Data-based choice of the number of pilot stages for plug-in bandwidth selection*

J.E. Chacón† and C. Tenreiro‡

July 8, 2011

Abstract

The choice of the bandwidth is a crucial issue for kernel density estimation. Among all the data-dependent methods for choosing the bandwidth, the direct plug-in method has shown a particularly good performance in practice. This procedure is based on estimating an asymptotic approximation of the optimal bandwidth, using two 'pilot' kernel estimation stages. Although two pilot stages seem to be enough for most densities, for a long time the problem of how to choose an appropriate number of stages has remained open. Here we propose an automatic (i.e., data-based) method for choosing the number of stages to be employed in the plug-in bandwidth selector. Asymptotic properties of the method are presented and an extensive simulation study is carried out to compare its small-sample performance with that of the most recommended bandwidth selectors in the literature.

KEYWORDS: bandwidth selection, density estimation, kernel method, plug-in rule

2000 AMS SUBJECT CLASSIFICATIONS: Primary 62G05, Secondary 62G07, 62G20

*This is an electronic version of an article published in Communications in Statistics – Theory and Methods (Vol. 42, 2013, 2200–2214). DOI:10.1080/03610926.2011.606486

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

‡CMUC, Department of Mathematics, University of Coimbra, Apartado 3008, 3001–454 Coimbra, Portugal. E-mail: tenreiro@mat.uc.pt

1 Introduction

In this paper we give a solution to an open problem posed by Park and Marron (1992), which is also highlighted in Wand and Jones (1995, p. 73). The background of the problem is kernel density estimation. Specifically, if $X_1, \ldots, X_n$ are independent copies of a real random variable X, having an absolutely continuous probability distribution P, with density f, the kernel estimator of f is defined as

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

where the kernel K is a real integrable function with $\int K = 1$, h is a positive real number, called the bandwidth or smoothing parameter, and we are using the notation $K_h(x) = K(x/h)/h$.

It is widely known (see, e.g., Silverman, 1986; Simonoff, 1996) that the performance of this estimator depends strongly on the choice of h. In this sense, the so-called optimal bandwidth h_{MISE} is the minimizer of the mean integrated squared error function, $\mathrm{MISE}(h) = \mathbb{E}[\mathrm{ISE}(h)]$, where $\mathrm{ISE}(h) = \int \{f_{nh}(x) - f(x)\}^2 dx$. Chacón, Montanero, Nogales and Pérez (2007) provide sufficient conditions for h_{MISE} to exist. A data-based bandwidth selector is just an estimator of the theoretically optimal bandwidth h_{MISE}.

For an arbitrary real function α, denote $R(\alpha) = \int \alpha(x)^2 dx$ and $\mu_p(\alpha) = \int x^p \alpha(x) dx$ for $p \in \mathbb{N}$. When a positive, symmetric and bounded kernel with a finite second-order moment $\mu_2(K)$ is used in (1), under some smoothness assumptions on f, it is possible to give an asymptotic approximation of h_{MISE}, namely

$$h_0 = c_1 \psi_4^{-1/5} n^{-1/5}, \tag{2}$$

where we are abbreviating $\psi_r = \int f^{(r)}(x) f(x) dx = \mathbb{E} f^{(r)}(X)$ for an even number r (see Wand and Jones, 1995) and $c_1 = [R(K)/\mu_2(K)^2]^{1/5}$. As the only unknown quantity in (2) is ψ_4, the problem of providing a bandwidth selector reduces to that of estimating ψ_4.

The kernel estimator of ψ_r for an arbitrary even r is given by

$$\hat{\psi}_r(g) = \frac{1}{n^2} \sum_{i,j=1}^{n} L_g^{(r)}(X_i - X_j) \tag{3}$$

(Hall and Marron, 1987; Jones and Sheather, 1991), where in this case the kernel L and the bandwidth g may be different from K and h. Although a better asymptotic performance can be obtained by taking for L a higher-order kernel (see Wand and Jones, 1995, p. 67–70), from a practical point of view, and in common with other studies in the literature (see, e.g., Marron and Wand, 1992, or Jones et al. 1996), no significant improvements

over the estimator based on a positive and symmetric kernel L are observed. Therefore we will adopt the usual approach of taking L in (3) to be the standard Gaussian density $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

As ψ_r is a real parameter, it is natural to use in (3) the bandwidth g minimizing the mean squared error of the estimator, $\mathrm{MSE}(g) = \mathbb{E}[\{\hat{\psi}_r(g) - \psi_r\}^2]$. Under some additional assumptions on f it is possible to obtain an asymptotic representation of the MSE function, namely $\mathrm{AMSE}(g)$, and the minimizer of this AMSE function is given by

$$g_{0,r} = \left(\frac{(-1)^{r/2+1} r!}{2^{(r-1)/2} (r/2)! \sqrt{\pi}\, \psi_{r+2} n} \right)^{1/(r+3)} \tag{4}$$

(Jones and Sheather, 1991; Wand and Jones, 1995, p. 70). In view of (4), it is clear that the problem becomes somehow a cyclic process, as the asymptotically best bandwidth for estimating ψ_r depends on ψ_{r+2}, another of these density functionals.

To overcome this problem, the usual solution is to use an ℓ-stage bandwidth selection procedure (see Tenreiro, 2003, and references therein), which consists in the following:

1. Provide a quick and simple estimate of $\psi_{r+2\ell}$. This may be achieved by using an estimate of the corresponding functional for some reference distribution. The normal distribution with zero mean and standard deviation σ is mostly used as a reference since in this case, following Wand and Jones (1995, p. 72), any functional ψ_s with even s can be written as

$$\psi_s^{\mathrm{NR}} \equiv \psi_s^{\mathrm{NR}}(\sigma) = \frac{(-1)^{s/2} s!}{(2\sigma)^{s+1} (s/2)! \sqrt{\pi}}, \tag{5}$$

   so that an easy estimate of $\psi_{r+2\ell}$ is given by $\hat{\psi}_{r+2\ell}^{\mathrm{NR}} = \psi_{r+2\ell}^{\mathrm{NR}}(\hat{\sigma})$ where $\hat{\sigma}$ denotes any scale estimate.

2. Estimate successively the ℓ density functionals

$$\psi_{r+2(\ell-1)}, \psi_{r+2(\ell-2)}, \ldots, \psi_{r+2}, \psi_r,$$

   with a kernel estimator. The bandwidth $g = \hat{g}_{0,r+2j}$ used in the kernel estimator $\hat{\psi}_{r+2j}(g)$ is just the one given by (4), with the unknown functional $\psi_{r+2(j+1)}$ replaced by its previously calculated estimate.

The final step of the above procedure will give us an estimate of ψ_r, which we will denote $\hat{\psi}_{r,\ell}$. In particular, for $r = 4$, replacing ψ_4 with $\hat{\psi}_{4,\ell}$ in (2) results in what is called the ℓ-stage plug-in bandwidth selector, $\hat{h}_{\mathrm{PI},\ell}$. In particular the normal scale rule, which consists

of replacing ψ_4 with $\hat{\psi}_4^{\mathrm{NR}}$ in (2), can be thought as being a zero-stage plug-in bandwidth selector.

Park and Marron (1992) observed that the influence on the plug-in selector of the arbitrary reference distribution used in the initial step diminishes as the number of stages increases. However, the cost of using additional estimation steps results in an increment of the variance of the bandwidth selector. Therefore, Park and Marron (1992) posed the following problem: how many kernel functional estimation stages should be used? It would be useful to have a method to select the correct (in some sense) number of steps, in order to balance the two aforementioned effects. This is the main goal of this paper.

The rest of the paper is organized as follows. In Section 2, we describe the behavior of plug-in bandwidth selectors depending on the number of pilot stages. In Section 3 we introduce a method for choosing the number of pilot stages from the data, which can be seen as a hybrid between cross-validation and direct plug-in bandwidths, and we describe its asymptotic behavior. In Sections 4 and 5 we describe the finite sample behavior of the proposed method. An extensive simulation study is carried out to compare its performance with the most recommended methods in the literature. The simulation results confirm that the new procedure performs quite well presenting a good overall performance for a wide set of density features. All the proofs are deferred to Section 6.

2 Asymptotic and finite sample behavior of multistage plug-in bandwidth selectors

Here we present some theoretical results and examples providing some insight into the problem of how to select the number of stages for the plug-in bandwidth selector.

First of all we should say that, asymptotically, all the multistage plug-in bandwidth selectors achieve the same order of convergence, as long as they use $\ell \geq 2$ pilot stages. This is a well-known result, which can be stated in the following way. As mentioned before, we set L in (3) to be the standard normal density.

Theorem 1 (Tenreiro, 2003)**.** *Assume that K is a positive, bounded and symmetric kernel with a finite second-order moment, f has bounded derivatives up to order $4+2\ell$, and there exists $\sigma_f \neq 0$ such that $\hat{\sigma} - \sigma_f = O_P(n^{-1/2})$, where $\hat{\sigma}$ is the scale distribution estimator in the multistage procedure. Then $\hat{h}_{\mathrm{PI},\ell}/h_0 - 1 = O_P(n^{-\alpha})$ with $\alpha = 2/7$ for $\ell = 1$ and $\alpha = 5/14$ for all $\ell \geq 2$. Moreover, for all $\ell \geq 2$ we have $n^{5/14}(\hat{h}_{\mathrm{PI},\ell}/h_0 - 1) \xrightarrow{d} N(0, \sigma^2_{\mathrm{PI}})$, where the asymptotic variance $\sigma^2_{\mathrm{PI}} = 2^{-9/14}3^{-2/7}7\pi^{1/7}\psi_0\psi_4^{-2}|\psi_6|^{9/7}/80$ is independent of ℓ.*

The previous result justifies the usual recommendation of using $\ell = 2$ (Aldershof, 1990;

Sheather and Jones, 1991; Park and Marron, 1992). However, from a nonasymptotic point of view, considerable improvements can be obtained in some cases if we allow for a higher number of pilot estimation stages.

To see this, let us consider the case where the kernel K is taken to be the standard normal density, and the density f is a mixture of normal densities, as in Marron and Wand (1992). For this kernel and class of densities there are fast and easy-to-implement formulas to compute the exact ISE of the kernel estimator, therefore, we can easily obtain a sample of size B of the random variable $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$ by using B artificially generated samples with density f. This way, we can explore the distribution of $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$ for several values of ℓ. Moreover, by averaging over the B samples we get an impression of the behaviour of $\mathrm{EISE}(\ell) = \mathbb{E}[\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})]$ as a function of ℓ. It is to be remarked here that the EISE function should not be mistaken for the MISE function (see Jones, 1991).

In Figure 1 we give plots showing the effect of the number of pilot stages both on the ISE and the EISE. This figure shows 15 graphs, corresponding to the 15 normal mixture densities in Marron and Wand (1992). In all cases we have set L in (3) to be the standard normal density. Additionally, we have taken the estimator proposed by Silverman (1986, p. 47) as the scale estimator $\hat{\sigma}$. In each graph we show 21 boxplots representing the distribution of the random variable $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$ for $\ell = 0, 1, 2, \ldots, 20$ based on $B = 1000$ simulated samples of size $n = 200$. Also, we include a polygonal line going through the sample mean values of these distributions, thus giving an approximation of $\mathrm{EISE}(\ell)$ for $\ell = 0, 1, 2, \ldots, 20$. The solid black circle is used then to point out the optimal number of stages in the EISE sense; that is, the number of stages minimizing the (approximation of the) EISE function.

Similar pictures were generated for sample sizes $n = 50$, 100, 400, 800 and 1600, but they are not included here to save space. Nevertheless, we include in Table 1 the EISE-optimal number of stages for these sample sizes for the 15 normal mixture densities considered.

In view of Table 1 and Figure 1 we can classify our 15 test densities into two groups:

1. When the true density is close to the normal one, the straightforward use of a normal reference estimate of ψ_4 in the formula of the asymptotically optimal bandwidth h_0 does a good job. This is the case mainly for densities #1, #2 and #5. We can also include in this group those densities having easy-to-identify features, such as #6, #7, #8 and #9, for which a low number of stages (less than 5) seems to be the reasonable choice. Finally, there are some hard-to-estimate densities that fall into this group only when the sample size is small, like densities #10 and #12, or #13 for small and moderate values of n and #11 for $n \leq 1600$ at least. The reason for the good performance of a low number of stages for such a combination of densities and

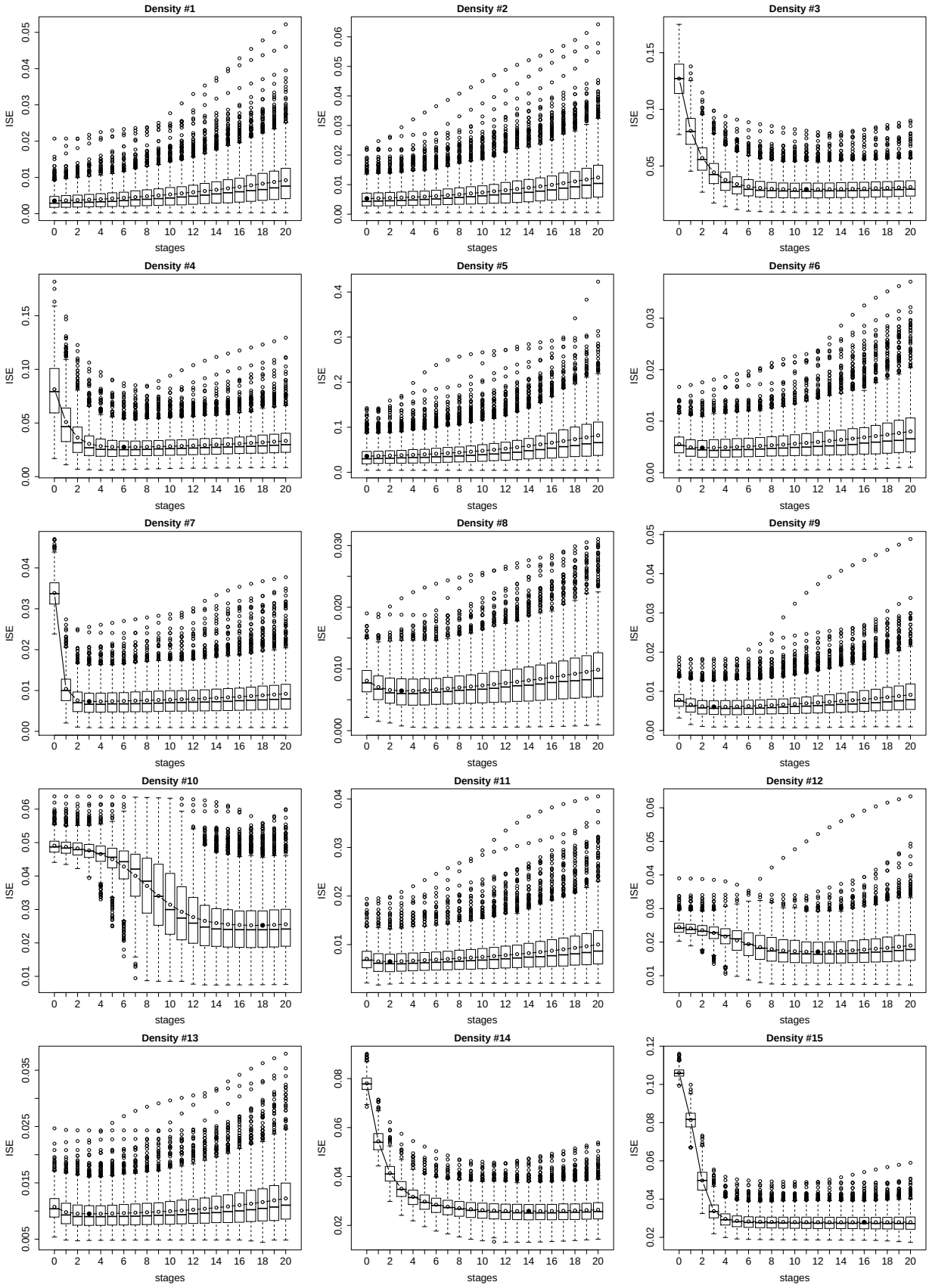

Figure 1: *Distribution of* $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$ *depending on the number of stages* ($n = 200$).

Sample size	Density number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$n = 50$	0	0	10	7	0	1	4	2	2	0	1	0	1	10	8
$n = 100$	0	0	11	7	0	2	4	3	3	16	2	10	2	12	8
$n = 200$	0	0	11	6	0	2	3	3	3	18	2	12	3	14	16
$n = 400$	0	0	11	6	0	3	3	4	4	13	2	11	4	16	26
$n = 800$	0	0	10	5	0	2	3	4	5	10	2	13	16	20	23
$n = 1600$	0	1	8	5	0	2	3	4	5	8	3	14	21	22	18

Table 1: *EISE-optimal number of stages.*

sample sizes is that they present distributional features that are not revealed until the sample size is above some threshold. For instance, Figures 6 and 7 in Marron and Wand (1992) show that it is difficult to distinguish between densities #6 and #11 for $n \leq 10^4$, and that is the reason for the similar number of optimal stages. It is reasonable to expect, however, that for larger values of n a larger number of pilot stages will be advisable for density #11, as happens with densities #10, #12 (for $n \geq 100$) and #13 (for $n \geq 800$).

2. The second group of densities comprises those for which using a multistage plug-in selector with a large number of pilot stages is highly advisable, in the sense that a big decrease of ISE is clearly noticeable from the 0-stage method to a certain number of stages (depending on each particular density), from which the ISE distribution stabilizes. In this group we include densities #3, #4, #14 and #15, and depending on the sample size also #10, #12, #13 (for moderate and large values of n) and #11 (for very large sample sizes). For these densities the EISE-optimal number of stages is high but, in most of those cases, using such a high number of stages does not represent a significant gain over using, say, a 10-stage method. However, for densities #10, #14 and #15 using a higher number of stages is advisable for moderate and large values of n.

The main conclusion after observing Figure 1 is that in some cases, especially for those densities in group 2 above, the plug-in method may improve considerably if we allow for a higher number of stages than the usual advice $\ell = 2$.

We finish this section by presenting a result that gives some important theoretical insight into the previously described finite sample behavior of the ℓ-stage plug-in bandwidth as a function of ℓ. Based on this result, and taking into account that the simple normal scale rule usually leads to a large bandwidth $\hat{h}_{\mathrm{PI},0}$ due to the fact that it is based on a very smooth

reference distribution family (see Terrell, 1990, Theorem 1), the ℓ-stage plug-in bandwidth selector can be seen as a correction for the zero-stage plug-in bandwidth whenever the number of pilot stages ℓ is properly chosen. This is discussed in the next section.

Theorem 2. *Under the conditions of Theorem 1, for a fixed $\bar{\ell} \in \mathbb{N}$ assume that f has bounded derivatives up to order $4 + 2\bar{\ell}$ and*

$$|\psi_{4+2\ell}| \geq |\psi_{4+2\ell}^{\mathrm{NR}}(\sigma_f)|, \quad \textit{for all} \quad \ell = 1, 2, \ldots, \bar{\ell}. \tag{6}$$

Then $P\big(\hat{h}_{\mathrm{PI},\bar{\ell}} \leq \hat{h}_{\mathrm{PI},\bar{\ell}-1} \leq \cdots \leq \hat{h}_{\mathrm{PI},1} \leq \hat{h}_{\mathrm{PI},0}\big) \to 1$, as $n \to \infty$.

Remark 1. Condition (6) is not very restrictive due to the smoothness properties of the normal distribution. However, it can be improved or even suppressed if for each $\ell = 1, 2, \ldots, \bar{\ell}$, an appropriate reference distribution family is used. This is the case when the reference distribution used in the multistep procedure is taken from the scale family of the beta distribution $\mathrm{Beta}(-1, 1, s/2 + 2, s/2 + 2)$ with $s = 4 + 2\ell$. Precisely, if we denote by $\psi_s^{\mathrm{BR}} \equiv \psi_s^{\mathrm{BR}}(\sigma)$ the value of the ψ_s functional corresponding to the member of the scale family of the distribution $\mathrm{Beta}(-1, 1, s/2 + 2, s/2 + 2)$ with standard deviation σ, then condition (6) becomes $|\psi_{4+2\ell}| \geq |\psi_{4+2\ell}^{\mathrm{BR}}(\sigma_f)|$ for all $\ell = 1, 2, \ldots, \bar{\ell}$, which is fulfilled by every density f (cf. Terrell, 1990, Theorem 1). Besides, as in the case of the normal reference distribution, explicit formulas for ψ_s^{BR} for even s are easy to obtain. In fact,

$$\psi_s^{\mathrm{BR}} = \frac{(-1)^{s/2}(s!)^2(s+1)(s+3)}{2^s((s/2)!)^2(s+5)^{(s+3)/2}\sigma^{s+1}},$$

where σ is the scale parameter. Some preliminary simulations were also conducted to analyze the behavior of the proposed plug-in bandwidth selector for the beta scale rule but no significant practical improvements over the normal scale rule were observed.

3 Data-based choice of the number of stages

The natural question which arises from the previous considerations is: how should we choose the number of pilot stages ℓ in practice? If we fix a minimum and a maximum number of pilot stages $\underline{\mathrm{L}}$ and $\overline{\mathrm{L}}$, say, choosing a stage ℓ among the set of possible pilot stages $\mathscr{L} = \{\underline{\mathrm{L}}, \underline{\mathrm{L}} + 1, \ldots, \overline{\mathrm{L}}\}$ is naturally equivalent to selecting one of the bandwidths

$$\hat{h}_{\mathrm{PI},\ell} = c_1 \hat{\psi}_{4,\ell}^{-1/5} n^{-1/5},$$

for $\ell \in \mathscr{L}$. As discussed in the previous section, for densities such as those in group 2, the plug-in method may improve considerably if we allow for a higher number of stages

than the usual advice $\ell = 2$. However, choosing a larger fixed number of pilot stages will lead to undersmoothing especially for group 1 densities, as explained by Theorem 2. This is an unattractive feature because, as is well known, the kernel density estimator is penalized much more by excessively small rather than by excessively large bandwidths. An alternative approach for choosing ℓ is described in this section. This approach enables us to obtain a bandwidth selector that could deal with a higher number of pilot stages without being strongly affected by undersmoothing.

Following Hall and Marron (1988) who used cross-validation as a method for choosing the kernel order for kernel density estimators, we propose here a similar technique for the practical choice of the number of pilot stages to be used in the plug-in bandwidth selector. The least-squares cross-validation criterion proposed by Rudemo (1982) and Bowman (1984), is given by

$$\mathrm{CV}(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum_{i \neq j} (\tfrac{n-1}{n} K_h * K_h - 2K_h)(X_i - X_j),$$

where $*$ denotes the convolution product. For a fixed $h > 0$, $\mathrm{CV}(h)$ is an unbiased estimator of $\mathrm{MISE}(h) - R(f)$, and the cross-validation bandwidth selector is given by the value $\hat{h}_{\mathrm{CV}}$ of $h > 0$ that minimizes $\mathrm{CV}(h)$. See Hall (1983), Stone (1984), Hall and Marron (1987b) and Park and Marron (1990) for some asymptotic properties of $\hat{h}_{\mathrm{CV}}$.

Using the previous criterion, our proposal is to take for the number of pilot stages the value $\hat{\ell} = \hat{\ell}(\underline{\mathrm{L}}, \overline{\mathrm{L}}; X_1, \ldots, X_n)$ defined by

$$\hat{\ell} = \mathrm{argmin}_{\ell \in \mathscr{L}} \mathrm{CV}(\hat{h}_{\mathrm{PI},\ell}). \tag{7}$$

This method for choosing the number of pilot stages leads to the data-based bandwidth $\hat{h}_{\mathrm{PI},\hat{\ell}}$, that can be seen as a hybrid between cross-validation and direct plug-in bandwidths.

Next we show that $\hat{h}_{\mathrm{PI},\hat{\ell}}$ inherits the asymptotic rates of convergence of the worst performing bandwidth of the set $\{\hat{h}_{\mathrm{PI},\ell}\}_{\ell \in \mathscr{L}}$. This is established in the next result as a direct consequence of Theorems 1 and 2. Although it is stated for the previously introduced data-dependent choice of ℓ, it is also valid for any other (measurable) rule $\hat{\ell}$ for choosing ℓ taking values in $\mathscr{L}$. Note that this result justifies the recommendation of using $\underline{\mathrm{L}} = 2$ in (7). The role of $\overline{\mathrm{L}}$ will be discussed later in detail.

Theorem 3. *Under the conditions of Theorem 1, if f has bounded derivatives up to order $4 + 2\overline{\mathrm{L}}$ we have $\hat{h}_{\mathrm{PI},\hat{\ell}}/h_0 - 1 = O_P(n^{-\alpha})$ with $\alpha = 2/7$ for $\underline{\mathrm{L}} = 1$ and $\alpha = 5/14$ for $\underline{\mathrm{L}} \geq 2$. Moreover, if $\underline{\mathrm{L}} \geq 2$ and condition (6) is fulfilled with $\bar{\ell} = \overline{\mathrm{L}}$, then $n^{5/14}(\hat{h}_{\mathrm{PI},\hat{\ell}}/h_0 - 1) \xrightarrow{d} N(0, \sigma^2_{\mathrm{PI}})$, where σ^2_{PI} is given in Theorem 1.*

Remark 2. The order of convergence $O_P(n^{-5/14})$ obtained in Theorem 3 is also shared by the two-stages direct plug-in method (see Theorem 1), by the two-stage solve-the-equation bandwidth selector method proposed by Sheather and Jones (1991) and by the improved Sheather and Jones method recently introduced by Liao, Wu and Lin (2010). However, we improve on the CV order of convergence $O_P(n^{-1/10})$ (see Hall and Marron, 1987b) by choosing h from a set of well-behaved plug-in bandwidths (instead of the whole $h > 0$ range). Therefore, it is expected that $\hat{h}_{\mathrm{PI},\hat{\ell}}$ presents less sample variability than the CV bandwidth.

4 Finite sample behavior of the proposed method

In order to gain some insight into the finite sample behavior of the bandwidth $\hat{h}_{\mathrm{PI},\hat{\ell}}$ as a function of $\overline{\mathrm{L}}$ for $\hat{\ell}$ given by (7), we consider one density from each one of the groups described in Section 2: we take density #1 from group 1 and density #15 from group 2. For each one of these densities, we compare in Figure 2 the empirical distributions of $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\hat{\ell}})$ for different values of $\overline{\mathrm{L}}$ based on 500 simulated samples of size $n = 400$. We take $\overline{\mathrm{L}} = 5, 10, 20, 40$. Moreover, aiming to illustrate the usefulness of the proposed cross-validation based procedure for selecting the number of pilot stages in relation to a fixed based approach, the empirical distributions of $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$, for $\ell = 5, 10, 20, 40$, are also shown. For comparative purposes we also include in the figure the ISE distributions of the standard two-stage direct plug-in bandwidth $\hat{h}_{\mathrm{PI},2}$ proposed by Sheather and Jones (1991) (labeled SJdpi) and the cross-validation bandwidth $\hat{h}_{\mathrm{CV}}$ (labeled CV). We adopt the previous recommendation of using $\underline{\mathrm{L}} = 2$ and we take for K the standard normal density.

The boxplots show that a smaller value for $\overline{\mathrm{L}}$ is recommended for densities from group 1 whereas a larger value for $\overline{\mathrm{L}}$ is most suitable for densities from group 2. This behavior, which is explained in large part by Theorem 2, is in accordance with the conclusions of Section 2. For both densities we see that the best results are observed when $\overline{\mathrm{L}}$ is close to the EISE-optimal number of stages given in Table 1. However, $\hat{h}_{\mathrm{PI},\hat{\ell}}$ is quite robust against the choice of $\overline{\mathrm{L}}$ whenever $\overline{\mathrm{L}}$ is larger, but not excessively larger, than the EISE-optimal number of stages. This last property, that is not shared by $\hat{h}_{\mathrm{PI},\ell}$, shows the usefulness of the proposed data-based method for choosing ℓ. In order to find a compromise between these two situations, in view of Table 1 we decided to take $\overline{\mathrm{L}} = 10$. By choosing such an intermediate value for $\overline{\mathrm{L}}$ we expect that the new data-based bandwidth might present a good overall performance for a wide range of density features. This is studied in the next section.

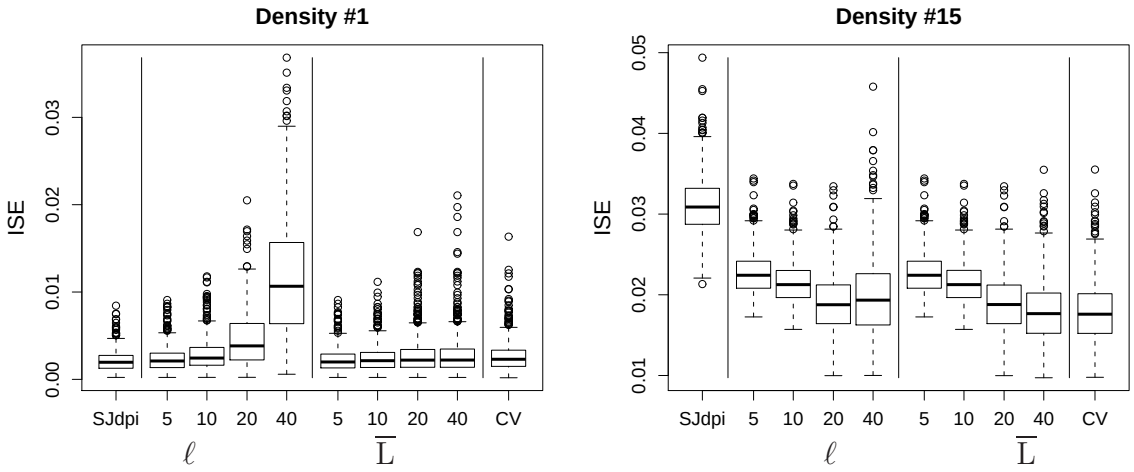

Figure 2: *Distribution of* $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$ *as a function of* ℓ *and of* $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\hat{\ell}})$ *as a function of* $\overline{\mathrm{L}}$ $(n = 400)$.

5 Simulation study

We performed a simulation study to compare the new procedure based on $\hat{h}_{\mathrm{PI},\hat{\ell}}$ with $\underline{\mathrm{L}} = 2$ and $\overline{\mathrm{L}} = 10$ (labeled CT) with some of the most successful bandwidth selection methods in the literature, namely the two-stage direct plug-in method and the two-stage solve-the-equation plug-in method proposed by Sheather and Jones (1991) (labeled SJste and SJdpi, respectively) and the classical least-squares cross-validation method (labeled CV). These methods have been shown to provide quite reasonable results in practice; see Cao, Cuevas and González-Manteiga (1994) or Jones et al. (1996), and references therein. A recently proposed solve-the-equation plug-in type method by Liao et al. (2010) (labeled LWL) that has revealed a promising behavior was also included in the study. In the implementation of CT, SJdpi and SJste the normal density was used as the reference distribution and we have taken the estimator proposed by Silverman (1986, p. 47) as the scale estimator. See Wand and Jones (1995, p. 71–75) for the implementation of the Sheather and Jones (1991) methods.

We use as test densities the same 15 normal mixture densities that we referred to in Section 2. Based on 500 samples of sizes $n = 100$, $n = 400$ and $n = 800$, from each test density in the study we plot in Figures 3, 4 and 5 the boxplots for the distributions of ISEs corresponding to each of the five bandwidth selection methods.

From the figures we see that the SJdpi, SJste and LWL methods are the best of the considered methods for the densities of group 1. However, for some of the densities of group 2 the SJ methods present a disappointing performance, especially the SJdpi method. The

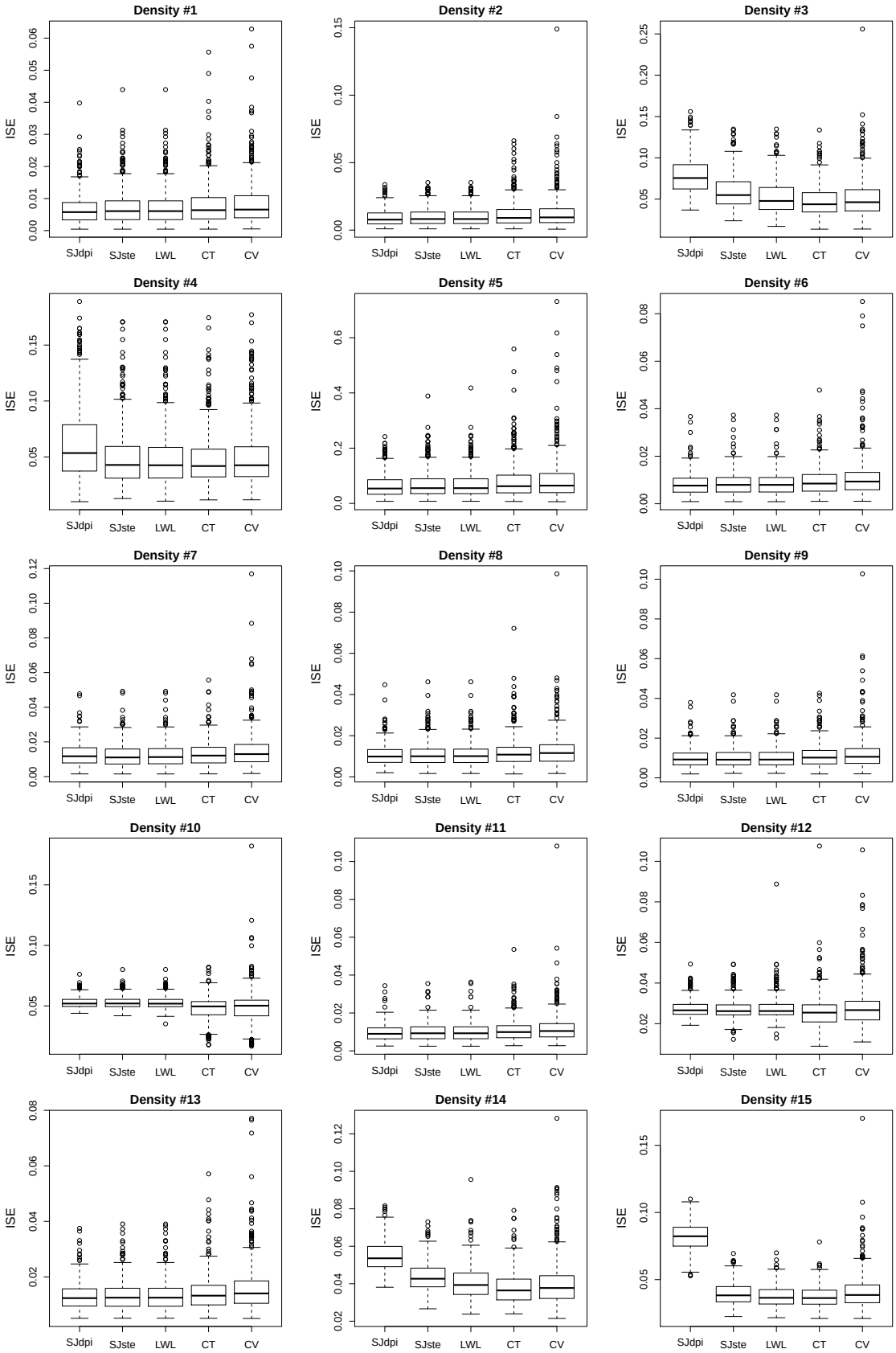

Figure 3: ISE *distribution for the bandwidth selectors methods* SJdpi, SJste, LWL, CT *and* CV ($n = 100$).

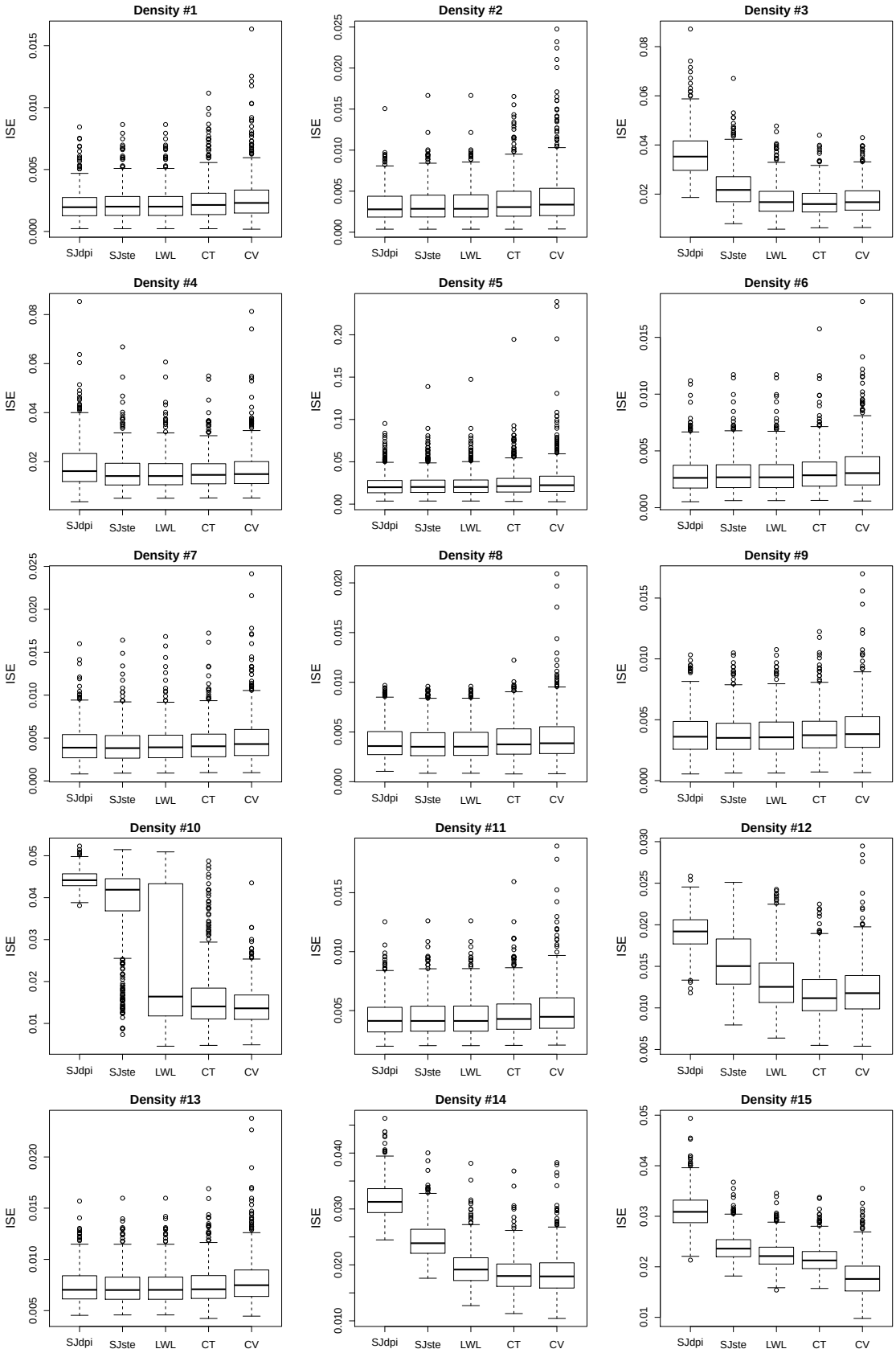

Figure 4: ISE *distribution for the bandwidth selectors methods* SJdpi, SJste, LWL, CT *and* CV ($n = 400$).

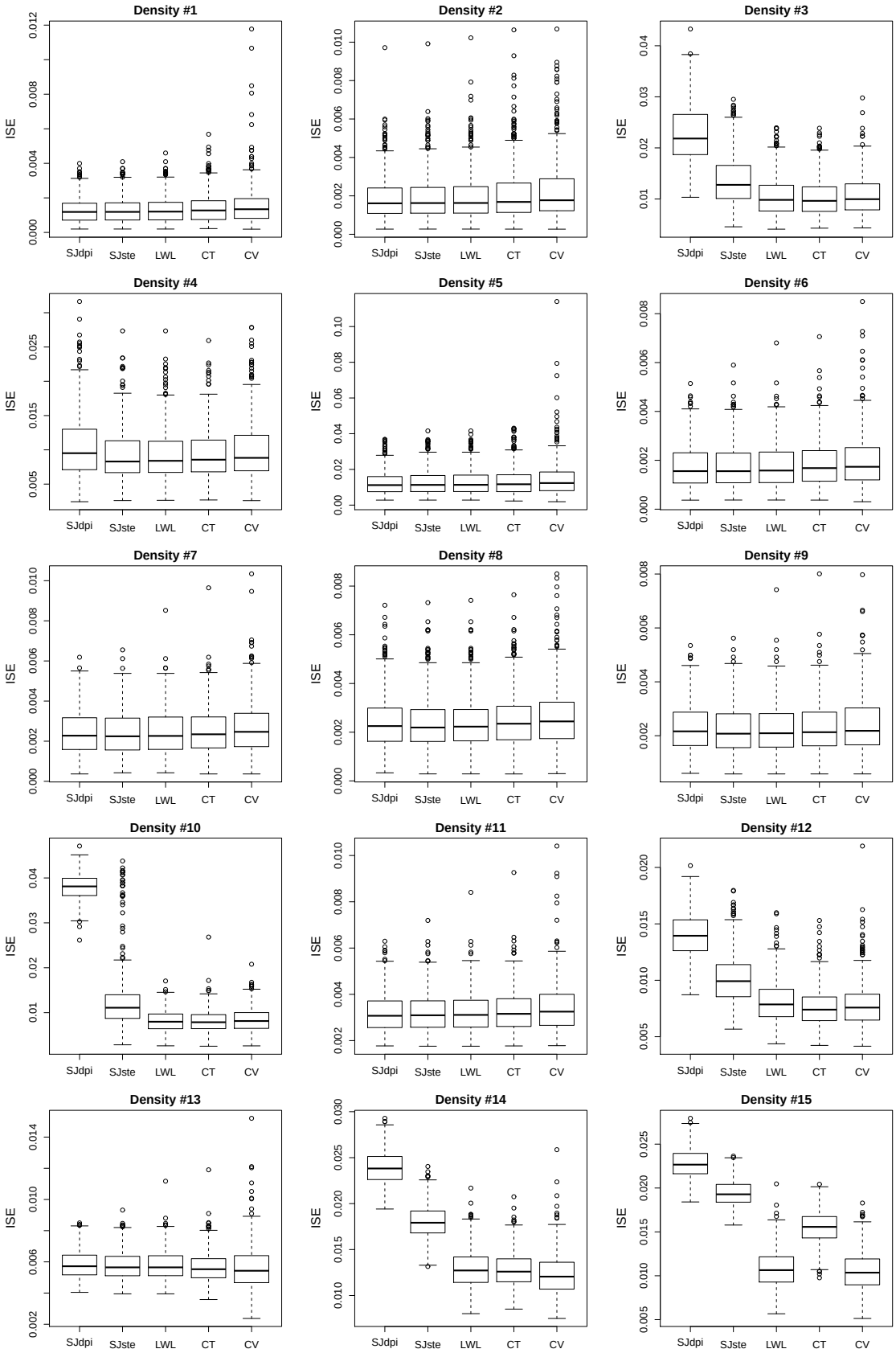

Figure 5: ISE *distribution for the bandwidth selectors methods* SJdpi, SJste, LWL, CT *and* CV ($n = 800$).

strong dependence of the SJdpi method on the normal reference distribution explains its relatively weak performance for the densities of group 2. The methods SJste and LWL, especially the latter one, are shown to be much more robust against the use of the normal density as reference distribution.

The new CT procedure presents a good overall performance for a wide range of density features: it is slightly more variable than the considered plug-in methods for densities of group 1, but not so variable as cross-validation, and it improves over both SJ methods for densities of group 2. This robust behavior, which is shared with the LWL method, is relevant for real data situations, where there is usually little prior information on the underlying density shape. This indicates that the new CT procedure should produce reliable bandwidths for most practical scenarios.

6 Proofs

We first obtain a non-asymptotic result that describes the behavior of the multistage plug-in bandwidths $\hat{h}_{\mathrm{PI},\ell}$ as a function of ℓ. Recall that the standard Gaussian density is used as the kernel for the estimator (3).

Lemma 1. *If for fixed $r \in \{0, 2, \dots\}$ and $\ell \in \{0, 1, \dots\}$ the sample $\mathscr{X} = \{X_1, \dots, X_n\}$ is such that $|\hat{\psi}_{r+2\ell,1}| \geq |\hat{\psi}^{\mathrm{NR}}_{r+2\ell}|$ then $|\hat{\psi}_{r,\ell+1}| \geq |\hat{\psi}_{r,\ell}|$. Therefore, if $\mathscr{X}$ is such that $|\hat{\psi}_{4+2\ell,1}| \geq |\hat{\psi}^{\mathrm{NR}}_{4+2\ell}|$, for all $\ell = 0, 1, \dots, \bar{\ell}$, then $\hat{h}_{\mathrm{PI},\bar{\ell}} \leq \hat{h}_{\mathrm{PI},\bar{\ell}-1} \leq \cdots \leq \hat{h}_{\mathrm{PI},2} \leq \hat{h}_{\mathrm{PI},1} \leq \hat{h}_{\mathrm{PI},0}$.*

Proof: For $r = 0, 2, \dots$, denote

$$\varphi_r(t) = \left(\frac{r!}{2^{(r-1)/2}(r/2)!\sqrt{\pi}\, n\, t} \right)^{1/(r+3)}, \quad t > 0,$$

so that we can write $g_{0,r} = \varphi_r(|\psi_{r+2}|)$ for the AMSE-optimal bandwidth of the kernel estimator (3). The ℓ-stage plug-in estimator $\hat{\psi}_{r,\ell}$ of ψ_r which involves the estimation of the ℓ density functionals $\psi_{r+2(\ell-1)}, \psi_{r+2(\ell-2)}, \dots, \psi_r$, can be written in a recursive way in terms of the i-stage plug-in estimators $\hat{\psi}_{r+2i,\ell-i}$ of ψ_{r+2i}, for $i = 1, \dots, \ell - 1$:

$$\begin{aligned} \hat{\psi}_{r,\ell} &= \hat{\psi}_r(\varphi_r(|\hat{\psi}_{r+2,\ell-1}|)), \\ \hat{\psi}_{r+2,\ell-1} &= \hat{\psi}_{r+2}(\varphi_{r+2}(|\hat{\psi}_{r+4,\ell-2}|)), \\ &\vdots \end{aligned}$$

$$\begin{aligned} \hat{\psi}_{r+2(\ell-2),2} &= \hat{\psi}_{r+2(\ell-2)}(\varphi_{r+2(\ell-2)}(|\hat{\psi}_{r+2(\ell-1),1}|)), \\ \hat{\psi}_{r+2(\ell-1),1} &= \hat{\psi}_{r+2(\ell-1)}(\varphi_{r+2(\ell-1)}(|\hat{\psi}^{\mathrm{NR}}_{r+2\ell}|)). \end{aligned}$$

Therefore,

$$|\hat{\psi}_{r,\ell}| = \Psi_r(\Psi_{r+2}(\ldots(\Psi_{r+2(\ell-1)}(|\hat{\psi}^{\mathrm{NR}}_{r+2\ell}|))))$$

and also

$$|\hat{\psi}_{r,\ell+1}| = \Psi_r(\Psi_{r+2}(\ldots(\Psi_{r+2(\ell-1)}(|\hat{\psi}_{r+2\ell,1}|)))),$$

where $\Psi_s = |\hat{\psi}_s| \circ \varphi_s$, for $s = 0, 2, \ldots$, is a function depending on the sample $\mathscr{X}$. Since $\mathscr{X}$ is such that $|\hat{\psi}_{r+2\ell,1}| \geq |\hat{\psi}^{\mathrm{NR}}_{r+2\ell}|$, and φ_s is a strictly decreasing function, in order to conclude it is enough to prove that $g \to |\hat{\psi}_s|(g)$ is a decreasing function. Using the positive-definiteness of $(-1)^{s/2}\phi^{(s)}$ we get $|\hat{\psi}_s|(g) = (-1)^{s/2}\hat{\psi}_s(g)$ for all $g > 0$, and then

$$\frac{d|\hat{\psi}_s|}{d\,g}(g) = -\frac{1}{n^2 g^{s+2}} \sum_{i,j=1}^{n} W\left(\frac{X_i - X_j}{g}\right) \leq 0,$$

for all $g > 0$ since $W(t) = (-1)^{s/2}((s+1)\phi^{(s)}(t) + t\phi^{(s+1)}(t))$ is also a positive-definite function on the real line, as it is the Fourier transform of $x \to x^{s+2}\phi(x)$.

□

Proof of Theorem 2: Taking into account that $\hat{\psi}_{4+2\ell,1} = \psi_{4+2\ell}(1 + o_P(1))$ (see Tenreiro, 2003) and $\hat{\psi}^{\mathrm{NR}}_{r+2\ell} = \psi^{\mathrm{NR}}_{r+2\ell}(\sigma_f)(1 + o_P(1))$, Theorem 2 follows easily from Lemma 1.

□

Proof of Theorem 3: Let us denote $\xi_{\hat{\ell}} = \hat{h}_{\mathrm{PI},\hat{\ell}}/h_0 - 1$, $\xi_{\underline{\mathrm{L}}} = \hat{h}_{\mathrm{PI},\underline{\mathrm{L}}}/h_0 - 1$ and $\xi_{\overline{\mathrm{L}}} = \hat{h}_{\mathrm{PI},\overline{\mathrm{L}}}/h_0 - 1$. For $\underline{\mathrm{L}} \geq 1$, the stated probability orders of convergence follow easily from Theorem 1 and the inequality $P(n^\alpha|\xi_{\hat{\ell}}| > M) \leq P(n^\alpha|\xi_{\underline{\mathrm{L}}}| > M) + P(n^\alpha|\xi_{\overline{\mathrm{L}}}| > M)$, which is valid for all $\alpha > 0$ and $M > 0$. Writing $\Omega_{\mathrm{L}} = \{\hat{h}_{\mathrm{PI},\overline{\mathrm{L}}} \leq \hat{h}_{\mathrm{PI},\underline{\mathrm{L}}}\}$, by Theorem 2 we have $P(\Omega_{\mathrm{L}}) \to 1$, as n goes to infinity. Therefore, the stated asymptotic normality follows from Theorem 1 since $\xi_{\overline{\mathrm{L}}} \leq \xi_{\hat{\ell}} \leq \xi_{\underline{\mathrm{L}}}$ for a sample in Ω_{L}, and $n^{5/14}\xi_{\overline{\mathrm{L}}}$ and $n^{5/14}\xi_{\underline{\mathrm{L}}}$ are both asymptotically normal $N(0, \sigma^2_{\mathrm{PI}})$, whenever $\underline{\mathrm{L}} \geq 2$.

□

Acknowledgments. The authors would like to thank the reviewers for the comments and suggestions. Part of this work was developed while the first author was visiting the second one at the University of Coimbra. He is most grateful to the CMUC (Centre for Mathematics, University of Coimbra) for partially funding his visit. Also, both authors have been partially supported by the Spanish Ministerio de Ciencia y Tecnología project MTM2010-16660 and C. Tenreiro was funded by the CMUC/FCT.

References

Aldershof, B. (1991) *Estimation of Integrated Squared Density Derivatives.* Ph.D. thesis, University of North Carolina, Chapel Hill.

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. *Comput. Statist. Data Anal.*, **17**, 153–176.

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

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

Hall, P. and Marron, J.S. (1987) Estimation of integrated squared density derivatives. *Statist. Probab. Lett.*, **6**, 109–115.

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

Hall, P. and Marron, J.S. (1988) Choice of kernel order in density estimation. *Ann. Statist.*, **16**, 161–173.

Jones, M.C. (1991) The roles of ISE and MISE in density estimation. *Statist. Probab. Lett.*, **11**, 511–514.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996) Progress in data-based bandwidth selection for kernel density estimation. *Comput. Statist.*, **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. *Statist. Probab. Lett.*, **11**, 511–514.

Liao, J.G., Wu, Y. and Lin, Y. (2010) Improving Sheather and Jones' bandwidth selector for difficult densities in kernel density estimation. *J. Nonparametr. Stat.*, **22**, 105–114.

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

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

Park, B.U. and Marron, J.S. (1992) On the use of pilot estimators in bandwidth selection. *J. Nonparametr. Stat.*, **1**, 231–240.

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

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **53**, 683–690.

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

Simonoff, J.S. (1996) *Smoothing Methods in Statistics*. Springer, New York.

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

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

Terrell, G.R. (1990) The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.*, **85**, 470–477.

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