Computational Statistics & Data Analysis 17 (1994) 153–176
North-Holland

A comparative study of several smoothing methods in density estimation

Ricardo Cao
Universidad de La Coruña, La Coruña, Spain

Antonio Cuevas
Universidad Autónoma de Madrid, Madrid, Spain

Wensceslao González Manteiga
Universidad de Santiago de Compostela, Santiago de Compostela, Spain

Received April 1992
Revised August 1992

Abstract: The theory of bandwidth choice in density estimation is developing very fast. Several methods (with plenty of varieties and subvarieties) have been recently proposed as an alternative to least squares cross-validation, the standard for years. This paper includes (a) A critical up-to-date review of the main methods currently available. The discussion provide some new insights on the important problem of estimating the minimization criteria and on the choice of pilot bandwidths in bootstrap-based methods. (b) An extensive simulation study of ten selected bandwidths. (c) A final discussion with some recommendations for practitioners. The conclusions are not easily summarized in a few words, because different cases have to be considered and important nuances must be pointed out. However, we could mention that the classical cross-validation bandwidths show, generally speaking, a relatively poor behavior (this is especially clear for the pseudo-likelihood method). On the other hand, although no selector appears to be uniformly better, the plug-in (in a similar version to that proposed by Sheather and Jones, J. Royal Statist. Soc. Ser. B 5 1991) and the (smoothed) bootstrap-based selectors show a fairly satisfactory performance which suggests that they could be the new standard methods for the problem of smoothing in density estimation. Interesting results are also obtained for a new type of bandwidths based on the number of inflection points.

Keywords: Bandwidth choice; Bootstrap bandwidths; Comparisons by simulation; Cross-validation bandwidths; Double kernel method; Fast Fourier Transform; Inflection points bandwidths; Kernel density estimators; Plug-in bandwidths.

Correspondence to: A. Cuevas, Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, Canto Blanco, 28049 Madrid, Spain.

SSDI 0167-9473(92)00066-D

1. Introduction

This paper is concerned with the problem of automatic (data-driven) choice of smoothing parameters in density estimation (see, e.g., Silverman, 1986, for introductory background). In spite of recent intensive research this issue remains very controversial and partially unsolved. Our aim is to provide a selective critical review of the main methods, including some very recent developments, and to compare them by means of an extensive simulation study. The current state of the problem seems to be particularly suitable for such a purpose, since the lines of research have evolved rapidly in the last two years and the present situation is rather confusing from the point of view of the users; today, there is no smoothing procedure widely accepted as a standard to be included in the software packages. After a period (say 1982–1988) of a clear preponderance of cross-validation methods (Rudemo, 1982; Bowman, 1984; Stone, 1984; Marron, 1985, 1987; Hall, 1987; Hall and Marron, 1987b; Scott and Terrell, 1987), the theory is currently expanding in several directions, which range from a revival of the classical plug-in bandwidths (Park and Marron, 1990; Hall and Marron, 1991; Sheather and Jones, 1991) to a strong development of bootstrap-motivated techniques (Taylor, 1989; Hall, 1990; Cao-Abad, 1990; Faraway and Jhun, 1990; Léger and Romano, 1990; Marron, 1990) simultaneously with the appearance of new procedures based on different ideas (Devroye, 1989; Cuevas and González-Manteiga, 1991).

In the next section we give an up-to-date review of the main methods and justify our choice of the bandwidths to be compared by simulation. The discussion includes new proposals as well as some remarks of theoretical interest which, as far as we know, have not been previously pointed out in the literature. In Section 3 we describe the structure of the simulation study and analyze its results. Some final remarks are included in Section 4. The numerical outputs are given in the Appendix.

2. A critical review of smoothing methods

We limit ourselves to the problem of estimating a univariate density function f by means of kernel estimators. Throughout this paper $\hat{f}(\cdot; h)$ will denote the usual kernel density estimator (Rosenblatt, 1956; Parzen, 1962),

$$\hat{f}(t; h) = \sum_{i=1}^{n} \frac{1}{n} K_h(t - X_i),$$

where $K_h(x) = h^{-1}K(x/h)$ and the kernel K is the standard Gaussian density, unless otherwise stated. The notation, $\hat{f}_h$ will be used, when convenient, instead of $\hat{f}(\cdot; h)$.

The subject we consider is the choice of the bandwidth (or smoothing parameter) h. Of course, in what follows h will generally depend on the sample

size n and on the sample values $X_1, \ldots, X_n$, but this dependence will not be made explicit, unless necessary.

Some notation: $c_K = \int K^2(t)\,dt$, $d_K = \int t^2 K(t)\,dt$, $MISE(h) = E\int(\hat{f}(t; h) - f(t))^2\,dt$, h_f will denote the (unobservable) optimum bandwidth which minimizes $MISE(h)$.

2.1. Cross-validation methods

There are essentially two varieties, pseudo-likelihood cross-validation (PL) and least squares cross-validation (LS). They are related by the use of a leave-one-out device in order to minimize an expression closely connected with some discrepancy measure between the estimate and the underlying density. We will also consider in this section the so-called *smoothed cross-validation*.

2.1.1. Pseudo-likelihood cross-validation

The PL method is based on the idea of considering the bandwidth h as a parameter which can be estimated by the maximum likelihood procedure. Since f is unknown, the likelihood function is estimated by

$$L(h) = \prod_{j=1}^{n} \hat{f}_j(X_j; h),$$

where $\hat{f}_j$ denotes the estimator $\hat{f}$ defined on the sub-sample $X_1, \ldots, X_{j-1}$, $X_{j+1}, \ldots, X_n$. The PL bandwidth is defined as the value h_{PL} maximizing $L(h)$. The basic idea of this method goes back to Habbema, Hermans and Van den Broek (1974) but the L_1-consistency of $\hat{f}(\cdot; h_{\mathrm{PL}})$ was established, under fairly general conditions, by Chow, Geman and Wu (1983) and Devroye and Györfi (1985). Hall (1987) proved that h_{PL} asymptotically minimizes, under some conditions, the Kullback–Leibler discrepancy between $\hat{f}$ and f. However, a major drawback of the PL method is its poor behavior in the heavy-tailed case (see Marron, 1985; Hall, 1987; Bowman, 1984, 1985). In this connection, a conclusive result has been proved by Broniatowski, Deheuvels and Devroye (1989) (see also Broniatowski, 1986): $\int |\hat{f}(t; h_{\mathrm{PL}}) - f| \to 0$ a.s. (in prob.) iff the extreme order statistics $X_{(1)}$ and $X_{(n)}$ are strongly (weakly) stable. Since many usual distributions (including those with heavier-than-exponential tails) do not satisfy this condition, the above result could be considered as a nearly definitive argument against the PL method. Moreover, Broniatowski, Deheuvels and Devroye (1989) prove also that, even in the case of normal tails, the speed of L_1-convergence is dissapointingly slow since $n^{\epsilon}\int |\hat{f}(t; h_{\mathrm{PL}}) - f| \to \infty$, *for every* $\epsilon > 0$.

Nevertheless, the PL bandwidth shows an interesting feature which, in our opinion, has not received enough attention: h_{PL} provides *universal consistency* (in L_1) with the only condition of compact support for f (see Devroye and Györfi, 1985, Ch. 6, Th. 4). Note, in particular, that no continuity or differentiability assumption is imposed on f. In the case of densities with non-compact support, Devroye and Györfi (1985, p. 153) suggest the simple idea of transform-

ing the data into the interval $[-1, 1]$ (by using a homeomorphism such that $x \to x/(1 + |x|)$), calculating h_{PL} for the transformed data, and transforming back the estimate obtained on $[-1, 1]$. The invariance of L_1 metric under monotone transformations ensures the consistency of this method *for all densities f*, which can be viewed as a property of robustness. We do not incorporate this idea in our study, since the choice of an adequate transformation would involve additional problems going beyond the scope of the present work. Anyway, we feel that h_{PL} has still some interest as a reference for comparisons (especially in the case of compact support).

2.1.2. Least-squares cross-validation

This method (Rudemo, 1982; Bowman, 1984; Stone, 1984) is perhaps better adapted to the usual L_2-framework in density estimation, since it is directly oriented to asymptotically minimize the mean integrated square error: the bandwidth h_{LS} is defined as the value minimizing the function

$$C(h) = \int \hat{f}^2(t; h)\, dt - 2n^{-1} \sum_{j=1}^{n} \hat{f}_j(X_j; h),$$

which is an unbiased estimate of $E\int(\hat{f}(t; h) - f(t))^2\, dt + \int f^2 = MISE(h) + \int f^2$. The asymptotic optimality of h_{LS} with respect to *MISE* has been established, under fairly general conditions, by Hall (1983) and Stone (1984). In addition to the intuitive appeal of this asymptotic optimality, the behavior of the LS bandwidth in the simulation experiments is slightly better than that of h_{PL}, especially in the heavy-tailed case. For these reasons h_{LS} has been in the last years the standard method of automatic smoothing in density estimation; it is included in our study as a classical reference. Nevertheless, the general performance of h_{LS} is far from satisfactory. First, the asymptotic optimality is achieved at a very slow convergence rate in such a way that enormous sample sizes are required in order to obtain an approximate optimality. In fact, Hall and Marron (1987a) have shown that $(h_{LS} - h_f)/h_f = O_p(n^{-1/10})$. Second (and closely related with the first point), the statistic h_{LS} presents an undesirably high variability; this has been clearly corroborated by our numerical results (see also Hall and Marron, 1987a).

Scott and Terrell (1987) have proposed an interesting modified version of the LS cross validation bandwidth, defined as the value h_{ST} minimizing a different (biased) estimate of $MISE(h)$, given by

$$BC(h) = \frac{c_K}{nh} + \frac{d_K^2}{2n^2h} \sum_{i<j} \phi\left(\frac{X_i - X_j}{h}\right),$$

where $\phi(x) = \int K''(u)K''(u + x)\, du$.

To appreciate the similarities and differences between this method (henceforth called *biased cross-validation*) and least-squares cross-validation, let us note that $C(h)$ and $BC(h)$ are both U-statistics but with different kernels. On the other hand, whereas $C(h)$ is aimed at the estimation (up to a constant) of

$MISE(h)$, the expression of $BC(h)$ is rather motivated by the asymptotic $MISE$, ($AMISE$) whose dominant part is known to be (see, e.g., Silverman, 1986, p. 39)

$$\frac{c_K}{nh} + \tfrac{1}{4}d_K^2 h^4 \int f''(x)^2 \, dx.$$

Thus, the definition of $BC(h)$ incorporates the first (constant) *variance term* (which corresponds to $\int \mathrm{var}(\hat{f}_h)$) of $AMISE(h)$ as well as an estimator of the second term (*bias term*), involving an improved estimate of the curvature (different from the natural one $\int(\hat{f}'')^2$). See Section 3.2 in Scott and Terrell (1987) for details; see also Setting 2.1.3 and 2.3 below.

An important advantage of biased cross-validation lies in the fact that h_{ST} presents a lesser variability than h_{LS}: e.g., Scott and Terrell (1987) show that the ratio of the asymptotic standard deviations of $h_{LS} - h_f$ and $h_{ST} - h_f$ is about 4.98 for the triweight kernel.

Let us observe, however, that for many usual kernels ϕ is continuous and $\lim_{x\to\infty} x\phi(x) = 0$; in particular, this holds for the Gaussian kernel, where $\phi(x) = \frac{1}{32}\sqrt{\pi}\ (x^4 - 12x^2 + 12)\exp(-\frac{1}{4}x^2)$. In these case $\lim_{h\to 0^+} BC(h) = \infty$, and $\lim_{h\to\infty} BC(h) = 0$. It is not difficult to prove that the least squares function $C(h)$ shows the same behavior; the basic difference is that, whenever $4c_K - d_K^2\phi(0) > 0$ (this is the case for the Gaussian kernel), $BC(h) > 0$, for all $h \in (0, \infty)$. Hence $BC(h)$ *has no finite global minimum*. Nevertheless, the plots of the function $BC(h)$ typically show a *local minimum* which is in fact the value h_{ST} used in our simulation results.

2.1.3. Smoothed cross-validation

It is well-known that $MISE(h)$ can be expressed in the form

$$MISE(h) = \int \mathrm{var}\big(\hat{f}(t; h)\big)\, dt + \int E^2\big[\hat{f}(t; h) - f(t)\big]\, dt$$

As we have mentioned in 2.1.2, the dominant part of the variance term is given by $(nh)^{-1}\int K^2$ (thus, it does not depend on f). As for the second, bias term $B(h)$, a natural estimator is

$$\hat{B}(h) = \int \left[\left(\int K_h(x-t)\hat{f}_g(t)\, dt\right) - \hat{f}_g(x)\right]^2 dx,$$

where $\hat{f}_g$ is an auxiliary (*pilot*) estimator of f. We thus have an estimator of $MISE(h)$ given by

$$SC(h) = (nh)^{-1}\int K^2(t)\, dt + \hat{B}(h).$$

This suggests to define the *smoothed cross-validation bandwidth* as the value h_{SC} which minimizes $SC(h)$. This idea has been first proposed by Hall, Marron and Park (1992) and later developed (with a slight technical difference) by Jones, Marron and Park (1991).

The name *smoothed cross-validation* is a bit misleading, since no leave-one-out procedure is used; in a way, h_{SC} is rather related to the bootstrap-based

bandwidths considered in Section 2.2 below; our simulation results confirm this idea.

The most interesting point in connection with h_{SC} is that this bandwidth provides (relative) convergence rates to the optimum h_f which are much better than the typical one ($n^{-1/10}$) for cross-validation (Hall and Marron, 1987a; see also Jones and Kappenman, 1990). Hall and Marron (1991a) (see also Marron, 1991) have proved that the best attainable rate is $n^{-1/2}$. This optimum *root n* rate can be achieved by smoothed cross-validation. More precisely, Jones, Marron and Park (1991) (a similar approach has been used by Hall, Marron and Park, 1992) have shown

$$\frac{h_{SC}-h_f}{h_f} = O_p(n^{-1/2}),$$

where h_{SC} is the minimizer of $SC(h)$, $g = Cn^p h^m$ with $m = -2$, $p = -\frac{23}{45}$ and C is a constant having a complicated expression depending on f. An interesting point is that the optimum *root n* rate is achieved in this case without using higher order kernels (which present the unpleasant feature of taking negative values). In our simulation study we approximate the value of C by using a normal reference distribution $N(0, \hat{\sigma}^2)$. Other choices for p and m are also considered in Jones, Marron and Park (1991). These authors also suggest that choices of type $g = Cn^p h^m$ would be useful as well for other smoothing methods (plug-in, bootstrap) requiring the use of pilot bandwidths but, as we will see below, such a dependence between g and h is not advisable in the bootstrap-based methods (at least in the case $m > 0$).

2.2. Bootstrap-based procedures

This is a very active line of research in the current work on automatic smoothing. From a methodological point of view, there are two remarkable points: first, the bootstrap is used here for a purpose different to its standard use, which is the approximation of sampling distributions. Second, the bootstrap methodology arises here just as a simple motivation for defining some bandwidths; in fact, as we will see, no resampling procedure is required in order to implement some bootstrap bandwidths.

The basic idea behind bootstrap smoothing is very simple: we estimate $MISE(h)$ by a bootstrap version of the form

$$MISE_*(h) = E_* \int \left(\hat{f}^*(t; h) - \hat{f}(t; g)\right)^2 dt, \tag{1}$$

where E_* denotes the expectation with respect to the bootstrap sample $X_1^*, \ldots, X_n^*$, g is some pilot bandwidth, $\hat{f}(t; g)$ is a density estimate which depends on the original sample $X_1, \ldots, X_n$, and $\hat{f}^*(t; h)$ is an estimate, based on $X_1^*, \ldots, X_n^*$. Then, we choose the value h minimizing $MISE_*(h)$. The basic differences among the various versions of this bootstrap methodology lie on the

choice of the auxiliary window g and on the procedure (smoothed or not) for generating the resampled data $X_1^*, \ldots, X_n^*$.

2.2.1. Smoothed bootstrap without pilot bandwidth

To our knowledge, the first (published) bootstrap proposal is due to Taylor (1989). He takes $g = h$ in (1) and computes the exact value of $MISE_*(h)$ in the case that the bootstrap sample (smoothed bootstrap) is drawn from the density $\hat{f}(\cdot; h)$. The result is (for the Gaussian kernel),

$$MISE_*(h) = \frac{1}{2n^2 h(2\pi)^{1/2}} \left[\sum_{i,j} \exp\left\{ -\frac{(X_i - X_j)^2}{8h^2} \right\} - \frac{4}{3^{1/2}} \sum_{i,j} \exp\left\{ -\frac{(X_j - X_i)^2}{6h^2} \right\} + 2^{1/2} \sum_{i,j} \exp\left\{ -\frac{(X_j - X_i)^2}{4h^2} \right\} + n2^{1/2} \right]. \quad (2)$$

Taylor (1989) minimizes, using numerical methods, expression (2). He finds in this way a bandwidth h_{BT} which shows a fairly good behavior in the simulation studies reported in the paper. Observe that Taylor's proposal avoids the use of a pilot bandwidth. Also, no resampling is required in order to approximate $MISE_*(h)$, because this quantity is exactly known. However, it is worth mentioning that, as pointed out by Cao-Abad (1990), $MISE_*(h) \to 0$ as $h \to \infty$. Therefore, $MISE_*(h)$ is an unsuitable estimator of $MISE(h)$ for smoothing purposes, since it has no finite minimum. This drawback is not directly observed in the simulations because, in practice, the numerical search of the minimum is restricted to a finite interval. This explains also the low variablility of h_{BT} (see Cao-Abad, 1990): the value of h_{BT} obtained by numerical optimization tends to coincide with the upper extremum of the interval of search. If this interval is not too large h_{BT} could casually provide good efficiency results through a moderate oversmoothing, but if the interval is enlarged, the efficiency deteriorates rapidly. It is also possible that the function (2) has a local (necessarily not global) minimum, similarly to the case (see Subsection 2.1.2) of the Scott–Terrell estimate $BC(h)$.

A choice of type $g = Cn^p h^m$ (see Jones, Marron and Park, 1991) presents a similar drawback whenever $m > 0$.

2.2.2. Smoothed bootstrap with pilot bandwidth

Faraway and Jhun (1990) have considered a smoothed bootstrap procedure (the bootstrap sample is taken from $\hat{f}_g$) where g is chosen by least-squares cross-validation from $X_1, \ldots, X_n$. These authors do not use an exact expression for $MISE_*(h)$. They approximate $MISE_*(h)$ by resampling; that is, B bootstrap samples are drawn from $\hat{f}_g$ and $MISE_*(h)$ is approximated by $BMISE(h) =$

$B^{-1}\sum_{j=1}^{B}\int(\hat{f}^{*}_{(j)}(t;\,h)-\hat{f}(t;\,g))^2\,\mathrm{d}t$, where $\hat{f}^{*}_{(j)}(t;\,h)$ denotes the value of the estimator for the j-th bootstrap sample. Faraway and Jhun (1990) give no asymptotic result for the resulting bandwidth h_{BF} (defined as the value of h which minimizes $BMISE(h)$) but in the simulations reported by these authors, h_{BF} performs almost uniformly better than h_{LS}.

A different approach is given by Cao-Abad (1990). In this case, the bootstrap sample in (1) is also drawn from the pilot density estimate $\hat{f}_g$. The choice of the auxiliary bandwidth g follows, in a natural way, as a consequence of some asymptotic expansions for $MISE(h)$ and $MISE_*(h)$ which provide a very useful insight on the nature of the problem. In short, the basic results are (the proofs can be found in Cao-Abad, 1990, pp. 77–177):

(a) It can be shown that the exact expression of $MISE_*(h)$ is

$$MISE_*(h)=n^{-1}h^{-1}c_K-n^{-1}\int\left(\int K(u)\hat{f}_g(x-hu)du\right)^2\mathrm{d}x+n^{-2}g^{-6}\times\sum_{i,j}\int\left(\int K(u)\big(K_g(x-hu-X_i)-K_g(x-X_i)\big)\,\mathrm{d}u\right)\mathrm{d}x. \quad (3)$$

Again, it is important to note that no resampling is needed because the explicit knowledge of $MISE_*(h)$. We define the bootstrap bandwidth h_{BC} as the minimizer in h of the function (3).

(b) Under standard regularity assumptions (which we omit), the following expressions hold

$$MISE(h)=4^{-1}d_K^2h^4\int f''(x)^2\,\mathrm{d}x+n^{-1}h^{-1}c_K-n^{-1}\int f^2(x)\,\mathrm{d}x+\mathrm{O}(n^{-1}h^2)+\mathrm{O}(h^6), \quad (4)$$

and

$$MISE_*(h)=4^{-1}d_K^2h^4\int\hat{f}_g''(x)^2\,\mathrm{d}x+n^{-1}h^{-1}c_K-n^{-1}\int\hat{f}_g(x)^2\,\mathrm{d}x+\mathrm{O}_p\big(n^{-1}h^2(n^{-1}g^{-3}+1)\big)+\mathrm{O}_p\big(n^{-1}h^4(n^{-1}g^{-5}+1)\big)+\mathrm{O}_p\big(h^6(n^{-1}g^{-7}+1)\big)+\mathrm{O}_p\big(h^8(n^{-1}g^{-9}+1)\big). \quad (5)$$

(c) From (4) and (5) it is clear that the pilot bandwidth g should be chosen in order to minimize the mean square error (MSE) of $\int\hat{f}_g''(x)^2\,\mathrm{d}x$ when viewed as an estimator of $\int f''(x)^2\,\mathrm{d}x$. After very laborious calculations, it turns out that the dominant part of the value g minimizing this MSE is

$$\left[\left(\int K''(u)^2\,\mathrm{d}u\right)d_K^{-1}\left(\int f'''(x)^2\,\mathrm{d}x\right)^{-1}n^{-1}\right]^{1/7}, \quad (6)$$

which is of exact order $n^{-1/7}$. By replacing this expression in (5) we obtain

$$MISE_*(h) = 4^{-1}d_K^2 h^4 \int \hat{f}_g''(x)^2 \, dx + n^{-1}h^{-1}c_K$$

$$- n^{-1}\int \hat{f}_g(x)^2 \, dx + O_p(n^{-1}h^2)$$

$$+ O_p(h^6) + O_p(n^{2/7}h^8). \tag{7}$$

(d) Now, it can be shown that the values h_f and h_{BC} minimizing $MISE(h)$ and $MISE_*(h)$ respectively, satisfy

$$\frac{h_{BC} - h_f}{h_f} = O_p(n^{-5/14}), \tag{8}$$

and

$$n^{6/5}g^{9/2}c_1(h_{BC} - h_f) \to N(0, 1) \quad \text{(weakly)}, \tag{9}$$

where c_1 is a constant depending on f and K.

(e) An alternative estimator of the curvature $\int f''(x)^2 \, dx$ (and hence a different bootstrap estimator of $MISE(h)$) can be obtained by taking into account the following idea proposed by Hall and Marron (1987c) (see also Jones and Sheather, 1991): if we expand the summatory under the integral in the naive estimator of the curvature $\int \hat{f}_g''(x)^2 \, dx$, we have, after a change of variable,

$$\int \hat{f}_g''(x)^2 \, dx = n^{-1}g^{-5}\int K''(u)^2 \, du$$

$$+ n^{-2}g^{-6}\sum_{i \neq j}\int K''\left(\frac{x - X_i}{g}\right)K''\left(\frac{x - X_j}{g}\right) dx. \tag{10}$$

The first summand in the right-hand side of (10) (corresponding to the *diagonal terms* with $i = j$ in $(\sum_{i,j}(1/ng)K''((x - X_i)/g)^2)$) contributes to the estimation of the curvature with a positive constant bias which can be avoided by simply deleting these diagonal terms. We thus obtain a new estimator of the curvature given by

$$C(g) = n^{-2}g^{-6}\sum_{i \neq j}\int K''\left(\frac{x - X_i}{g}\right)K''\left(\frac{x - X_j}{g}\right) dx.$$

Then, a new estimator of $MISE(h)$, denoted by $MMISE_*(h)$, is defined from expression (3) where $\int f_g''(x)^2 \, dx$ is replaced by $C(g)$.

(f) The modified bootstrap bandwidth h_{BM} minimizing $MMISE_*(h)$ satisfies

$$\frac{h_{BM} - h_f}{h_f} = Op(n^{-4/13}) \tag{11}$$

and

$$n^{6/5}g^{9/2}c_1(h_{\mathrm{BM}} - h_f) \to \mathrm{N}(\tfrac{3}{2}, 1) \quad \text{(weakly)}. \tag{12}$$

In this case, the dominant part of the minimizer in g of the mean square error of $C(g)$ is

$$\left[\frac{\frac{1}{2}a \int f(x)^2 \, \mathrm{d}x \int \left(\int K''(u)K''(u+v) \, \mathrm{d}u \right)^2 \mathrm{d}v}{\left(nd_K \int f''(x)^2 \, \mathrm{d}x \right)^2} \right]^{1/13}.$$

By comparing (11) and (12) with (8) and (9) we observe a rather surprising fact: the bandwidth h_{BM}, which is based on an improved estimator of the curvature (we have just removed a constant bias), shows a worse performance that the bandwidth h_{BC}, based on the naive estimator. This behavior (which has been corroborated by our preliminary simulations) can be explained as follows: the positive constant bias arising in the estimator $\int \hat{f}_g(x)^2 \, \mathrm{d}x$ is useful in order to compensate another bias term which appears in the derivation of the asymptotic distribution. To be more precise, we obtain the result

$$n^{6/5}g^{9/2}c_1(h_{\mathrm{BC}} - h_f) - ng^{13/2}c_2 + g^{-1/2}c_3 \to \mathrm{N}(0, 1) \quad \text{(weakly)},$$

where c_1, c_2 and c_3 are constants depending on f and K. It turns out that the bias $ng^{13/2}c_2 + g^{-1/2}c_3$ vanishes for the optimum value of g given by (6), whereas the modified bootstrap bandwidth h_{BM} retains an asymptotic bias as indicated in (12). A similar phenomenon has been pointed out by Sheather and Jones (1991) in a different setting (see Subsection 2.3 below).

For these reasons we have preferred to include h_{BC}, rather than h_{BM}, in our simulation study. The unknown value $\int f'''(x)^2 \, \mathrm{d}x$ in (6) is replaced by the corresponding expression in $\mathrm{N}(0, \hat{\sigma}^2)$ whose variance is estimated from the data.

2.2.3. Non-smoothed bootstrap

Hall (1990) has proposed a non-smoothed approach for selecting the bootstrap samples. That is, the values X_i^* are drawn from the empirical distribution. In this case the choice $g = h$ is not suitable for the purpose of estimating $MISE(h)$ by $MISE_*(h)$ since it necessarily underestimates the bias (which can be a substantial component of $MISE(h)$). Indeed, as Hall (1990) points out, the bootstrap bias $B_*(h) = E_*(\hat{f}^*(t; h) - \hat{f}(t; h))$ vanishes identically. To avoid this problem he proposes to draw bootstrap samples $X_1^*, \ldots, X_m^*$ of size $m < n$, to take $h = g(m/n)^{1/5}$, and to minimize (with respect to g) the bootstrap MISE (1). More specifically, Hall suggests taking the local minimum g_0 nearest to $m^{-1/5}$. We will denote by h_{BH} the corresponding value of h, $h_{\mathrm{BH}} = g_0(m/n)^{1/5}$. Let us recall (see, e.g., Silverman, 1986) that the first order term of the optimum bandwidth h_f for kernels of order two is proportional to $n^{-1/5}$; this justifies the relationship $h = g(m/n)^{1/5}$. As for the choice of m, we will follow Hall's suggestion $m \simeq n^{1/2}$ in our simulation study.

2.3. Plug-in methods

Since Parzen's (1962) pioneering work it is known that the value of h_f which minimizes the *MISE* has (asymptotically) the form

$$h_0 = d_K^{-2/5} c_K^{1/5} \left(\int f''(t)^2 \, dt \right)^{-1/5} n^{-1/5}. \tag{13}$$

This suggests to define a data-driven bandwidth of the form

$$\hat{h} = d_K^{-2/5} c_K^{1/5} \hat{R}^{-1/5} n^{-1/5}, \tag{14}$$

where $\hat{R}$ is an estimator of the curvature $R(f'') := \int f''(t)^2 \, dt$. It should be noted that, given the asymptotic character of (13), the motivation of these type of *plug-in* bandwidths is mostly based on asymptotic arguments. This can be considered as a drawback from the point of view of practitioners.

In view of (14) it could be said that every estimator of the curvature provides a different plug-in bandwidth. Early versions of this idea can be found, for instance, in the classical book of Tapia and Thompson (1978). In the last two years the plug-in methodology has received considerable attention in the literature. Hall, Sheather, Jones and Marron (1991) have proved that the optimum convergence rate $n^{1/2}$ (see the discussion in Subsection 2.1.3 above) can be achieved by using plug-in bandwidths, but this approach requires the unappealing use of higher order kernels. In the case of nonnegative kernels, Park and Marron (1990) have studied a plug-in selector which shows an excellent performance in both simulation and theoretical aspects. It is based on estimating the curvature in a similar way to that commented in Subsection 2.2.2: that is, by deleting the diagonal non-stochastic terms (with $i = j$) in the summatory of the natural estimate $\hat{R} = \int f_g''(t)^2 \, dt$. In our simulation study we include an improved version of this plug-in bandwidth, proposed by Sheather and Jones (1991), whose definition is

$$h_{\mathrm{PI}} = c_K^{1/5} d_K^{-2/5} \left(\hat{S}(g) \right)^{-1/5} n^{-1/5}, \tag{15}$$

where $\hat{S}(g)$ is a kernel estimate of the curvature,

$$\hat{S}(g) = n^{-2} g^{-5} \sum_{i,j} K^{iv}\left(\frac{X_i - X_j}{g} \right),$$

and the auxiliary window g is defined by

$$\left(\frac{2K^{iv}(0)}{d_K} \right)^{1/7} \hat{T}^{-1/7} n^{-1/7},$$

$\hat{T}$ being an estimator of $T = \int f'''(t)^2 \, dt$. In our case $\hat{T}$ is defined in a parametric way, by assuming normality (single-stage approach), whereas Sheather and Jones (1991) use a nonparametric estimator (as those discussed in Hall and Marron,

1987c); this, in turn, requires the use of a further pilot bandwidth which is finally given by a normal scale model estimate (two-stage approach). In Jones and Sheather (1991) can be found an approximate expression (which is of order $n^{-5/7}$) for the Mean Square Relative Error [i.e., $E(h_{SJ}/h_0 - 1)^2$] of this two-stage plug-in bandwidth h_{SJ}.

The precise motivation of the selector (15) involves technical arguments which will not be repeated here. The basic idea has been already outlined in Subsection 2.2.2: it consists in re-incorporating the diagonal terms in the estimation of the curvature and using the bandwidth g to approximately cancel the (positive) bias due to the diagonal terms with the (negative) leading smoothing bias term. As Sheather and Jones (1991) claim, their plug-in bandwidth seems to show a very good performance in several aspects. For this reason we incorporate it (in a slightly simplified, single stage, version) to our study.

2.4. Methods based on qualitative assumptions. Other methods

2.4.1. The IP-method

In many cases it is not unrealistic to assume a previous knowledge of some qualitative, nonparametric information about the target density f. For instance, the number of modes or that of inflection points of f could be known in advance. Then, a natural idea would be to incorporate this information in the nonparametric estimator via the smoothing parameter; this would represent a good example of the *tailor-designed* techniques mentioned by Devroye (1987). So far, this idea has received little attention in the literature. Cuevas and González-Manteiga (1991) have considered the case in which the target density is *piecewise convex–concave* having q inflection points, where q is assumed to be known (at least in principle). According to the general idea indicated above, a natural data-driven bandwidth can be defined by

$$h_{IP} = \inf\{h\colon \hat{f}(\cdot\,; h) \text{ has at most } q \text{ inflection points}\}. \tag{16}$$

A closely related approach has been proposed earlier by Silverman (1981, 1983) for the problem of testing multimodality. In particular, this author has proved that (in the case of Gaussian kernel) the number of sign changes as t varies in $\partial^m \hat{f}(t; h)/\partial t^m$ is a right-continuous decreasing function of h; from this property, it is straightforward to obtain an algorithm for computing the value h_{IP} with any required precision.

Cuevas and González-Manteiga (1991) have proved the almost sure uniform consistency of $\hat{f}(\cdot\,; h_{IP})$ to f (provided that $h_{IP} \to 0$ a.s.). The case of heavy-tailed f leads to inconsistency since h_{IP} does not converge to zero. In practice, this problem can be avoided by restricting the search in (16) to a (large enough) finite interval of values for t. Recently, Mammen (1990b) (see also Mammen, Marron and Fisher, 1992) has shown that, under some standard conditions (which include the restriction to a finite interval), $h_{IP} = O_p(n^{-1/7})$. Hence $MISE(h_{IP}) = O_p(n^{-4/7})$, whereas the optimum order of convergence is

$O_p(n^{-4/5})$. This result seems somewhat paradoxical: it turns out that the use of additional information about f results in a loss of efficiency with respect to the optimum. In fact, the paradox is only apparent since it is well-known that the *MISE*-optimum bandwidth does not necessarily estimate the shape (defined by the derivatives) of the target density. The estimation of derivatives of higher order requires more and more oversmoothing with respect to the optimum (see Silverman, 1978). So, the loss in efficiency of h_{IP} can be considered as the price to be paid for obtaining *nice* estimates, oriented not only to efficiency (mainly in the supremum norm) but also to qualitative aims (estimation of inflection points and derivatives, bump-hunting,...). The numerical comparisons between h_{IP} and other smoothing methods should be understood from this perspective.

Cuevas and González-Manteiga (1991) have also proposed to extend the applicability of the IP-bandwidth by estimating the number q of inflection points in the case that this number is unknown. Our simulation study includes the bandwidths h_{IP}, where q is known, and h_{EP}, where q is estimated from the data; to be concrete, q is the number of the inflection points of the estimator $\hat{f}(\cdot; g)$, where g is the same pilot bandwidth used in the calculation of h_{BC}.

2.4.2. The double kernel method

This method, based on a radically new principle, has been proposed by Devroye (1989). It is especially adapted to the L_1-approach, which is itself a stimulating novelty, given the lack of papers about automatic L_1-smoothing (see, however, Hall and Wand, 1988). The basic idea is as follows: let $\hat{g}(\cdot; h)$ another kernel density estimate, based on the same data as $\hat{f}(\cdot; h)$, but with a different kernel, L. The smoothing parameter h_{DK} is chosen in order to minimize $\int |\hat{f}(t; h) - \hat{g}(t; h)| \, dt$. The main restriction imposed on the auxiliary kernel L is that its (generalized) characteristic function should not coincide with that of K in an open neighbourhood of the origin. In practice, this leads to take kernels K and L of a different order.

This method presents a number of interesting properties. First, it provides *L_1-universally consistent estimators* (see Subsection 2.1.1 above); second, the estimators $\hat{f}(\cdot; h_{\mathrm{DK}})$ fulfil a property of optimality in L_1 for a large class of "nice" densities; third, the quantity $\int |\hat{f}(t; h_{\mathrm{DK}}) - \hat{g}(t; h_{\mathrm{DK}})| \, dt$ is a good estimate of the L_1-error $\int |\hat{f}(t; h_{DK}) - f(t)| \, dt$. As a matter of fact, Devroye (1989) points out that the auxiliary estimator $\hat{g}(\cdot; h_{\mathrm{DK}})$ may be better than $\hat{f}(\cdot; h_{\mathrm{DK}})$ although, as we mention above, the kernel L is usually of order greater than 2 (provided that K is of order 2) and so $\hat{g}(\cdot; h_{\mathrm{DK}})$ could take negative values.

In the simulations below we study $\hat{f}(\cdot; h_{\mathrm{DK}})$. As for the kernels K, L, we use (following a Devroye's, 1989, suggestion) $K(x) = \frac{3}{4}(1 - x^2)^+$ ($A^+ \equiv$ positive part of A) and $L(x) = \frac{75}{16}(1 - x^2)^+ - \frac{105}{32}(1 - x^4)^+$. This is the only case in which we use a kernel K different from the Gaussian one. The reason is that the indicated pair (K, L) fulfils all the restrictions required for the consistency and optimality of the method, whereas the use of the Gaussian kernel will require a more complicated choice for L.

3. Simulations

3.1. General description of the study

We will compare by simulation the following bandwidths: h_{PL}, h_{LS}, h_{ST}, h_{SC}, h_{BH}, h_{BC}, h_{PI}, h_{IP}, h_{EP}, h_{DK}. The corresponding definitions and the motivations for this choice can be found in Section 2 above.

We will consider seven different populations: a standard normal (N(0, 1)), a beta distribution with parameters 18 and 12 ($\beta(18, 12)$), a mixture (denoted by 0.9N) consisting of a standard normal with probability 0.9 and a normal with zero mean and variance 4 with probability 0.1, a mixture giving probability $\frac{1}{2}$ to a standard normal and $\frac{1}{2}$ to a normal N(6, 2) (0.5N), a standard lognormal distribution (*logn*(0, 1)) and two Student-*t* distributions with 5 and 10 degrees of freedom (t_5 and t_{10}).

We first approximate the mean integrated square error

$$MISE(h) = E\int \left(\hat{f}(t, h) - f(t)\right)^2 \mathrm{d}t,$$

the mean integrated absolute error

$$MIAE(h) = E\int |\left(\hat{f}(t, h) - f(t)\right)| \,\mathrm{d}t,$$

and the mean uniform absolute error

$$MUAE(h) = E\left(\sup_t |\, \hat{f}(t, h) - f(t)\,|\right),$$

by drawing 500 random samples of size $n = 100$ and averaging the L_1, L_2 or L_∞ norm for 1000 equispaced different values of h ranging from 0.015 to 5.01. The first step of the program ends with the numerical minimization of these functions, providing some approximations for their minima; h_{MISE}, h_{MIAE} and h_{MUAE}.

At this point we should remark that the use of a fast Fourier transform (FFT) method, to evaluate the kernel density estimator, made the program run quite fast. In order to save CPU time, we use the FFT to compute efficiently the estimator based on 1000 different bandwidths for each fixed random sample. Later on, we average these "functions" in h over the 500 trials. All these computations are made twice: for the Gaussian kernel and for the one (Epanechnikov kernel) used in the double kernel procedure. Thus, two different functions and minimizers are obtained for each criterion.

The second (and last) step in the program consists of drawing 500 samples, selecting the bandwidth $\hat{h}$ according to every method and approximating by Monte Carlo the mean square errors of $M(\hat{h})$ with respect to the optimal value $\min_{h>0} M(h)$, where M is any of the functions *MISE*, *MIAE* and *MUAE*. These quantities provide a very expressive measure of the efficiency and variability of

the different bandwidths, by informing about how far their respective measure errors are (in average) of the minimum possible values for each criterion.

Most of the bandwidths under study are obtained by minimizing some function. In many cases these functions are approximated in a very efficient way, using a FFT algorithm similar to that presented in Silverman (1986) for h_{LS}. This is the case of h_{LS}, h_{ST}, h_{SC}, h_{BC} and h_{BH}, while h_{PL} and h_{DK} are computed using the FFT for the density estimator itself. Both inflection-points bandwidths use the FFT to calculate the second derivative of the estimator and then count the number of sign changes. The value $\hat{S}(g)$, used in h_{PI}, is also computed by FFT.

The numerical outputs are presented in nine tables included in the Appendix. The main results are in Tables 1–7 which give the values of $E[M(\hat{h}) - \min_{h>0} M(h)]^2$ for the seven target densities under study. Each table has three rows, with headings L_2, L_1 and L_∞, corresponding respectively to the cases where $M(h)$ is $MISE(h)$, $MIAE(h)$ and $MUAE(h)$. The headings PL, LS,... of the ten columns of each table obviously correspond to the different possible choices for $\hat{h}$: h_{PL}, h_{LS},...

Table 8 gives the values of the optimum bandwidths (using Epannechnikov and Gaussian kernels) with respect to each criterion. Finally, Table 9 provides the average bandwidths (over 500 trials) obtained with the different smoothing methods.

3.2. *What the numbers say: discussion and comments*

We are now faced with the problem of summarizing the numerical results (see the Appendix) in order to get a general perspective and draw some practical conclusions. In this respect, a natural idea is to elaborate some kind of ranking for the ten bandwidths from the results obtained in Section 3.2. The trouble is that there does not exist a unique, obvious criterion to rank the bandwidths; moreover, any ordering would necessarily represent a simplification overlooking some important features of the results. So, the rankings given below have only an orientative value and *by no means should be considered as definitive, formal classifications of the selectors*. In order to provide a more complete picture, we will give two rankings (A and B) for each distance (L_2, L_1 and L_∞), elaborated by using two different classification criteria. Ranking A has been prepared according to the following rules: for each distance the bandwidths are ordered taking into account the efficiency results given in Tables 1–7. For each table, the worst bandwidth gets 10 penalization points, the second worst selector gets 9 points and so on. Moreover when the efficiency of some bandwidth is 5^n ($n \geqslant 1$) times worse than that of the previous one, it gets n extra penalization points which are, of course, inherited by the subsequent bandwidths. The final score for each selector is the sum of those obtained in the seven target densities under study.

In Ranking B the scores do not indicate penalizations: the best bandwidth in each table gets 10 points, the second one, 9 and so on. Moreover, if the results

in the table are of orders 10^{p_0}, $10^{p_1}, \ldots, 10^{p_k}$, with $1 \leqslant k \leqslant 10$ and $p_0 > p_1 > \cdots > p_k$, those bandwidths with the order p_j get j extra points. As in Ranking A, the final score is the sum of those corresponding to the seven tables. However, unlike Ranking A, the higher scores correspond to the best performances.

The results are as follows (scores are indicated between parentheses):

Ranking A.
L_2 : BC(25) PI(30) SC(32) EP(32) ST(33) BH(39) DK(44) LS(51) IP(57) PL(75)
L_1 : PI(26) ST(29) BC(32) EP(34) SC(34) LS(39) DK(40) BH(41) IP(63) PL(66)
L_∞ : EP(23) BH(27) IP(30) BC(31) PI(38) SC(39) ST(41) DK(44) LS(54) PL(69)

Ranking B.
L_2 : BC(71) PI(67) SC(63) ST(63) EP(63) BH(57) DK(49) LS(39) IP(32) PL(20)
L_2 : PI(66) ST(64) BC(60) SC(58) EP(57) DK(51) LS(51) BH(51) IP(24) PL(21)
L_∞ : EP(66) BH(62) IP(58) BC(56) PI(48) SC(47) ST(47) DK(42) LS(30) PL(21)

Both rankings are surprisingly similar and lead basically to the same conclusions. The results of these rankings as well as the direct inspection of the tables given in the Appendix, suggest the following general remarks:

(*a*) The selector h_{PL} is the only one showing a consistently bad behavior. Of course this is not new (see Subsection 2.1.1 above) but it is perhaps a little surprising the extent of the differences between h_{PL} and the remaining bandwidths.
(*b*) Roughly speaking, all the selectors show a similar performance for the metrics L_1 and L_2, whereas L_∞ presents some pecularities.
(*c*) Taking into account the efficiency as well as the "reliability" (regular behavior for different densities and criteria), the selectors h_{BC} and h_{PI} turn out to be the best with respect to the integral metrics L_1 and L_2. In the second place, there is a group of four bandwidths (h_{SC}, h_{ST}, h_{DK} and h_{EP}) having very similar performances. Occasionally, h_{LS} and h_{BH} provide good results as well but they show a more irregular behavior.
(*d*) As far as the supremum norm is concerned, the most remarkable fact is the good performance of the new selector h_{EP}.

In order to make more detailed comments it is useful to classify the seven target densities under study into three groups: (i) Symmetric (or almost symmetric) thin-tailed densities: standard normal, mixture 0.9N and $\beta(18, 12)$; (ii) Asymmetric densities with thin (or medium) tails: mixture 0.5N and lognormal; (iii) Heavy-tailed densities: t_5 and t_{10}.

3.2.1. Cross-validation

While the PL method has become discredited years ago, h_{LS} still remains as the standard selector in density estimation. In view of the results of the present study and the most recent developments of the theory, such a privileged position

does not seem longer justified. The results reflected in Tables 1–7 are uneven: the behavior of h_{LS} is quite satisfactory for densities of group (ii), but really bad in groups (i) and (iii). The global performance of h_{LS}, as given in Rankings A and B, is rather poor.

The results for h_{ST} are clearly better but still uneven. The efficiency of h_{ST} is moderate or high for densities of groups (ii) and (iii) and low in group (i).

The smoothed cross-validation selector h_{SC} shows a fairly good behavior. Its results are quite close to those of the smoothed bootstrap bandwidth h_{BC}; this is not very surprising since h_{SC} is the minimizer of $SC(h)$ (see Subsection 2.1.3) which can be viewed as a bootstrap estimator of $MISE(h)$ of type *Variance + Bootstrap estimator of bias*. Interestingly, the efficiency of h_{SC} is always worse than that of h_{BC} with the only exception of the L_1-results for the densities t_5 and t_{10} (Tables 6 and 7). We could point out two possible reasons to account for the superiority of h_{BC}. First, the performance of h_{SC} for finite sample size could be perhaps damaged by a suboptimal choice of the constant C in the pilot bandwidth $g = Cn^p h^m$; other possibilities could be tried. Second, the estimator $MISE(h_{BC})$ could be better than $MISE(h_{SC})$ even in the case that h_{SC} were preferable to h_{BC} as an estimator for the MISE-optimum bandwidth; this would happen, for instance, if the bias of h_{BC} were lesser on the side (right or left) in which the U-shaped function $MISE(h)$ increases more rapidly.

3.2.2. Plug-in

The good performance of the plug-in bandwidth h_{PI} arises clearly as one of the most remarkable results of our study. Similar conclusions have been previously reported by Park and Marron (1990) and Sheather and Jones (1991). The results of h_{PI} are quite homogeneous for the different densities and especially good for those of types (ii) and (iii). Obviously, the basic point in this recuperation of the old plug-in methodology has been a careful choice of the pilot bandwidth g. In this connection there is probably considerable room for research and improvement.

3.2.3. Bootstrap bandwidths

If we had to declare an only bandwidth as the "winner" in our simulation study, the choice would probably correspond to the smoothed bootstrap bandwidth h_{BC}. It seems particularly suitable for the L_2 criterion, with densities of type (iii) but, in general, the most attractive feature of this selector is the reliability; observe that, in the worst case, h_{BC} is placed in an intermediate position in the partial rankings for each target density.

On the contrary, the bandwidth h_{BH}, based on non-smoothed bootstrap, has an uneven performance, showing poor results (except for the supremum norm) in the cases of medium or heavy-tailed densities (groups (ii) and (iii)). However, h_{BH} is clearly the winner in the group (i).

By the way, let us remark that our results suggest an interesting conclusion in the context of bootstrap theory: the smoothed bootstrap has proved to be clearly better than the standard, non-smoothed version of this technique for the important problem of bandwidth choice.

3.2.4. Double kernel

Although the performance of h_{DK} is not spectacular, it is good enough as to deserve further research (using different sample sizes and other target densities). In general, h_{DK} is a reliable bandwidth which seems to be especially competitive in the estimation of *difficult* (i.e., discontinuous, non-smooth,...) target densities. In Quintela (1992) can be found further simulation results for exponential and double exponential densities in which h_{DK} shows a fairly satisfactory behavior (especially in the exponential case).

3.2.5. IP-bandwidths

As we indicate in Subsection 2.4.1, any comparison between h_{IP} and the remaining selectors should take into account that the IP method sacrifies a certain amount of efficiency in order to get some qualitative aims. Bearing this in mind, the global performance of h_{IP} can be described as reasonably good, especially with respect to the supremum norm.

The case of h_{EP} is quite different: this selector does not use prior information nor imposes shape restrictions on the estimators. Hence, it can be directly compared on equal terms with the other bandwidths. We believe that the results obtained by h_{EP} are one of the most stimulating contributions of this work. This selector is the best for the supremum norm and also behaves quite satisfactorily with respect to the integral norms.

4. Final remarks

4.1. Other simulation studies

Other comparative studies on bandwidth choice in density estimation are Park and Turlach (1992) and Jones, Marron and Sheather (1992). In spite of the different setups of these papers and the present work, some general common conclusions can be drawn. Thus, the classical least squares cross-validation window is outperformed by other bandwidths and the plug-in methodology seems to offer an interesting alternative.

4.2. Cross-validation revisited

Some interesting modifications of least squares cross-validation have been recently proposed: Chiu (1991) suggests a modified cross-validation window with a lesser variability; the Fourier transform methodology used by this author offers a valuable tool for bandwidth selection.

Stute (1992) proposes a new bandwidth defined as the minimizer of a biased estimator of $MISE(h)$, which is motivated via Hajek projections.

The proposal of Feluch and Koronacki (1992) is based on the idea of excluding the pairs, (X_i, X_j), of observations *too close* in the cross-validation estimate $C(h)$ (see Subsection 2.1.2 above).

The problem of local minima in the least squares cross-validation function has been considered by Hall and Marron (1991b).

Finally, let us mention the comprehensive study of Hall and Johnstone (1992) which provides new and deeper insights on the problem of automatic smoothing.

4.3. Hard densities

The case of estimating *hard* (multimodal, non-smooth, discontinuous,...) densities is excluded from our simulation study. In the paper by Jones, Marron and Sheather (1992) some of these densities are considered. Another interesting reference in this area is Wand, Marron and Ruppert (1991). Devroye's (1989) double kernel method arises as a very promising selector in the estimation of hard densities.

4.4. Sample size

In our simulations we have considered the (unique) sample size $n = 100$. This choice deserves a few words of explanation. First, nonparametric density estimation should be used, as a rule, in the case of moderate or large sample sizes, since its theoretical basis relies mostly on asymptotic results. Second, the magnitude of the sample size under study should be of common use in the practical applications with real data sets. We believe that our choice $n = 100$ fulfils reasonably both requirements. On the other hand, our partial results with $n = 50$ and $n = 1000$ lead to similar conclusions to those obtained for $n = 100$.

Appendix: Tables

Table 1
Standard normal

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	5.09E−6	3.80E−5	2.45E−6	1.60E−6	5.90E−8	8.74E−7	2.03E−6	6.86E−6	3.88E−6	7.13E−7
L_1	4.51E−4	1.99E−3	2.12E−4	1.35E−4	4.46E−6	7.01E−5	1.74E−4	5.62E−4	9.19E−4	6.14E−5
L_∞	3.34E−4	1.93E−3	2.80E−4	2.02E−4	1.43E−5	1.24E−4	2.45E−4	5.07E−4	6.47E−5	8.06E−5

Table 2
Beta distribution with parameters 18 and 12

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	7.53E−4	2.56E−3	8.20E−4	5.60E−4	4.14E−5	3.55E−4	7.00E−4	1.11E−3	1.76E−4	2.47E−4
L_1	4.56E−4	1.47E−3	4.91E−4	3.26E−4	1.62E−5	1.98E−4	4.15E−4	6.74E−4	4.23E−4	1.33E−4
L_∞	5.74E−2	1.94E−1	8.21E−2	6.16E−2	7.80E−3	4.26E−2	7.35E−2	8.40E−2	4.00E−3	2.46E−2

Table 3
Mixture 0.9N(0, 1)+0.1N(0, 4)

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	6.02E−5	2.57E−5	1.21E−6	6.90E−7	8.54E−8	3.35E−7	9.31E−7	5.62E−6	1.32E−4	2.18E−7
L_1	2.24E−3	1.78E−3	1.52E−4	8.83E−5	2.60E−5	4.26E−5	1.19E−4	5.88E−4	1.52E−2	3.25E−5
L_∞	1.21E−3	1.29E−3	1.46E−4	9.48E−5	4.37E−6	5.30E−5	1.21E−4	3.99E−4	1.34E−3	2.73E−5

Table 4
Mixture 0.5N(0, 1)+0.5N(6, 2)

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	1.69E−2	4.04E−6	2.62E−7	1.63E−5	1.83E−4	1.11E−5	3.23E−6	2.21E−6	4.84E−6	4.84E−6
L_1	5.25E−1	7.34E−4	8.83E−5	7.96E−3	5.67E−2	5.74E−3	1.98E−3	5.24E−4	2.28E−3	2.28E−3
L_∞	2.18E−1	1.53E−4	5.90E−6	2.38E−4	1.59E−3	1.69E−4	5.29E−5	1.03E−4	6.46E−5	6.46E−5

Table 5
Lognormal density

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	3.83E−3	1.65E−5	8.60E−4	2.14E−3	4.90E−2	1.40E−3	9.89E−4	1.84E−4	2.82E−2	4.29E−3
L_1	4.06E−2	6.54E−4	1.07E−2	2.17E−2	7.48E−1	1.31E−2	8.85E−3	8.28E−4	4.00E−1	4.84E−2
L_∞	5.61E−2	6.10E−4	6.16E−3	1.70E−2	1.18E−1	1.30E−2	1.05E−2	2.23E−3	9.07E−2	2.68E−2

Table 6
Student's t_5

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	1.17E−2	4.73E−6	5.33E−8	2.03E−8	1.29E−6	1.26E−8	2.86E−8	6.65E−7	3.23E−5	1.44E−7
L_1	7.35E−2	5.72E−5	2.96E−5	6.46E−5	1.22E−3	9.46E−5	3.98E−5	1.76E−4	1.91E−2	4.56E−4
L_∞	1.63E−1	4.91E−4	2.09E−4	1.80E−4	1.08E−4	1.64E−4	1.94E−4	1.89E−4	1.63E−4	1.27E−4

Table 7
Student's t_{10}

	PL	LS	ST	SC	BH	BC	PI	DK	IP	EP
L_2	1.58E−2	6.85E−6	1.81E−8	4.11E−9	1.84E−7	3.76E−9	7.37E−9	7.01E−7	1.18E−5	1.09E−7
L_1	9.26E−2	9.65E−5	5.72E−5	6.88E−5	6.10E−4	1.18E−4	5.34E−5	3.60E−4	6.32E−3	4.20E−4
L_∞	2.05E−1	4.37E−4	1.21E−4	1.07E−4	5.20E−5	9.17E−5	1.16E−4	1.14E−4	3.00E−5	6.85E−5

Table 8
Optimum bandwidths

		N(0, 1)	0.9N	Beta(18, 12)	0.5N	Log $n(0, 1)$	t_5	t_{10}
L_2	Gauss	4.40E−1	4.65E−1	4.00E−2	5.80E−1	1.75E−1	8.75E−1	7.45E−1
	Epan.	9.55E−1	1.01E+0	8.50E−2	1.255E+0	3.60E−1	1.805E+0	1.575E+0
L_1	Gauss	4.20E−1	4.55E−1	3.50E−2	5.60E−1	2.15E−1	5.65E−1	4.80E−1
	Epan.	9.15E−1	9.90E−1	8.00E−2	1.215E+0	4.45E−1	1.225E+0	1.05E+0
L_∞	Gauss	4.80E−1	5.00E−1	4.50E−2	6.10E−1	1.35E−1	2.315E+0	1.89E+0
	Epan.	1.04E+0	1.095E+0	9.50E−2	1.305E+0	2.75E−1	4.905E+0	4.02E+0

Table 9
Average bandwidths

	N(0, 1)	0.9N	Beta(18, 12)	0.5N	log $n(0, 1)$	t_5	t_{10}
PL	4.21E−1	5.57E−1	3.61E−2	1.83E−3	3.25E−1	2.62E−3	2.01E−3
LS	2.82E−1	3.24E−1	2.47E−2	4.73E−1	1.77E−1	6.89E−1	6.26E−1
ST	3.13E−1	3.58E−1	2.76E−2	5.52E−1	4.90E−1	7.72E−1	6.99E−1
SC	3.23E−1	3.71E−1	2.85E−2	1.01E+0	6.01E−1	8.31E−1	7.24E−1
BH	4.10E−1	4.88E−1	3.59E−2	1.46E+0	9.31E+0	1.17E+0	9.48E−1
BC	3.42E−1	3.94E−1	3.01E−2	9.62E−1	5.33E−1	8.71E−1	7.67E−1
PI	3.15E−1	3.62E−1	2.78E−2	8.52E−1	4.84E−1	7.98E−1	7.05E−1
DK	7.86E−1	8.50E−1	6.99E−2	1.08E+0	4.98E−1	1.76E+0	1.66E+0
IP	5.21E−1	7.58E−1	4.57E−2	7.87E−1	2.24E+0	1.98E+0	1.35E+0
EP	3.91E−1	4.48E−1	3.46E−2	7.87E−1	7.60E−1	1.02E+0	8.78E−1

Acknowledgements

We are indebted to Luc Devroye and Steve Marron for valuable comments. The suggestions of the referee and the associate editor are also gratefully acknowledged. This work has been partially supported by DGICYT Spanish Grant PB91-0794 and also (for the second author) by Grant PB91-0014.

References

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

Bowman, A.W., A comparative study of some kernel-based nonparametric density estimators, *J. Statist. Comput. Simul.* **21** (1985) 313–327.

Broniatowski, M., Convergence L_1 presque sure de l'estimateur de la densité obtenu par validation croissée, *C.R. Acad. Sc. Paris*, **303** Ser. I, **10** (1986) 487–490.

Broniatowski, M., P. Deheuvels and L. Devroye, On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate, *Ann. Statist.* **17** (1989) 1070–1086.

Cao-Abad, R., Applicaciones y nuevos resultados del método bootstrap en la estimación no paramétrica de curvas, Ph. D. dissertation (University of Santiago de Compostela, 1990).

Chiu, S.T., Bandwidth selection for kernel density estimation, *Ann. Statist.* **19** (1991) 1883–1905.
Chow, Y.S., S. Geman and L.D. Wu, Consistent cross-validated density estimation, *Ann. Statist.* **11** (1983) 25–38.
Cuevas, A. and W. González-Manteiga, Data-driven smoothing based on convexity properties, in: G.G. Roussas et al., eds. *Nonparametric Functional Estimation and Related Topics* (Kluwer Academic Publishers, Dordrecht, 1991) 225–240.
Devroye, L., *A Course in Density Estimation* (Birkhäuser, Boston, 1987).
Devroye, L., The double kernel method in density estimation, *Ann. Inst. Henri Poincaré* **25** (1989) 533–580.
Devroye, L. and L. Györfi, *Nonparametric Density Estimation: The L_1-View* (Wiley, New York, 1985).
Faraway, J.J. and M. Jhun, Bootstrap choice of bandwidth for density estimation, *J. Amer. Statist. Assoc.* **85** (1990) 1119–1122.
Feluch, W. and J. Koronacki, A note on modified cross-validation in density estimation, *Comp. Statist. Data Anal.* **13** (1992) 143–151.
Habbema, J.D.F., J. Hermans and K. van den Broeck, A stepwise discrimination analysis program using density estimation, in: G. Bruckmann, ed., *Compstat 1974: Proceedings in Computational Statistics* (Physica Verlag, Vienna, 1974) 101–110.
Hall, P., Large-sample optimality of least-squares cross-validation in density estimation, *Ann. Statist.* **11** (1983) 1156–1174.
Hall, P., On Kullback–Leibler loss and density estimation, *Ann. Statist.* **15** (1987) 1491–1519.
Hall, P., Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems, *J. Multiv. Anal.* **32** (1990) 177–203.
Hall, P. and I. Johnstone, Empirical functionals and efficient smoothing parameter selection, *J.R. Statist. Soc. B* **54** (1992) 475–530.
Hall, P. and J.S. Marron, On the amount of noise inherent in bandwidth selection for a kernel density estimator, *Ann. Statist.* **15** (1987a) 163–181.
Hall, P. and J.S. Marron, Extent to which least-squares cross-validation minimise integrated square error in nonparametric density estimation, *Probab. Th. Rel. Fields* **74** (1987b) 567–581.
Hall, P. and J.S. Marron, Estimation of integrated squared density derivatives, *Statist. Probab. Lett.* **6** (1987c) 109–115.
Hall, P. and J.S. Marron, Lower bounds for bandwidth selection in density estimation, *Probab. Th. Rel. Fields* **90** (1991a) 143–173.
Hall, P. and J.S. Marron, Local minima in cross-validation functions, *J.R. Statist. Soc. B* **53** (1991b) 245–252.
Hall, P., S.J. Marron and B. Park, Smoothed cross-validation, *Probab. Th. Rel. Fields* (1992) 1–20.
Hall, P., S.J. Sheather, M.C. Jones and J.S. Marron, On optimal data-based bandwidth selection in kernel density estimation, *Biometrika* **78** (1991) 263–269.
Hall, P. and M. Wand, Minimizing L_1-distance in nonparametric density estimation, *J. Multiv. Anal.* **26** (1988) 59–88.
Jones, M.C. and R.F. Kappenman, On a class of kernel density estimate bandwidth selectors, Manuscript (1990).
Jones, M.C., J.S. Marron and B.U. Park, A simple root n bandwidth selector, *An.. Statist.* **19** (1991) 1919–1932.
Jones, M.C., J.S. Marron and S.J. Sheather, Progress in data-based selection for kernel density estimation, Manuscript (1992).
Jones, M.C. and S.J. Sheather, Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statist. Prob. Letters* **11** (1991) 511–514.
Léger, C. and J.P. Romano, Bootstrap choice of tuning parameters, *Ann. Inst. Statist. Math.* **42** (1990) 709–735.
Mammen, E., A short note on optimal bandwidth selection for kernel estimators, *Statist. Prob. Letters* **9** (1990a) 23–25.

Mammen, E., On qualitative smoothness of kernel density estimates, Manuscript (1990b).
Mammen, E., J.S. Marron and N.I. Fisher, Some asymptotics for multimodality tests based on kernel density estimates, *Probab. Th. Rel. Fields* **91** (1992) 115–132.
Marron, J.S., An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** (1985) 1011–1023.
Marron, J.S., A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15** (1987) 152–162.
Marron, J.S., Bootstrap bandwidth selection. In: Le Page, P. and Billand, L., eds., Exploring the Limits of Bootstrap, 249–262 (1992).
Marron, J.S., Root *n* bandwidth selection, in: G.G. Roussas et al., eds., *Nonparametric Functional Estimation and Related Topics* (Kluwer Academic Publishers, Dordrecht, 1991) 251–260.
Park, B.U. and J.S. Marron, Comparison of data-driven bandwidth selectors, *J. Amer. Statist. Assoc.* **85** (1990) 66–72.
Park, B.U. and B.A. Turlach, Practical performance of several data driven bandwidth selectors, manuscript (1992).
Parzen, E., On estimation of a probability density function and mode, *Ann. Math. Statist.* **33** (1962) 1065–1076.
Quintela, A., Cálculo del parámetro de suavización en la estimación no paramétrica de curvas con datos dependientes (Ph. D. dissertation. University of Santiago de Compostela, 1992).
Rosenblatt, M., Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.* **27** (1956) 832–837.
Scott, D.W. and G.R. Terrell, Biased and unbiased cross-validation in density estimation, *J. Amer. Statist. Assoc.* **82** (1987) 1131–1146.
Sheather, S.J. and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *J. Roy. Statist. Soc. Ser. B* **53** (1991) 683–690.
Silverman, B.W., Using kernel density estimates to investigate multimodality, *J. Roy. Statist. Soc. B* **43** (1981) 97–99.
Silverman, B.W., Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *Ann. Statist.* **6** (1978) 177–184.
Silverman, B.W., Some properties of a test for multimodality based on kernel density estimates. In: J.F.C. Kingman and G.E.H. Reuter, eds., *Probability, Statistics and Analysis* (Cambridge University Press, 1983) 248–259.
Silverman, B.W., *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).
Stone, C.J., An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist.* **12** (1984) 1285–1297.
Stute, W., Modified cross-validation in density estimation, *J. Statist. Plann. Inf.* **30** (1992) 293–305.
Tapia, R.A. and J.R. Thompson, *Nonparametric Probability Density Estimation* (Johns Hopkins University Press, Baltimore, MD, 1978).
Taylor, C.C., Bootstrap choice of the smoothing parameter in kernel density estimation, *Biometrika* **76** (1989) 705–712.
Wand, M.P., J.S. Marron and D. Ruppert, Transformations in density estimation, *J. Amer. Statist. Assoc.* **86** (1991) 343–361.