A COMPARISON OF KERNEL DENSITY ESTIMATES

Alain Berlinet and Luc Devroye

Institut de Mathématiques
Université des Sciences et Techniques du Languedoc
Place Eugène Bataillon
34095 Montpellier Cedex
France

and

School of Computer Science
McGill University
Montreal, Canada H3A 2A7

Abstract. In the double kernel density estimate, the smoothing parameter h is chosen so as to minimize the L_1 distance between two kernel density estimates having identical smoothing factors but different kernels. This method is known to be consistent for any density and to be asymptotically optimal for a certain smooth class of densities. We propose a plug-in modification of the estimate and introduce various other data-based bandwidth estimates. Finally, a simulation study is presented in which the new bandwidth selectors are compared with a host of well-known methods.

Keywords and phrases. Density estimation, kernel estimate, convergence, smoothing factor, simulation, nonparametric estimation, double kernel.

1991 Mathematics Subject Classifications: Primary 62G05.

Table of contents.

1. Introduction.

The purpose of this paper is to demonstrate the usefulness of the double kernel estimate introduced by Devroye (1989b) as an alternative for choosing the smoothing factor h in the Akaike-Parzen-Rosenblatt density estimate. We consider an i.i.d. sample $X_1, \ldots, X_n$ drawn from a univariate density f, and estimate f by

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

where K is the kernel (a function integrating to one), $K_h(x) = (1/h)K(x/h)$, and $h > 0$ is the smoothing factor (Akaike, 1954; Parzen, 1962; Rosenblatt, 1956). The fundamental problem in kernel density estimation is that of the joint choice of h and K in the absence of a priori information regarding f. Watson and Leadbetter (1963) show that the choice of h and K should not be split into two independent subproblems. Also, the choice of K largely depends upon the smoothness of f (Devroye, 1992). In this work, we take a less ambitious goal: we fix K to be a bona fide density, so that f_{nh} is a density as well. Standard asymptotic theory in L_2 (Bartlett, 1963; Epanechnikov, 1969) and L_1 (Devroye and Györfi, 1985) shows that for smooth densities, the asymptotically optimal nonnegative kernel is given by

$$K(x) = \frac{3}{4}(1 - x^2)_+ \ .$$

We take K as our kernel in the simulations that follow. This kernel is inadmissible in the expected L_2 norm. By that we mean that there exists another kernel L and corresponding density estimate g_{nh} such that, with the same h in both estimates,

$$\mathbf{E} \int (g_{nh} - f)^2 \leq \mathbf{E} \int (f_{nh} - f)^2$$

for all n, all h and all densities. This follows from the expressions given in Watson and Leadbetter (see Cline, 1988): it suffices to choose L such that its Fourier transform is $\max(0, \psi(t))$, where ψ is the characteristic function for K:

$$\psi(t) = \frac{3(\sin t - t \cos t)}{t^3} \ .$$

However, L takes negative values, and hence, the comparison of g_{nh} with f_{nh} is not considered "fair" by some. This interesting anomaly can also be put another way: if we use K and pick h such that $\limsup n^{2/5} \mathbf{E} \int |f_{nh} - f| < \infty$, then there exists another kernel L and another sequence h' such that the kernel estimate $g_{nh'}$ with (L, h') is asymptotically infinitely superior:

$$\mathbf{E} \int |g_{nh'} - f| = o(n^{-2/5}) \ .$$

For this existence result, see section 7.5 of Devroye, 1987. It suffices to take a symmetric kernel L integrating to one, having compact support, possessing a zero second moment. We cannot in general tell how to choose h'. This is frustrating, because nobody likes to work with the knowledge that there is something better out there. However, it is also a blessing, as we will use this property to our advantage to design an automatic smoothing factor selector.

All global smoothing factors can be written in the general form $H = H_n(X_1, \ldots, X_n)$. A selection method is thus nothing but a sequence of functions $\{H_n, n \geq 1\}$. One of the major open problems is to establish the existence or non-existence of a selection method such that

$$\frac{\mathbf{E} \int |f_{nH} - f|}{\mathbf{E} \inf_h \int |f_{nh} - f|} \to 1$$

for all f. For some selection methods, this property is known to hold for a subclass of nice densities. If a universally optimal $\{H_n(.)\}$ sequence does not exist, then any inventor of a selection method should tell us what class of densities the selection method is designed for. At this moment, many selection methods have been proposed in the literature. Each one is geared towards a given class of densities. A fair comparison of such methods is indeed very difficult if not impossible. The best we can do is to compare different selectors on a wide class of densities with varying shapes, numerous combinations of skewness and kurtosis, many grades of tail heaviness, several brands of infinite peaks and discontinuities, and various degrees of smoothness. Without the variety in the testbed of densities, our study would be worthless.

In the double kernel method, one takes two different kernels K and L whose characteristic functions do not coincide on any open neighborhood of the origin. The kernel estimate with smoothing factor h and kernel K is denoted by f_{nh}, while for kernel L, we will write g_{nh}. The smoothing factor that will be employed in practice is H, where

$$H = \arg\min_{h>0} \int |f_{nh} - g_{nh}| \ .$$

There are two fundamental properties that make this estimate useful. First of all, for any density f, the estimate is consistent:

$$\mathbf{E} \int |f_{nH} - f| \to 0 \ .$$

This feature distinguishes it from many other bandwidth selectors, which fail to yield consistent estimates in all cases unless the bandwidth is unnaturally restricted to a deterministic interval. Note that the minimization above is performed over the entire positive halfline.

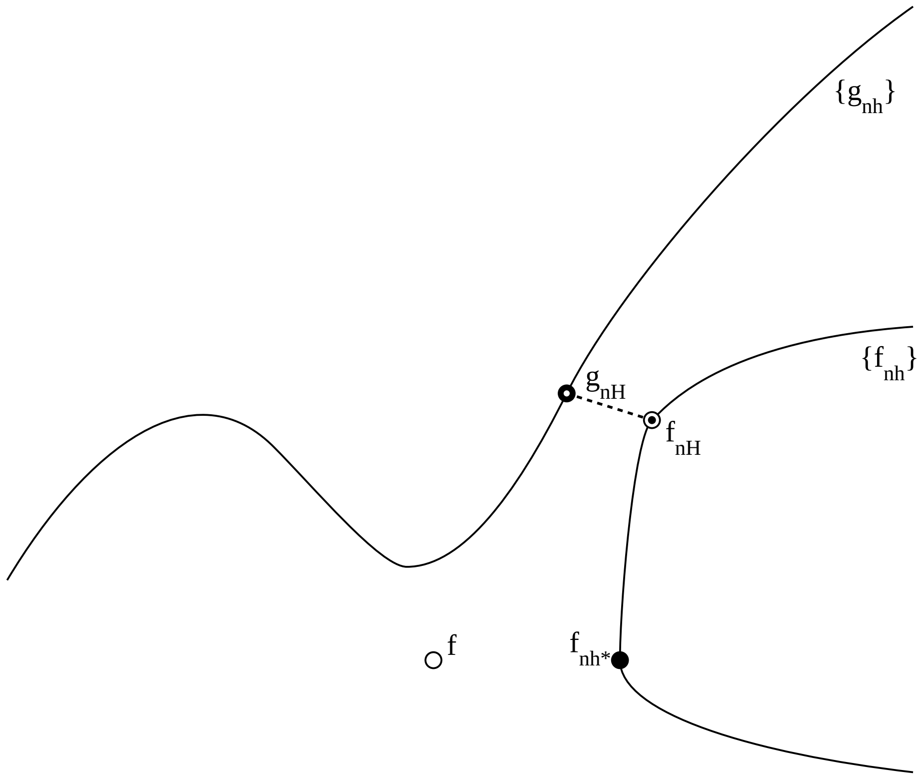

FIGURE 1. We show two families of density estimates in the set of all densities. The double kernel bandwidth minimizes the L_1 distance between f_{nh} and g_{nh}.

kernel	reference
$2K - K * K$	Stuetzle and Mittal (1979)
$(3K + xK')/2$	Schucany and Sommers (1977)
	Jones (1990)
$2\phi_1 - \phi_2$	Su-Wong, Prasad and Singh (1982)
$(105/64)(1-5x^2+7x^4-3x^6)\ ,\ \|x\| \leq 1$	Müller (1984)
$(1/2)(3-x^2)\phi_1$	Wand and Schucany (1990)
	Deheuvels (1977a,b)
$(75/16)(1-x^2)-(105/32)(1-x^4)\ ,\ \|x\| \leq 1$	Devroye (1989b)
	Gasser, Müller and Mammitzsch (1985)
	Scott (1992)

TABLE 1. Table of fourth-order kernels. The top two entries represent general rules for obtaining fourth-order kernels from a standard kernel K. ϕ_a represents the normal density with variance a. More examples are given in the references and in Berlinet (1991), Singh (1979), Bosq and Lecoutre (1987), Hall and Marron (1987c), and Fan and Hu (1992).

The second property goes to the heart of the matter. Assume that K is a symmetric positive kernel with $\int xK = 0$ and that L is a symmetric kernel with $\int xL = \int x^2 L = \int x^3 L = 0$, $\int x^4 L \neq 0$. Such kernels are called fourth-order kernels. From Berlinet (1993) it turns out that higher order kernels can be grouped into hierarchies generated by probability densities. More precisely a general form for a fourth-order kernel is

$$\left(\sum_{i=0}^{3} P(i,0)P(i,x) + \varphi(x) \right) \Psi(x)$$

where Ψ is a density with finite eighth moment, $\{P(i,x)\ ,\ 0 \leq i \leq 3\}$ are orthonormal polynomials in $L_2(\Psi)$, $P(i,x)$ being of exact degree i, φ is any function orthogonal to the space of polynomials of degree at most 3 in $L_2(\Psi)$ (thus φ can be taken equal to 0 to minimize computing time). See section 12 for the choice of our second kernel.

Assume that both K and L are symmetric, bounded, and have compact support. Also, both K and L must be L_1 Lipschitz (that is, $\int |K_1 - K_h|$ is bounded by $C(h-1)$ for some constant C and all $h > 1$, and similarly for L). In that case, $\mathbf{E} \int |g_{nh} - f| = o(\mathbf{E} \int |f_{nh} - f|)$ when f is smooth enough: more precisely, when f is absolutely continuous with derivative f', which in turn is absolutely continuous, and when

$$\int \sqrt{\sup_{|y|\le 1} f(x+y)}\, dx < \infty$$

(a tail condition on f), then the following property holds true:

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |f_{nH} - f|}{\inf_h \mathbf{E}\int |f_{nh} - f|} \le \frac{1+\epsilon}{1-\epsilon} ,$$

where

$$\epsilon = 4\sqrt{\int L^2 / \int K^2}$$

(Devroye, 1989b). The upper bound can be pushed as close to one as desired by stretching L out. One may obtain a limit of one if L is fixed and we replace g_{nh} by $g_{nh'}$, where $h'/h \to \infty$ in a prescribed manner. For the limited simulations we are performing, this is hardly worthwhile.

The sheer simplicity of the estimate, and its versatility—there are infinitely many pairs K and L one may choose from—should make this an attractive alternative. The greatest drawback is that the method is numerically slow, as we need to minimize a multimodal function, whose values are computed as integrals.

It is necessary to limit the scope of the paper. We are deliberately not considering local bandwidth selectors or variable kernel methods, although some of these have proven track records. One should also keep in mind that we may always transform the data, apply a fixed kernel estimate such as the estimates discussed in this paper, and then retransform the kernel estimate (see chapter 9 of Devroye and Györfi, 1985). This has the effect of introducing variable bandwidths.

We cannot begin this paper without mentioning the formidable comparative simulations carried out by Cao, Cuevas and Gonzalez-Manteiga (1994). Their study deals with smooth densities, but is more far-reaching than our study as ten different selectors are compared with respect to several error criteria. The authors conclude that the time-honored plug-in method is exceptionally good. Some modifications of the L_2 cross-validation method are not far behind, and the double kernel method typically ends up third or fourth out of ten methods. The latter method should be fine-tuned however to obtain optimal results—we will show how. Another conclusion of the Spanish study is that the double kernel method never performs poorly—it is very robust. This result is corroborated by the theoretical properties of the double kernel estimate. With the aforementioned fine-tuning, the double kernel method routinely beats other selectors that are designed for other error criteria (such as L_2 cross-validation). However, a new L_1 plug-in method introduced below also works well in practice. But perhaps the biggest selling points for the double kernel estimate are that it is genuinely automatic—no parameters have to be picked— and that it always works—the estimate is universally consistent.

The paper is organized as follows: in section 2 we give a quick survey of the smoothing factor selection literature, and take the opportunity to define some bandwidths that will be used in the simulations. In section 2.6.2, new plug-in versions of the double kernel estimate are proposed. Methods can be classified according to several criteria. Many believe that scale is important, as measured by the collection of values $\{|X_i - X_j|\}$. This is false. A density is only a tool for computing probabilities. Hence good bandwidth design should be based on probabilities. The double kernel method and its derivatives do just that. The spacings methods of sections 2.7.2 and 2.7.3 too use only integrals of density estimates over intervals. In section 3, we present our simulation results. The following quantities will be defined as we move along. They denote the various bandwidths that will be compared with one another, and are collected here for easy reference.

$h_{\mathtt{ref,L1}}$	plug-in, $L1$-based, reference method, larger constant
$h_{\mathtt{ref,l1}}$	plug-in, $L1$-based, reference method, smaller constant
$h_{\mathtt{ref,L2}}$	plug-in, $L2$-based, reference method
$h_{\mathtt{DH,L1}}$	plug-in, $L1$-based, reference method
	(DH refers to Deheuvels and Hominal)
$h_{\mathtt{DH,L2}}$	plug-in, $L2$-based, reference method
$h_{\mathtt{ms,L1}}$	plug-in, maximal smoothing, L_1 version
$h_{\mathtt{ms,L2}}$	plug-in, maximal smoothing, L_2 version
$h_{\mathtt{pi,l1}}$	plug-in, L_1-based, pilot $h_{\mathtt{ref,L1}}$
$h_{\mathtt{pi,L1}}$	plug-in, L_1-based, pilot $h_{\mathtt{dk,2}}$, stretch 1.50
$h_{\mathtt{pi,L2}}$	plug-in, L_2-based, Sheather and Jones
$h_{\mathtt{cv}}$	L_2 cross-validation
$h_{\mathtt{sh}}$	spacings method based on Sherman statistic
$h_{\mathtt{gr}}$	spacings method based on Greenwood statistic
$h_{\mathtt{pr}}$	projection method
$h_{\mathtt{op}}$	optimal bandwidth
$h_{\mathtt{dk,1}}$	double kernel method, second kernel stretch 1.20
$h_{\mathtt{dk,2}}$	double kernel method, second kernel stretch 1.44
$h_{\mathtt{dk,3}}$	double kernel method, second kernel stretch 1.73
$h_{\mathtt{dk,4}}$	double kernel method, second kernel stretch 2.07

2. Methods of selecting the bandwidth.

2.1. <u>L2 cross-validation</u>. Rudemo (1984) and Bowman (1984) proposed picking h so as to minimize an estimate of

$$\int (f_{nh}-f)^2 - \int f^2 = \int f_{nh}^2 - 2\int f f_{nh} \ .$$

An unbiased estimate of this is given by

$$M_{nh} = \int f_{nh}^2 - \frac{2}{n(n-1)} \sum_{i\neq j} K_h(X_i - X_j) \ .$$

The smoothing factor for which M_{nh} is minimal is called the L_2 cross-validation estimate. Asymptotically equivalent criteria have been proposed by many. An example includes

$$\int f_{nh}^2 - \frac{2}{n}\sum_{i=1}^{n} f_{nhi}(X_i) \ ,$$

where f_{nhi} is the kernel estimate with X_i deleted. The optimality of the L_2 cross-validation estimate H was established in Hall (1983), Burman (1985) and Stone (1984). From the latter paper, we retain that

$$\frac{\int (f_{nH}^2 - f)^2}{\inf_h (f_{nh}-f)^2} \to 1 \text{ a.s.}$$

under the sole condition that f is bounded. The L_2 cross-validation method is too variable, leading often to undersmoothing and sometimes to oversmoothing (Hall and Marron, 1987a,b; Scott and Terrell, 1987; Hall, Marron and Park, 1992). See also Marron (1987). Hall and Marron (1991) found that the L_2 criterion that is minimized typically shows many local minima. Devroye (1989d) points out that for any constant $a > 1$, one can find a density f such that with probability tending to one, $H \leq n^{-a}$. The smoothing factor is thus much too small, leading to a divergent estimator. The densities in this class of counterexamples all have infinite peaks.

A modified criterion, the biased cross-validation estimate, was proposed by Scott and Terrell (1987). As pointed out in Cao et al (1994), it has a global minimum at $h = \infty$, so that one has to apply a further modification to insure consistency.

A related modification due to Stute (1992) minimizes

$$\frac{\int K^2}{nh} + \frac{1}{n(n-1)} \sum_{i\neq j} M_h\left(X_i - X_j\right) \ ,$$

where

$$M \equiv K * K - K - \sigma^2 K'' \ ,$$

σ^2 is $\int x^2 K$, and K is assumed to be three times differentiable. Technical reasons lead to recommend the Gram-Charlier type kernel

$$K(t) = \frac{t^4/8 - 3t^2/4 + 11/8}{\sqrt{2\pi}} e^{-t^2/2} \; .$$

Smoothed cross-validation was developed by Jones, Marron and Park (1990), and Hall, Marron and Park (1992). The bandwidths achieve the optimal rate with respect to the optimal bandwidth for the mean integrated square error (Jones and Kappenman, 1990; Marron, 1991; Hall and Marron, 1990). This optimality property is achieved for a small subclass of densities. One minimizes the criterion

$$\frac{\int K^2}{nh} + \int (f_{nh'} - K_h * f_{nh'})^2 \; ,$$

where h' is a pilot bandwidth. The second term is a natural estimate of the bias term in the mean integrated square error. Various choices for h' have been suggested in the cited papers. For example, Cao et al (1994) take $h' = Cn^{-23/45}/h^2$, where C is a constant depending upon a normal reference distribution. The potential of this method remains largely untapped, as there are many ways of picking h'.

Jones and Kappenman (1992) observed that most of the cross-validation methods may be cast in the same light. They all asymptotically minimize

$$\frac{\int K^2}{nh} + \frac{1}{n^2} \sum_{i \neq j} M_h(X_i - X_j)$$

for some function M. The following table is partially borrowed from Jones and Kappenman (1992), who take K normal in experimental comparisons.

M	reference
$2K - K * K$	L_2 cross-validation (Bowman, 1984; Rudemo, 1984)
$\frac{\sigma^4}{4}(K * K)''''$	biased cross-validation (Scott and Terrell, 1987)
$\frac{\sigma^4}{4}K''''$	biased cross-validation (Jones and Kappenman, 1992)
$K * K * K * K - 2K * K * K + K * K$	presmoothed cross-validation (Hall, Marron and Park, 1989)
$K * K - K - \sigma^2 K''/2 + (1/24)(6\sigma^4 - \mu_4)K''''$	complete cross-validation (Jones and Kappenman, 1992)
$K * K - K - \sigma^2 K''$	Stute's modified cross-validation (Stute, 1992)

TABLE 2. Table of cross-validation methods listed according to the choice of M. The symbols σ^2 and μ_4 are used for the second and fourth moment of the symmetric kernel K.

2.2. The L2 plug-in method. The plug-in method for obtaining an L_2-optimal smoothing factor was introduced by Woodroofe (1970), who obtained an asymptotically optimal expression for the optimal h as a function of f and n, and, in a second step, estimated the unknown functional of f (in this case, $\int f''^2$) from the data in a nonparametric manner using a pilot bandwidth. For a similar idea, see Nadaraya (1974) and Deheuvels and Hominal (1980). To minimize $\mathbf{E}\int(f_n - f)^2$ when f is sufficiently smooth and K is a nonnegative kernel, the asymptotically optimal h has the following form:

$$h = \left(\frac{A(K)}{n\int f''^2}\right)^{1/5} ,$$

where $A(K) = (\alpha/\beta)^2$, $\alpha = \sqrt{\int K^2}$ and $\beta = \int x^2 K(x)\,dx$. This formula is at the heart of the plug-in method. The kernel K asymptotically minimizing $\mathbf{E}\int(f_n - f)^2$ among nonnegative kernels is the Bartlett or Epanechnikov kernel $(3/4)(1-x^2)_+$ (Bartlett, 1963; Epanechnikov, 1969). With this kernel, the formula reduces to

$$h = \left(\frac{15}{n\int f''^2}\right)^{1/5} .$$

See for example Watson and Leadbetter (1963), Rosenblatt (1971), or Deheuvels (1977a,b). Ways of estimating the unknown factor $\int f''^2$ in the formula are reviewed in the following subsections.

2.2.1. The reference density method. In the L_2 setting, one computes $\int f''^2$ for a reference density such as the normal (μ, σ). In the latter case for example, the (asymptotically) optimal h is given by

$$h = \sigma \left(\frac{40\sqrt{\pi}}{n} \right)^{1/5} = 2.345 \ldots \sigma \, n^{-1/5} \; .$$

The unknown parameters (such as σ in the above example) are estimated by a generally accepted method from the data, leading in turn to an estimate for $\int f''^2$. For the normal reference density, Deheuvels (1977a,b) suggests using

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{1}$$

instead of σ^2. A robust method advocated by many uses the interquartile estimate

$$\widehat{\sigma} = \frac{X_{[3n/4]} - X_{[n/4]}}{F^{\mathrm{inv}}(3/4) - F^{\mathrm{inv}}(1/4)} = \frac{X_{[3n/4]} - X_{[n/4]}}{1.35 \ldots} \; , \tag{2}$$

where F is the standard normal distribution function. One really needs a scale estimate that is less sensitive to outliers than averages and more accurate than quantile-based quick-and-dirty estimates. Janssen, Marron, Veraverbeke and Sarle (1992) tackle this problem head-on, and make several interesting suggestions, some of which were implemented by Jones, Marron and Sheather (1992).

More versatility could be created by considering a large reference family such as Pearson's or Johnson's that covers all possible combinations of skewness and kurtosis (see Devroye, 1986, for descriptions). We are not aware of any attempt along these lines in the literature, except for a passage in Scott (1992, p. 56–57) where lognormal and t families were considered as reference densities. The reference density method with a normal reference density led us to include the following bandwidths in our simulations.

- $h_{\mathtt{DH,L2}} = 2.345 \, \widehat{\sigma} \, n^{-1/5}$, where $\widehat{\sigma}$ is (1).
- $h_{\mathtt{ref,L2}} = 2.345 \, \widehat{\sigma} \, n^{-1/5}$, where $\widehat{\sigma}$ is (2).

2.2.2. The maximum smoothing principle. For any density for which the asymptotic formula for h given above is valid, we have, according to Terrell (1990), for the

Epanechnikov kernel,

$$h \leq 3(3/7n)^{1/5}\sigma = 2.532362\ldots\sigma\, n^{-1/5} \ ,$$

where σ is the standard deviation. See also Scott and Terrell (1985). Terrell presents arguments in favor of using the upper bound in some situations. The bandwidth used in our experiments is:

- $h_{\texttt{ms,L2}} = 2.532362\,\widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is defined in (2).

2.2.3. <u>Two-stage plug-in methods</u>. Park and Marron (1990), Park (1989), Hall and Marron (1990) and Sheather and Jones (1991) provide modern versions of the plug-in estimate for the L_2 criterion. See also Hall and Marron (1987a,b), and Hall, Sheather, Jones and Marron (1991). In all these papers, various methods are evaluated for estimating $\int f''^2$ for use in the asymptotic formula for the optimal h. We include in our experiments the method of Sheather and Jones (1991), which performed very well in the studies of Cao et al (1994), Park and Turlach (1992), and Jones, Marron and Sheather (1992). In the last paper, one also finds comparisons with related bandwidth selectors suggested by Engel, Herrmann and Gasser (1992). Sheather and Jones require an estimate of $\int f''^2$ in the formula

$$h = \left(\frac{15}{n\int f''^2}\right)^{1/5} \ .$$

They suggest using

$$\rho = \frac{1}{n^2 h'^5}\sum_{i,j} L''''\left(\frac{X_i - X_j}{h'}\right) = \frac{1}{n^2}\sum_{i,j}(L_h)''''\,(X_i - X_j) \ ,$$

where h' is yet another bandwidth, and L is a smooth kernel, for which we will take standard normal, as in Cao et al (1994). Theoretical considerations suggest that the optimal h' here is given by the formula

$$\left(\frac{2L''''(0)}{n\int f'''^2 \int x^2 L}\right)^{1/7} \ .$$

Cao et al (1994) suggest estimating $\int f'''^2$ by the reference density method based upon the normal density. Mimicking them, we estimate $\int f'''^2$ by

$$\frac{15}{16\sqrt{\pi}\,\widehat{\sigma}^7} \ ,$$

where $\widehat{\sigma}$ is the robust interquartile estimate of the standard deviation. Replacement shows then that

$$h' = \widehat{\sigma} \times \left(\frac{32}{5n\sqrt{2}}\right)^{1/7} \ .$$

The resulting bandwidth is called

$$h_{\texttt{pi,L2}} = \min\left(h_{\texttt{ms,L2}}, \left(\frac{15}{n\rho}\right)^{1/5}\right) .$$

It is easy to show that for all densities, $h_{\texttt{pi,L2}} \to 0$ and $n\,h_{\texttt{pi,L2}} \to \infty$ in probability whenever L'''' is uniformly bounded. This implies that $\mathbf{E}\int |f_{nH} - f| \to 0$ for all densities f.

It is worth mentioning that Parseval's identity has been employed by some to aid in the estimation of the optimal h. For example, Chiu (1991) has a plug-in method that is based upon the empirical characteristic function

$$\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j} .$$

Let Λ be the smallest positive t such that $|\varphi_n(t)|^2 \le 3/n$ (where the constant 3 is a design parameter). Then use

$$h = \left(\frac{A(K)}{Cn}\right)^{1/5} ,$$

where

$$C = \frac{1}{\pi}\int_0^{\Lambda} t^4(|\varphi_n(t)|^2 - 1/n)\,dt .$$

Devroye (1993) showed that this is not consistent whenever the density has a large infinite peak. A modification of Chiu's estimate was proposed in Chiu (1992).

2.2.4. <u>Iterative plug-in methods</u>. Scott and Factor (1981) and Scott, Tapia and Thompson (1977) estimate $\int f''^2$ by $\int g''^2_{n,h}$, where $g_{n,h}$ in turn is a kernel estimate, now with a kernel having two derivatives. The smoothing factor in $g_{n,h}$ should be chosen according to the same asymptotic formula, which suggests that one should take h as a solution of the equation

$$h = \left(\frac{15}{n\int g''^2_{n,h}}\right)^{1/5} .$$

Scott and Factor propose an iterative solution to this that is based upon Newton's method. However, their method does not always converge.

Park and Marron (1990) have a modern version of the iterative method. This approach is continued in Park (1989) and Sheather and Jones (1991). In most cases, one has to restrict h to an interval that is bounded away from 0 and ∞.

2.3. <u>L1 plug-in methods</u>. Devroye and Györfi (1985) consider the class $\mathcal{F}$ of all densities f with compact support, such that f is absolutely continuous, f' is absolutely continuous and there exists a version of f'' that is bounded and continuous on the real line. Define

$$\alpha = \sqrt{\int K^2} \ , \ \beta = \int x^2 K(x)\, dx$$

and $A(K) = \alpha^{4/5}\beta^{1/5}$. We also introduce the function $\psi(u) \stackrel{\text{def}}{=} \mathbf{E}|N-u|$, where N is a normal $(0,1)$ random variable. If $f \in \mathcal{F}$ and

$$\lim_{n\to\infty} h = 0 \ , \ \lim_{n\to\infty} nh = \infty \ ,$$

then

$$\left| \mathbf{E}J_{nh} - \alpha \int \sqrt{\frac{f}{nh}}\psi\left(\sqrt{nh^5}\frac{\beta|f''|}{2\alpha\sqrt{f}}\right)\right| \leq o(h^2) + o(1/\sqrt{nh})$$

(Devroye and Györfi, 1985). As noted by Hall and Wand (1988), this implies the following. For $f \in \mathcal{F}$,

$$n^{2/5} \inf_h \mathbf{E}J_{nh} \to 2^{-1/5}A(K)Q(f) \ ,$$

where

$$Q(f) \stackrel{\text{def}}{=} \inf_{u>0} \int \frac{\sqrt{f}}{u^{1/5}}\psi\left(\frac{u|f''|}{\sqrt{f}}\right) \ .$$

A generalization of this result that is valid even if $f \notin \mathcal{F}$, e.g., when f is the isosceles triangular density or the Laplace density, is given in Devroye and Wand (1993). For $f \in \mathcal{F}$, we note among other things that the asymptotically optimal formula for h is given by $h = (c^2/n)^{1/5}$, where

$$c \stackrel{\text{def}}{=} \arg\min_{u>0} \int \frac{\alpha\sqrt{f}}{\beta u}\psi\left(\frac{u^5\beta|f''|}{\alpha\sqrt{f}}\right) \ .$$

Needless to say, this is a cumbersome formula to work with. An adaptive method generalizing Woodroofe's method for L_2 was developed in Hall and Wand (1988). It rests on good pointwise estimates of f'' and $\sqrt{f}$. Devroye and Györfi (1985) elected to pick h so as to minimize a simple but more manageable upper bound for the expected L_1 error: for $f \in \mathcal{F}$, if

$$B(f) = \left(\int \sqrt{f}\right)^{4/5} \left(\int |f''|\right)^{1/5} \ ,$$

then

$$\inf_{u>0} \psi(u)/u \stackrel{\text{def}}{=} \gamma = 1.028493\ldots \leq \frac{Q(f)}{B(f)} \leq 5(8\pi)^{-2/5} = 1.3768102\ldots \ .$$

The choice of h for which we have

$$n^{2/5}\mathbf{E}J_{nh} \to 2^{-1/5}A(K) \times 1.3768102\ldots B(f)$$

is given in Devroye and Györfi (1985, p. 107): for the Epanechnikov kernel, with $\alpha = \sqrt{3/5}$ and $\beta = 1/5$, this yields

$$h = \left(\frac{\sqrt{15} \int \sqrt{f}}{\sqrt{2\pi} \int |f''|} \right)^{2/5} n^{-1/5} .$$

This h is never very far from the true optimal h. We note the following:

A. For the normal $(0, \sigma^2)$ density, the latter h can be written as

$$h = \sigma \left(\frac{15 e \sqrt{2\pi}}{8n} \right)^{1/5} = 1.6644 \ldots \sigma\, n^{-1/5} .$$

Hall and Wand (1988, Table 4.1) report that the optimal h asymptotically varies as

$$h = 2.279 \ldots \sigma\, n^{-1/5} .$$

B. In any case, the h yielding the bound on $B(f)$ is bounded from above as follows:

$$h \le \left(\frac{98415\pi^4}{65536n} \right)^{1/5} \sigma = 2.71042 \ldots \sigma\, n^{-1/5}$$

(Devroye and Györfi, 1985, p. 113). This does not mean that the true optimal h cannot be larger of course, but it is neverthless a useful bound. In a sense, it is the L_1 counterpart of a similar inequality for the L_2-optimal h noted recently by Terrell (1990).

The reference density method with a normal reference density led us to include the following bandwidths in our simulations.

- $h_{\mathtt{ref,L1}} = 2.279\, \widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is defined in (2).
- $h_{\mathtt{DH,L1}} = 2.279\, \widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is defined in (1). `DH` is a mnemonic for Deheuvels and Hominal.
- $h_{\mathtt{ms,L1}} = 2.71042\, \widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is given in (2).
- $h_{\mathtt{ref,l1}} = 1.6644\, \widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is defined by (2).

2.4. <u>Discussion of plug-in methods</u>. The formulae at the basis of most plug-in methods are valid under certain conditions on the density that are difficult to verify in practice. For example, the standard formulae for L_1 and L_2 plug-in smoothing factors are not valid for uniform or exponential densities. If the formulae were valid, one should still remember that they are only valid asymptotically, with no guarantees regarding the applicability for finite n.

Even if we accept that n is large enough such that the asymptotics may kick in, using a formula designed for L_2 provides us with little clues as to its suitability for L_1. Nevertheless, as the L_2 plug-in methods are very popular, it is necessary to see how they perform even if L_1 is the criterion that is considered.

Even if we accept the formula and its validity, one typically needs additional guarantees in order to insure the convergence of the estimates of factors such as $\int |f''|$ or $\int f''^2$.

On the other hand, nearly all comparative simulations indicate that plug-in methods are competitive. In their favor, one might argue that small samples from arbitrarily ill-behaved densities are all but indistinguishable from same-sized samples drawn from smooth small-tailed densities, for which the plug-in formulae are approximately valid. Finally, one should not forget that plug-in methods do not require any optimization at all. This may be important when designing real-time software.

2.5. <u>The bootstrap method</u>. In most bootstrap-based methods, one picks h so as to minimize

$$\mathbf{E}^* \int (f^*_{nh}(x) - f_{nh'}(x))^2 \, dx \ ,$$

where h' is some pilot bandwidth, f^*_{nh} is the kernel estimate with bandwidth h based upon a bootstrap sample $X_1^*, \ldots, X_n^*$, and $\mathbf{E}^*$ denotes expected value with respect to this bootstrap sample. The choice of h' and the bootstrap sample distribution have been the subject of various recent research projects. Note that $\mathbf{E}^*$ can be explicitly computed, so that the bootstrap sample is used in the formal definition only—one never has to actually generate it.

Taylor (1989) takes $h' \equiv h$ and draws the bootstrap sample from f_{nh} with a normal kernel K. His criterion has a minimum at infinity, so that a modification is necessary to insure a meaningful choice. Further theoretical properties were obtained by Mihoubi (1992).

Faraway and Jhun (1990) take h' by L_2 cross-validation. They also do not minimize $\mathbf{E}^*$ as above, but prefer to use an average over a large number of bootstrap samples to approximate $\mathbf{E}^*$.

Cao-Abad (1990) draws the bootstrap estimate from $f_{nh'}$. The value of $\mathbf{E}^*$ is then easily determined. In the L_2 sense, the best theoretical value for h' is given by

$$h' = \left(\frac{\int K''^2}{n \int f'''^2 \int t^2 K} \right)^{1/7} .$$

The unknown integral in this expression is replaced by the corresponding expression for the normal density with variance estimated from the data.

Hall (1990) draws a bootstrap sample of size $m < n$ from the empirical distribution, and sets $h' = h(n/m)^{1/5}$. One suggestion tried out in Cao et al (1994) is $m \sim \sqrt{n}$. He takes the minimum of $\mathbf{E}^*$ closest to $n^{-1/5}$ to avoid degenerate solutions. For a survey and more discussion of bootstrap methods in density estimation, we refer to Marron (1992).

2.6. The double kernel estimate.

2.6.1. General setting. In the double kernel method, we take two different kernels K and L whose characteristic functions do not coincide on any open neighborhood of the origin. The kernel estimate with smoothing factor h and kernel K is denoted by f_{nh}, while for kernel L, we will write g_{nh}. The smoothing factor that will be employed in practice is H, where

$$H = \arg\min_{h>0} \int |f_{nh} - g_{nh}| .$$

Connection with the bootstrap method. When $L = 2K - K * K$, it is easy to see that H is identical to the H obtained if we had taken $L = K * K$. This has an intriguing interpretation, as $g_{nh} = f_{nh} * K_h$ in the latter case: H minimizes

$$\int |f_{nh} - f_{nh} * K_h| .$$

The density $f_{nh} * K_h$ is that of a sample drawn from f_{nh} (as one would draw in a smoothed bootstrap), in which each observation receives an additional perturbation in the form of hW, where W has density K. In other words, we are minimizing the distance between the density of $X_N + hW$ and that of $X_N + hW + hW'$, where N is a random integer between 1 and n, and W, W' are i.i.d. perturbations with density K. This sort of criterion is closely linked to the criteria proposed in the bootstrap literature.

A STABILITY CRITERION. Continuing in the same vein, we note that formally, if μ_n is the standard empirical measure, $f_{nh} = \mu_n * K_h$. We are thus looking for the operator $*K_h$ that yields the most stable solution: one application of the operation yields $f_{nh} = \mu_n * K_h$, while two applications yields $g_{nh} = \mu_n * K_h * K_h$, which is by definition very close to f_{nh}.

CHOICE OF THE SECOND KERNEL. It is computationally interesting to work with kernels that are piecewise polynomials of low order. For this reason, we picked the Epanechnikov kernel as kernel K :

$$K(x) = \frac{3}{4}(1 - x^2) \ , \ |x| \leq 1 \ .$$

We also constructed a fourth-order piecewise quadratic kernel L. To get such a kernel which moreover would be symmetric and continuous, two ways are possible. Either fix for L a polynomial form on some intervals or work within the general framework given in the introduction :

$$L(x) = (1 + P(0)P(x))\,\Psi(x)$$

where Ψ is an even density and P is an even polynomial of degree 2, with norm equal to 1 in $L_2(\Psi)$, and orthogonal to the monomials 1 and x. In both cases this leads to systems of non-linear equations. A solution is given by

$$L(x) = \begin{cases} \frac{7-31x^2}{4} & , |x| \leq 1/2 \\ \frac{x^2-1}{4} & , 1/2 \leq |x| \leq 1 \\ 0 & , 1 \leq |x| \end{cases} .$$

This kernel is the kernel of order 4 in the hierarchy generated by the density

$$\Psi(x) = \begin{cases} \frac{31}{4C} & , |x| \leq 1/2 \\ \frac{31(x^2-1)}{4C(7-31x^2)} & , 1/2 \leq |x| \leq 1 \\ 0 & , |x| \geq 1 \ , \end{cases}$$

where $C = 8.57444\ldots$. Ψ appears as a continuous regularization of the naive uniform kernel on $[-1/2, 1/2]$ and is very "close" to it. Thus the performance of our kernel can be expected to be close to the ones of the fourth-order "minimum variance" kernel since the minimum variance hierarchy (or Legendre hierarchy) is generated by the uniform kernel (Berlinet, 1993).

The asymptotic variance of a kernel estimate is proportional to $W(K) = \int K^2$ and the asymptotic MISE for a kernel of order $r \geq 2$ is proportional to

$$T_r(K) = \left(\int K^2\right)^r |\int x^r K(x)\,dx| \ .$$

The minimizer of $T_r(K)$ over the set of square integrable kernels of order r with $r-2$ sign changes on $\mathbb{R}$ is the kernel K_r of order r in the Epanechnikov hierarchy, i.e., the hierarchy generated by our kernel K (Berlinet, 1993; Granovsky and Müller, 1989). The kernel K_4 minimizing $T_4(K)$ is given by

$$K_4(x) = \begin{cases} (45 - 150x^2 + 105x^4)/32 & \text{if } |x| \le 1 \ ; \\ 0 & \text{otherwise.} \end{cases}$$

The kernel H_4 of order 4 with support $[-1, 1]$ minimizing $W(K)$ is given by

$$H_4(x) = \begin{cases} (9 - 15x^2)/8 & \text{if } |x| \le 1 \ ; \\ 0 & \text{otherwise.} \end{cases}$$

It is discontinuous at the ends of $[-1, 1]$. In the table below we list the values of $W(K)$ and $T_4(K)$ for H_4, K_4 and L.

kernel K	$W(K)$	$T_4(K)$	$W(H_4)/W(K)$	$T_4(K_4)/T_4(K)$
H_4	$\frac{9}{8}$	$\frac{19683}{143360}$	1	$0.84\ldots$
K_4	$\frac{5}{4}$	$\frac{625}{5376}$	0.9	1
L	$\frac{47}{30}$	$\frac{4879681}{37800000}$	$0.71\ldots$	$0.90\ldots$

TABLE 3. This table shows that L is nearly optimal in the sense of optimal MISE.

In this simulation study we will use four kernels defined from L by rescaling:

$$L_{2l}(x) = \frac{1}{2l} L\left(\frac{x}{2l}\right) \ ,$$

with $l = (1.2), (1.2)^2, (1.2)^3$ and $(1.2)^4$. We denote the double kernel smoothing factor by $h_{\mathtt{dk,1}}, h_{\mathtt{dk,2}}, h_{\mathtt{dk,3}}$ and $h_{\mathtt{dk,4}}$ respectively. The theory tells us that for large n, the scale factor of L should exceed that of K. This is why we do not consider the case $l \le 1$.

DOUBLE KERNEL-DOUBLE H METHOD. If K and L are a pair of kernels of second and fourth order respectively, we may define the double kernel-double h method by

$$(H, H') = \arg\min \int |f_{nh} - g_{nh'}| \ ,$$

where the kernel estimates are based upon the same data but different kernels K and L respectively. The optimization is not a sinecure, of course, but we believe that this method is asymptotically optimal in the sense that $\mathbf{E} \int |f_{nH} - f| \sim \inf_h \mathbf{E} \int |f_{nh} - f|$ for all smooth densities with a small tail. For small sample sizes, h' tends to hover around the value that makes $\int |K_h - L_{h'}|$ smallest, and thus, h'/h tends to remain fairly constant. The effect of the double optimization is only felt at larger sample sizes. For this reason, this method is not included in our simulation experiment : we will only minimize the L_1

distance between estimates by considering h' of the form $2lh$ with l as above. At this point, a better understanding of the theoretical properties of the double kernel-double h estimate is needed.

2.6.2. <u>A modified double kernel estimate</u>. We introduce two versions of the plug-in method for use in an L_1 context. Both are based upon the approximately asymptotically optimal formula for h given in subsection 2.3:

$$h = \left(\frac{\sqrt{15/(2\pi)} \int \sqrt{f}}{\int |f''|} \right)^{2/5} n^{-1/5} .$$

This formula comes with the standard caveats for plug-in methods. We define

$$h_{\texttt{pi,l1}} = \min \left(\left(\frac{\sqrt{15/(2\pi)} A}{B} \right)^{2/5} n^{-1/5} \ , \ h_{\texttt{ms,L1}} \right) \ , \tag{3}$$

where A is an estimate of $\int \sqrt{f}$, and B is an estimate of $\int |f''|$. All rests on a pilot bandwidth $h' = h_{\texttt{ref,L1}}$, a robust reference density bandwidth. With h' taken as the double kernel bandwidth with as second kernel L_3 (with stretch $l = 1.5$ in the notation of the previous section), the eventual plug-in estimate will be called $h_{\texttt{pi,L1}}$. To estimate the two unknown functionals, we employ once again nonparametric estimates: $\int \sqrt{f}$ is estimated by

$$A = \int \sqrt{f_{nh'}} \ , \tag{4}$$

where $f_{nh'}$ is a kernel estimate with Epanechnikov kernel K. The Taylor series expansion that leads to the asymptotic formula in the first place would suggest that $\int |f''|$ may be estimated by

$$B = \frac{2 \int |f_{nh''} - g_{nh''}|}{h''^2 \int x^2 K} \ , \tag{5}$$

where $f_{nh''}$ and $g_{nh''}$ are kernel estimates with kernels K and L respectively, and L is as in the double kernel method, with stretch $l = 1.5$. and the bandwidth h'' is defined by

$$h'' = h' \max \left(1, \left(\frac{R}{\tau} \right)^{2/5} \right) \ ,$$

where

$$R = \frac{A \sqrt{\int (K - L)^2}}{\sqrt{nh'} \int |f_{nh'} - g_{nh'}|}$$

and $\tau = 1/10$ (we will see that the value of the threshold τ is not very critical). We note that $h'' \geq h'$. The choice of our procedure is justified below. Note that the computation of the plug-in bandwidth requires some additional integration work, but the method is designed to be very robust. This is borne out in the results.

THEOREM. *The plug-in method with bandwidth (3), with (A) and (B) given by (4) and (5), and with K and L as in the double kernel method is universally consistent, i.e.,* $\mathbf{E}\int |f_{nH} - f| \to 0$ *for all* f.

PROOF. By a general theorem of Devroye and Györfi (1985), $H \to 0$ in probability and $nH \to \infty$ in probability imply $\mathbf{E}\int |f_{nH} - f| \to 0$ for all f. Applied to $H = h_{\mathtt{pi,l1}}$, the former condition follows from $h_{\mathtt{pi,l1}} \leq h_{\mathtt{ms,L1}}$, which tends to 0 in probability. The latter condition holds if $n^2 A/B \to \infty$ in probability. Since $h' \to 0$ in probability and $nh' \to \infty$ almost surely, we note that $f_{nh'} \to f$ almost surely at almost all x by general pointwise convergence theorems (Devroye, 1987). By Fatou's lemma, we have, almost surely,

$$\liminf_{n\to\infty} \int \sqrt{f_{nh'}} \geq \int \liminf_{n\to\infty} \sqrt{f_{nh'}} = \int \sqrt{f} > 0 \ .$$

Also,

$$B \leq \frac{2\int |K-L|}{h''^2 \int x^2 K} \ ,$$

so that $n^2 A/B \to \infty$ in probability whenever $nh'' \to \infty$ in probability. By definition, $h'' \geq h'$, and we are done. □

JUSTIFICATION. In general, for h large enough,

$$\begin{aligned}
\int |f_{nh} - g_{nh}| &\approx \int |f_{nh} - f| \\
&\approx \int |f * K_h - f| \\
&\approx (1/2)h^2 \int |f''| \int x^2 K \ .
\end{aligned}$$

These approximations can be made precise of course. The first one uses the fact that $\int |g_{nh} - f|$ is much smaller than $\int |f_{nh} - f|$ as L is a higher-order kernel. This assumes some degree of smoothness on f. Also, we assume that h is in such a range that bias is much larger than variation. The last approximation may be found in Devroye (1987, p. 110) or Devroye and Györfi (1985, p. 209). Under the said large bias assumption, $\int |f''|$ is approximately estimated by

$$\frac{2\int |f_{nh} - g_{nh}|}{h^2 \int x^2 K} \ .$$

This is B. It suffices to find a reasonable test for the validity of the large bias assumption. We begin with B evaluated at $h = h'$, but we also realize that this h' may be too small. To see this, a rough estimate of the variation term

$$\int |(f_{nh} - g_{nh}) - f * (K_h - L_h)|$$

is given by

$$\frac{\int \sqrt{f}\sqrt{\int (K-L)^2}}{\sqrt{nh'}} \approx \frac{A\sqrt{\int (K-L)^2}}{\sqrt{nh'}} .$$

If this is less than τ times the bias estimate $(1/2)h'^2 \int |f''| \int x^2 K$, (i.e., when $R < \tau$), then h' is indeed large enough, and we may safely accept B as an estimate of $\int |f''|$. Otherwise, we set $h'' = h'(R/\tau)^{2/5}$. With this choice, the bias estimate (being quadratic in h') becomes about $(R/\tau)^{4/5}$ times as large. The variation term on the other hand becomes about $(R/\tau)^{1/5}$ times smaller. The new ratio of variation over bias is thus about τ, which is a parameter we control. Therefore, the new h'' is large enough for B to be acceptable.

2.7. Other methods.

2.7.1. The projection method. Semiparametric methods may work well for certain families of densities. To illustrate this point, we will include in our study one projection method. The basic principle here is to start from a class $\mathcal{F}$ of densities (such as all Pearson densities; or all normal densities; or all log-concave densities). The class should not contain K even as a limit. We determine the pair (h, f) as follows:

$$(h, f) = \arg\min_{h>0; f\in\mathcal{F}} \int |f - f_{nh}| .$$

This idea has been around (Devroye, 1989a). We also discussed it with Cuevas and Gonzalez-Manteiga. This determines the projection of $\{f_{nh}, h > 0\}$ onto the class $\mathcal{F}$ as well as the closest kernel estimate to the class. As this sort of strategy warrants a separate study altogether, we will consider a simplified version. First we determine from the data the median m and the estimated standard deviation $s = (X_{[3n/4]} - X_{[n/4]})/1.35$ (which is the interquartile estimate assuming that the data are normally distributed). Let f_0 be the normal density with mean m and standard deviation s. Determine $H = h_{\mathtt{pr}}$ by

$$h_{\mathtt{pr}} \stackrel{\text{def}}{=} \arg\min_{h>0} \int |f_0 - f_{nh}| .$$

This will be called the simplified projection estimate. What matters for now is that we have at least one member of this family under consideration to detect a certain trend in the performance. One could relate this also to the well-known minimum-distance method, although there one usually considers distances from the empirical measure, and one grabs the best closest density (according to some criterion) in a given nonparametric class of densities (whereas we restrict our estimate to be of the kernel form). For more details of minimum-distance estimation, consult Vapnik and Stefanyuk (1978), Stefanyuk (1979), Reiss (1986), and Gajek (1989).

As the target is the normal density, the projection estimate should prove useful when the true density is bell-shaped. The following inequality highlights the utility of this approach:

$$\begin{aligned}\int |f - f_{nH}| &\leq \int |f - f_0| + \int |f_0 - f_{nH}| \\ &\leq \int |f - f_0| + \int |f_0 - f_{nh^*}| \quad (h^* \text{ minimizes } \textstyle\int |f - f_{nh}|) \\ &\leq 2 \int |f - f_0| + \int |f - f_{nh^*}| \\ &= 2 \int |f - f_0| + \inf_h \int |f - f_{nh}| \ .\end{aligned}$$

Thus, if f_0 is close to f, we have a formidable performance guarantee. With positive kernels, we know that asymptotically, for any f and h,

$$\inf_h \mathbf{E} \int |f_{nh} - f| \geq 0.86\, n^{-2/5}$$

(Devroye and Györfi, 1985). Ignoring the $\mathbf{E}$ for a moment, the inequality above says that if we can pinpoint a parametric family in which one density is within $0.215\, n^{-2/5}$ of f, then $\int |f - f_{nH}|$ is roughly speaking within 50% of its optimal value, whatever it may be. The inequality thus links what we know about f (which is reflected in our choice of f_0) with actual performance, providing a continuous bridge between parametric and nonparametric.

2.7.2. The spacings method. The idea of using spacings to select parameters has been explored by many researchers, both in a finite parameter setting (Cheng and Amin, 1983; Ranneby, 1984) and in a more general context (Roeder, 1990). One of the methods included in our simulation is based upon first principles from hypothesis testing. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics for the data sequence $X_1, \ldots, X_n$, drawn from density f. It is well-known that the spacings

$$D_1 = \int_{-\infty}^{X_{(1)}} f, \ldots, D_n = \int_{X_{(n-1)}}^{X_{(n)}} f, D_{n+1} = \int_{X_{(n)}}^{\infty} f$$

are distributed as uniform spacings. These are at the basis of many spacings tests (Pyke, 1965). Of these tests, we are mainly interested in those that provide us with an L_1 flavour. For example, the Kendall-Sherman statistic

$$K_n = \sum_{i=1}^{n+1} \left| D_i - \frac{1}{n+1} \right|$$

was suggested by Kendall in the discussion of Greenwood (1946). Under the null hypothesis, the limit distribution was obtained by Sherman (1950). Further studies have been undertaken recently by El abdin Ras (1989) who showed, among other things, that for all $\epsilon > 0$,

$$\mathbf{P}\{|K_n - 2/e| \geq \epsilon\} \leq 4e^{-7(n+1)\epsilon^2/120} \ .$$

To apply this in our setting, we combine it with a cross-validation idea. We define the leave-two-out double spacings

$$D_i'' = \int_{X_{(i-1)}}^{X_{(i)}} f_{n,h,i-1,i}(x)\, dx$$

where $X_{(0)} = -\infty$, $X_{(n+1)} = \infty$, and $f_{n,h,i-1,i}$ is the kernel estimate with smoothing factor h, with $X_{(i-1)}$ and $X_{(i)}$ deleted. We compute the statistic

$$K_n'' = \sum_{i=1}^{n+1} \left| D_i'' - \frac{1}{n+1} \right| \ .$$

Observe that for fixed n,

$$\lim_{h\to\infty} K_n'' = 2\frac{n-1}{n+1} \text{ and } \lim_{h\to 0} K_n'' = 1 \ .$$

Define

$$h_{\mathtt{sh}} \stackrel{\text{def}}{=} \arg\min_{h>0} K_n'' \ .$$

It is not difficult to show that for any density, $h_{\mathtt{sh}} \to 0$ almost surely, and that for any sequence h with $h \to 0$ and $nh \to \infty$, $K_n'' \to 2/e$ almost surely. Therefore, almost surely, $h_{\mathtt{sh}}$ is well-defined and stays away from 0 and ∞.

There are no guarantees that $n\, h_{\mathtt{sh}} \to \infty$ as required for consistency. In practice, $h_{\mathtt{sh}}$ often undersmooths. For h large enough, K_n'' is very close to the true bias, $\int |f - f * K_h|$. Therefore, there may be merit in considering the

$$h = \arg\min\left(K_n'' + V_n\right) \ ,$$

where V_n is an appropriate estimate of the variation, $\mathbf{E}\int |f * K_h - f_{nh}|$. (Read, however, Mammen (1990) and Jones (1991a) regarding the minimization of expected values.) A simple proposal for V_n is given below. Nevertheless, we won't include this modification in our simulation.

ESTIMATION OF THE VARIATION. If we split the data into two parts, $X_1, \ldots, X_m$, $X_{m+1}, \ldots, X_n$, with $m = \lfloor n/2 \rfloor$, then the variation

$$W_n = \mathbf{E} \int |f_{nh} - \mathbf{E} f_{nh}|$$

can be estimated by

$$V_n \stackrel{\text{def}}{=} \frac{1}{2} \int |f_{mh} - f'_{mh}| \ ,$$

where f_{mh} is the kernel estimate with kernel K, smoothing factor h, and data $X_1, \ldots, X_m$, and f'_{mh} is the kernel estimate with kernel K, smoothing factor h, and data $X_{m+1}, \ldots, X_n$.

To study the properties of V_n, note first of all that if one X_i changes value, then V_n changes by at most $2\int |K|/n$. Thus, by McDiarmid's inequality (1989), for $\epsilon > 0$,

$$\mathbf{P}\{|V_n - \mathbf{E} V_n| > \epsilon\} \le 2 e^{-n\epsilon^2/(2\int^2 |K|)} \ .$$

In other words, $V_n - \mathbf{E} V_n$ is of the order of $1/\sqrt{n}$ or less. We know that

$$W_n \sim \sqrt{\frac{2}{\pi}} \times \frac{\sqrt{\int K^2} \int \sqrt{f}}{\sqrt{nh}}$$

under some conditions on f. For other f, we have $\sqrt{nh}\, W_n \to \infty$. For the former class of densities, it is easy to see that we also have

$$\mathbf{E} V_n \sim \sqrt{\frac{2}{\pi}} \times \frac{\sqrt{\int K^2} \int \sqrt{f}}{\sqrt{nh}} \ .$$

Therefore, we may use V_n as a good estimate of W_n.

2.7.3. <u>A method based upon the Greenwood statistic</u>. Start with the inter-point integrals

$$D_1 = \int_{-\infty}^{X_{(1)}} f_n, \ldots, D_n = \int_{X_{(n-1)}}^{X_{(n)}} f_n, D_{n+1} = \int_{X_{(n)}}^{\infty} f_n \ .$$

If f_n is replaced by f above, then $D_1, \ldots, D_{n+1}$ would be distributed as the spacings of a uniform sample. This implies that $(n+1)D_1, \ldots, (n+1)D_{n+1}$ are asymptotically distributed as exponential random variables. Ideally, then,

$$U_1, \ldots, U_{n+1} \stackrel{\text{def}}{=} e^{-(n+1)D_1}, \ldots, e^{-(n+1)D_{n+1}}$$

are approximately distributed as an i.i.d. uniform $[0,1]$ random sequence. What we need now is a test statistic that measures departure from uniformity. Among the myriad of possibilities, we picked Greenwood's test statistic (Greenwood, 1946; Pyke, 1965)

$$G_n = (n+2)\sum_{i=1}^{n+2}(U_{(i)} - U_{(i-1)})^2 \ ,$$

where $U_{(i)}, 1 \leq i \leq n+1$ are the order statistics for $U_1, \ldots, U_{n+1}$, $U_{(0)} = 0$, and $U_{(n+2)} = 1$. Formally, we define the bandwidth

$$h_{\mathtt{gr}} = \arg\min_{h>0} G_n \ .$$

2.7.4. <u>Maximum likelihood cross-validation</u>. The number $h > 0$ maximizing

$$\prod_{i=1}^{n} f_{nhi}(X_i) \ ,$$

where f_{nhi} is the kernel estimate based upon a sample of size $n-1$ with X_i deleted from $X_1, \ldots, X_n$, is called the maximum likelihood cross-validation method. It was introduced by Duin (1976) and Habbema, Hermans and van den Broek (1974), and was later modified by Marron (1985). Convergence conditions were established by Chow, Geman and Wu (1983) and Devroye and Györfi (1985). Unfortunately, when the distribution has tails that decrease exponentially quickly or slower, the estimator is not consistent. This phenomenon was first observed by Schuster and Gregory (1981), while necessary and sufficient conditions of convergence are given by Broniatowski, Deheuvels and Devroye (1989). For the size of the smoothing factor, see Hall (1982) and van Es (1988, 1989). The estimate tends to minimize the Kullback-Leibler distance between f_n and f, and has no direct relationship to the L_1 error. A universally consistent estimate can be obtained by transforming the data to $[-1,1]$ via a monotone transformation like $x :\to x/(1+|x|)$, applying the maximum likelihood cross-validation method, and re-transforming the data (Devroye and Györfi, 1985). In most studies carried out to date, and in particular in the study of Cao et al (1994), the maximum-likelihood cross-validation method performed very poorly. For this reason, it is not included in our simulation experiment.

3. Comparisons and simulations.

General surveys of bandwidth selectors are given in Devroye and Györfi (1985), Titterington (1985), Marron (1987, 1988, 1989a), Izenman (1991), Jones, Marron and Sheather (1992), and Turlach (1993).

Cao, Cuevas and González-Manteiga (1994) consider L_1, L_2 and L_∞ error criteria, and provide us with a wealth of practical information. Few other studies offer practical experiments with the L_1 criterion. An example is Bean and Tsokos (1982), who are mainly concerned with penalized or smoothed maximum-likelihood estimation. Various L_2 cross-validation and L_2-based plug-in methods are compared from an L_1 point of view on six normal mixture test densities in Park and Turlach (1992).

Scott and Factor (1981) compare maximum likelihood cross-validation and the method of Scott, Tapia and Thompson (1977) on normal and normal mixture data with L_2 as a criterion. Bowman (1985) looks at many methods, among which maximum likelihood cross-validation, L_2 cross-validation, the normal reference density method, and a goodness-of-fit method. The densities included normal, normal mixture, t(5), Cauchy, chi-square (6) and beta (2,2). The normal reference method was declared the overall winner under the L_2 criterion. Abdous (1990) too only considers L_2 errors, and takes the normal mixture family as well as the power exponential family as prototypes. Also included is the chi-square distribution with 4 degrees of freedom, the t(5) density, and the normal density. The L_2 plug-in method of Deheuvels (1977a,b) wins against L_2 cross-validation and maximum likelihood cross-validation.

Further L_2 simulation studies were conducted by various research groups in the late eighties. Typical test densities include the normal density, normal mixtures, the t and gamma families. See for example Kappenman (1987), Marron (1989a), and Faraway and Jhun (1990), where a semi-automatic bootstrap method is compared with other methods. Jones, Marron and Sheather (1992) offer an L_2-based comparison on a battery of 15 unimodal and multimodal densities. They also test various scale estimates.

Nearly every bandwidth selection paper in the nineties offers a limited simulation. Notable examples are Park and Marron (1990), and Jones and Kappenman (1992). The latter L_2 study too favors various plug-in methods, and shows the feasability of modified L_2 cross-validation estimates such as those listed in Table 2. Other recent simulation studies include Park and Turlach (1992) described above, Sheather (1993), and Kim, Park and Marron (1993).

3.1. <u>The comparative simulation: a benchmark test</u>. In our simulation, it is not our intention to duplicate the study of Cao, Cuevas and González-Manteiga (1994), who compared ten different bandwidth selectors on many different densities under the L_p

criterion with $p \in \{1, 2, \infty\}$. Our goal is more modest: we will only look at L_1, we don't average over so many simulation runs, and we include only one representative of each selection methodology in our study. On the other hand, the double kernel method is studied in more detail than before, and our test bank of densities comprises a smorgasbord of densities of varied shapes. In all the tests, K is the Epanechnikov kernel.

We take the following view with regards to the criterion. First of all, what matters in practice is

$$J_{nH} = \int |f_{nH} - f| \ .$$

We will compare this with the best possible error,

$$Q_n = \inf_h \int |f_{nh} - f| \ ,$$

which measures the quality of the sample (hence the choice of the symbol Q_n). To partially offset the variablity in Q_n and J_{nH}, one might look at things like $J_{nH} - Q_n$, $(J_{nH} - Q_n)/Q_n$ or J_{nH}/Q_n. This will be done in our simulations. Especially the last two quantities are convenient as they allow us to compare performances across different densities on a more or less absolute scale. Note that we do not attach a lot of importance to $\mathbf{E} \int |f_{nh} - f|$ per se, as the $\mathbf{E}$ averages over many data sets, and this clearly is not something one would have in practice.

For a fair comparison, all the kernels are the same—we pick Epanechnikov's kernel because of its optimality property among positive kernels.

3.2. <u>The test densities</u>. Twenty-eight test densities are included in our simulation. Random variate generation is trivial in all cases—see Devroye (1986) for a general description of non-uniform random variate generation. Throughout, we have $n = 100$. The group of densities contains several smooth bell-shaped ones such as the normal, logistic, lognormal, Maxwell, Cauchy, inverse exponential, and extreme-value densities. These have varying tail sizes and asymmetries. We add five densities with an infinite peak at the origin (numbers 8, 14, 15, 18 and 19). Again, the densities differ in peak sizes and skewness. Note that of these, only density 15 is in L_2. Three continuous densities with discontinuous first derivatives were included: the Laplace density, the beta $(2, 2)$ density and the isosceles triangle. The discontinuity occurs either at the peak or near the extrema of the support. Next, we throw in a uniform density, a uniform mixture, and an exponential density, in the hope of testing the robustness in the presence of simple discontinuities. A Pareto and asymmetric Pareto distribution are introduced as well to test the performance in the presence of very big tails, not unlike those obtained with economic or linguistic data.

The first 20 densities are unimodal. The latter eight densities are multimodal. The normal mixtures 21 through 24 were suggested to us by Steve Marron, who earlier (Marron, 1989a) obtained interesting simulations with normal mixtures, that show the richness of this class of densities. Since the collection of normal mixtures is a dense subset of all densities in the L_1 sense, one may effectively restrict oneself to such types of simulations. However, the denseness also tells us about the enormous variety of possible test densities. Others may argue that in blatantly multimodal settings, one would resort to some form of variable bandwidth kernel estimation, and that it is useless to test uncompetitive fixed kernel estimates such as the ones dealt with here. Densities 22 through 24 are taken from Marron and Wand (1992). The marronite density (number 21) is included to test the robustness with respect to well-separated modes of varying scales.

1. The uniform density on $[0, 1]$.
2. The standard exponential density $f(x) = e^{-x}, x > 0$.
3. Maxwell's density $f(x) = xe^{-x^2/2}, x > 0$.
4. The Laplace density $f(x) = (1/2)e^{-|x|}$.
5. The logistic density $f(x) = e^{-x}/(1 + e^{-x})^2$.
6. The Cauchy density $f(x) = (1/\pi)(1 + x^2)^{-1}$.
7. The extreme value distribution. The distribution function is $F(x) = \exp(-\exp(-x))$.
8. The infinite peak distribution, having density $f(x) = 1/(2\sqrt{x})$ on $[0, 1]$.
9. The asymmetric Pareto distribution with parameter 3/2: it has density $f(x) = 1/(2x^{3/2})$ on $[1, \infty)$.
10. The symmetric Pareto distribution with parameter 3/2: it has density $f(x) = 1/(4(1 + |x|)^{3/2})$ on the real line.
11. The standard normal density.
12. The standard lognormal density: $f(x) = (1/x\sqrt{2\pi}) \exp(-(\log x)^2/2)$ on $[0, \infty)$.
13. A uniform mixture: 50% weight is put on a uniform $[-1/2, 1/2]$ distribution, and 50% weight on a uniform $[-5, 5]$ distribution.
14. The Matterhorn: an incredibly peaked density defined as the density of $Se^{-2/U}$, where S is a random sign, and U is uniformly distributed on $[0, 1]$. The density

has support on $[-1/e^2, 1/e^2]$ and is given by $f(x) = 1/(|x|(\log(|x|)^2))$.

15. The density of UV, the product of two independent uniform $[0,1]$ random variables: $f(x) = -\log(x)$ on $[0,1]$.
16. The isosceles triangular density: $f(x) = (1-|x|)_+$.
17. The beta $(2,2)$ density $f(x) = 6x(1-x)$, $0 \le x \le 1$.
18. The chi-square density with one degree of freedom: $f(x) = (1/\sqrt{2\pi x})e^{-x/2}$, $x > 0$.
19. The normal cubed distribution: the distribution of N^3, where N is a standard normal random variable.
20. The inverse exponential distribution: the distribution of $1/E^2$, where E is a standard exponential random variable. The distribution function is $F(x) = e^{-1/\sqrt{x}}$.
21. The marronite density: if $\phi(\mu, \sigma)$ denotes the normal density with mean μ and standard eviation σ, define
$$f = \frac{1}{3}\phi(-20, 1/4) + \frac{2}{3}\phi(0,1) \ .$$
22. The skewed bimodal density: another normal mixture (density # 8 in Marron and Wand, 1992), with
$$f = \frac{3}{4}\phi(0,1) + \frac{1}{4}\phi(1.5, 1/3) \ .$$
23. The claw density: a normal mixture (density # 10 in Marron and Wand, 1992), with
$$f = \frac{1}{2}\phi(0,1) + \frac{1}{10}\phi(-1, 0.1) + \frac{1}{10}\phi(-0.5, 0.1) + \frac{1}{10}\phi(0, 0.1) + \frac{1}{10}\phi(0.5, 0.1) + \frac{1}{10}\phi(1, 0.1) \ .$$
24. The smooth comb: a normal mixture (density # 14 in Marron and Wand, 1992), with
$$\begin{aligned} f = {} & \frac{32}{63}\phi\left(-\frac{31}{21}, \frac{32}{63}\right) + \frac{16}{63}\phi\left(\frac{17}{21}, \frac{16}{63}\right) + \frac{8}{63}\phi\left(\frac{41}{21}, \frac{8}{63}\right) \\ & + \frac{4}{63}\phi\left(\frac{53}{21}, \frac{4}{63}\right) + \frac{2}{63}\phi\left(\frac{59}{21}, \frac{2}{63}\right) + \frac{1}{63}\phi\left(\frac{62}{21}, \frac{1}{63}\right) \ . \end{aligned}$$
25. The caliper: The density of $S(X + 0.1)$, where S is a random sign, and X has density $f(x) = 4(1 - x^{1/3})$ on $[0,1]$.
26. The trimodal uniform density: $f = 0.5 f_{[-1,1]} + 0.25 f_{[20,20.1]} + 0.25 f_{[-20.1,-20]}$, where $f_{[a,b]}$ denotes the uniform density on $[a,b]$.

27. The sawtooth density: the density of $N + X$, where N is uniformly distributed in $\{-9, -7, -5, -3, -1, 1, 3, 5, 7, 9\}$, and X has the isosceles triangular density on $[-1, 1]$.

28. The bilogarithmic peak: $f(x) = -(1/2)\log(x(1-x))$ on $[0, 1]$. This is the only density with two separated infinite peaks, and an outspoken U-shape in the middle. It also is the mixture of two logarithmic peak densities.

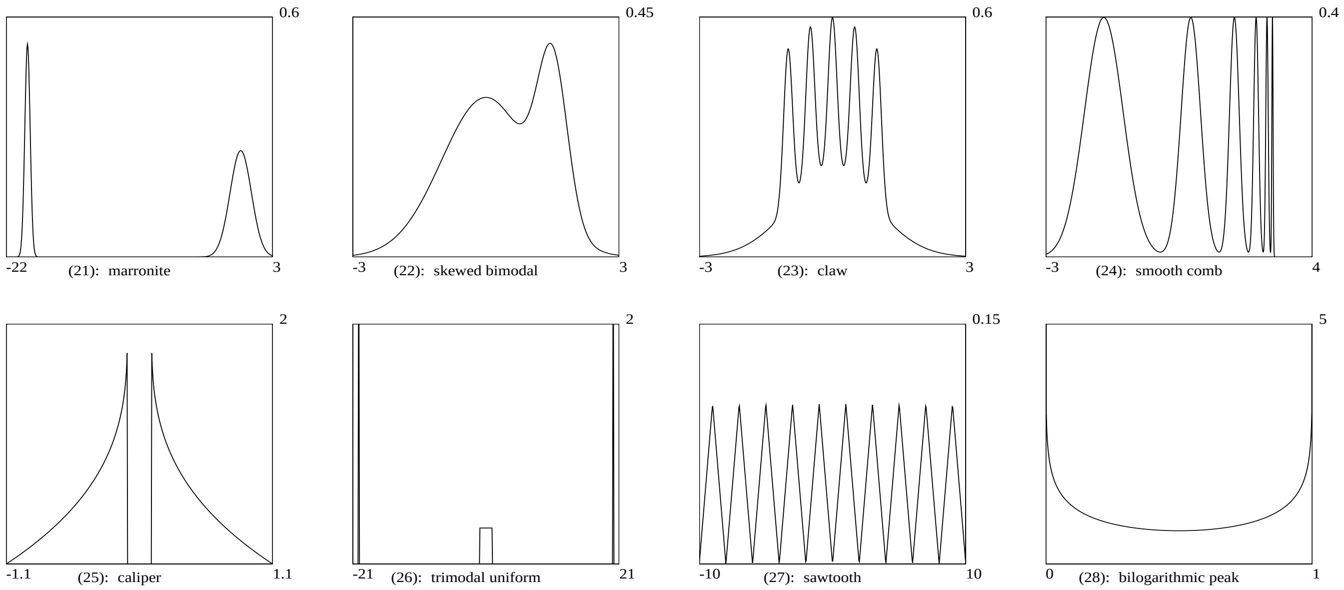

FIGURE 2. The multimodal densities in our collection.

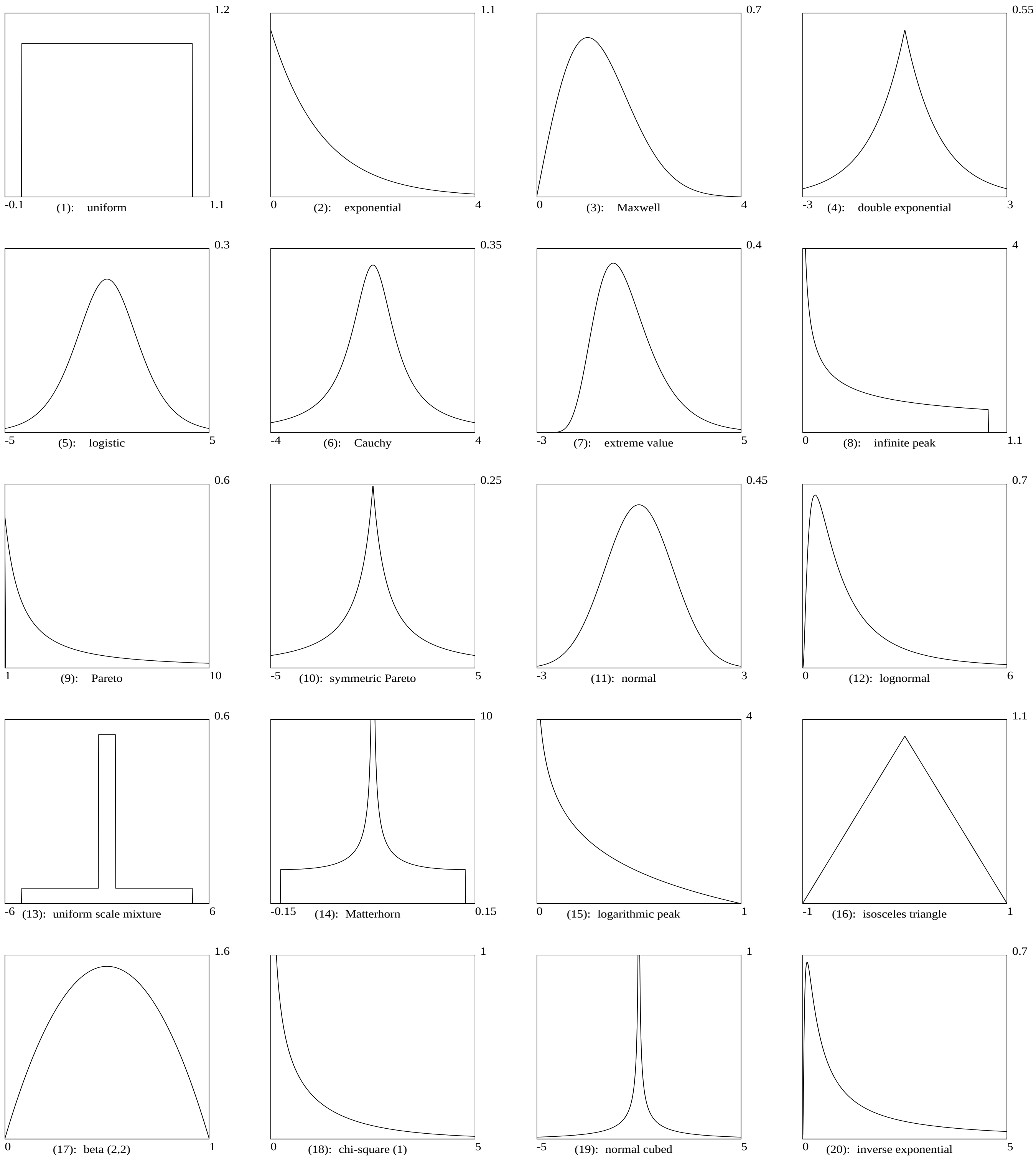

Figure 3. The unimodal densities in our collection.

3.3. <u>The comparative simulation: the results</u>. For each of the 20 densities, we generated 20 samples of size 100 each, and tried 17 different bandwidth selectors. All programs were written in PASCAL. The computation of $\int |.|$ needed in various places was done with great care as standard numerical integration routines are unsatisfactory under the extreme circumstances encountered here, especially when h is extremely small or very large. For example, if we have two density functions f and g, and if we can identify a finite number of intervals $A_j = (a_j, b_j)$ for the set

$$\{f > g\} = \cup_{j=1}^{k} A_j$$

(by solving $f = g$), then we have

$$\int |f - g| = 2 \sum_{j=1}^{k} (F(b_j) - F(a_j) - G(b_j) + G(a_j)) \ ,$$

where F and G are the distribution functions for f and g respectively. This sort of property aids tremendously in getting precise numerical results. Densities with infinite peaks and large tails are easy to deal with in this setting, while numerical integration is known to be problematic.

The following quantities are estimated for each density:

A. The average L_1 error, i.e., the average value of $\int |f - f_{nH}|$, where H is the (random) bandwidth. In one case, $h_{\texttt{op}}$, we take for H the optimal bandwidth:

$$h_{\texttt{op}} \stackrel{\text{def}}{=} \arg\min_{h>0} \int |f - f_{nh}| \ .$$

B. The average relative L_1 error, i.e., the average value of

$$P_n = \frac{\int |f - f_{nH}|}{\inf_{h>0} \int |f - f_{nh}|} - 1 \ .$$

C. The probability that the relative L_1 error P_n exceeds 0.1: $\mathbf{P}\{P_n > 0.1\}$.

D. The probability that the relative L_1 error P_n exceeds 0.5: $\mathbf{P}\{P_n > 0.5\}$.

E. The average rank of each method, where methods are ranked from 1 to 17 according to L_1 error on every run.

F. The probability of oversmoothing with respect to the optimal bandwidth.

G. The average value of the bandwidth.

H. The maximal value of P_n observed over the runs.

Our results still have a lot of residual variability. This may be reduced by averaging over collections ("baskets") of densities, such as the log-concave densities, the continuous densities, etcetera. Such averaging also counteracts cheating, as it is much harder to fine-tune design parameters in bandwidths to work well uniformly over large classes of sets. In fact, density averaging is like calculating stock market indices.

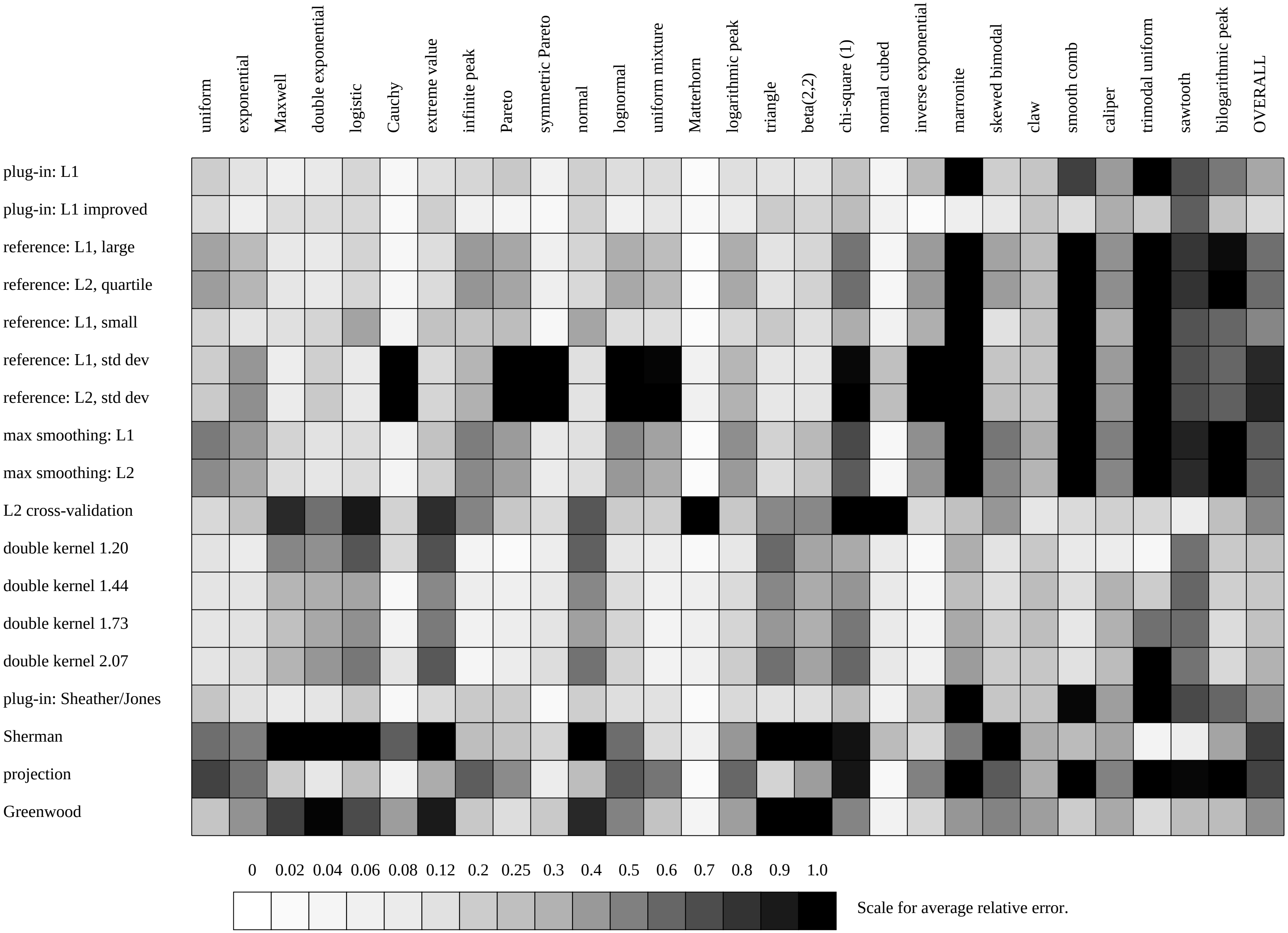

FIGURE 4. The average relative error (P_n) is shown for all densities.

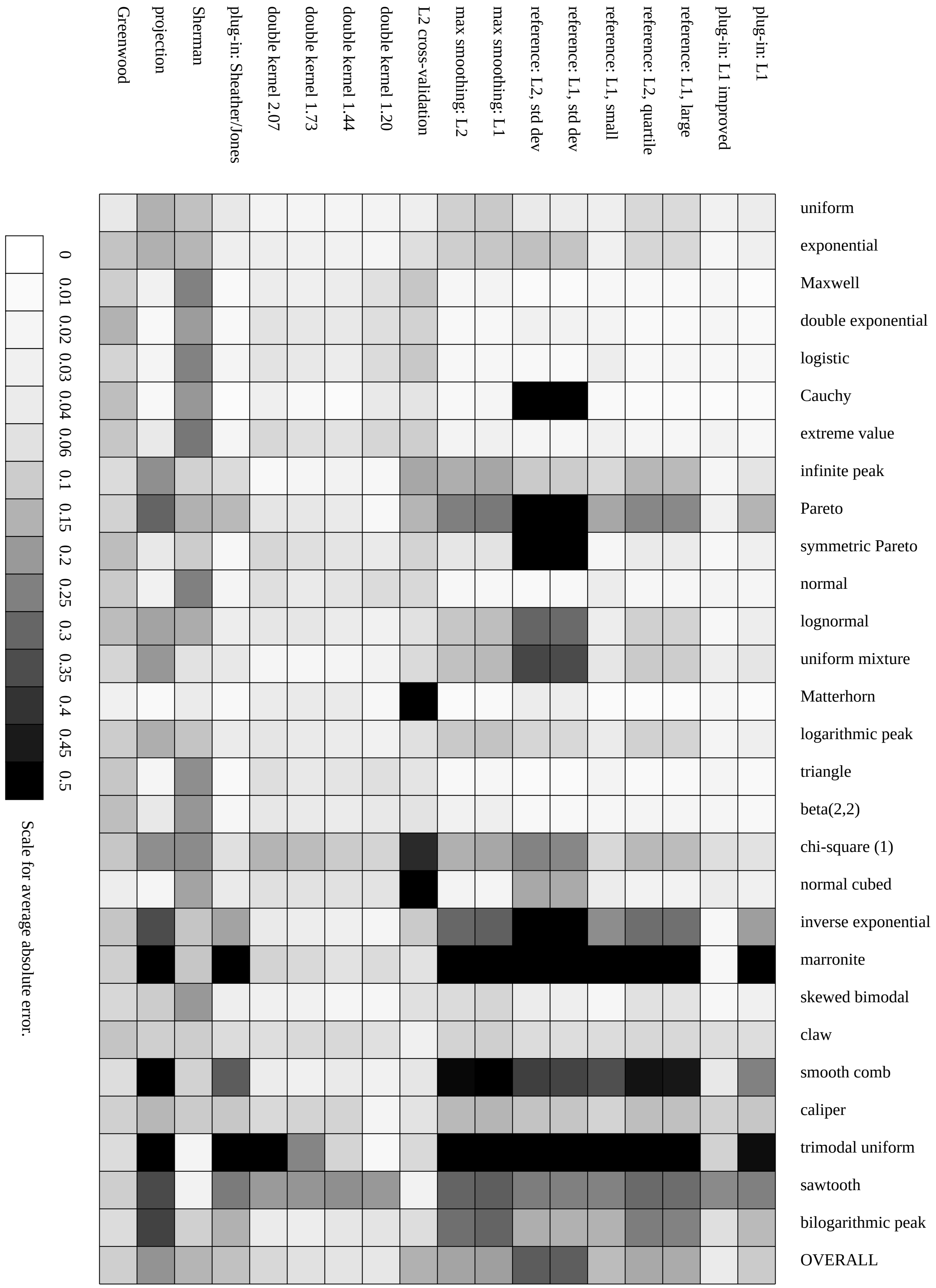

FIGURE 5. The average absolute error, defined as the average of $\int |f_{nH} - f|$ minus $\inf_h \int |f_{nh} - f|$, is shown for all densities.

3.4. Bounds for minimization. When minimizing either $J_{nh} = \int |f_{nh} - f|$ or $J'_{nh} = \int |f_{nh} - g_{nh}|$ with respect to h, we are faced with a multimodal optimization problem over an unbounded interval. In all cases, it is possible to quickly detect a finite interval $[a, b]$ to which we may restrict the search. It is possible to find simple functions $\chi(h)$ and $\xi(h)$ with the following property: $\chi(h) \downarrow$, $\chi(0) = 2$, $\xi(h) \uparrow 2$ as $h \uparrow \infty$, and

$$J_{nh} \geq \max(\chi(h), \xi(h)) \ .$$

The constant a is then easily determined as the largest number of the form $h_{\texttt{ref,l1}} / 2^i$ with the property that $\chi(a) > J_{n\, h_{\texttt{ref,l1}}}$, and b as the smallest number of the form $h_{\texttt{ref,l1}} \times 2^i$ with the property that $\xi(a) > J_{n\, h_{\texttt{ref,l1}}}$. This procedure works with any starting point, not just $h_{\texttt{ref,l1}}$. For J'_{nh}, the same thing is valid, except that the limits of the functions χ and ξ are $\int |K - L|$, not 2. The following functions are valid for J_{nh} when f is unimodal with mode at m. We let s be the upper bound of the support of the kernel K (one, for the Epanechnikov kernel). In what follows, F and F_n are the distribution functions for f and f_{nh} respectively, $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. We also assume that K has a mode at zero, and define $u \leq v$ as the two roots

$$u = \inf\{x : x \leq m; f(x) \geq K(0)/h\} \ , v = \sup\{x : x \geq m; f(x) \geq K(0)/h\} \ ,$$

These numbers are on both sides of the mode of f.

$$\begin{aligned}\chi(h) = \max \Big(& 2(F(m - 2nhs) + 1 - F(m + 2nhs)) \ , \\ & \sum_{i=0}^{n} 2(F(X_{(i+1)} - hs) - F(X_{(i)} + hs))_+ \Big) \ ;\end{aligned}$$

$$\begin{aligned}\xi(h) = \max \Big(& 2(F(v) - F(u) - (v - u)K(0)/h) \ , \\ & 2(F_n(X_{(1)}) + 1 - F_n(X_{(n)})) - 2(F(X_{(1)}) + 1 - F(X_{(n)})) \Big) \ .\end{aligned}$$

For J'_{nh}, we cannot use the unimodality of f, and are therefore somewhat more restricted. Let $\widehat{\mu}$ be the sample mean, and let C be the Lipschitz constant for $K - L$. Then

$$\xi(h) = \int |K - L| - \frac{1}{n} \sum_{i=1}^{n} \min \left(2 \int |K - L| \ , \ \frac{(2h + |X_{(i)} - \widehat{\mu}|) C |X_{(i)} - \widehat{\mu}|}{h^2} \right) \ ;$$

$$\chi(h) = \frac{\int |K - L|}{n} \left\{ \sum_{i=2}^{n-1} I_{[X_{(i-1)} + 2h \leq X_{(i)} \leq X_{(i+1)} - 2h]} + I_{[X_{(1)} + 2h \leq X_{(2)}]} + I_{[X_{(n-1)} + 2h \leq X_{(n)}]} \right\} \ .$$

These bounds are used in all our computations of $\inf J_{nh}$ and $\inf J'_{nh}$. The minimization is also simplified because J_{nh} and J'_{nh} satisfy the following simple smoothness property:

$$|J_{nh} - J_{nh'}| \leq \int |K_h - K_{h'}| \leq \frac{C|h - h'|}{\max(h, h')} \ ,$$

where C is some finite constant depending upon the kernel. Also,

$$|J'_{nh} - J'_{nh'}| \leq \int |(K-L)_h - (K-L)_{h'}| \leq \frac{C|h-h'|}{\max(h,h')} \ ,$$

where C is some finite constant depending upon $K - L$.

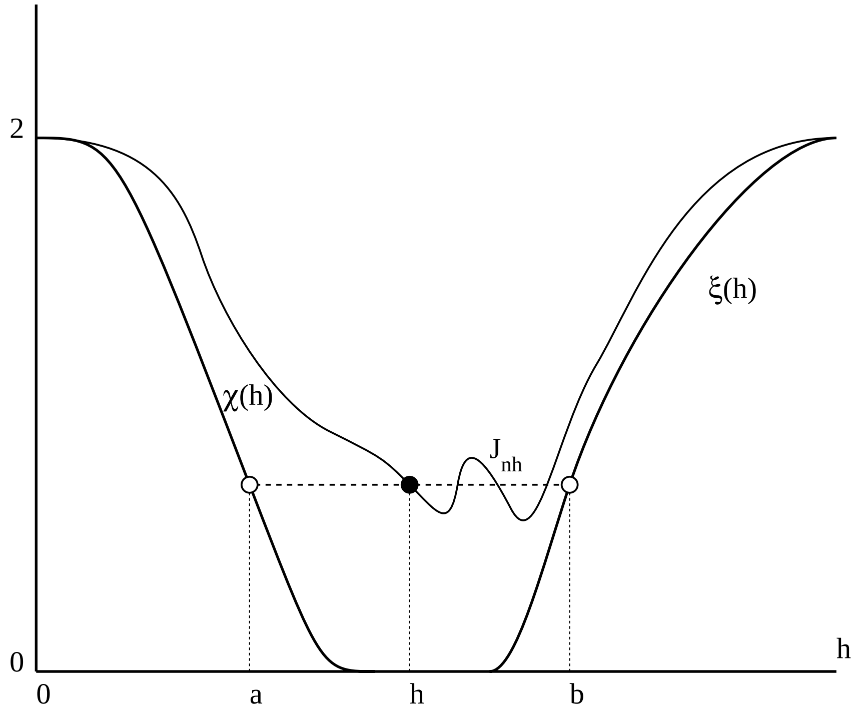

FIGURE 6. A hypothetical L_1 error J_{nh} is shown as a function of h. Also shown are the lower bounds $\chi(h)$ and $\xi(h)$. The figure illustrates how one computation of J_{nh} directly leads to an interval $[a,b]$ that contains the overall minimum, and yet stays bounded away from 0 and ∞.

3.5. Interpretation of the results. The variability of the results may be measured by the ratio of the worst relative error over the average relative error, although some may argue that this criterion itself is too "variable". As a measure of general trends, it will do. We found the reference methods and the plug-in methods to be amazingly stable in this respect.

For every density, we call "admissible" a method whose average relative error and average absolute error are not both simultaneously dominated by those of another method.

$h_{\texttt{ref,L1}}$	double exponential	Matterhorn	
$h_{\texttt{ref,l1}}$			
$h_{\texttt{ref,L2}}$			
$h_{\texttt{DH,L1}}$	logistic		
$h_{\texttt{DH,L2}}$	isosceles triangle	normal	beta (2,2)
$h_{\texttt{ms,L1}}$			
$h_{\texttt{ms,L2}}$			
$h_{\texttt{pi,l1}}$	extreme value	chi square (1)	Maxwell
$h_{\texttt{pi,L1}}$	exponential	Cauchy	marronite
	logarithmic peak	lognormal	inverse exponential
	skewed bimodal		
$h_{\texttt{pi,L2}}$	symmetric Pareto	Cauchy	
$h_{\texttt{cv}}$	claw		
$h_{\texttt{sh}}$	sawtooth		
$h_{\texttt{gr}}$			
$h_{\texttt{pr}}$	normal cubed		
$h_{\texttt{dk,1}}$	smooth comb	caliper	trimodal uniform
	Pareto		
$h_{\texttt{dk,2}}$	uniform		
$h_{\texttt{dk,3}}$	uniform	uniform mixture	bilogarithmic peak
$h_{\texttt{dk,4}}$	infinite peak		

TABLE 4. For every bandwidth selector, we list the test densities for which the selector is admissible according to average absolute and relative errors.

Table 4 shows why density estimation is fascinating—every method seems to "like" certain types of densities. The L_1-based plug-in methods are admissible in the above sense for 16 out of the 28 densities. Of these 16, 10 are densities for which the rate $n^{-2/5}$ is not achievable because of either a big tail or a discontinuity. We provide a method–by–method discussion.

A. REFERENCE DENSITY METHODS. Except when faced with simple smooth unimodal densities such as the normal or beta (2,2), reference density methods are predictably overshadowed. This is especially noticeable for multimodal densities. Also, $h_{\texttt{ms,L1}}$ and $h_{\texttt{ms,L2}}$ are almost always too large.

B. L_2 CROSS-VALIDATION. This is among the good methods in some multimodal settings when many minor peaks are present (as for the claw and sawtooth densities). This may be a bit a case of good luck as the bandwidth is typically much too small. For some peaked densities (Matterhorn, chi square (1), normal cubed), its performance is dismal, perhaps showing at even this small sample size that the method is inconsistent.

C. PROJECTION METHOD. This uncompetitive toy method behaves very much as the reference density methods.

D. SPACINGS METHODS. When we minimize the maximal value of the average absolute error over all densities, the bandwidth $h_{\mathtt{gr}}$ comes out well ahead of the other methods. Its theoretical properties may show some surprising universal strengths. Nevertheless, $h_{\mathtt{gr}}$ routinely oversmooths because it tries to put some of the mass of f_{nH} in each interval between the data points. The bandwidth $h_{\mathtt{sh}}$ undersmooths nearly all the time. Both $h_{\mathtt{gr}}$ and $h_{\mathtt{sh}}$ are good but not excellent in multimodal situations, as they are based on scale-less principles.

E. THE L_2 PLUG-IN METHOD. $h_{\mathtt{pi,L2}}$ is excellent for the collection of unimodal densities, thus confirming what is known from other studies. In multimodal situations, our version performs poorly because we are using a reference density pilot bandwidth. It is not clear how the pilot bandwidth problem may be resolved without eventually resorting to a universal method such as $h_{\mathtt{cv}}$, $h_{\mathtt{gr}}$ or $h_{\mathtt{dk,1}}$.

F. THE DOUBLE KERNEL METHOD. In absolute terms, only the sawtooth density brought $h_{\mathtt{dk,1}}$ to its knees. In relative terms, it also lost ground on very smooth small-tailed densities. The method shines under special circumstances—infinite peaks, large tails, tricky multimodal densities. As a result, averaged over the collection, its performance is better than any method discussed thus far. The bandwidths are usually a bit small and show some variation. Also, $h_{\mathtt{dk,4}}$ and $h_{\mathtt{dk,3}}$ are less competitive. Particular success was achieved for the Cauchy, infinite peak, Pareto, Matterhorn, inverse exponential, skewed bimodal, smooth comb, caliper, trimodal uniform, and bilogarithmic densities.

G. L_1 PLUG-IN METHODS. Finally, $h_{\mathtt{pi,l1}}$ shows the same pattern as $h_{\mathtt{pi,L2}}$ across the board. Its modification $h_{\mathtt{pi,L1}}$ however combines the stability of the plug-in methods with the robustness of the double kernel method to produce the overall winner for the average relative error. On some densities (lognormal, inverse exponential, marronite) it leaves the competition far behind. On average, it is the best method. It also has the lowest maximal value of average absolute error over the collection of densities (in a virtual tie with $h_{\mathtt{dk,1}}, h_{\mathtt{dk,2}}, h_{\mathtt{dk,3}}$).

Table 5 gives the performances, averaged over the set of 28 densities. This includes the average L_1 error, the average relative error, the average worst relative error, the estimate of the probability that the relative error exceeds 10% and 50%, and the average rank. The L_1-based plug-in methods are first with respect to all criteria. Among the double kernel methods, $h_{\mathtt{dk,2}}$ is consistently best. The most important entries in the table are those for the average error, and the average rank. They clearly confirm studies performed by others (Cao et al, 1994) that indicate the power of plug-in methods. According to every criterion, $h_{\mathtt{pi,L1}}$ is best on average. It is always followed by the double kernel method or $h_{\mathtt{pi,l1}}$. For the average error, the new method $h_{\mathtt{gr}}$ beats $h_{\mathtt{pi,L2}}$ and $h_{\mathtt{cv}}$.

	average error	average relative error	worst relative error	probability {relative error > 10 pct}	probability {relative error > 50 pct}	average rank	average smoothing factor	probability {oversmoothing}
optimal	0.327	0.000	0.000	0.000	0.000	0.000	0.593	0.000
plug-in: L1	0.429	0.345	0.784	0.583	0.185	6.957	1.397	0.775
plug-in: L1 improved	0.366	0.146	0.613	0.391	0.075	5.764	0.794	0.617
reference: L1, large	0.491	0.563	1.137	0.680	0.275	9.539	2.247	0.826
reference: L2, quartile	0.495	0.576	1.157	0.685	0.280	10.416	2.312	0.846
reference: L1, small	0.459	0.472	1.069	0.610	0.216	8.107	1.641	0.671
reference: L1, std dev	0.643	0.843	1.501	0.789	0.446	11.226	2.72505e+06	0.903
reference: L2, std dev	0.647	0.858	1.526	0.798	0.453	12.255	2.80397e+06	0.912
max smoothing: L1	0.516	0.652	1.260	0.758	0.348	12.789	2.672	0.898
max smoothing: L2	0.505	0.614	1.198	0.724	0.312	11.625	2.496	0.876
L2 cross-validation	0.480	0.472	1.496	0.639	0.298	9.612	0.291	0.107
double kernel 1.20	0.375	0.237	0.811	0.467	0.164	7.553	0.496	0.323
double kernel 1.44	0.379	0.220	0.931	0.467	0.137	7.148	0.618	0.433
double kernel 1.73	0.386	0.239	0.963	0.480	0.137	7.292	0.620	0.367
double kernel 2.07	0.404	0.301	1.055	0.525	0.166	8.185	0.673	0.308
plug-in: Sheather/Jones	0.448	0.421	0.935	0.594	0.196	7.685	1.490	0.760
Sherman	0.473	0.762	1.845	0.787	0.466	12.012	0.200	0.074
projection	0.538	0.741	1.523	0.724	0.453	13.076	3.280	0.826
Greenwood	0.422	0.437	1.722	0.639	0.258	9.658	1.306	0.667

Table 5. A summary of the results, averaged over 28 test densities and 20 repetitions each, with $n = 100$.

4. Complementary discussions.

4.1. Fine-tuning parameters. One of the aims of this paper is to present new L_1-based plug-in kernel estimates. We do not claim that this research is finished. We are encouraged though by the results of our experiments. Both plug-in estimates have parameters that may be further fine-tuned. To do this, we ran two more full tests on the 20 unimodal densities, with 20 repetitions per density, $n = 100$. We do not have the space here to report on the full extent of these experiments. Instead, we only give the summary table for the 400 tests.

The tests involve $h_{\texttt{pi,l1}}$ and $h_{\texttt{pi,L1}}$, and the parameter picked here is the threshold τ, which was 1/10 in our main round of simulations. Here, τ is varied to prove that its actual value does not matter much. The criterion shows great robustness with respect to the threshold parameter.

method	average L_1 error	average P_n	$\mathbf{P}\{P_n > 0.1\}$	$\mathbf{P}\{P_n > 0.5\}$	average rank	probability oversmoothing
$h_{\texttt{op}}$	0.3255					
$h_{\texttt{pi,l1}}, \tau = 1$	0.3601	0.1616	0.4000	0.0600	5.3725	0.4075
$h_{\texttt{pi,l1}}, \tau = 0.5625$	0.3563	0.1319	0.3300	0.0425	4.8500	0.4450
$h_{\texttt{pi,l1}}, \tau = 0.36$	0.3541	0.1108	0.2850	0.0275	4.7125	0.5100
$h_{\texttt{pi,l1}}, \tau = 0.25$	0.3551	0.1046	0.3250	0.0250	4.6850	0.5750
$h_{\texttt{pi,l1}}, \tau = 0.1837$	0.3574	0.1055	0.3600	0.0225	4.7525	0.6425
$h_{\texttt{pi,l1}}, \tau = 0.1406$	0.3602	0.1107	0.3775	0.0200	5.0450	0.6875
$h_{\texttt{pi,l1}}, \tau = 0.1111$	0.3634	0.1179	0.4150	0.0225	5.4675	0.7200
$h_{\texttt{pi,l1}}, \tau = 0.09$	0.3665	0.1252	0.4400	0.0225	5.9475	0.7500
$h_{\texttt{pi,l1}}, \tau = 0.0744$	0.3694	0.1326	0.4625	0.0250	6.4950	0.7700
$h_{\texttt{pi,l1}}, \tau = 0.0625$	0.3723	0.1401	0.4900	0.0250	7.2050	0.7925

TABLE 6. The summary results for $h_{\texttt{pi,l1}}$ are shown when the threshold parameter τ is varied. Depending upon the criterion, the best τ is somewhere in the range $0.36 \ldots 0.14$.

method	average L_1 error	average P_n	$\mathbf{P}\{P_n > 0.1\}$	$\mathbf{P}\{P_n > 0.5\}$	average rank	probability oversmoothing
$h_{\mathtt{op}}$	0.3255					
$h_{\mathtt{pi,L1}}, \tau = 1$	0.3678	0.2001	0.4725	0.0975	6.8575	0.2550
$h_{\mathtt{pi,L1}}, \tau = 0.5625$	0.3615	0.1698	0.3975	0.0775	6.1800	0.3225
$h_{\mathtt{pi,L1}}, \tau = 0.36$	0.3572	0.1471	0.3475	0.0600	5.7050	0.3650
$h_{\mathtt{pi,L1}}, \tau = 0.25$	0.3541	0.1301	0.3100	0.0450	5.3450	0.4025
$h_{\mathtt{pi,L1}}, \tau = 0.1837$	0.3517	0.1173	0.2925	0.0350	5.0825	0.4325
$h_{\mathtt{pi,L1}}, \tau = 0.1406$	0.3500	0.1082	0.2800	0.0350	4.8900	0.4725
$h_{\mathtt{pi,L1}}, \tau = 0.1111$	0.3492	0.1029	0.2950	0.0325	4.9175	0.5275
$h_{\mathtt{pi,L1}}, \tau = 0.09$	0.3489	0.1007	0.3150	0.0275	5.0375	0.5700
$h_{\mathtt{pi,L1}}, \tau = 0.0744$	0.3492	0.1008	0.3200	0.0250	5.2625	0.6150
$h_{\mathtt{pi,L1}}, \tau = 0.0625$	0.3498	0.1024	0.3400	0.0225	5.6600	0.6450

TABLE 7. The summary results for $h_{\mathtt{pi,L1}}$ are shown when the threshold parameter τ is varied. Depending upon the criterion, the best τ is somewhere in the range $0.14 \dots 0.09$.

4.2. Catastrophic behavior. Our experiments are too limited to properly illustrate several important issues in density estimation. Most software users will undoubtedly be abhorred by possible catastrophic behavior of an estimate. Foremost among this is the consistency: is there a nonempty subclass $\mathcal{F}$ of densities for which

$$\inf_{f \in \mathcal{F}} \limsup_{n \to \infty} \mathbf{E} \int |f_{nH} - f| > 0 \ ?$$

All methods that rely somewhere on a scale factor computed as an average (such as $h_{\mathtt{DH,L1}}, h_{\mathtt{DH,L2}}$) fail this test whenever the scale estimate diverges (i.e., when f has a long tail). Many estimates we did not consider (including most bootstrap estimates) are ill-defined as the criterion to be minimized would yield $H = \infty$. Strictly speaking, they are not consistent. The maximum likelihood method is inconsistent whenever the tail of the distribution is at least as big as an exponential tail (Broniatowski, Deheuvels and Devroye, 1989). As pointed out in Devroye (1989d), the choice $h_{\mathtt{cv}}$ is inconsistent when the densities have too large infinite peaks. The choices $h_{\mathtt{sh}}$ and $h_{\mathtt{pr}}$ are also inconsistent for certain densities. The double kernel and plug-in bandwidths of this paper are universally consistent.

Another important point, also discussed in Jones, Marron and Sheather (1992), is that some methods do not pass a bimodality test. To put it simply, let g be a fixed unimodal density on $[0,1]$, and consider the family of bimodal densities

$$f(x) = pg(x) + (1-p)g(x-\delta) \ ,$$

where $\delta > 1$. Create an infinite family of samples from f as follows: start with n i.i.d. pairs drawn from (Y, U), where Y has density g and U is uniform $[0,1]$. Define

$$X = \begin{cases} Y & \text{if } U < p \\ Y + \delta & \text{otherwise} \end{cases} .$$

Then X has density f. Fix n. A kernel density estimate f_{nH} does not pass the bimodality test if for some g, almost surely,

$$\sup_{p,\delta} \int |f_{nH} - f| = 2$$

for the given sample. This would happen if as $\delta \to \infty$, we have $H \to \infty$. Densities that fail the bimodality test are typically based upon the reference density method in one step of the definition. These can be made to perform arbitrarily poorly in the sense given above. As such, the parameters $h_{\texttt{ref,L1}}$, $h_{\texttt{ref,l1}}$, $h_{\texttt{ref,L2}}$, $h_{\texttt{DH,L1}}$, $h_{\texttt{DH,L2}}$ are inadmissible. The same is true for $h_{\texttt{ms,L1}}$ and $h_{\texttt{ms,L2}}$. Plug-in methods invariably require the estimation of certain functionals. This typically forces one to solve another nonparametric estimation problem. A pilot bandwidth is introduced, which in turn depends upon an unknown functional. One may continue this chain, but eventually it has to come to an end (for a simulation that involves a variable number of layers in this chain, see Park and Marron, 1992). If a reference method is used at the end of the chain, then bimodal examples may be constructed that for sufficiently large n make the whole procedure useless. Absolute methods are those that end the estimation chain by appealing to an absolute principle, such as minimization by L_2 or L_∞ cross-validation, or the double kernel method. Only those will be totally immune against bimodal separation viruses. $h_{\texttt{pi,l1}}$ and $h_{\texttt{pi,L2}}$ are not immune. Among the tested bandwidths, only $h_{\texttt{dk,1}}$, $h_{\texttt{dk,2}}$, $h_{\texttt{dk,3}}$, $h_{\texttt{dk,4}}$, $h_{\texttt{pi,L1}}$, $h_{\texttt{sh}}$ and $h_{\texttt{cv}}$ are absolute and pass our bimodality test.

Robustness may be measured in many ways. Perhaps the most trivial way of measuring it is by what happens if we move one data point to different locations: we say that the density estimate is sensitive to one point if

$$\sup_{x_1} \int |f_{nH} - f| = 2$$

almost surely, where $H = H(x_1, X_2, \ldots, X_n)$. This would occur for example if with probability one, $\inf_{x_1} H(x_1, X_2, X_3, \ldots, X_n) = 0$ (as in the case of $h_{\texttt{cv}}$) or $\sup_{x_1} H(x_1, X_2, X_3, \ldots, X_n) = \infty$ (as in the case of $h_{\texttt{DH,L1}}$ or $h_{\texttt{DH,L2}}$). This idea may be generalized to insensitivity with respect to an ϵ-fraction of the sample.

While the above three criteria are very disconcerting for some estimates, we should not lose track of the original goal—to design an asymptotically optimal bandwidth selector. There is still an open question as to whether such a selector exists for all densities. So, we will water down things a bit by considering the class $\mathcal{N}$ of nice densities, that is, all densities on $[0,1]$ that have infinitely many continuous bounded derivatives on the real line. We say that H is expedient if

$$\sup_{f\in\mathcal{N}} \limsup_{n\to\infty} \frac{\mathbf{E}\int |f_{nH}-f|}{\inf_h \mathbf{E}\int |f_{nh}-f|} < \infty \ .$$

This criterion says that we come within a finite constant of the optimal performance for large n, uniformly over all nice densities. All L_2-based methods, including $h_{\mathtt{pi,L2}}$, fail this test. While it is true that for all nice densities, the optimal L_1 and L_2 choices for h differ by a constant factor only, the ratio is not uniformly bounded. Unfortunately, only $h_{\mathtt{pi,l1}}$, $h_{\mathtt{pi,L1}}$, and the double kernel choices $h_{\mathtt{dk,1}}, h_{\mathtt{dk,2}}, h_{\mathtt{dk,3}}, h_{\mathtt{dk,4}}$ are expedient. We hasten to add that we would like the supremum in the definition of expediency to be one. This is not the case for any method discussed in this paper. Possible remedies for this problem include a plug-in method as suggested in Hall and Wand (1988), or a modified double kernel method as discussed above.

method	consistent?	bimodality test?	insensitive to one point?	L_1 expedient?
robust reference methods				
$(h_{\texttt{ref,l1}}, h_{\texttt{ref,L1}}, h_{\texttt{ref,L2}})$	yes	yes	no	no
standard reference methods				
$(h_{\texttt{DH,L1}}, h_{\texttt{DH,L2}})$	no	no	no	no
maximal smoothing				
$(h_{\texttt{ms,L1}}, h_{\texttt{ms,L2}})$	yes	yes	no	no
L_1 plug-in				
$(h_{\texttt{pi,l1}}, h_{\texttt{pi,L1}})$	yes	no/yes	yes	yes
L_2 plug-in				
$(h_{\texttt{pi,L2}})$	yes	no/yes	yes	no
double kernel method				
$(h_{\texttt{dk,1}}, h_{\texttt{dk,2}}, h_{\texttt{dk,3}}, h_{\texttt{dk,4}})$	yes	yes	yes	yes
L_2 cross-validation				
$(h_{\texttt{cv}})$	no	yes	no	no
spacings methods				
$(h_{\texttt{sh}},\ h_{\texttt{gr}})$	no/yes	yes	yes	no
projection method				
$(h_{\texttt{pr}})$	no	no	yes	no
L_2 bootstrap	no	no	no	no
L_∞ maximum likelihood	no	no	no	no

TABLE 8. The table shows which methods pass the consistency, bimodality, insensitivity to one point, and expediency criteria.

Acknowledgements. We thank Antonio Cuevas, Ricardo Cao-Abad and Wenceslao Gonzalez-Manteiga for making their experimental results available to us. We also thank Steve Marron, who suggested including densities 21–24 in our simulation. Research of the authors was sponsored by Université Montpellier II, NSERC (grant A3456) and FCAR (grant 90-ER-0291). The simulation was carried out during a visit of Luc Devroye at the Unité de Biométrie at the Université de Montpellier.

References.

B. Abdous, "Adapting the classical kernel density estimator to data," *Computational Statistics and Data Analysis*, vol. 9, pp. 169–178, 1990.

H. Akaike, "An approximation to the density function," *Annals of the Institute of Statistical Mathematics*, vol. 6, pp. 127–132, 1954.

M. S. Bartlett, "Statistical estimation of density functions," *Sankhya Series A*, vol. 25, pp. 245–254, 1963.

S. J. Bean and C. P. Tsokos, "Developments in nonparametric density estimation," *International Statistical Review*, vol. 48, pp. 267–287, 1980.

A. Berlinet, "Reproducing kernels and finite order kernels," in: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 3–18, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

A. Berlinet, "Hierarchies of higher order kernels," *Probability Theory and Related Fields*, vol. 94, pp. 489–504, 1993.

D. Bosq and J.-P. Lecoutre, *Théorie de l' Estimation Fonctionnelle*, Economica, Paris, 1987.

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

A. W. Bowman, "A comparative study of some kernel-based non-parametric density estimators," *Journal of Statistical Computation and Simulation*, vol. 21, pp. 313–327, 1985.

J. Bretagnolle and C. Huber, "Estimation des densités: risque minimax," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 47, pp. 119–137, 1979.

M. Broniatowski, P. Deheuvels, and L. Devroye, "On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate," *Annals of Statistics*, vol. 17, pp. 1070–1086, 1989.

P. Burman, "A data dependent approach to density estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 69, pp. 609–628, 1985.

R. Cao, A. Cuevas, and W. González-Manteiga, "A comparative study of several smoothing methods in density estimation," *Computational Statistics and Data Analysis*, vol. 17, pp. 153–176, 1994.

R. Cao-Abad, "Aplicaciones y nuevos resultados del método bootstrap en la estimación no paramétrica de curvas," Ph.D. Dissertation, University of Santiago de Compostela, Spain, 1990.

R. C. H. Cheng and N. A. K. Amin, "Estimating parameters in continuous univariate distributions with a shifted origin," *Journal of the Royal Statistical Society*, vol. B45, pp. 394–403, 1983.

S.-T. Chiu, "Bandwidth selection for kernel density estimation," *Annals of Statistics*, vol. 19, pp. 1883–1905, 1991.

S.-T. Chiu, "An automatic bandwidth selector for kernel density estimation," *Biometrika*, vol. 79, pp. 771–782, 1992.

Y. S. Chow, S. Geman, and L. D. Wu, "Consistent cross-validated density estimation," *Annals of Statistics*, vol. 11, pp. 25–38, 1983.

D. B. H. Cline, "Admissible kernel estimators of a multivariate density," *Annals of Statistics*, vol. 16, pp. 1421–1427, 1988.

D. B. H. Cline, "Optimal kernel estimation of densities," *Annals of the Institute of Statistical Mathematics*, vol. 42, pp. 287–303, 1990.

P. Deheuvels, "Estimation non paramétrique de la densité par histogrammes generalisés," *Revue de Statistique Appliquée*, vol. 25, pp. 5–42, 1977a.

P. Deheuvels, "Estimation nonparamétrique de la densité par histogrammes generalisés," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 22, pp. 1–23, 1977b.

P. Deheuvels and P. Hominal, "Estimation automatique de la densité," *Revue de Statistique Appliquée*, vol. 28, pp. 25–55, 1980.

L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896–904, 1983.

L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

L. Devroye, *A Course in Density Estimation*, Birkhäuser, Boston, 1987.

L. Devroye, "The kernel estimate is relatively stable," *Probability Theory and Related Fields*, vol. 77, pp. 521–536, 1988a.

L. Devroye, "Asymptotic performance bounds for the kernel estimate," *Annals of Statistics*, vol. 16, pp. 1162–1179, 1988b.

L. Devroye, "Nonparametric density estimates with improved performance on given sets of densities," *Statistics (Mathematische Operationsforschung und Statistik)*, vol. 20, pp. 357–376, 1989a.

L. Devroye, “The double kernel method in density estimation,” *Annales de l’Institut Henri Poincaré*, vol. 25, pp. 533–580, 1989b.

L. Devroye, “A universal lower bound for the kernel estimate,” *Statistics and Probability Letters*, vol. 8, pp. 419–423, 1989c.

L. Devroye, “On the non-consistency of the L2 cross-validated kernel density estimate,” *Statistics and Probability Letters*, vol. 8, pp. 425–433, 1989d.

L. Devroye, “A note on the usefulness of superkernels in density estimation,” *Annals of Statistics*, vol. 20, pp. 2037–2056, 1992.

L. Devroye, “On the non-consistency of Chiu’s estimate,” *Statistics and Probability Letters*, vol. 0, pp. 0–0, 1993.

L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

L. Devroye and M. P. Wand, “On the effect of density shape on the performance of its kernel estimate,” *Statistics*, vol. 24, pp. 215–233, 1993.

R. P. W. Duin, “On the choice of smoothing parameters for Parzen estimators of probability density functions,” *IEEE Transactions on Computers*, vol. C-25, pp. 1175–1179, 1976.

J. Engel, E. Herrmann, and T. Gasser, “An iterative bandwidth selector for kernel estimation of densities and derivatives,” Technical Report, University of Heidelberg, 1992.

V. A. Epanechnikov, “Nonparametric estimation of a multivariate probability density,” *Theory of Probability and its Applications*, vol. 14, pp. 153–158, 1969.

M. Falk, “Bootstrap optimal bandwidth selection for kernel density estimation,” *Journal of Statistical Planning and Inference*, vol. 30, pp. 13–22, 1992.

J. Fan and T. C. Hu, “Bias correction and higher order kernel functions,” *Statistics and Probability Letters*, vol. 13, pp. 235–243, 1992.

J. Fan and J. S. Marron, “Best possible constant for bandwidth selection,” *Annals of Statistics*, vol. 20, pp. 2057–2070, 1992.

J. Faraway and M. Jhun, “Bootstrap choice of bandwidth for density estimation,” *Journal of the American Statistical Association*, vol. 85, pp. 1119–1122, 1990.

L. Gajek, “Estimating a density and its derivatives via the minimum distance method,” *Probability Theory and Related Fields*, vol. 80, pp. 601–617, 1989.

T. Gasser, H.-G. Müller, and V. Mammitzsch, "Kernels for nonparametric curve estimation," *Journal of the Royal Statistical Society, Series B*, vol. 47, pp. 238–252, 1985.

B. L. Granovsky and H.-G. Müller, "On the optimality of a class of polynomial kernel functions," *Statistics and Decisions*, vol. 7, pp. 301–312, 1989.

M. Greenwood, "The statistical study of infections," *Journal of the Royal Statistical Society*, vol. A109, pp. 85–110, 1946.

J. D. F. Habbema, J. Hermans, and K. Vandenbroek, "A stepwise discriminant analysis program using density estimation," in: COMPSTAT *1974*, ed. G. Bruckmann, pp. 101–110, Physica Verlag, Wien, 1974.

P. Hall, "Cross-validation in density estimation," *Biometrika*, vol. 69, pp. 383–390, 1982.

P. Hall, "Large-sample optimality of least squares cross-validation in density estimation," *Annals of Statistics*, vol. 11, pp. 1156–1174, 1983.

P. Hall, "Asymptotic theory of minimum integrated square error for multivariate density estimation," in: *Multivariate Analysis VI*, ed. P. R. Krishnaiah, pp. 289–309, North-Holland, Amsterdam, 1985.

P. Hall, "Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems," *Journal of Multivariate Analysis*, vol. 32, pp. 177–203, 1990.

P. Hall, T. J. DiCiccio, and J. P. Romano, "On smoothing and the bootstrap," *Annals of Statistics*, vol. 17, pp. 692–704, 1989.

P. Hall and I. M. Johnstone, "Empirical functionals and efficient smoothing parameter selection," *Journal of the Royal Statistical Society*, vol. B54, pp. 475–530, 1992.

P. Hall and J. S. Marron, "On the amount of noise inherent in bandwidth selection of a kernel density estimator," *Annals of Statistics*, vol. 15, pp. 163–181, 1987a.

P. Hall and J. S. Marron, "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation," *Probability Theory and Related Fields*, vol. 74, pp. 567–581, 1987b.

P. Hall and J. S. Marron, "Lower bounds for bandwidth selection in density estimation," *Probability Theory and Related Fields*, vol. 90, pp. 149–173, 1990.

P. Hall and J. S. Marron, "Local minima in cross-validation functions," *Journal of the Royal Statistical Association*, vol. B53, pp. 245–252, 1991.

P. Hall, J. S. Marron, and B. Park, "Smoothed cross-validation," *Probability Theory and Related Fields*, vol. 92, pp. 1–20, 1992.

P. Hall, S. J. Sheather, M. C. Jones, and J. S. Marron, "On optimal data-based bandwidth selection in kernel density estimation," *Biometrika*, vol. 78, pp. 263–269, 1992.

P. Hall and M. P. Wand, "Minimizing L_1 distance in nonparametric density estimation," *Journal of Multivariate Analysis*, vol. 26, pp. 59–88, 1988.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

A. J. Izenman, "Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, pp. 205–225, 1991.

P. Janssen, J. S. Marron, N. Veravereke, and W. Sarle, "Scale measures for bandwidth selection," Technical Report, University of Limburg, Belgium, 1992.

M. C. Jones, "Changing kernels' orders," Technical Report, Department of Statistics, The Open University, Milton Keynes, U.K., 1990.

M. C. Jones, "The roles of ISE and MISE in density estimation," *Statistics and Probability Letters*, vol. 12, pp. 51–56, 1991b.

M. C. Jones, "Prospects for automatic bandwidth selection in extensions to basic kernel density estimation," in: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 241–250, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991b.

M. C. Jones and R. F. Kappenman, "On a class of kernel density estimate bandwidth selectors," *Scandinavian Journal of Statistics*, vol. 19, pp. 337–349, 1992.

M. C. Jones, J. S. Marron, and B. U. Park, "A simple root n bandwidth selector," *Annals of Statistics*, vol. 19, pp. 1919–1932, 1991.

M. C. Jones, J. S. Marron, and S. J. Sheather, "Progress in data-based bandwidth selection for kernel density estimation," Mimeo series 2088, Department of Statistics, University of North Carolina, 1992.

M. C. Jones and S. J. Sheather, "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives," *Statistics and Probability Letters*, vol. 11, pp. 511–514, 1991.

W. C. Kim, B. U. Park, and J. S. Marron, "Asymptotically best bandwidth selectors in kernel density estimation," *Statistics and Probability Letters*, vol. 0, pp. 0–0, 1993.

W. C. Kim, B. U. Park, and J. S. Marron, "Asymptotically best bandwidth selectors in kernel density estimation," *Statistics and Probability Letters*, vol. 0, pp. 0–0, 1993.

E. Mammen, "A short note on optimal bandwidth selection for kernel estimators," *Statistics and Probability Letters*, vol. 9, pp. 23–25, 1990.

J. S. Marron, "An asymptotically efficient solution to the bandwidth problem of kernel density estimation," *Annals of Statistics*, vol. 13, pp. 1011–1023, 1985.

J. S. Marron, "Will the art of smoothing ever become a science?," *Contemporary Mathematics*, vol. 59, pp. 169–178, 1986.

J. S. Marron, "A comparison of cross-validation techniques in density estimation," *Annals of Statistics*, vol. 15, pp. 152–162, 1987.

J. S. Marron, "Automatic smoothing parameter selection: a survey," *Empirical Economics*, vol. 13, pp. 187–208, 1988.

J. S. Marron, "Automatic smoothing parameter selection: a survey," in: *Semiparametric and Nonparametric Economics*, ed. A. Ullah, pp. 65–86, Heidelberg, 1989a.

J. S. Marron, "Comments on a data based bandwidth selector," *Computational Statistics and Data Analysis*, vol. 8, pp. 155–170, 1989b.

J. S. Marron, "Root n bandwidth selection," in: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 251–260, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

J. S. Marron, "Bootstrap bandwidth selection," in: *Exploring the Limits of the Bootstrap*, ed. R. LePage and L. Billard, pp. 249–262, John Wiley, New York, 1992.

J. S. Marron and M. P. Wand, "Exact mean integrated square error," *Annals of Statistics*, vol. 20, pp. 712–736, 1992.

C. McDiarmid, "On the method of bounded differences," in: *Surveys in Combinatorics 1989*, vol. 141, pp. 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge, 1989.

A. Mihoubi, "Bootstrap et validation croisée en estimation non paramétrique de la densité," Thèse de doctorat, Université Paris VI, 1992.

H.-G. Müller, "Smooth optimum kernel estimators of densities, regression curves and modes," *Annals of Statistics*, vol. 12, pp. 766–774, 1984.

E. A. Nadaraya, "On the integral mean square error of some nonparametric estimates for the density function," *Theory of Probability and its Applications*, vol. 19, pp. 133–141, 1974.

B.-U. Park, “On the plug-in bandwidth selectors in kernel density estimation,” *Journal of the Korean Statistical Society*, vol. 18, pp. 107–117, 1989.

B.-U. Park and J. S. Marron, “Comparison of data-driven bandwidth selectors,” *Journal of the American Statistical Association*, vol. 85, pp. 66–72, 1990.

B.-U. Park and J. S. Marron, “On the use of pilot estimators in bandwidth selection,” *Journal of Nonparametric Statistics*, vol. 1, pp. 231–240, 1992.

B.-U. Park and B. A. Turlach, “Practical performance of several data driven bandwidth selectors (with discussion),” *Computational Statistics*, vol. 7, pp. 251–270, 1992.

B.-U. Park and B. A. Turlach, “Practical performance of several data driven bandwidth selectors (with discussion),” *Computational Statistics*, vol. 7, pp. 251–270, 1992.

E. Parzen, “On the estimation of a probability density function and the mode,” *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

R. Pyke, “Spacings,” *Journal of the Royal Statistical Society Series B*, vol. 7, pp. 395–445, 1965.

B. Ranneby, “The maximum spacings method: an estimation method,” *Scandinavian Journal of Statistics*, vol. 11, pp. 93–112, 1984.

Z. El abdin Ras, “Contribution à la Théorie des Espacements,” Thèse de Doctorat d’État ès Sciences Mathématiques, Université Pierre et Marie Curie, 1989.

R.-D. Reiss, “Sharp rates of convergence of minimum penalized distance estimators,” *Sankhya, Series A*, vol. 48, pp. 59–68, 1986.

K. Roeder, “Density estimation with confidence sets exemplified by superclusters and voids in the galaxies,” *Journal of the American Statistical Association*, vol. 85, pp. 617–624, 1990.

M. Rosenblatt, “Remarks on some nonparametric estimates of a density function,” *Annals of Mathematical Statistics*, vol. 27, pp. 832–837, 1956.

M. Rosenblatt, “Global measures of deviation for kernel and nearest neighbor density estimates,” in: *Proceedings of the Heidelberg Workshop*, pp. 181–190, Springer Lecture Notes in Mathematics 757, Springer-Verlag, Berlin, 1979.

M. Rudemo, “Empirical choice of histograms and kernel density estimators,” *Scandinavian Journal of Statistics*, vol. 9, pp. 65–78, 1982.

W. R. Schucany and J. P. Sommers, “Improvement of kernel type density estimators,” *Journal of the American Statistical Association*, vol. 72, pp. 420–423, 1977.

E. F. Schuster and G. G. Gregory, "On the nonconsistency of maximum likelihood nonparametric density estimators," in: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ed. W. F. Eddy, pp. 295–298, Springer Verlag, New York, N.Y., 1981.

D. W. Scott, *Multivariate Density Estimation*, John Wiley, New York, 1992.

D. W. Scott and L. E. Factor, "Monte Carlo study of three data-based nonparametric probability density estimators," *Journal of the American Statistical Association*, vol. 76, pp. 9–15, 1981.

D. W. Scott, R. A. Tapia, and J. R. Thompson, "Kernel density estimation revisited," *Nonlinear Analysis*, vol. 1, pp. 339–372, 1977.

D. W. Scott and G. R. Terrell, "Biased and unbiased cross-validation in density estimation," *Journal of the American Statistical Association*, vol. 82, pp. 1131–1146, 1987.

S. J. Sheather, "The performance of six popular bandwidth selection methods on some real data sets (with discussion)," *Computational Statistics*, vol. 0, pp. 0–0, 1993.

S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society*, vol. B53, pp. 683–60, 1991.

B. Sherman, "A random variable related to the spacing of sample values," *Annals of Mathematical Statistics*, vol. 21, pp. 339–361, 1950.

R. S. Singh, "Mean squared errors of estimates of a density and its derivatives," *Biometrika*, vol. 66, pp. 177–180, 1979.

A. R. Stefanyuk, "Convergence rate of a class of probability density estimates," *Automation and Remote Control*, vol. 40, pp. 1706–1711, 1979.

C. J. Stone, "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, vol. 12, pp. 1285–1297, 1984.

W. Stuetzle and Y. Mittal, "Some comments on the asymptotic behavior of robust smoothers," in: *Proceedings of the Heidelberg Workshop*, ed. T. Gasser and M. Rosenblatt, pp. 191–195, Springer Lecture Notes in Mathematics 757, Springer-Verlag, Heidelberg, 1979.

W. Stute, "Modified cross-validation in density estimation," *Journal of Statistical Planning and Inference*, vol. 30, pp. 293–305, 1992.

H. Y. Su-Wong, B. Prasad, and R. S. Singh, "A comparison between two kernel estimators of a probability density function and its derivatives," *Scandinavian Actuarial Journal*, vol. 0, pp. 216–222, 1982.

J. W. H. Swanepoel, "On the construction of nonparametric density function estimators using the bootstrap," *Communications in Statistics: Theory and Methods*, vol. 15, pp. 1399–1415, 1986.

G. R. Terrell, "The maximal smoothing principle in density estimation," *Journal of the American Statistical Association*, vol. 85, pp. 470–477, 1990.

G. R. Terrell and D. W. Scott, "Oversmoothed nonparametric density estimates," *Journal of the American Statistical Association*, vol. 80, pp. 209–214, 1985.

D. M. Titterington, "Common structure of smoothing techniques in statistics," *International Statistical Review*, vol. 53, pp. 141–170, 1985.

B. A. Turlach, "Bandwidth selection in kernel density estimation: a review," Technical Report, Université Catholique de Louvain, 1993.

B. van Es, "Aspects of Nonparametric Density Estimation," Ph.D. Dissertation, University of Amsterdam, The Netherlands, 1988.

B. van Es, "Likelihood cross-validation bandwidth selection for nonparametric kernel density estimators," Technical Report 89-10, Faculty of Technical Mathematics and Informatics, TU Delft, The Netherlands, 1989.

V. N. Vapnik and A. R. Stefanyuk, "Nonparametric methods for reconstructing probability densities," *Automation and Remote Control*, vol. 39, pp. 1127–1140, 1978.

M. P. Wand and L. Devroye, "How easy is a given density to estimate?," *Computational Statistics and Data Analysis*, vol. 16, pp. 311–323, 1993.

M. P. Wand and W. R. Schucany, "Gaussian-based kernels," *Canadian Journal of Statistics*, vol. 18, pp. 197–204, 1990.

G. S. Watson and M. R. Leadbetter, "On the estimation of the probability density," *Annals of Mathematical Statistics*, vol. 34, pp. 480–491, 1963.

M. Woodroofe, "On choosing a delta sequence," *Annals of Mathematical Statistics*, vol. 41, pp. 1665–1671, 1970.

G. A. Young, "Alternative smoothed bootstraps," *Journal of the Royal Statistical Society*, vol. B52, pp. 477–484, 1990.