Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 08–59

A DATA-BASED METHOD FOR CHOOSING THE NUMBER OF PILOT STAGES FOR PLUG-IN BANDWIDTH SELECTION

JOSÉ E. CHACÓN AND CARLOS TENREIRO

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 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.
AMS Subject Classification (2000): Primary 62G05, Secondary 62G07, 62G20.

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 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) that the performance of this estimator depends strongly on the choice of h. In this sense,

Received December 12, 2008.
This research has been partially supported by the Spanish Ministerio de Ciencia y Tecnología project MTM2006-06172 (first author) and by the CMUC (Centre for Mathematics, University of Coimbra)/FCT.

the so-called optimal bandwidth h_{MISE} is the minimizer of the mean integrated squared error function, $\text{MISE}(h) = \mathbb{E}[\text{ISE}(h)]$, where $\text{ISE}(h) = \int \{f_{nh}(x) - f(x)\}^2 dx$. Chacón *et al.* (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 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, 1987a; Jones and Sheather, 1991), where in this case the kernel L and the bandwidth g may be different from K and h. We say that L is a kernel of order ν if $\mu_p(L) = 0$ for $p = 1, 2, \ldots, \nu - 1$ and $\mu_\nu(L) \neq 0$. A method for constructing a kernel G_ν of arbitrary even order ν based on the Gaussian one, $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$, is given in Wand and Schucany (1990). Such a class of higher-order kernels is the one that will be used henceforth for the estimation of ψ_r, so that $L = G_\nu$ in (3), with

$$G_\nu(x) = \sum_{s=0}^{\nu/2-1} \frac{(-1)^s}{2^s s!} \phi^{(2s)}(x).$$

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, $\text{MSE}(g) = \mathbb{E}[\{\hat{\psi}_r(g) - \psi_r\}^2]$. Under some additional assumptions on f it is possible to obtain an asymptotic representation of the MSE function, namely $\text{AMSE}(g)$, and the minimizer of this AMSE function is given by

$$g_{0,r} = \left| \frac{\nu! G_\nu^{(r)}(0)}{\mu_\nu(G_\nu) \psi_{r+\nu} n} \right|^{1/(r+\nu+1)} \tag{4}$$

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.

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. (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.

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.

Wand, M.P. and Schucany, W.R. (1990) Gaussian-based kernels. *Canad. J. Statist.*, **18**, 197–204.

José E. Chacón
Departamento de Matemáticas, Universidad de Extremadura, Spain
E-mail address: jechacon@unex.es
URL: http://kolmogorov.unex.es/∼jechacon/

Carlos Tenreiro
CMUC, Department of Mathematics, University of Coimbra, 3001–454 Coimbra, Portugal
E-mail address: tenreiro@mat.uc.pt
URL: http://www.mat.uc.pt/∼tenreiro/

(Jones and Sheather, 1991). 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 $\psi_{r+\nu}$, 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 $\psi_{r+\ell\nu}$. 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 $\psi_{r+\ell\nu}$ is given by $\hat{\psi}^{\mathrm{NR}}_{r+\ell\nu} = \psi^{\mathrm{NR}}_{r+\ell\nu}(\hat{\sigma})$ where $\hat{\sigma}$ denotes any standard deviation estimate.

(2) Estimate successively the ℓ density functionals

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

with a kernel estimator. The bandwidth $g = \hat{g}_{0,r+j\nu}$ used in the kernel estimator $\hat{\psi}_{r+j\nu}(g)$ is just the one given by (4), with the unknown functional $\psi_{r+(j+1)\nu}$ 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) observe 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) pose 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 behaviour 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. In Section 4 an extensive simulation study is carried out to compare the performance of the proposed method with the most recommended ones in the literature. All the proofs are deferred to Section 5.

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

Here we will 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 are equivalent, as long as they use $\ell \geq 2$ pilot stages. This is a well known result, which can be stated in the following way.

Theorem 1 (Tenreiro, 2003)**.** *Assume that f has bounded derivatives up to order $4+\ell\nu$ and there exists $\sigma_f \neq 0$ such that $\hat\sigma - \sigma_f = O_P(n^{-1/2})$, where $\hat\sigma$ is the standard deviation estimate in the multistage procedure.*

a) *If $\nu = 2$ 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 $\ell \geq 2$.*

b) *If $\nu = 4$ then $\hat h_{\mathrm{PI},\ell}/h_0 - 1 = O_P(n^{-\alpha})$ with $\alpha = 4/9$ for $\ell = 1$ and $\alpha = 1/2$ for all $\ell \geq 2$.*

c) *If $\nu \geq 6$ then $\hat h_{\mathrm{PI},\ell}/h_0 - 1 = O_P(n^{-1/2})$ for all $\ell \geq 1$.*

The previous result justifies the usual recommendation of using $\ell = 2$ when $\nu = 2$ (Aldershof, 1991; 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 behavior 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

and also

$$|\hat\psi_{r,\ell+1}| = \Psi_r(\Psi_{r+\nu}(\dots(\Psi_{r+(\ell-1)\nu}(|\hat\psi_{r+\ell\nu,1}|)))),$$

where $\Psi_s = |\hat\psi_s| \circ \varphi_s$, for $s = 0, 2, \dots$, is a function depending on the sample $\mathcal{X}$. Since $\mathcal{X}$ is such that $|\hat\psi_{r+\ell\nu,1}| \geq |\hat\psi^{\mathrm{NR}}_{r+\ell\nu}|$, 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 $(-1)^{s/2}G_\nu^{(s)}$ 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 $W(t) = (-1)^{s/2}((s+1)G_\nu^{(s)}(t) + tG_\nu^{(s+1)}(t))$ is also a positive-definite function on the real line, as it is the Fourier transform of $x \to x^{s+\nu}\phi(x)/(2^{\frac{\nu}{2}-1}(\frac{\nu}{2}-1)!)$.

□

Proof of Theorem 3: Writing $\Omega_{\mathrm{L}} = \{\hat h_{\mathrm{PI,L}} \leq \hat h_{\mathrm{PI,L-1}} \leq \dots \leq \hat h_{\mathrm{PI},2} \leq \hat h_{\mathrm{PI},1}\}$, from Theorem 1 and Lemma 1, we have $\mathrm{P}(\Omega_L) \to 1$ as n goes to infinity. The conclusion follows now easily from Theorem 1 since $\hat h_{\mathrm{PI,L}}/h_0 - 1 \leq \hat h_{\mathrm{PI},\hat\ell(\mathrm{L})}/h_0 - 1 \leq \hat h_{\mathrm{PI},1}/h_0 - 1$ for a sample in Ω_{L}.

□

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

Hall, P. and Marron, J.S. (1987a) 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.

where

$$n^{4/5}\sup_{h\in H_n}|r_n(h)| = o_P(1).$$

Consequently

$$\mathrm{ISE}(\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})}) = \min_{h\in H_n}\mathrm{ISE}(h) + o_P(n^{-4/5}).$$

Since

$$\min_{h\in H_n}\mathrm{ISE}(h) \geq \min_{h>0}\mathrm{ISE}(h),$$

in order to conclude it is enough to show that

$$n^{4/5}\min_{h>0}\mathrm{ISE}(h) = C + o_P(1),$$

for some positive constant C. This is a consequence of Theorem 2.2 of Hall and Marron (1987b) and Theorem 2 of Hall (1984). In fact, if $h_{\mathrm{ISE}} = \mathrm{argmin}_{h>0}\mathrm{ISE}(h)$ we have

$$\begin{aligned} n^{4/5}\min_{h>0}\mathrm{ISE}(h) &= n^{4/5}\mathrm{ISE}(h_{\mathrm{ISE}}) \\ &= n^{4/5}\mathrm{ISE}(h_{\mathrm{MISE}}) + O_P(n^{-1/5}) \\ &= 5R(K)^{4/5}\mu_2(K)^{2/5}R(f'')^{1/5}/4 + o_P(1). \end{aligned}$$

□

Proof of Lemma 1: For $r = 0, 2, \ldots$, denote

$$\varphi_r(t) = \left(\frac{\nu!|G_\nu^{(r)}(0)|}{|\mu_\nu(G_\nu)|nt}\right)^{1/(r+\nu+1)}, \quad t > 0,$$

so that we can write $g_{0,r} = \varphi_r(|\psi_{r+\nu}|)$ 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 $\psi_{r+(\ell-1)\nu}, \psi_{r+(\ell-2)\nu}, \ldots, \psi_r$, can be written in a recursive way in terms of the i-stage plug-in estimators $\hat{\psi}_{r+i\nu,\ell-i}$ of $\psi_{r+i\nu}$, for $i = 1, \ldots, \ell-1$:

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

Therefore,

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

to the 15 normal mixture densities in Marron and Wand (1992). In all cases we have set L to be the standard normal density, so that $\nu = 2$. 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.

The same picture was generated for sample sizes $n = 100$ and $n = 400$, but they will not be included here to save space. Nevertheless, we include in Table 1 the EISE-optimal number of stages for these three sample sizes along the 15 normal mixture densities considered. This shows that, with the only exception of density #15, there are not drastic changes in the EISE-optimal number of stages with the sample size. This supports the usual recommendation that the number of stages does not need to be chosen depending on the sample size.

Sample	Density number														
size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$n = 100$	0	0	10	7	0	2	3	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	10	6	0	2	3	4	4	15	2	12	4	17	26

TABLE 1. *EISE-optimal number of stages.*

In view of Figure 1 we can classify our 15 test densities into three groups:

(1) There is a group of densities for which 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 those densities whose shape is close to #1 (the normal one), like #2, or for those densities for which the sample size $n = 200$ is perhaps too small to try to estimate their more complicated features. The latter is the case for densities #5, #6, #8, #9, #11 and #13.

(2) There is another different group of densities for which using a multistage plug-in selector 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

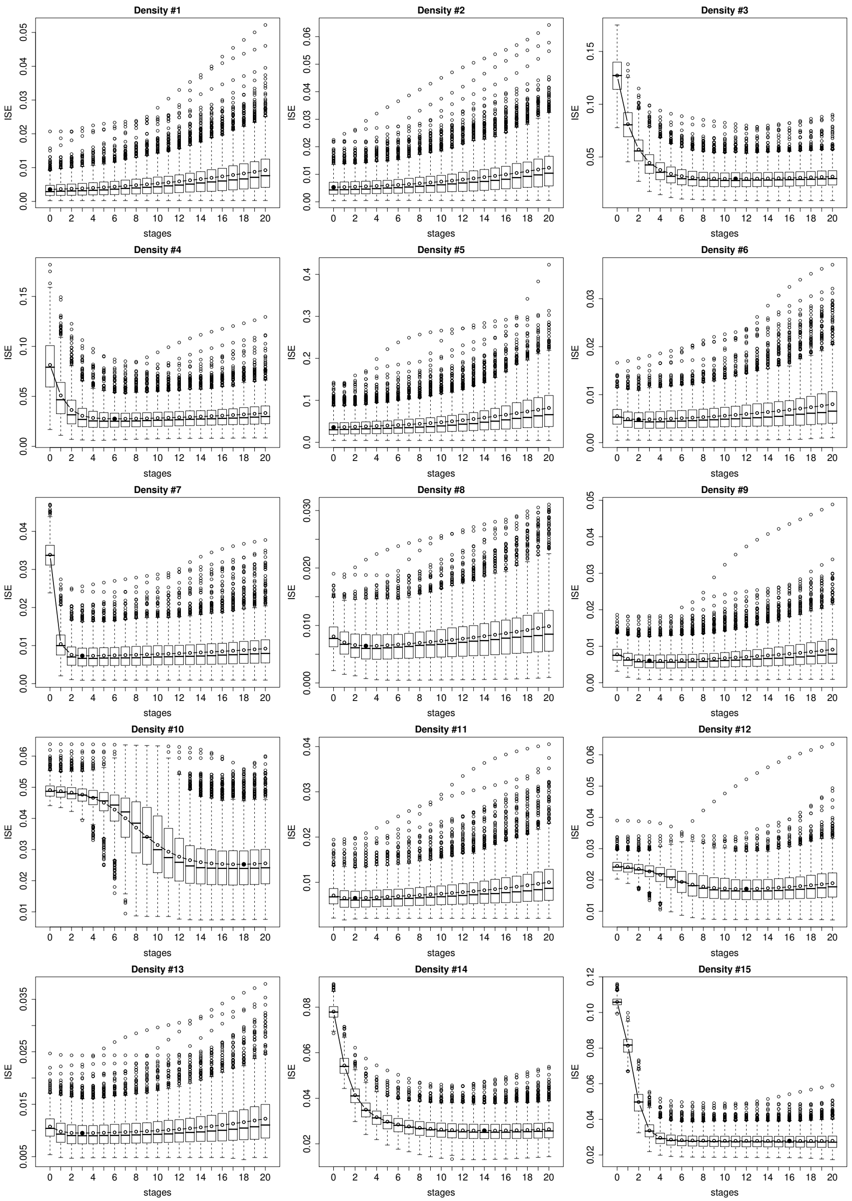

FIGURE 1. *Distribution of* $\mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})$ *depending on the number of stages* ($n = 200$).

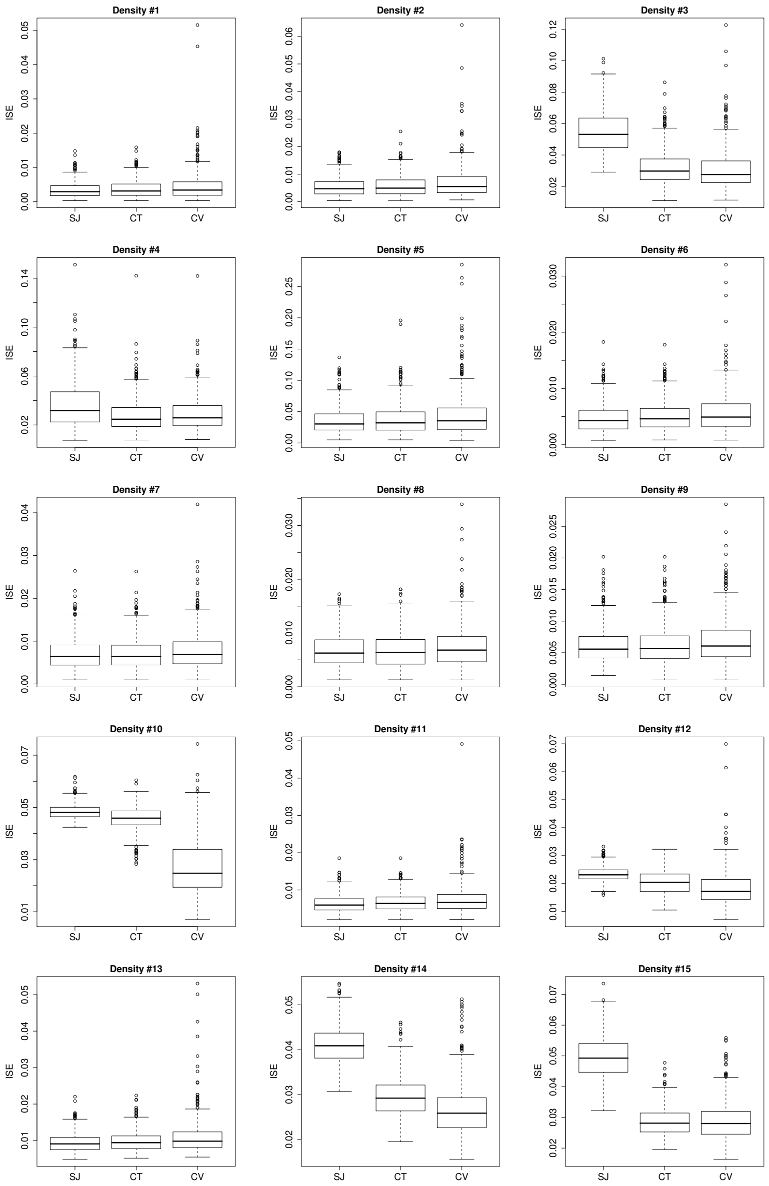

FIGURE 2. *ISE distribution for the bandwidth selectors* $\hat{h}_{\mathrm{PI},2}$, $\hat{h}_{\mathrm{PI},\hat{\ell}(5)}$ *and* $\hat{h}_{\mathrm{CV}}$ ($n = 200$).

proposed by Sheather and Jones (labelled SJ) and the least-squares cross-validation method (labelled CV). These two methods have been shown to provide quite reasonable results in practice; see Cao, Cuevas and González-Manteiga (1994) or Jones, Marron and Sheather (1996), and references therein.

Based on the observation that for most densities there seems not to represent a significant gain to consider more than 5 or 6 pilot stages, we will select the number of stages among $\{0, 1, \ldots, 5\}$ in the new CT proposal.

We will use as test densities the same 15 normal mixture densities that we referred to in Section 2. Based on 500 samples of size $n = 200$ for each test density in the study, we plot in Figure 2 the boxplots for the distributions of ISEs corresponding to each of the three bandwidth selection methods.

Due to the fact that the cross-validation choice is made on a set of "good bandwidths" in the case of the CT method, the reduction of the variability of the cross-validation procedure, as indicated in Theorem 3, is also clear from the practical point of view.

Besides, the CT bandwidth shows some kind of adaptive behavior, in the sense that its performance gets close to the one which is best between SJ and CV. For instance, for density #8 SJ is better than CV due to its smaller variability and CT is quite close to SJ (although slightly more variable), whereas for density #3, say, CV is clearly better than SJ because the latter is too biased and in this case CT gets quite close to CV. The exceptions to this general behavior are densities #10, #12 and #14, where the advantage of CV over CT is noticeable. However, this is not unexpected in view of Table 1, as for these hard-to-estimate densities maybe L = 5 can be regarded as being too few pilot estimation stages. Anyway, in all these cases CT is generally better than SJ. Therefore, if we were to make a single recommendation for bandwidth selection, we would take the new CT method.

5. Proofs

Proof of Theorem 2: Let $H_n = \{\hat{h}_{\mathrm{PI},\ell} : \ell = 0, 1, \ldots, \mathrm{L}\}$. From Theorem 1 we have

$$\mathrm{P}(H_n \subset [\epsilon n^{-1/5}, \lambda n^{-1/5}]) \to 1, \tag{8}$$

for some $0 < \epsilon < \lambda < \infty$. Therefore, from Theorem 1 of Hall (1983) we get

$$\mathrm{CV}(h) = \mathrm{ISE}(h) - \frac{2}{n}\sum_{i=1}^{n} f(X_i) + R(f) + r_n(h),$$

densities #3, #4, #7, #14 and #15. For these densities the EISE-optimal number of stages may be high but, in most of those cases, using such a high number of stages does not represent a significant gain over using, say, 5 or 6 stages.

(3) The third pattern that may be observed in Figure 1 corresponds to densities #10 and #12. In those cases, the EISE-optimal number of stages is high, but the distribution of the ISE does not show an abrupt descent in the first stages, as in the previous case.

The main conclusion after observing Figure 1 is that in some cases (specially for those densities in groups 2 and 3 above) the plug-in method may improve considerably if we allow for a higher number of stages than the usual advice $\ell = 2$.

We also performed some preliminary simulations to analyze the behavior of the plug-in selector using a pilot kernel of order $\nu = 4$ but, in common with other studies in the literature (see, e.g., Marron and Wand, 1992, or Jones, Marron and Sheather, 1996), despite its theoretical superiority there were no significant improvement in practice over the plug-in selector with $\nu = 2$. Therefore, we will not include higher-order kernels further in the simulations below, although we still allow higher-order kernels for the sake of completeness in our theoretical analysis.

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? If we fix a maximum number of pilot stages L, say, choosing a stage ℓ among the set of possible pilot stages $\{0, 1, \ldots, \mathrm{L}\}$ is naturally equivalent to choosing one of the bandwidths

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

for $\ell = 0, 1, \ldots, \mathrm{L}$, where $\hat{\psi}_{4,0} = \hat{\psi}_4^{\mathrm{NR}}$. Following Hall and Marron (1988) who used cross-validation as a method for choosing the kernel order for kernel density estimators, we propose here the same 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 an unbiased estimator of $\mathrm{MISE}(h) - R(f)$ given by

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

where $*$ denotes the convolution product. The least-squares cross-validation method involves choosing $h > 0$ to minimize $\mathrm{CV}(h)$. See Hall (1983)

and Stone (1984) for some optimal properties of the least-squares cross-validation bandwidth. Our proposal is to take for the number of pilot stages the value $\hat{\ell}(\mathrm{L}) = \hat{\ell}(X_1, \ldots, X_n; \mathrm{L})$ defined by

$$\hat{\ell}(\mathrm{L}) = \mathrm{argmin}_{\ell \in \mathcal{L}} \mathrm{CV}(\hat{h}_{\mathrm{PI},\ell}), \tag{6}$$

where $\mathcal{L} = \{0, 1, \ldots, \mathrm{L}\}$. This method for choosing the number of pilot stages leads to the data-based bandwidth $\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})}$, that can be seen as a hybrid between cross-validation and direct plug-in bandwidths.

In the following, as a consequence of the results of Hall (1983, 1984) and Hall and Marron (1987b), we establish the large sample optimality of the cross-validation method for choosing the number of pilot stages by proving that $\hat{\ell}(\mathrm{L})$ works as well as the 'optimal' choice for ℓ in the following asymptotic sense.

Theorem 2. *Under the conditions of Theorem 1 assume that:*

a) *K is a compactly supported, symmetric density function which is two-times differentiable with Hölder continuous second derivative, that is, $|K''(x) - K''(y)| \leq c|x - y|^\delta$, for some $c, \delta > 0$ and all $x, y \in \mathbb{R}$.*
b) *The distribution function F of f satisfies $\int (F(x)(1 - F(x)))^{1/2} dx < \infty$.*

Therefore

$$\frac{\mathrm{ISE}(\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})})}{\min_{\ell \in \{0,1,\ldots,\mathrm{L}\}} \mathrm{ISE}(\hat{h}_{\mathrm{PI},\ell})} \xrightarrow{P} 1.$$

Hereafter, we show that if we take $\mathcal{L} = \{1, \ldots, \mathrm{L}\}$ in (6), the data-based bandwidth $\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})}$ inherits the asymptotic rates of convergence of the multistage bandwidths $\hat{h}_{\mathrm{PI},\ell}$, $\ell = 1, \ldots, \mathrm{L}$, presented in Theorem 1, thus reducing the cross-validation variability and leading to better asymptotics. The key point to achieve this goal is given in the following non-asymptotic result which strongly depends on the Gaussian kernel family used in the multistage procedure. This result also gives some important insight into the finite sample behaviour of the ℓ-stage plug-in bandwidth as a function of ℓ described in Figure 1.

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

The next result is a direct consequence of Theorem 1 and the previous lemma. Although it is stated for the previously introduced cross-validatory choice of ℓ with $\mathcal{L} = \{1, \ldots, \mathrm{L}\}$, it is also valid for any other (measurable) rule $\hat{\ell}$ for choosing ℓ taking values in $\{1, \ldots, \mathrm{L}\}$. Also remark that better asymptotic rates of convergence could be obtained for $\nu \leq 4$ by restricting the choice of ℓ to the set $\{2, \ldots, \mathrm{L}\}$.

Theorem 3. *Under the conditions of Theorem 1 assume that f has bounded derivatives up to order $4 + \mathrm{L}\nu$ and*

$$|\psi_{4+\ell\nu}| \geq |\psi^{\mathrm{NR}}_{4+\ell\nu}(\sigma_f)|, \quad \textit{for all} \quad \ell = 1, 2, \ldots, \mathrm{L}. \tag{7}$$

For $\hat{\ell}(\mathrm{L})$ defined by (6) with $\mathcal{L} = \{1, \ldots, \mathrm{L}\}$ we have:

a) *If $\nu = 2$ then $\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})}/h_0 - 1 = O_P(n^{-2/7})$.*
b) *If $\nu = 4$ then $\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})}/h_0 - 1 = O_P(n^{-4/9})$.*
c) *If $\nu \geq 6$ then $\hat{h}_{\mathrm{PI},\hat{\ell}(\mathrm{L})}/h_0 - 1 = O_P(n^{-1/2})$.*

Although condition (7) is not very restrictive due to the smoothness properties of the normal distribution, it can be improved or even suppressed if for each $\ell = 1, 2, \ldots, \mathrm{L}$, 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 $s = 4 + \nu\ell$. Precisely, if we denote by $\psi^{\mathrm{BR}}_s \equiv \psi^{\mathrm{BR}}_s(\sigma)$ 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 (7) becomes $|\psi_{4+\ell\nu}| \geq |\psi^{\mathrm{BR}}_{4+\ell\nu}(\sigma_f)|$ for all $\ell = 1, 2, \ldots, \mathrm{L}$, 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^{BR}_s for even s are easy to obtain. In fact,

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

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.

4. Simulation study

We performed a simulation study to compare the new procedure (labelled CT) with the two most successful bandwidth selection methods in the literature, namely the direct plug-in method with fixed number of stages $\ell = 2$