}$ the set of accumulation points of the sequence (g_{MSE}) . Take $0 < \lambda \in \Lambda_{f, L}$ and (g_{n_k}) a subsequence of (g_{MSE}) such that $\lambda = \lim_{k \rightarrow \infty} g_{n_k}$. Writing $B(g; n)$ and $\text{MSE}(g; n)$ for $B(g)$ and $\text{MSE}(g)$, respectively, from equalities (10) and (11) we get that, for fixed $g > 0$,

$$\lim_{n \rightarrow \infty} \text{MSE}(g; n) = \lim_{n \rightarrow \infty} B(g; n)^2 = \{R_{M, g}(f) - \psi\}^2,$$

so that using Lemma 1 and Theorem 3, we obtain

$$\begin{aligned} 0 &= \lim_{g \rightarrow 0} \{R_{M, g}(f) - \psi\}^2 = \lim_{g \rightarrow 0} \lim_{k \rightarrow \infty} \text{MSE}(g; n_k) \\ &\geq \lim_{k \rightarrow \infty} \text{MSE}(g_{n_k}; n_k) \geq \lim_{k \rightarrow \infty} B(g_{n_k}; n_k)^2 = \{R_{M, \lambda}(f) - \psi\}^2. \end{aligned}$$

Therefore

$$\Lambda_{f, L} \subset \{\lambda \geq 0 : R_{M, \lambda}(f) = \psi\}$$

Taking into account that $\varphi_M(t) = (2 - \varphi_L(t))\varphi_L(t) \leq 1$, for all $t \in \mathbb{R}$, then reasoning as in the proof of Theorem 4 in Chacón and Tenreiro (2012) we conclude that

$$\Lambda_{f,L} \subset \left[0, \min \left(\frac{S_M}{C_f}, \frac{T_M}{D_f} \right) \right].$$

As $S_M = S_L$ and $T_M = T_L$, we finally get

$$0 \leq \limsup_{n \rightarrow \infty} g_{\text{MSE}} \leq \min \left(\frac{S_L}{C_f}, \frac{T_L}{D_f} \right),$$

which concludes the proof. ■

In the following result we study the differentiability of the functions $g \mapsto R_{\varphi,g}(f)$ and $g \mapsto T_{\varphi,\phi,g}(f)$ defined in Section A.2, and we provide asymptotic expansions for them and the corresponding derivatives when $g \rightarrow 0$.

Lemma 2. *Under assumptions (D1) and (D2), let φ and ϕ be bounded and integrable functions such that $|m|_\ell(\varphi), |m|_\ell(\phi) < \infty$ and $m_k(\varphi) = m_k(\phi) = 0$, for all $k = 1, \dots, \ell - 1$, for some even integer ℓ such that $2 \leq \ell \leq 2\nu$.*

(a) *The function $g \mapsto R_{\varphi,g}(f)$ is twice differentiable with*

$$R_{\varphi,g}(f) = R(f) m_0(\varphi) + (-1)^{\ell/2} g^\ell R(f^{(\ell/2)}) m_\ell(\varphi) / \ell! + o(g^\ell),$$

$$dR_{\varphi,g}(f)/dg = (-1)^{\ell/2} g^{\ell-1} R(f^{(\ell/2)}) m_\ell(\varphi) / (\ell-1)! + o(g^{\ell-1})$$

and

$$d^2 R_{\varphi,g}(f)/dg^2 = (-1)^{\ell/2} g^{\ell-2} R(f^{(\ell/2)}) m_\ell(\varphi) / (\ell-2)! + o(g^{\ell-2}).$$

(b) *The function $g \mapsto T_{\varphi,\phi,g}(f)$ is differentiable with*

$$T_{\varphi,\phi,g}(f) = \eta_0 m_0(\varphi) m_0(\phi) + (-1)^{\ell/2} g^\ell \eta_\ell (m_\ell(\varphi) m_0(\phi) + m_0(\varphi) m_\ell(\phi)) / \ell! + o(g^\ell)$$

and

$$dT_{\varphi,\phi,g}(f)/dg = (-1)^{\ell/2} g^{\ell-1} \eta_\ell (m_\ell(\varphi) m_0(\phi) + m_0(\varphi) m_\ell(\phi)) / \ell! + o(g^{\ell-1}),$$

where $\eta_\ell = \int_{\mathbb{R}} f^{(\ell/2)}(z) (f^2)^{(\ell/2)}(z) dz$.

Proof: Taking into account assumption (D2), the functions $\bar{f} * f$ and $h = \bar{f} \odot \bar{f} \odot f$ have continuous derivatives up to order 2ν . From the expressions (15) and (16) for $R_{\varphi,g}(f)$ and $T_{\varphi,\phi,g}(f)$, respectively, standard Taylor's expansions and the dominated convergence theorem lead to

$$\begin{aligned} R_{\varphi,g}(f) &= \sum_{i=0}^{\ell-1} \frac{g^i}{i!} (\bar{f} * f)^{(i)}(0) \int_{\mathbb{R}} u^i \varphi(u) du \\ &\quad + \frac{g^\ell}{(\ell-1)!} \int_{\mathbb{R}} \int_0^1 u^\ell \varphi(u) (1-t)^{\ell-1} (\bar{f} * f)^{(\ell)}(ugt) dt du \\ &= R(f) m_0(\varphi) + g^\ell (\bar{f} * f)^{(\ell)}(0) m_\ell(\varphi) / \ell! + o(g^\ell) \end{aligned}$$

and

$$\begin{aligned}
& T_{\varphi, \phi, g}(f) \\
&= \sum_{0 \leq i+j \leq \ell-1} \frac{g^{i+j}}{i!j!} \frac{\partial^{i+j} h}{\partial x^i \partial y^j}(0,0) \int_{\mathbb{R}} u^i \varphi(u) du \int_{\mathbb{R}} v^j \phi(v) dv \\
&+ \sum_{i+j=\ell} \frac{\ell g^\ell}{i!j!} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_0^1 u^i \varphi(u) v^j \phi(v) (1-t)^{\ell-1} \frac{\partial^\ell h}{\partial x^i \partial y^j}(tug, tvg) dt du dv \\
&= \eta_0 m_0(\varphi) m_0(\phi) + \frac{g^\ell}{\ell!} \left(m_\ell(\varphi) m_0(\varphi) \frac{\partial^\ell h}{\partial x^\ell}(0,0) + m_0(\varphi) m_\ell(\phi) \frac{\partial^\ell h}{\partial y^\ell}(0,0) \right) + o(g^\ell).
\end{aligned}$$

Moreover, from the differentiation theorem under the integral sign we have

$$\begin{aligned}
dR_{\varphi, g}(f)/dg &= \int_{\mathbb{R}} u \varphi(u) (\bar{f} * f)'(ug) du, \\
d^2 R_{\varphi, g}(f)/dg^2 &= \int_{\mathbb{R}} u^2 \varphi(u) (\bar{f} * f)''(ug) du
\end{aligned}$$

and

$$dT_{\varphi, \psi, g}(f)/dg = \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(u) \phi(v) \left(u \frac{\partial h}{\partial x}(ug, vg) + v \frac{\partial h}{\partial y}(ug, vg) \right) du dv.$$

By using again Taylor's expansions and the dominated convergence theorem we get

$$\begin{aligned}
dR_{\varphi, g}(f)/dg &= \frac{g^{\ell-1}}{(\ell-1)!} (\bar{f} * f)^{(\ell)}(0) m_\ell(\varphi) + o(g^{\ell-1}), \\
d^2 R_{\varphi, g}(f)/dg^2 &= \frac{g^{\ell-2}}{(\ell-2)!} (\bar{f} * f)^{(\ell)}(0) m_\ell(\varphi) + o(g^{\ell-2}),
\end{aligned}$$

and

$$dT_{\varphi, \phi, g}(f)/dg = \frac{g^{\ell-1}}{(\ell-1)!} \left(m_\ell(\varphi) m_0(\varphi) \frac{\partial^\ell h}{\partial x^\ell}(0,0) + m_0(\varphi) m_\ell(\phi) \frac{\partial^\ell h}{\partial y^\ell}(0,0) \right) + o(g^{\ell-1}).$$

Parts (a) and (b) follow now from the fact that $(\bar{f} * f)^{(\ell)}(0) = (-1)^{\ell/2} R(f^{(\ell/2)})$ and $\frac{\partial^\ell h}{\partial x^\ell}(0,0) = \frac{\partial^\ell h}{\partial y^\ell}(0,0) = (-1)^{\ell/2} \int_{\mathbb{R}} f^{(\ell/2)}(z) (f^2)^{(\ell/2)}(z) dz$. \blacksquare

Proof of Theorem 5: (a) From expansion (14) and taking for g the bandwidth $g = c n^{1/(2\nu+1)}$, with $c > 0$, we get

$$\lim n^{4\nu/(2\nu+1)} (\text{MSE}(g_0) - 4n^{-1} \text{Var } f(X_1)) = U(c),$$

with

$$U(c) = \{c^{-1} R(L) + c^{2\nu} R(f^\nu) m_\nu(L)^2 / (\nu!)^2\}^2.$$

Therefore, as $\text{MSE}(g_{\text{MSE}}) \leq \text{MSE}(g_0)$, we have

$$\limsup n^{4\nu/(2\nu+1)} (\text{MSE}(g_{\text{MSE}}) - 4n^{-1} \text{Var } f(X_1)) \leq U(c) < \infty. \quad (17)$$

Moreover, using the fact that $g_{\text{MSE}} \rightarrow 0$, from expansion (14) we also get

$$\begin{aligned} & n^{4\nu/(2\nu+1)} (\text{MSE}(g_{\text{MSE}}) - 4n^{-1} \text{Var } f(X_1)) \\ &= \left\{ (n^{1/(2\nu+1)} g_{\text{MSE}})^{-1} R(L) + (n^{1/(2\nu+1)} g_{\text{MSE}})^{2\nu} R(f^{(\nu)}) m_\nu(L)^2 / (\nu!)^2 \right\}^2 \\ & \quad + o \left((n^{1/(2\nu+1)} g_{\text{MSE}})^{-1} + (n^{1/(2\nu+1)} g_{\text{MSE}})^{2\nu-1} + (n^{1/(2\nu+1)} g_{\text{MSE}})^{4\nu} \right), \end{aligned}$$

which contradicts (17) if $\liminf n^{1/(2\nu+1)} g_{\text{MSE}} = 0$ or $\limsup n^{1/(2\nu+1)} g_{\text{MSE}} = \infty$. This completes the proof of part (a).

(b) Taking into account that inequality (17) is true for all $c > 0$, we also have

$$\limsup n^{4\nu/(2\nu+1)} (\text{MSE}(g_{\text{MSE}}) - 4n^{-1} \text{Var } f(X_1)) \leq U(c_0(f, L)), \quad (18)$$

where

$$c_0(f, L) = \left(\frac{R(L)(\nu!)^2}{2\nu R(f^{(\nu)}) m_\nu(L)^2} \right)^{1/(2\nu+1)} \quad (19)$$

is the minimizer of the function $c \mapsto U(c)$. By using standard arguments, inequality (18) enables us to conclude that $c_0(f, L)$ is the only accumulation point of the bounded sequence $(n^{1/(2\nu+1)} g_{\text{MSE}})$, that is, g_{MSE} is asymptotically equivalent to $g_0 = c_0(f, L) n^{-1/(2\nu+1)}$ (see Chacón, 2004, pp. 44–46). This concludes the proof of part (b) as it is well-known that the sequence g_0 is asymptotically equivalent to g_{MISE} (see Chacón et al., 2007b, pp. 293–294).

(c) From Lemma 2 and equality (10) the function $B(g)$ is twice differentiable with

$$\begin{aligned} B''(g) &= -2n^{-1} g^{-3} R(L) - 2\nu(2\nu-1) g^{2\nu-2} R(f^{(\nu)}) m_\nu(L)^2 / (\nu!)^2 \\ & \quad + O(n^{-1} g^{\nu-2}) + o(g^{2\nu-2}). \end{aligned} \quad (20)$$

Therefore, a Taylor's expansion leads to

$$0 = \text{MISE}'(g_{\text{MISE}}) = -B'(g_{\text{MISE}}) = -B'(g_{\text{MSE}}) - (g_{\text{MISE}} - g_{\text{MSE}}) B''(\tilde{g}),$$

from which we deduce that

$$n^{1/(2\nu+1)} (g_{\text{MSE}}/g_{\text{MISE}} - 1) = c_0(f, L) n g_{\text{MSE}} B'(g_{\text{MSE}}) (n \tilde{g}^3 B''(\tilde{g}))^{-1} (1 + o(1)), \quad (21)$$

for some \tilde{g} between g_{MISE} and g_{MSE} , and $c_0(f, L)$ given by (19). Taking into account that \tilde{g} is asymptotically equivalent to g_{MSE} , from equation (20) we obtain

$$n \tilde{g}^3 B''(\tilde{g}) = -(2\nu+1) R(L) (1 + o(1)). \quad (22)$$

On the other hand, from Lemma 2 and equality (11) we know that the function $V(g)$ is differentiable with

$$\begin{aligned} V'(g) &= -2n^{-2} g^{-2} R(f) R(M) - 8n^{-1} g^{2\nu-1} \{ \eta_{2\nu} - 2\nu R(f^{(\nu)}) R(f) \} m_\nu(L)^2 / (\nu!)^2 \\ & \quad + o(n^{-2} g^{-2} + n^{-1} g^{2\nu-1}) \end{aligned}$$

and

$$n^2 g_{\text{MSE}}^2 V'(g_{\text{MSE}}) = D(f, L)(1 + o(1)),$$

with $D(f, L) = -2R(f)(R(M) - 4R(L)) - 4\nu^{-1}\eta_{2\nu}R(f^{(\nu)})^{-1}R(L)$. Therefore, by using equation $\text{MSE}'(g_{\text{MSE}}) = 2B(g_{\text{MSE}})B'(g_{\text{MSE}}) + V'(g_{\text{MSE}}) = 0$, and the fact that $ng_{\text{MSE}}B(g_{\text{MSE}}) = -(2\nu + 1)(2\nu)^{-1}R(L)(1 + o(1))$, which can be deduce from equation (13), we get

$$ng_{\text{MSE}}B'(g_{\text{MSE}}) = \nu(2\nu + 1)^{-1}D(f, L)R(L)^{-1}(1 + o(1)). \quad (23)$$

The stated result with $C = -\nu(2\nu + 1)^{-2}c_0(f, L)D(f, L)R(L)^{-2}$ follows now from (21), (22) and (23). \blacksquare

B Additional information on the simulation study

Here, a more detailed account of the simulation results summarized in Section 3 of the main text is given. Let us recall that the test densities included in the study are the 15 normal mixture densities introduced in Marron and Wand (1992), plus the 10-modal normal mixture described in Loader (1999), which will be referred to as Density #16.

These test densities are all not equally hard to estimate. To measure how easy is a given density f to estimate we can resort to the functional

$$Q(f) = \inf_{u>0} u^{-1} \int_{\mathbb{R}} f(x)^{1/2} \rho\{u^5 f''(x) f(x)^{-1/2}\} dx$$

introduced in Wand and Devroye (1993), with $\rho(t) = \mathbb{E}|Z - t|$, where Z is a random variable with a standard normal distribution. These authors also showed that this functional is approximately minimised for the Beta(5.3, 5.3) distribution, yielding a value of about 1.92. The value of $Q(f)$ for each of the test densities in the simulation study is given in Table 2.

Density #	1	2	3	4	5	6	7	8
$Q(f)$	1.99	2.16	4.36	4.2	3.26	2.29	2.59	2.55
Density #	9	10	11	12	13	14	15	16
$Q(f)$	2.62	4.18	7.08	5.11	3.82	6.2	5.32	4.99

Table 2: Density-estimation-difficulty functional $Q(f)$.

From Table 2, the test densities can be categorised into three groups: the first one comprises densities 1, 2, 6, 7, 8 and 9, which can be considered as easy-to-estimate densities, with $Q(f) < 2.7$; densities 3, 4, 5, 10 and 13 could be seen as medium-estimation-difficulty densities, since they have $3.2 < Q(f) < 4.4$; and densities 11, 12, 14, 15 and 16 may be cast as difficult-to-estimate densities, with $Q(f) > 4.9$.

Figures 1 and 2 (for $n = 100$ and $n = 1000$, respectively) contain boxplots showing the distribution of the relative errors for the estimators $\hat{\psi}_{\text{CT}}$, $\hat{\psi}_{\text{SH}}$ and $\hat{\psi}_{\text{JS}}$ in the study, for

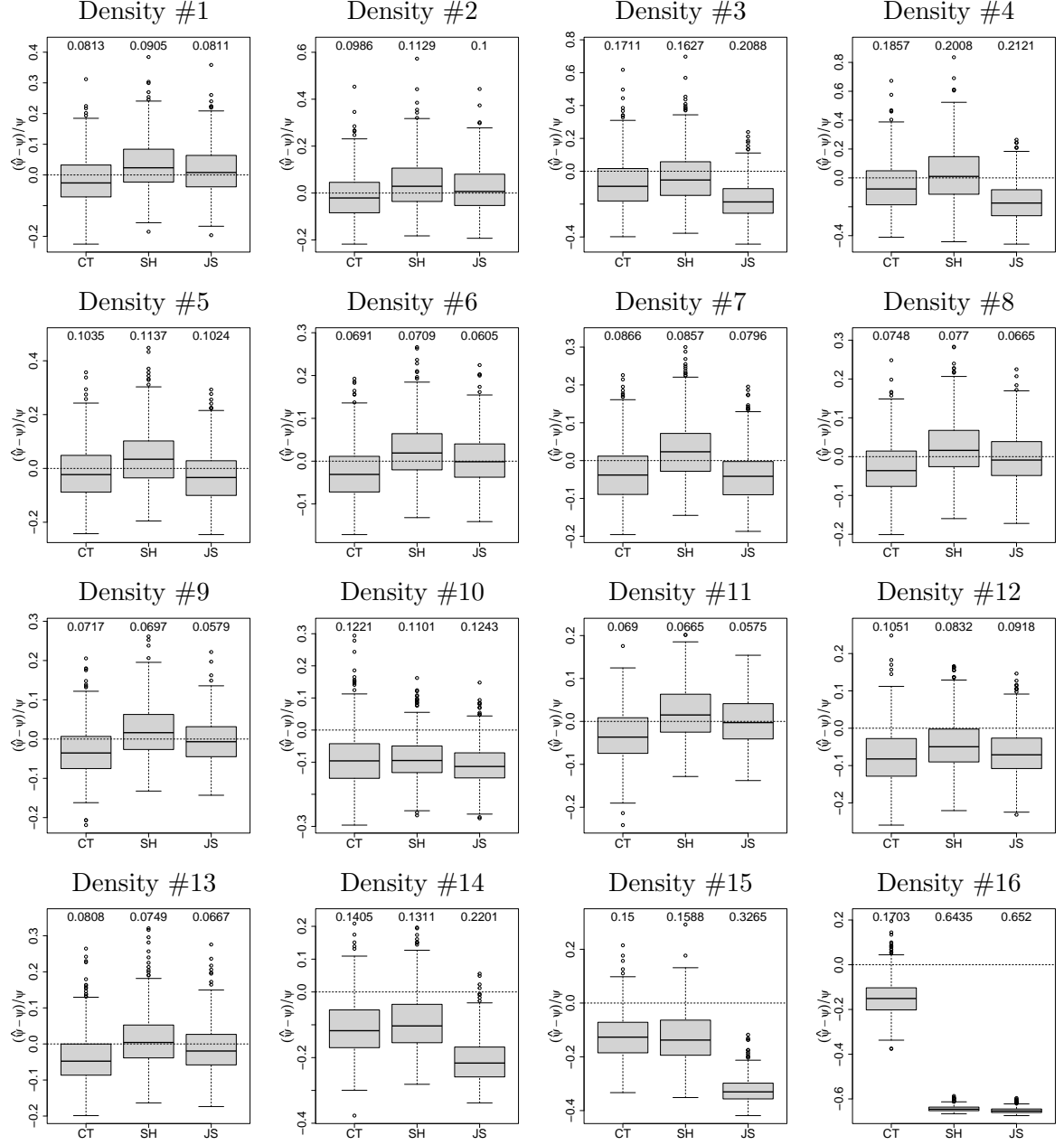


Figure 1: *Distribution boxplots for the relative error of the estimators $\hat{\psi}_{CT}$, $\hat{\psi}_{SH}$ and $\hat{\psi}_{JS}$ for $n = 100$.*

each of the test densities considered. On the top of each boxplot there is also a number, showing the sample relative root mean squared error of the corresponding estimator for the test density under consideration, along the $B = 500$ simulation runs, as defined in the main text.

These results reveal that, for the group of easy-to-estimate densities, the plug-in-type

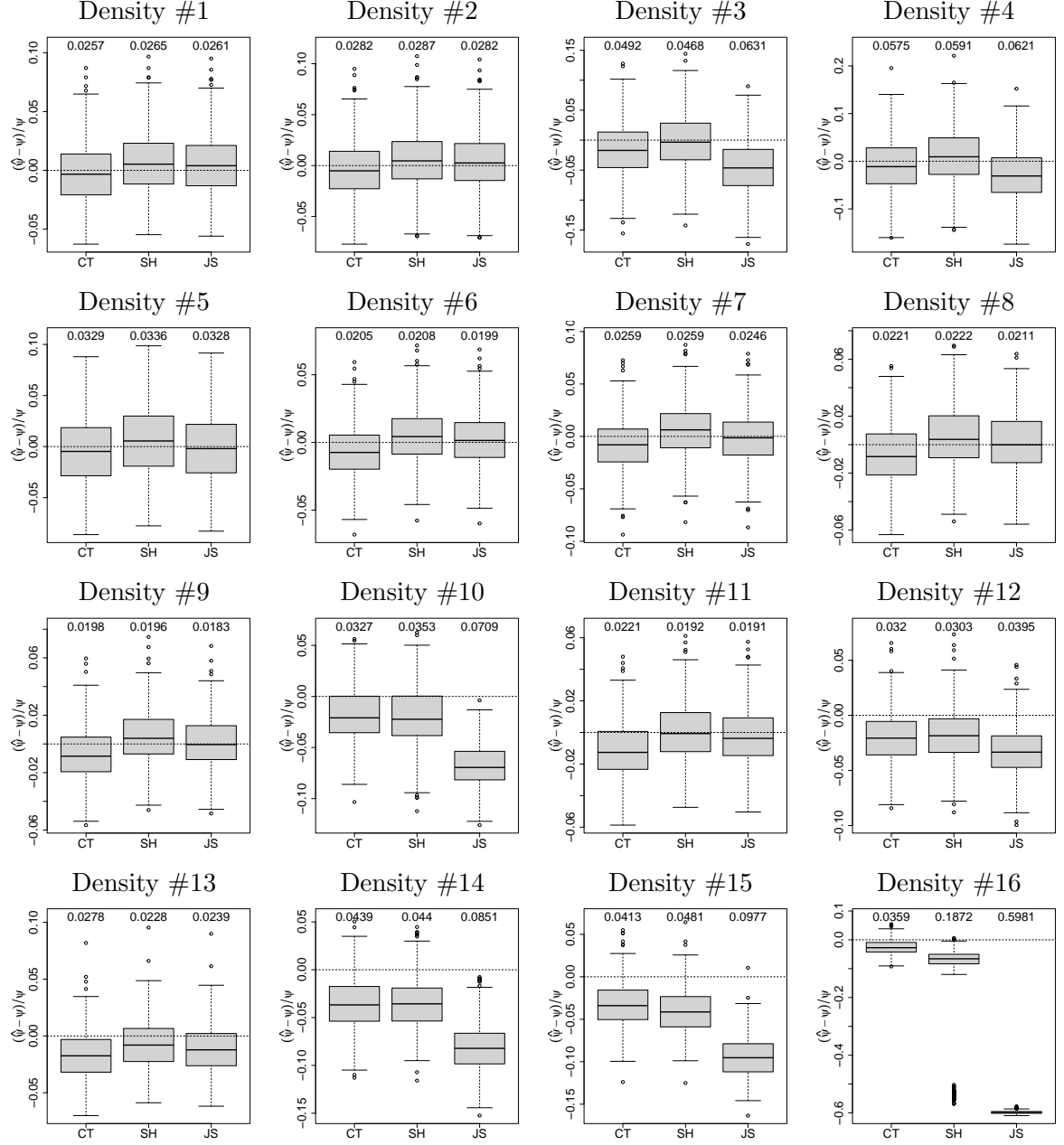


Figure 2: Same as Figure 1, but for $n = 1000$.

estimator $\hat{\psi}_{JS}$ performs quite satisfactorily. But it does not appear so competitive for some of the other density models, especially for densities 3, 4, 10, 12, 14, 15 and 16. The classical alternative $\hat{\psi}_{SH}$, based on a solve-the-equation rule, seems to adapt better to the more difficult-to-estimate scenarios. However, it also fails abysmally for Density #16, more markedly with sample size $n = 100$.

The breakdown of both estimators $\hat{\psi}_{JS}$ and $\hat{\psi}_{SH}$ for this model #16 is caused by the use of a quite inadequate reference distribution at their starting steps. This is an unfortunate

issue that the new estimator $\hat{\psi}_{CT}$ does not present, since it does not depend on the choice of a reference distribution. This feature pays off extremely well for this model, since $\hat{\psi}_{CT}$ appears to have a clear advantage over the two other estimators for this Density #16. But in fact, the new proposal shows a very competitive performance along the whole group of test densities, ranking either first or very close to the first one for all density models, as reflected in the summary statistics reported in the main text.

References

- 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. (2007a). A note on kernel density estimation at a parametric rate. *J. Nonparametr. Stat.*, **19**, 13–21.
- Chacón, J.E., Montanero, J., Nogales, A.G. and Pérez, P. (2007b). On the existence and limit behavior of the optimal bandwidth in kernel density estimation. *Statist. Sinica*, **17**, 289–300.
- Chacón, J.E., Tenreiro, C. (2012). On the equivalence of exact and asymptotically optimal bandwidths for kernel estimation of density functionals. *Methodology and Computing in Applied Probability*, **14**, 523–548.
- Lee, A.J. (1990). *U-statistics, theory and practice*. New York: Marcel Dekker.
- Loader, C. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics*, **27**, 415–438.
- Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics*, **20**, 712–736.
- Wand, M.P. and Devroye, L. (1993). How easy is a given density to estimate? *Computational Statistics and Data Analysis*, **16**, 311–323.