UNIVERSAL SMOOTHING FACTOR SELECTION IN DENSITY ESTIMATION: THEORY AND PRACTICE

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6
luc@cs.mcgill.ca

Abstract. In earlier work with Gabor Lugosi, we introduced a method to select a smoothing factor for kernel density estimation such that, for *all densities* in all dimensions, the L_1 error of the corresponding kernel estimate is not larger than $3+\epsilon$ times the error of the estimate with the optimal smoothing factor plus a constant times $\sqrt{\log n/n}$, where n is the sample size, and the constant only depends on the complexity of the kernel used in the estimate. The result is nonasymptotic, that is, the bound is valid for each n. The estimate uses ideas from the minimum distance estimation work of Yatracos. We present a practical implementation of this estimate, report on some comparative results, and highlight some key properties of the new method.

Keywords and phrases. Density estimation, kernel estimate, convergence, smoothing factor, minimum distance estimate, asymptotic optimality, simulation study.

1991 Mathematics Subject Classifications: Primary 62G05.

Running Head: universal smoothing factor

The author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291.

1. Introduction.

We are given an i.i.d. sample $X_1, \ldots, X_n$ drawn from an unknown density f on $\mathbb{R}^d$, and consider the Akaike-Parzen-Rosenblatt density estimate

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

where $K : \mathbb{R}^d \to \mathbb{R}$ is a fixed kernel with $\int K = 1$, $K_h(x) = (1/h^d)K(x/h)$, and $h > 0$ is the smoothing factor (Akaike, 1954; Parzen, 1962; Rosenblatt, 1956). In this paper, we focus on density estimation without restrictions on the densities. 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. However, the choice of K will only be of secondary interest in this paper.

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 $\{H_n, n \geq 1\}$. If we let $\mathcal{F}$ denote the class of all densities on $\mathbb{R}^d$, and let f_{nH} denote a kernel estimate with data-based bandwidth H, we look at

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

and its non-asymptotic counterpart

$$\sup_{f \in \mathcal{F}} \frac{\mathbf{E} \int |f_{nH} - f|}{\inf_h \mathbf{E} \int |f_{nh} - f|}$$

The boundedness of these suprema shows that the bandwidth selector works well for all f, without exception.

Recently, Devroye and Lugosi (1996a) introduced a data-dependent smoothing factor H for which

$$\sup_{f} \limsup_{n \to \infty} \frac{\mathbf{E} \int |f_{nH} - f|}{\inf_h \mathbf{E} \int |f_{nh} - f|} \leq 3 ,$$

whenever the kernel K is nonnegative, Lipschitz, and of compact support. The estimate of that paper requires various parameter choices which in turn are used to define the procedure for finding H. A related estimate was proposed in Devroye and Lugosi (1996b) that comes with explicit non-asymptotic performance guarantees. Both estimates will be revisited in this paper.

We define various classes of bandwidth selectors as follows:

A. Universally consistent bandwidths. Bandwidths for which for all f,

$$\lim_{n \to \infty} \mathbf{E} \int |f_{nH} - f| = 0 .$$

Bandwidths that are not universally consistent are called inconsistent.

B. SUITABLE BANDWIDTHS. Bandwidths for which for all f,

$$\limsup_{n\to\infty} \frac{\mathbf{E}\int |f_{nH} - f|}{\inf_h \mathbf{E}\int |f_{nh} - f|} \le C(f) < \infty$$

for some finite constant $C(f)$. Bandwidths outside this class are said to be unsuitable.

C. UNIVERSALLY SUITABLE BANDWIDTHS. Suitable bandwidths for which the constant $C(f)$ is universally bounded for all f. Smoothing factors not in this class are called "not universally suitable".

D. ASYMPTOTICALLY OPTIMAL BANDWIDTHS. Suitable bandwidths for which for all f, $C(f) = 1$.

One can classify bandwidths into one of the four nested classes. We are interested in this paper in universally suitable bandwidths for all dimensions d. We will therefore only briefly review bandwidths that are not in this class. More complete surveys may be found in Devroye and Györfi (1985), Marron (1987, 1988, 1989a), Izenman (1991), Jones, Marron and Sheather (1992), Park and Turlach (1992), Titterington (1985), Turlach (1993), Cao, Cuevas and González-Manteiga (1994), Berlinet and Devroye (1994), and Wand and Jones (1995).

ASYMPTOTICAL OPTIMALITY AND AN OPEN PROBLEM. We will not focus on asymptotically optimal bandwidths, simply because we do not know if this class is nonempty. This remains one of the most compelling open problems in the field. It should be noted however that there are many published bandwidths that are asymptotically optimal for given subclasses. For example, if f is restricted to a class of univariate densities in which only a translation and scale parameter is unknown, using $h = a_n\hat{\sigma}$ for a function a_n (depending upon the family), where $\hat{\sigma}$ is a data-based estimate of the scale factor, will do (see Deheuvels (1977a, 1977b) or Deheuvels and Hominal, 1980). The smoothing factor h can also be based upon a plug-in of estimates of unknown functionals into a given formula. This method has the given property if the supremum is taken over classes of univariate densities restricted by smoothness and small tails (Hall and Wand, 1988). The double kernel estimate (Devroye, 1989) satisfies the property mentioned above when the supremum is restricted as in the work of Hall and Wand. Except for trivially restricted classes of densities, none of the L_2 cross-validated estimates in the literature (see Rudemo (1982), Bowman (1984) or Stone (1984) for the early papers on this) possesses the property mentioned above. □

☞ RELATED KERNEL ESTIMATES. 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. □

2. The first bandwidth.

Split the data set into a test set of size $m \ll n$ and a remainder. The first bandwidth of Devroye and Lugosi (1996a) uses Yatracos' minimum distance projection of the empirical measure based upon these m points to the class of densities defined by the kernel estimates based on the remaining $n-m$ points to find an optimal h (Yatracos, 1985). This is attractive because it directly relates a density estimate to the standard empirical measure.

Let $m < n$ be a positive integer, let K be a nonnegative kernel with $\int K(x)dx = 1$, and let $\mathcal{F}_n$ be the class of densities

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

with $h \in [a_n, b_n]$, where the nonnegative numbers a_n, b_n will be specified later such that the optimal smoothing factor eventually falls in $[a_n, b_n]$ for all densities. Next we cover the class $\mathcal{F}_n$ by finitely many densities as follows: let $\delta_n > 0$ be a parameter to be specified later, let $h_1 = a_n$, and $h_i = h_{i-1}(1+\delta_n)$ for all $i = 2, \ldots, N$, where N is the largest integer with $a_n(1+\delta_n)^{N-1} \leq b_n$. The finite class of densities $\{f_{n-m,h_i} : i = 1, \ldots, N\}$ is denoted by $\mathcal{G}_n$. Our estimate will be drawn from this finite class!

Let μ_m be the empirical measure defined by the rest of the data points: $X_{n-m+1}, \ldots, X_n$, i.e., for any Borel set $A \subseteq \mathbb{R}^d$,

$$\mu_m(A) = \frac{1}{m} \sum_{i=n-m+1}^{n} I_A(X_i),$$

where I_A denotes the indicator function of A. As is well-known, the L_1 distance is equivalent to the twice the total variation distance. If we are to use the empirical measure, we would thus be tempted to select h so as to minimize the total variation

$$T \stackrel{\text{def}}{=} \sup_A \left| \int_A f_{n-m,h} - \mu_m(A) \right| .$$

As μ_m is an atomic measure, $T \equiv 2$ for all h. Following a clever idea of Yatracos (1985), we take the supremum instead over a specially picked rich class of subsets $\mathcal{A}$, defined as the family of sets

$$\{x : f_{n-m,h_i}(x) > f_{n-m,h_j}(x)\}, \quad i, j \leq N.$$

The estimate f_n is defined to be that $f_{n-m,h_i} \in \mathcal{G}_n$ for which

$$\sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,h_i} - \mu_m(A) \right|$$

is minimal. If the minimum is not unique, we choose among the minimizing densities according to a prespecified rule, e.g., we choose the one with smallest index. Note that in any case, our estimate optimizes over a given finite set, and is thus defined with computional efficiency in mind.

☞ CHOICE OF THE PARAMETERS. It helps at this stage to pin down choices for a_n, b_n and δ_n. The choice is determined by the choice of the kernel K. We assume the following: a kernel is said to be *elegant* if it is nonnegative, if it is Lipschitz of constant C (i.e., $|K(x) - K(y)| \le C\|x - y\|$ for all x, y), and if $K = 0$ outside $[-1,1]^d$. Then define $a_n = e^{-n}$, $b_n = e^n$, $\delta_n = c/\sqrt{n}$ for a fixed constant c. This class contains the standard Deheuvels kernel that is optimal in $\mathbb{R}^d$ and is of the form $C'(1 - \|x\|^d)_+$. For more general classes, we will show how to take the parameters in remarks below. □

☞ A COMPUTATIONAL REMARK. In most univariate cases, the sets A above are finite unions of intervals. The number of such intervals can be rigorously controlled if the kernel is polynomial on a compact set (such as with the celebrated Epanechnikov-Bartlett kernel $3/4(1-x^2)_+$). The computations are much more involved for $d > 1$, unless K is the indicator function of a unit square. The class $\mathcal{A}$ has N^2 members. A quick calculation shows that

$$N - 1 \le \frac{\log(b_n/a_n)}{\log(1+\delta_n)} \le \frac{n(2+\delta_n)}{\delta_n} = n + 2n^{3/2}/c \ .$$

A lower bound on the number of integrals over sets A (if we were to naively minimize) would be of the order of n^3. However, clever shortcuts are possible. □

☞ THE SET $\mathcal{A}$. The set $\mathcal{A}$ cannot be replaced by the set of all rectangles of $\mathbb{R}^d$. This class is simply not rich enough, and Lemma 2 below would not be valid. □

Theorem 1 (Devroye and Lugosi, 1996a). *Let K be an elegant kernel. Let a_n, b_n be such that $na_n \to 0$ and $b_n \to \infty$. Assume that $\delta_n = c/\sqrt{n}$ for some constant c and that $\log(b_n/a_n) \le c' n^a$ for some finite $c', a > 0$. If*

$$\frac{m}{n} \to 0 \quad \text{and} \quad \frac{m}{n^{4/5}\log n} \to \infty \quad \text{as } n \to \infty,$$

then the estimate f_n defined above satisfies

$$\sup_f \limsup_{n\to\infty} \frac{\mathbf{E}\int |f_n - f|}{\inf_h \mathbf{E}\int |f_{n,h} - f|} \le 3.$$

This result is valid for any multivariate density. It may be possible to improve the constant in the bound. With a bit of work, one may also be able to replace f_n in the result by f_{nH}, where H is the smoothing factor used in f_n. The difference here is that f_{nH} uses all n data points, while f_n is the kernel estimate based on H and $X_1, \ldots, X_{n-m}$.

One may argue that the selected smoothing factor is not scale-invariant. This is easily taken care of by letting M_n denote the median of the $\binom{n-m}{2}$ distances $\|X_i - X_j\|$, $1 \le i, j \le n - m$, and setting $a_n = M_n e^{-n}$ and $b_n = M_n e^n$. As M_n is almost surely bounded away from 0 and infinity, one can verify that the Theorem holds for this choice of interval.

For convenience we assumed that the kernel K is nonnegative. It is well known, however, that some kernels taking negative values provide smaller L_1 errors for smooth densities. The above theorem is easily extended to such kernels at the expense of further restrictions on the growth of m, depending on the order of the kernel.

Finally, there is quite a bit of freedom in the choice of all the parameters. For example, δ_n does not have to tend to zero at the rate $1/\sqrt{n}$. A practical implementation is described in the next section.

The universality of Theorem 1 can only be achieved thanks to combinatorial arguments. Error analysis based on Taylor series expansions of f are simply out of the question, as such expansions may not exist. For the proof of Theorem 1, we refer to Devroye and Lugosi (1996a). However, we will state four lemmas, which each contribute key elements to the proof.

LEMMA 1 (DEVROYE AND LUGOSI, 1996A). *For each density f and elegant kernel K, $C' = C2^d\sqrt{d} + d$, the estimate $f_n (\in \mathcal{G}_n)$ satisfies*

$$\int |f_n - f| \le 3 \inf_{h\in[a_n,b_n]} \int |f_{n-m,h} - f| + 5C'c\delta_n + 4 \sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A f \right| .$$

The middle term accounts for a Lipschitz effect between nearby kernel estimates. The last term resembles a total variation distance between an empirical measure and a density. If $\mathcal{A}$ were the class of all Borel sets, this term would take the value 4. Fortunately, $\mathcal{A}$ is finite, so that we may bound the expected value of the last term by

$$\frac{8\sqrt{\log c'' + (2a+1)\log n} + 1}{\sqrt{2m}},$$

provided $\log(b_n/a_n) \le c'n^a$ for positive constants c', a. Thus, the right hand side in Lemma 1 is easily bounded. Now the range outside $[a_n, b_n]$ is uninteresting because of the following result (a version in which h is a random variable is given by Broniatowski, Devroye and Deheuvels, 1989).

LEMMA 2 (DEVROYE, 1983). *Assume $K \ge 0$. If $\mathbf{E}\int |f_{nh} - f| \to 0$ for some density f and some sequence h, then $h \to 0$ and $nh^d \to \infty$. Conversely, if $h \to 0$ and $nh^d \to \infty$, then $\mathbf{E}\int |f_{nh} - f| \to 0$ for all densities f.*

Finally, the ratio result in Theorem 1 follows from the following two lemmas.

LEMMA 3 (DEVROYE AND PENROD, 1984). *Assume $K \ge 0$. Then*

$$\inf_f \liminf_{n\to\infty} n^{2/5} \inf_h \mathbf{E} \int |f_{nh} - f| \ge 0.86.$$

We note that Lemma 3 was only proved in the cited paper for $d = 1$. However, if f and f_{nh} are a density and a kernel density estimate on $\mathbb{R}^d$, and if g and g_{nh} denote the marginal densities for f and f_{nh} (with respect to any fixed component), then,

$$\mathbf{E}\int |g_{nh} - g| \le \mathbf{E}\int |f_{nh} - f|$$

(Devroye and Györfi, 1985). Interestingly, g_{nh} itself is a valid univariate kernel estimate with as kernel the marginal density of the original kernel. Therefore, a universal lower bound for $d = 1$ of the type shown in Lemma 5 then applies equally for all dimensions d.

Lemma 4 (Devroye and Lugosi, 1996a). *Let K be a bounded kernel. Define*

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

If $m > 0$ is a positive integer such that $2m \le n$, then

$$1 \le \frac{\inf_h \mathbf{E} J_{n-m,h}}{\inf_h \mathbf{E} J_{n,h}} \le 1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} \ .$$

Theorem 1 deals with $f_n = f_{n-m,H}$, not f_{nH}. In other words, the estimate does not use the full sample. In fact, in Theorem 1, we may replace f_n by f_{nH} without harm.

3. An implementation of the first bandwidth.

For one-dimensional densities, we propose a practical implementation. Without loss of generality, we consider the Epanechnikov-Bartlett kernel $K(x) = (3/4)(1-x^2)_+$ where $(.)_+$ denotes $\max(., 0)$. Our method (with bandwidth called h_{d1}) has the following parameter settings:

A. $m = \lfloor n^{0.85} \rfloor$. Note that $m/n^{4/5} \to \infty$ as required. For $n = 100$, $m = 50$.

B. $t = 1.5$: $t > 1$ is a threshold to be discussed further on. As t approaches 1, the algorithm slows down but will give increasingly accurate results.

C. $k = 10$: k is the number of intervals considered in a global search for the optimum.

The algorithm may be described in a few steps:

1. Split the data into two sets, $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{X} = (X_1, \ldots, X_{n-m})$ and $\mathcal{Y} = (X_{n-m+1}, \ldots, X_n)$.
2. Compute h and H: h is the minimum distance between consecutive points in $\mathcal{X}$, $H = X_{(\lceil 3(n-m)/4 \rceil)} - X_{(\lfloor (n-m)/4 \rfloor)}$, and $X_{(j)}$ is the j-th smallest of the values in $\mathcal{X}$. Observe that $h \le H$, and that for $n-m$ large enough, the optimal bandwidth is almost surely in the range $[h, H]$. Define $\xi^* = \sqrt{hH}$. Define $T = H/h$.
3. While $T > t$ do:

3.1. Define $\delta = T^{1/k}$. Let $\mathcal{H}$ be the set of $k+1$ candidate bandwidths $h\delta^i$ for $0 \le i \le k$. Observe that the first and last bandwidths are h and H respectively.

3.2. For all $\xi \neq \xi' \in \mathcal{H}$ compute $A(\xi, \xi') = \{f_{n-m,\xi} > f_{n-m,\xi'}\}$. The collection of these sets is called $\mathcal{A}$. Observe that $|\mathcal{A}| = k(k-1)$.

3.3. For all $\xi \in \mathcal{H}$ compute $J(\xi) = \max_{A \in \mathcal{A}} \left| \int_A f_{n-m,\xi} - \frac{1}{m} \sum_{x \in \mathcal{Y}} I_{x \in A} \right|$.

3.4. Let ξ^* be that value in $\mathcal{H}$ for which $J(\xi)$ is minimal.

3.5. Set $(h, H) = (\xi^*/\delta, \xi^*\delta)$. Set $T = \delta^2$.

4. Return ξ^*.

The choice of the parameters is motivated by computational considerations. Assume a nice unimodal density with peak value M and with spread σ, where spread is measured as the difference between third and first quartiles. Then h is of the order of $1/(Mn^2)$ and H is initially close to σ, so that at the outset, $T \approx \sigma M n^2$. After i iterations, we have $T \approx (\sigma M n^2)^{2i/k}$. We stop as soon as this drops below the threshold t. This occurs roughly when

$$i \approx \frac{k}{2} \times \frac{\log(t)}{\log(\sigma M n^2)} .$$

The influence of σM is moderated by a logarithm. For the uniform density, $\sigma M = 1/2$. If we fill in the other parameter choices and set $n = 100$, we obtain

$$i \approx \frac{5 \log 1.7}{\log 5000} ,$$

that is, $i \approx 0.31$. In fact, it is likely that only one while loop is executed, which in turn requires computation proportional to $k^2 n$ because $|\mathcal{A}| = k(k+1)/2$ and because each integral takes time bounded by n if the data are storedproperly. Discounting the dependence upon the distribution, the time grows roughly as

$$k^2 n \max\left(1, \frac{k}{\log n}\right) .$$

Taking k very small does help, but reduces the quality of the solution, as the optimization is done over a rougher grid.

4. The second bandwidth.

The second bandwidth of Devroye and Lugosi (1996b) does not use a finite interval $[a_n, b_n]$ for the optimization, and thus eliminates the need to pick these parameters. The second difference is that the class $\mathcal{A}$ is infinite. This eliminates the need for a third parameter, δ_n, but renders the optimization a bit more problematic. Introduce the class $\mathcal{R}_k$ of kernels of the form

$$K'(x) = \sum_{i=1}^{k} \alpha_i I_{A_i}(x),$$

where I_A denotes the indicator function of a set A, $k < \infty$, $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, and $A_1, \ldots, A_k$ are Borel sets in $\mathbb{R}^d$ with the following property: the intersection of an infinite ray $\{x : x = tx_0, t \geq 0\}$, anchored at the origin, with any A_i is an interval. Examples of such $A_i's$ include all convex sets and all star-shaped sets (a set A is star-shaped if $x \in A$ implies $\lambda x \in A$ for all $\lambda \in [0,1]$). The A_i's need not be disjoint. However, if the A_i's are disjoint rectangles, the sum looks a bit like a Riemann approximation of a function. Thus, kernels of the type given here are called *Riemann kernels* of parameter k. Denote the class of all such functions by $\mathcal{R}_k$. The most important examples include the uniform densities on ellipsoids, balls, and hypercubes.

We first select k and $K' \in \mathcal{R}_k$ such that

$$\int |K - K'| \leq \frac{1}{n} .$$

Note that this is always possible if K is Riemann integrable. The size k as a function of n will be discussed further on. A kernel estimate with kernel K' is piecewise constant and thus easy to work with in simulations.

The second and last choice is that of a parameter $m \leq n/2$ that will be used to split the data set into a small test set of size m and a large main sample of size $n - m$. Define the kernel estimates

$$f'_{n-m,h}(x) = \frac{1}{n-m} \sum_{i=1}^{n-m} K'_h(x - X_i)$$

for all $h > 0$. Let μ_m be the empirical measure defined by the rest of the data points: $X_{n-m+1}, \ldots, X_n$, that is, for any Borel set $A \subseteq \mathbb{R}^d$,

$$\mu_m(A) = \frac{1}{m} \sum_{i=n-m+1}^{n} I_A(X_i).$$

Let H be that smoothing factor for which the quantity

$$\sup_{A \in \mathcal{A}} \left| \int_A f'_{n-m,h} - \mu_m(A) \right|$$

is minimal over $h \in (0, \infty)$, where $\mathcal{A}$ is a special collection of sets to be defined below. If the minimum is not unique, we choose among the minimizing densities according to a prespecified rule, for example, we choose the smallest one. Observe that since $f'_{n-m,h}$ is piecewise constant and $K' \in \mathcal{R}_k$, a minimum always exists.

As $\mu_m(A)$ is close to $\int_A f$ for all A, one may expect that $\int_A f'_{n-m,h}$ is close to $\int_A f$ as well if $\mathcal{A}$ is not too large. If $\mathcal{A}$ is the class of all Borel sets, the criterion to be minimized is equal to 2 for all h and becomes useless. If $\mathcal{A}$ is too small, the closeness of $\int_A f'_{n-m,h}$ to $\int_A f$ does not imply the closeness of $f'_{n-m,h}$ to f. Thus, a compromise must be struck. Based on ideas from Yatracos (1985), for each $u, v > 0$, we define the set $A_{u,v}$ by

$$A_{u,v} = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{n-m} K'_u(x - X_i) \geq \sum_{i=1}^{n-m} K'_v(x - X_i) \right\} = \left\{ x : f'_{n-m,u}(x) \geq f'_{n-m,v}(x) \right\} .$$

We call the class of sets

$$\mathcal{A} = \{ A_{u,v} : u > 0, v > 0 \}$$

a *Yatracos class*. This class becomes very rich, yet remains reasonably simple (even though it has an infinite number of members).

Finally, our estimate is

$$f_n \stackrel{\text{def}}{=} f_{n-m,H} .$$

Note that we have replaced K' by K again. The kernel K' is no longer needed. We may also use $f_n = f_{n,H}$ and refer to Devroye and Lugosi (1996) for analysis of this situation.

Let K be Riemann integrable kernel, and let n be a positive integer. The *kernel complexity of precision* $1/n$ of K is defined by

$$\kappa_n = \min \left\{ k : \text{ there exists a } \; K' \in \mathcal{R}_k \text{ such that } \int |K - K'| \le \frac{1}{n} \right\},$$

that is, κ_n is the smallest integer k such that there exists a Riemann kernel with parameter k whose L_1 distance from K is at most $1/n$. Clearly, if K is Riemann integrable, then $\kappa_n < \infty$ for all n.

THEOREM 2 (DEVROYE AND LUGOSI, 1996B). *Let K be a bounded (but not necessarily nonnegative) kernel, and $m \le n/2$. If κ_n is the kernel complexity of K of precision $1/n$, then there exists a Riemann kernel K' of parameter κ_n such that if K' is used in the estimate described in the previous section, then for all densities f,*

$$\mathbf{E} \int |f_n - f| \le 3 \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} \right) \inf_h \mathbf{E} \int |f_{nh} - f| + 4\sqrt{\frac{\log\left(4e^8(m^2+1)(1+2\kappa_n m^2(n-m))^2\right)}{2m}} + \frac{4}{n}.$$

COROLLARY 1. If we take $m = \lfloor n/2 \rfloor$, then

$$\mathbf{E} \int |f_n - f| \le 43 \inf_h \mathbf{E} \int |f_{nh} - f| + c\sqrt{\frac{\log(n\kappa_n)}{n}},$$

where c is a universal constant, independent of f and K.

COROLLARY 2. If $m = o(n)$, $m/(n^{4/5}\log n) \to \infty$, and $\kappa_n = O(n^\alpha)$ for some finite α, then

$$\mathbf{E} \int |f_n - f| \le (3 + o(1)) \inf_h \mathbf{E} \int |f_{nh} - f| + o(n^{-2/5}) \; .$$

As $\liminf_{n\to\infty} n^{2/5} \inf_h \mathbf{E} \int |f_{nh} - f| > 0$ for any f, $K \ge 0$ and d (see Devroye and Györfi, 1985), we have

$$\sup_f \limsup_{n\to\infty} \frac{\mathbf{E}\int |f_n - f|}{\inf_h \mathbf{E}\int |f_{nh} - f|} \le 3 \; .$$

This universal asymptotic bound is shared with the first bandwidth. However, Theorem 2 differs because it is entirely non-asymptotic. Every factor on the right-hand-side of the inequality is explicit and easy to control. Note that traditional Taylor series expansions to compute or bound errors in function approximations are no longer useful.

The arguments are entirely combinatorial. Below, we briefly indicate the key building blocks in the proof.

LEMMA 5 (DEVROYE AND LUGOSI, 1996B). *For each n, m, and for all f,*

$$\int |f_n - f| \le 3 \inf_h \int |f_{n-m,h} - f| + 4 \sup_{A\in\mathcal{A}} \left| \int_A f - \mu_m(A) \right| + 4 \int |K - K'|.$$

The first term on the right-hand side of the inequality of Lemma 1 may be bounded by the following result:

LEMMA 6 (DEVROYE AND LUGOSI, 1996A). *Let K be a bounded kernel. If $m > 0$ is a positive integer such that $2m \le n$, then*

$$1 \le \frac{\inf_h \mathbf{E} \int |f_{n-m,h} - f|}{\inf_h \mathbf{E} \int |f_{n,h} - f|} \le 1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} \ .$$

Therefore,

$$\inf_h \mathbf{E} \int |f_{n-m,h} - f| \le \inf_h \mathbf{E} \int |f_{n,h} - f| \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} \right) \ .$$

A suitable upper bound for $\sup_{A\in\mathcal{A}} \left| \int_A f - \mu_m(A) \right|$ may be obtained vi the inequality of Vapnik and Chervonenkis (1971) (see also Devroye, 1982) for uniform deviations of the empirical measure μ_m over the Yatracos class of sets $\mathcal{A}$. For $\epsilon > 0$, we have in fact

$$\mathbf{P} \left\{ \sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A f \right| > \epsilon \middle| X_1, \ldots, X_{n-m} \right\} \le 4e^8(m^2+1)(1+2km^2(n-m))^2 e^{-2m\epsilon^2},$$

when K' is a Riemann kernel with parameter k.

5. Kernel complexity.

In this section we obtain bounds for κ_n, the kernel complexity of precision $1/n$ appearing in the theorem, for several examples of kernels. Note that the theorem has the form

$$\mathbf{E} \int |f_n - f| \le 3 \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} \right) \inf_h \mathbf{E} \int |f_{nh} - f| + O\left(\sqrt{\frac{\log n}{m}} \right)$$

whenever $\kappa_n = O(n^\alpha)$ for some $\alpha < \infty$. Such kernels are *polynomially Riemann approximable.* All kernels that we have found in papers are in this class.

☞ Uniform kernels. If $K(x) = I_A(x)$ for a star-shaped set A, then obviously $\kappa_n = 1$ for all $n > 1$.

☞ Isosceles triangular density. If $K(x) = (1 - |x|)_+$, then elementary calculation shows that for all n, $\kappa_n \le n + 1$.

☞ Symmetric unimodal kernels. As a first main example, consider symmetric unimodal densities (i.e., $K \ge 0$ and $\int K = 1$) on the real line. Let β be the last positive value for which $\int_\beta^\infty K \le 1/(4n)$. Partition $[0, \beta]$ and $[-\beta, 0]$ into $N = \lceil 4nK(0)\beta \rceil$ equal intervals. On each interval, let K' be constant with value equal to the average of K over that interval. Let $\gamma = \beta + \int_\beta^\infty K/K(\beta)$, and set $K'(x) = K(\beta)$ on $[\beta, \beta + \gamma]$ and $[-\beta - \gamma, -\beta]$. Note that $\int K' = 1$, $\int |K - K'| \le 1/n$, and that K' is Riemann with parameter $k = 2N + 2 \le 8nK(0)\beta + 10$. Thus, $\kappa_n \le 8nK(0)\beta + 10$.

Example 1: bounded compact support densities. If $K(x) \le aI_{[-b,b]}(x)$ and K is symmetric, nonnegative, and unimodal (such as the Epanechnikov-Bartlett kernel), then $\kappa_n \le 8nab + 10$.

Example 2: the normal density. When $K(x) = \left(\sqrt{2\pi}\right)^{-1} e^{-x^2/2}$, we have $K(0) = \left(\sqrt{2\pi}\right)^{-1}$. Since for $\beta \ge 1$,

$$\int_\beta^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \le \frac{1}{\sqrt{2\pi}} \frac{1}{\beta} e^{-\beta^2/2} \le \frac{1}{\sqrt{2\pi}} e^{-\beta^2/2},$$

we may take $\beta = \sqrt{2 \log(4n/\sqrt{2}\pi)}$. Thus, for all $n > 1$,

$$\kappa_n \le \frac{8n\sqrt{\log n}}{\sqrt{\pi}} + 10.$$

Example 3: the Cauchy density. Take $K(x) = 1/(\pi(1 + x^2))$. Note that $K(0) = 1/\pi$, and that $\beta = \pi/(4n)$ will do. Therefore,

$$\kappa_n \le \frac{32n^2}{\pi^2} + 10.$$

Example 4: densities with polynomial tails. Note that if K is a symmetric unimodal density, and $|K(x)| \le c/(1 + |x|^{\gamma+1})$ for some $c < \infty$, $\gamma > 0$, then $\kappa_n = O\left(n^{1+1/\gamma}\right)$. In fact, for most cases of interest, $\kappa_n = O\left(n^\alpha\right)$ for some finite constant $\alpha > 0$. This remains so even for d dimensions.

☞ Kernels of bounded variation. If K is symmetric and a difference of two monotone functions, that is, $K = K_1 - K_2$, $K_1 \downarrow 0$, $K_2 \downarrow 0$ on $[0, \infty)$, then each K_1, K_2 may be approximated as above. Thus, in particular, if K is of *bounded variation*, and $|K(x)| \le c/(1 + |x|^{\gamma+1})$ for some $c < \infty$, $\gamma > 0$, then we may approximate with $\kappa_n = O\left(n^{1+1/\gamma}\right)$. Nearly every one-dimensional kernel falls in this class.

☞ Product kernels. If $K = K_1 \times \cdots \times K_d$ is a product of d univariate kernels, and if we approximate K_i with K_i' with parameter $\kappa_{nd}^{(i)}$ for all i (where $\kappa_{nd}^{(i)}$ is the kernel complexity of K_i of precision nd), and form $K' = K_1' \times \cdots \times K_d'$, then K' is a weighted sum of indicators of product sets, and it is Riemann with parameter not exceeding $\prod_{i=1}^d \kappa_{nd}^{(i)}$. Furthermore,

$$\begin{aligned}\int |K - K'| &\le \int |K_1 \times \cdots \times K_{d-1} \times K_d - K_1 \times \cdots \times K_{d-1} \times K_d'| \\ &\quad + \ldots \\ &\quad + \int |K_1 \times K_2' \cdots \times K_d' - K_1' \times K_2' \cdots \times K_d'| \\ &\le d\left(\frac{1}{nd}\right) \\ &= \frac{1}{n}.\end{aligned}$$

Thus, it suffices to replace κ_n throughout by $\prod_{i=1}^d \kappa_{nd}^{(i)}$, and only worry about univariate kernel approximations.

☞ Kernels that are functions of $\|x\|$. Assume that $K(x) = M(\|x\|)$, where M is a bounded nonnegative monotone decreasing function on $[0, \infty)$. Then we may approximate M by a stepwise constant function M', and use the Riemann kernel $K'(x) = M'(\|x\|)$ in the estimate as an approximation of K. Clearly,

$$\int |K(x) - K'(x)| dx = \int_0^\infty c_d u^{d-1} |M(u) - M'(u)| du,$$

where c_d is d times the volume of the unit ball in $\mathbb{R}^d$. We may define M' as follows. Let β be the largest positive number for which $\int_\beta^\infty c_d u^{d-1} M(u) du \le 1/(2n)$. Partition $[0, \beta]$ into $N = \lceil 2nc_d M(0)\beta^d \rceil$ equal intervals. On each interval, let M' equal to the average of M over that interval. Let $\gamma = \beta + \int_\beta^\infty c_d u^{d-1} M(u) du / M(\beta)$, and set $M'(u) = M(\beta)$ on $u \in [\beta, \beta + \gamma]$, and let $M'(u) = 0$ for $u > \gamma$. Clearly $\int K' = 1$, and that K' is Riemann with parameter $k = N + 1 \le 2nc_d K(0)\beta^d + 2$. Moreover,

$$\int |K(x) - K'(x)| dx = \int_0^\beta c_d u^{d-1} |M(u) - M'(u)| du + \int_\beta^\infty c_d u^{d-1} |M(u) - M'(u)| du$$

$$
\begin{aligned}
&\le \frac{1}{2n} + c_d\beta^{d-1}\int_0^\beta |M(u)-M'(u)|du \\
&\le \frac{1}{2n} + c_d\beta^{d-1}\frac{M(0)\beta}{N} \\
&\le \frac{1}{n}.
\end{aligned}
$$

Thus,

$$\kappa_n \le 2nc_dM(0)\beta^d + 2 \ .$$

☞ The multivariate standard normal kernel. We may apply the bound of the previous paragraph to the multivariate normal density. First note that it suffices to take $\beta = 2\sqrt{2\log n}$. From this, we deduce that the kernel complexity is

$$\kappa_n = O(n\log^{d/2} n) \ .$$

6. Improvements and new methods.

The estimate probably improves if we average h over several or all subsets of subsamples of size m drawn from $X_1,\ldots,X_n$. So, rotating the held out sample may stabilize the bandwidth.

The optimization is time-consuming and requires further investigation. This compels us to see if perhaps simple iterative methods exist that give acceptable results in reasonable time. Define $A = [f_{n-m,h} > f_{n-m,h'}]$, $B = [f_{n-m,h} < f_{n-m,h'}]$. Note that

$$\int_A (f_{n-m,h} - f_{n-m,h'}) = \frac{1}{2}\int |f_{n-m,h} - f_{n-m,h'}| \ .$$

Define quality indices

$$
\begin{aligned}
Q(h) &= \max\left(\left|\int_A f_{n-m,h} - \mu_m(A)\right| \ , \left|\int_B f_{n-m,h} - \mu_m(B)\right|\right) \ , \\
Q(h') &= \max\left(\left|\int_A f_{n-m,h'} - \mu_m(A)\right| \ , \left|\int_B f_{n-m,h'} - \mu_m(B)\right|\right) \ .
\end{aligned}
$$

The recursive algorithm.

Compute $[a,b]$, a range for h picked as for the first bandwidth. Set $h \leftarrow \sqrt{ab}$.

Repeat forever:

1.

$$h' = \begin{cases} he^{\sigma N} & \text{if } h \in [a,b] \\ a^{1-U}b^U & \text{if } h \notin [a,b] \end{cases}$$

where $\sigma > 0$, N is normal $(0,1)$, and U is uniform $[0,1]$. A fixed value $\sigma = 1.6$ was used in the experiments that follow.

2. Set
$$h = \begin{cases} h & \text{if } Q(h) \le Q(h') \\ h' & \text{otherwise .} \end{cases}$$

What is the asymptotic behavior of h and of $\int |f_{nh} - f|$ as we continue doing this? Preliminary experiments reported below suggest that the method is relatively robust.

THE ITERATED BANDWIDTH. Let $h_{\mathrm{dl,it}}$ be the bandwidth obtained for $m \approx n^{0.85}$ ($m = 50$ when $n = 100$) after $4\sqrt{n}$ iterations (40 when $n = 100$).

THE ROTATED BANDWIDTH. If the above method is applied with the held out sample rotated, then the geometric average of the bandwidths will be denoted by $h_{\mathrm{dl,rot}}$.

7. The double kernel estimate

In this section, we familiarize the user with the double kernel estimate (Devroye, 1989b) and return temporarily to dimension $d = 1$. 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
$$K(x) = \frac{3}{4}(1 - x^2)_+ \ .$$
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,
$$\mathbf{E}\int (g_{nh} - f)^2 \le \mathbf{E}\int (f_{nh} - f)^2$$
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.

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.

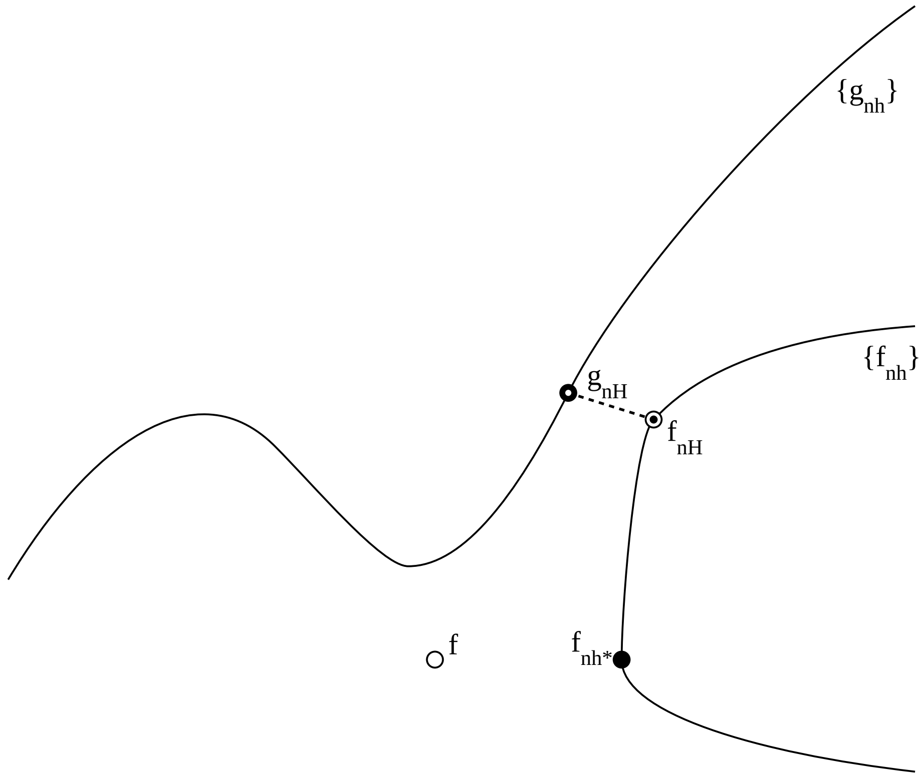

FIGURE 1. We show two families of density estimates in the set of all densities. The double kernel bandwidth minimizes the L_1 distance between f_{nh} and g_{nh}.

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. Examples include Müller's kernel $(105/64)(1-5x^2+7x^4-3x^6)$, $|x| \leq 1$ (Müller, 1984), the Gasser-Müller-Mammitzsch kernel $(75/16)(1-x^2)-(105/32)(1-x^4)$, $|x| \leq 1$ (Gasser, Müller and Mammitzsch, 1985; see also Scott, 1992 and Devroye, 1989b), and the simple kernel $(9-15x^2)/8$, $|x| \leq 1$ (Berlinet and Devroye, 1994). There are simple ways of constructing fourth-order kernels from standard second-order kernels K: Stuetzle and Mittal (1979) suggest twicing: $2K - K * K$. Schucany and Sommers (1977) propose the kernel $(3K + xK')/2$ (see also Jones, 1990). If ϕ_a represents the normal density with variance a, then one could also use $2\phi_1 - \phi_2$ (Su-Wong, Prasad and Singh, 1982) or $(1/2)(3-x^2)\phi_1$ (Wand and Schucany, 1990; Deheuvels, 1977a,b).

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

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

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

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

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.

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. We also note that it is unknown if the double kernel estimate is suitable, let alone universally suitable. As shown above, it is suitable uniformly over large subclasses of densities.

☛ 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}. □

8. Survey of other univariate bandwidths.

8.1. L1 plug-in methods. 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

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

(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 by Hall and Wand (1988) based on good pointwise estimates of f'' and $\sqrt{f}$ was shown to yield asymptotic optimality for a subclass of $\mathcal{F}$. 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

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

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

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

This h is often, but not always, close to the true optimal h. A bandwidth obtained by estimating $\int\sqrt{f}$ and $\int|f''|$ and plugging the estimates back into (1) is called an L1 plug-in method.

8.2. L1 reference density method. If (1) (or a similar asymptotic formula for h) is applied based upon a parametric assumption of f, we obtain the reference density method. For example, if we (usually, incorrectly) assume that f is the normal (μ, σ^2) density, the bandwidth in (1) can be written as

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

Hall and Wand (1988, Table 4.1) report that the optimal h for this family of densities varies asymptotically as

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

The parameter σ is easily estimated by ordinary statistical methods. For the normal reference density, Deheuvels (1977a,b) suggests using

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3}$$

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

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

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) in an L2 setting.

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 leads to the following bandwidths in our simulations.

- $h_{\text{ref,L1}} = 2.279\,\widehat{\sigma}\,n^{-1/5}$, where $\widehat{\sigma}$ is defined in (4).
- $h_{\text{DH,L1}} = 2.279\,\widehat{\sigma}\,n^{-1/5}$, where $\widehat{\sigma}$ is defined in (3). DH is a mnemonic for Deheuvels and Hominal.
- $h_{\text{ref,l1}} = 1.6644\,\widehat{\sigma}\,n^{-1/5}$, where $\widehat{\sigma}$ is defined by (4).

8.3. L2 plug-in methods. 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:

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

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

$$h = \left(\frac{15}{n\int f''^2}\right)^{1/5} \tag{6}$$

(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 above abound: see Park and Marron (1990), Park (1989), Hall and Marron (1990), Sheather and Jones (1991), Hall and Marron (1987a,b), and Hall, Sheather, Jones and Marron (1991). 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 suggest estimating $\int f''^2$ by

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

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

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

where

$$h_{\mathrm{ms,L2}} = 3(3/7n)^{1/5}\widehat{\sigma} = 2.532362\ldots\widehat{\sigma}n^{-1/5}$$

is a safe "maximal" bandwidth (Terrell, 1990; Scott and Terrell, 1985), and $\widehat{\sigma}$ is as in (3). It is easy to show that for all densities, $h_{\mathrm{pi,L2}} \to 0$ and $n\,h_{\mathrm{pi,L2}} \to \infty$ in probability whenever L'''' is uniformly bounded. This implies that $h_{\mathrm{pi,L2}}$ is universally consistent.

Other L2 plug-in methods were developed by Chiu (1991), Scott and Factor (1981), Scott, Tapia and Thompson (1977), Park and Marron (1990), Park (1989) and Sheather and Jones (1991).

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.

8.4. L2 reference density methods. The reference density method with a normal reference density used in (5) leads to the formula

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

This suggests the bandwidths

- $h_{\mathrm{DH,L2}} = 2.345\, \widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is (3).
- $h_{\mathrm{ref,L2}} = 2.345\, \widehat{\sigma}\, n^{-1/5}$, where $\widehat{\sigma}$ is (4).

8.5. 5. L2 cross-validation. 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 volatile,, leading often to undersmoothing (Hall and Marron, 1987a,b; Scott and Terrell, 1987; Hall, Marron and Park, 1992; 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.

Modifications proposed later include biased cross-validation (Scott and Terrell (1987)), Stute's modified cross-validation (Stute, 1992), smoothed cross-validation (Jones, Marron and Park (1990)), presmoothed cross-validation (Hall, Marron and Park (1992)), and the method of Jones and Kappenman (1992). All methods essentially minimize

$$\frac{\int K^2}{nh} + \frac{1}{n^2} \sum_{i \neq j} M_h(X_i - X_j) \tag{7}$$

for some function M. We conjecture that for any pair (M, K), the minimizer of (7) is not universally consistent.

8.6. The bootstrap method. In the bootstrap, 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: see Taylor (1989), Mihoubi (1992), Faraway and Jhun (1990), Cao (1990), Hall (1990), Cao et al. (1994) and Marron (1992). None of the bootstrap methods deals directly with the L_1 error and for this reason, the method is not included in this study.

8.7. Other methods. 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). Two bandwidth selectors based upon statistics related to spacings are studied and compared by Berlinet and Devroye (1994).

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.

Other interesting estimates based upon a Kolmogorov-Smirnov type criterion or penalized likelihoods were proposed by Eggermont and LaRiccia (1995, 1996).

8.8. Double kernel-double h method (Berlinet and Devroye, 1994). 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. The theoretical properties of this method remain largely unknown.

9. Practical implementation of the double kernel method.

It is computationally interesting to work with kernels that are piecewise polynomials of low order. For this reason, we suggest the double kernel pair

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

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

In our 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 $h_{\text{dk},1}, h_{\text{dk},2}, h_{\text{dk},3}$ and $h_{\text{dk},4}$ 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 $l \le 1$.

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. The optimization is greatly facilitated by two algorithmic tricks:

☞ First trick: quick detection of 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 $h_{\mathtt{ref,l1}}/2^i$ with the property that $\chi(a) > J_{n\,h_{\mathtt{ref,l1}}}$, and b as the smallest number of the form $h_{\mathtt{ref,l1}} \times 2^i$ with the property that $\xi(a) > J_{n\,h_{\mathtt{ref,l1}}}$. This procedure works with any starting point, not just $h_{\mathtt{ref,l1}}$. 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 $\inf J'_{nh}$.

☞ SECOND TRICK: AVOID LOCAL MINIMA BY PROFITING FROM LIPSCHITZ CONTINUITY . The minimization is also simplified because J_{nh} and J'_{nh} satisfy the following simple Lipschitz condition:

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

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

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

where C is some finite constant depending upon $K-L$. Therefore, the minimization can be carried out on a grid designed for a certain accuracy.

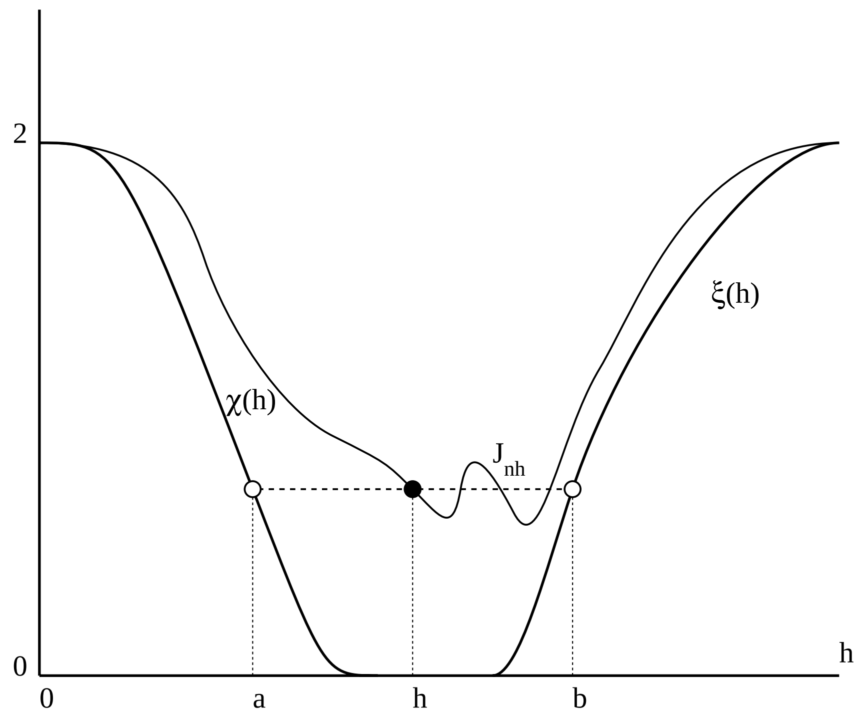

FIGURE 2. A hypothetical L_1 error J_{nh} is shown as a function of h. Also shown are the lower bounds $\chi(h)$ and $\xi(h)$. The figure illustrates how one computation of J_{nh} directly leads to an interval $[a,b]$ that contains the overall minimum, and yet stays bounded away from 0 and ∞.

10. A modified double kernel method.

Berlinet and Devroye (1994) 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 the section on the L_1 plug-in method:

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

Let K, L, f_{nh} and g_{nh} be as in the previous section. Here g_{nh} uses $L_{1.5}$, L with a stretch of 1.5. Also h_{dk} represents the double kernel bandwidth based upon $(K, L_{1.5})$. In algorithmic format, the bandwidths $h_{\mathrm{pi,l1}}$ and $h_{\mathrm{pi,L1}}$ are defined as follows. The former will be referred to as the L_1 plug-in method. The latter will be called the improved L_1 plug-in method.

$$\begin{aligned}
&h' \leftarrow h_{\mathrm{ref,L1}}(h_{\mathrm{dk}}) \\
&A = \int \sqrt{f_{nh'}} \\
&R = \frac{A\sqrt{\int (K-L)^2}}{\sqrt{nh'} \int |f_{nh'} - g_{nh'}|} \\
&h'' = h' \max\left(1, (10R)^{2/5}\right) \\
&B = \frac{2\int |f_{nh''} - g_{nh''}|}{h''^2 \int x^2 K} \\
&\widehat{\sigma} \text{ is defined by (3)} \\
&h_{\mathrm{ms,L1}} = 2.71042\ldots \widehat{\sigma} n^{-1/5} \\
&h_{\mathrm{pi,l1}}(h_{\mathrm{pi,L1}}) = \min\left(\left(\frac{\sqrt{15/(2\pi)}A}{B} \right)^{2/5} n^{-1/5} \ , \ h_{\mathrm{ms,L1}} \right)
\end{aligned}$$

REMARKS. A is an estimate of $\int \sqrt{f}$, and B is an estimate of $\int |f''|$. $h_{\mathrm{ms,L1}}$ is a safe "maximal" bandwidth derived on page 113 of Devroye and Györfi (1985). The coefficient $2.71042\ldots$ is computed for the Epanechnikov kernel and is equal to $(98415\pi^4/65536)^{1/5}$. Note also that both bandwidths are universally consistent (Berlinet and Devroye, 1994). Finally, both bandwidths are rather robust in practice.

11. Comparisons and simulations.

The extensive comparative simulations carried out by Cao, Cuevas and González-Manteiga (1994) reveal 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. After their simulation, Berlinet and Devroye (1994) proposed the modified double kernel estimate, a hybrid between L1 plug-in and double kernel methods, and found this modification to be excellent against 18 methods for 28 different test densities. Another conclusion of the Spanish study is that the double kernel method never performs poorly—it is very robust.

In the determination of bandwidths, some 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, the L1 plug-in method, the spacings method, the modified double kernel method and the bandwidths of Devroye and Lugosi (1996a, 1996b) do just that.

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).

Define

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

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 quantities such as $J_{nH} - Q_n$, $(J_{nH} - Q_n)/Q_n$ or J_{nH}/Q_n. Especially the last two quantities are convenient as they allow a comparison 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.

The twenty-eight test densities are those from Berlinet and Devroye (1994). Part of the results given here are borrowed from that study. 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 with varying tail sizes and asymmetries, five densities with an infinite peak at the origin, many discontinuous densities and continuous densities with discontinuous first derivatives, as well as eight multimodal densities with varying modal structures.

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: $f = 0.5f_{[-1,1]} + 0.25f_{[20,20.1]} + 0.25f_{[-20.1,-20]}$, 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.

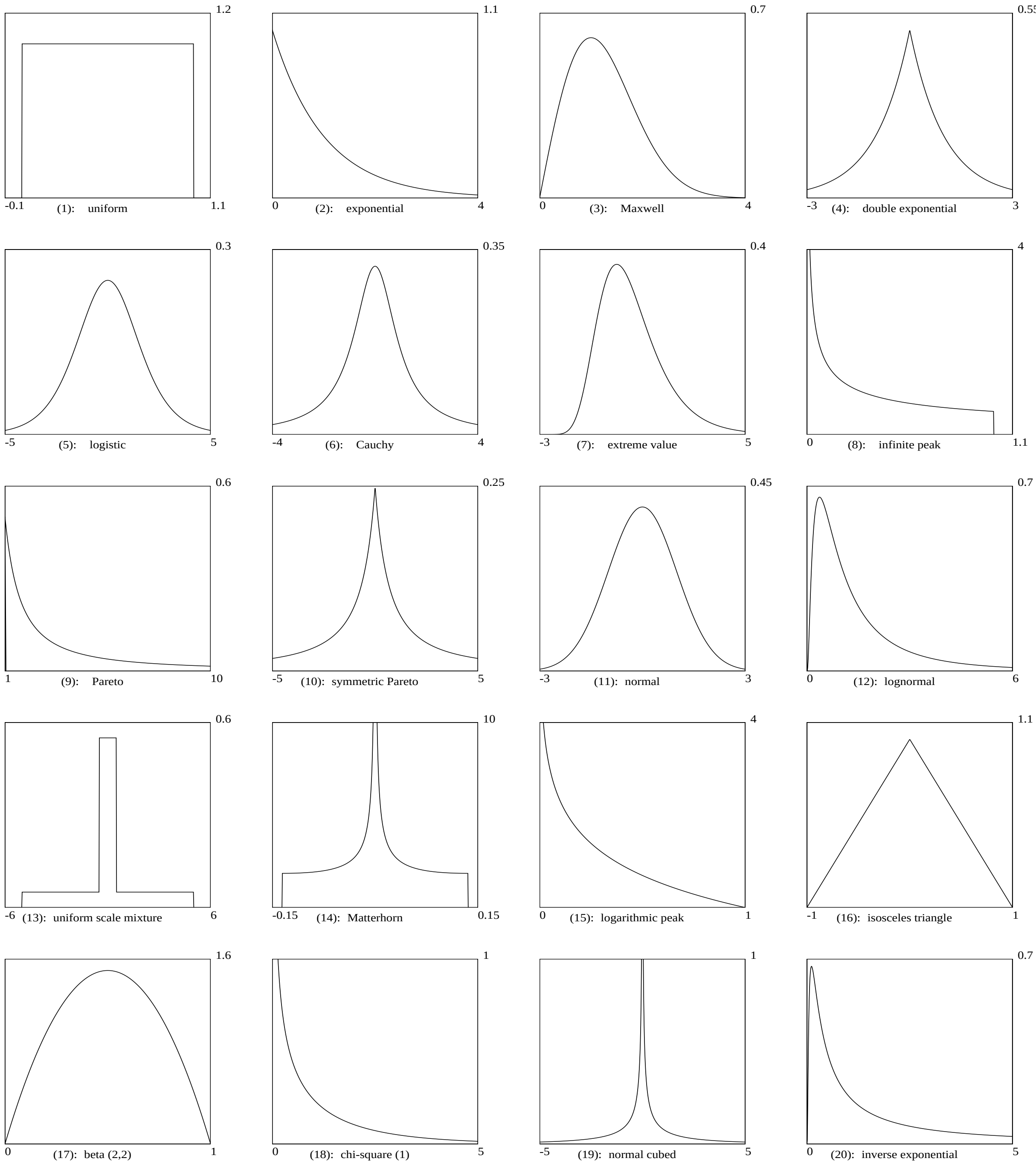

FIGURE 3. The unimodal densities in our collection.

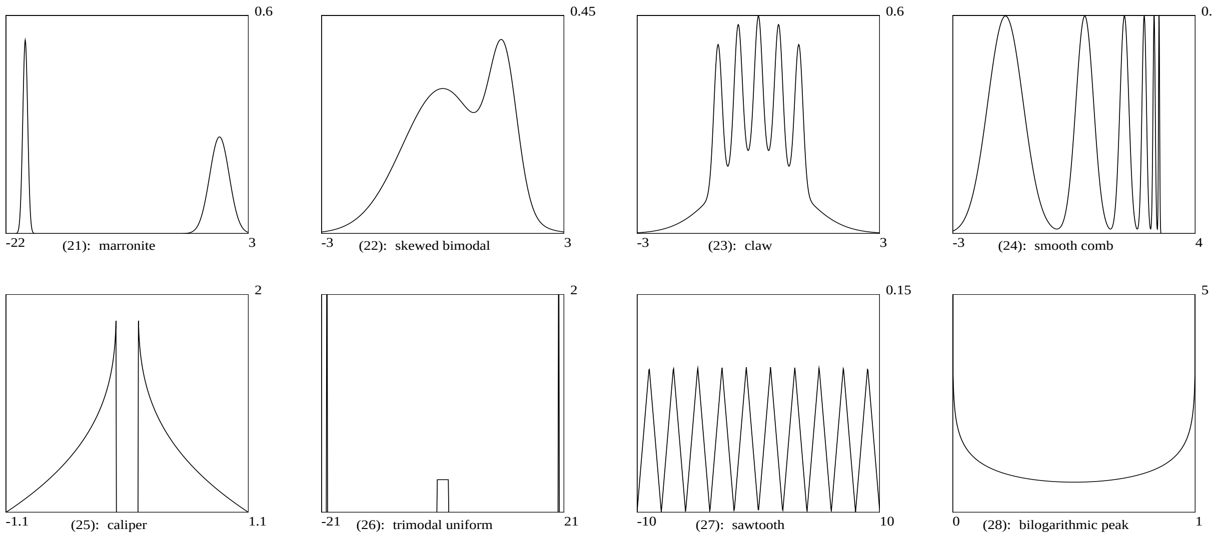

FIGURE 4. The multimodal densities in our collection.

For each of the 28 densities, Berlinet and Devroye (1994) generated 20 samples of size 100 each, and tried 17 different bandwidth selectors. By reporting results for the basket of densities, it is rather difficult to fine-tune bandwidths for all of them at once. For some densities $n = 100$ is a reasonable sample size, while for others it obviously is too small. Thus, it really is not a drawback to perform simulations for one value of n provided the basket is big enough. The program was written in PASCAL and then translated into C by the filter p2c. 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_{op}, we take for H the optimal bandwidth:

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

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 maximal value of P_n observed over the runs.

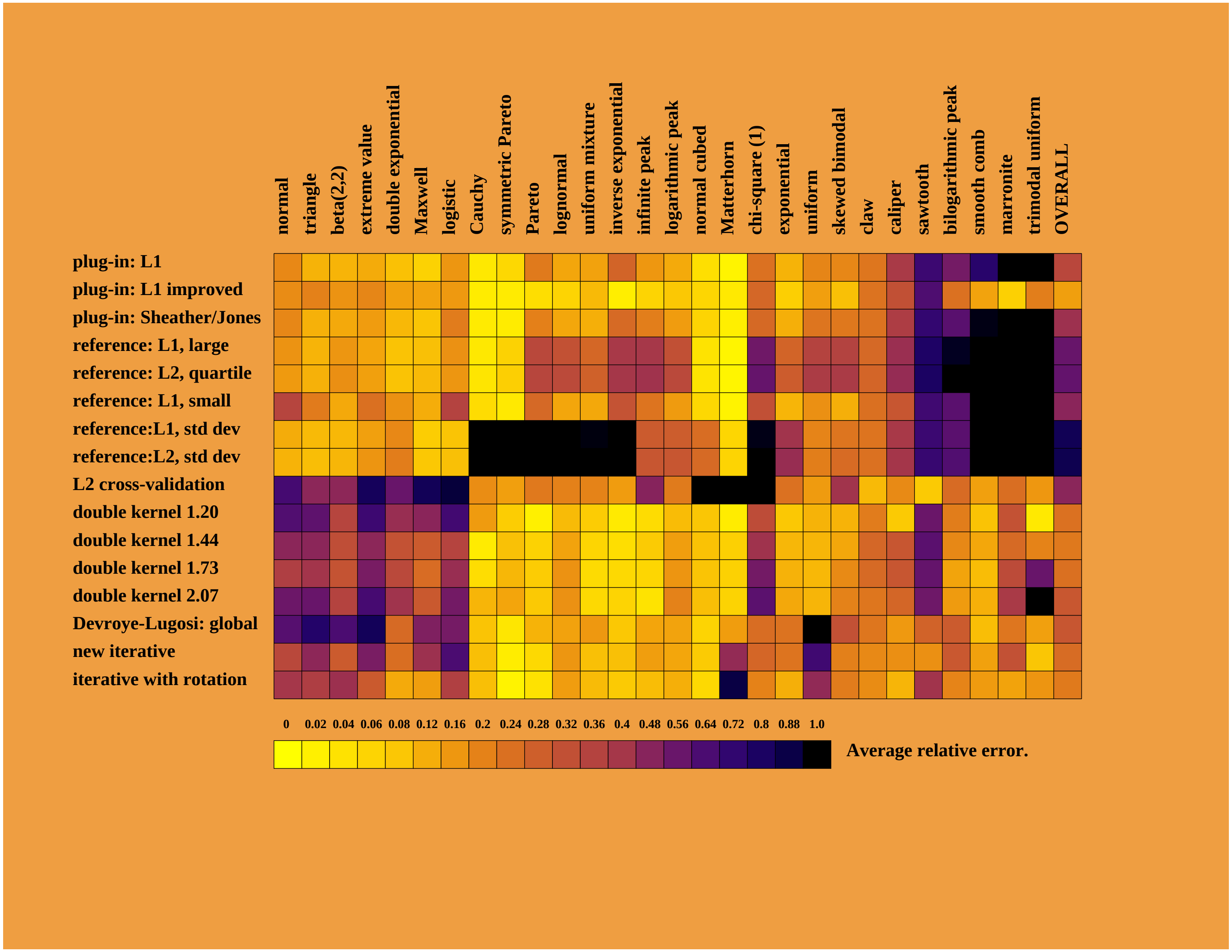

FIGURE 5. The average relative error (P_n) is shown for all densities.

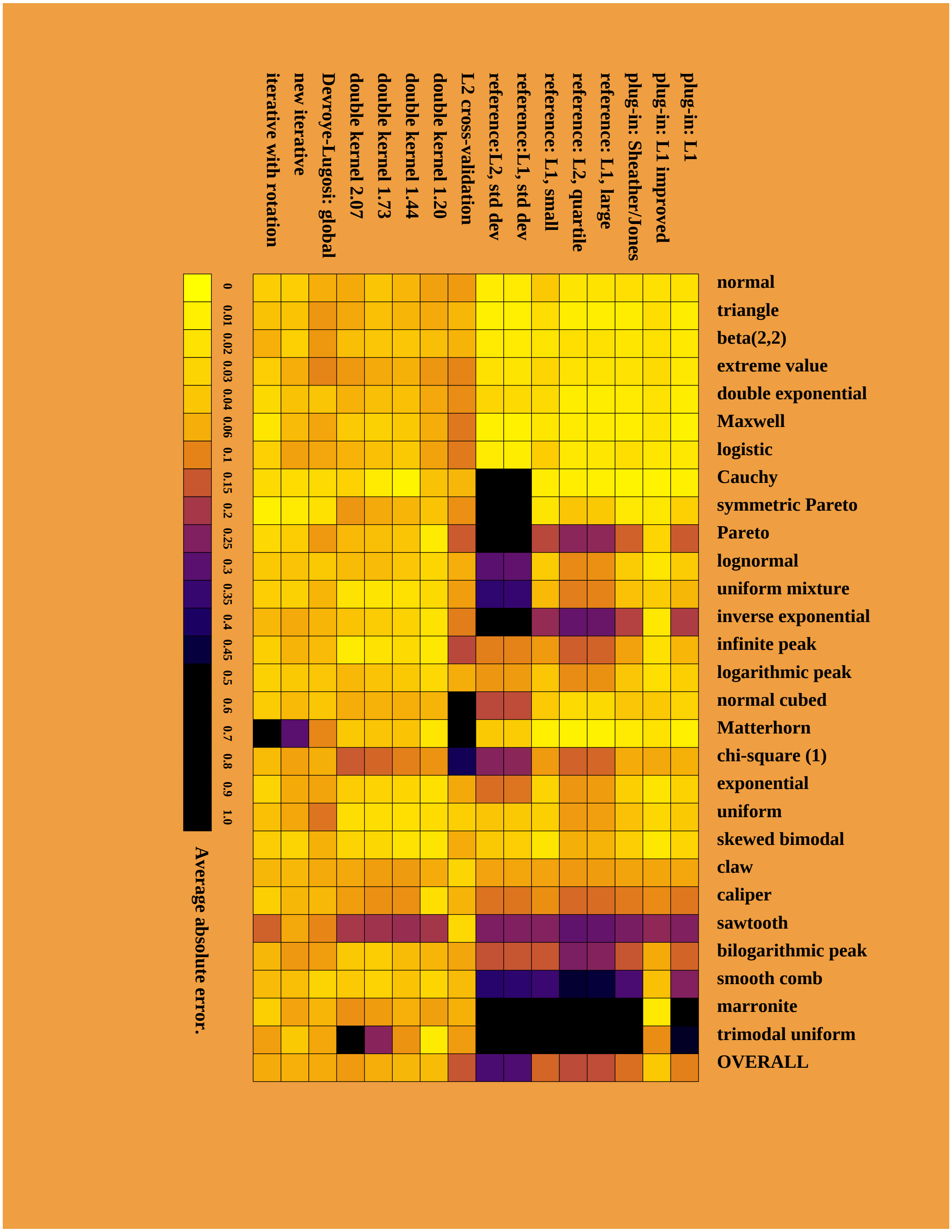

FIGURE 6. The average absolute error, defined as the average of $\int |f_{nH} - f|$ minus $\inf_h \int |f_{nh} - f|$, is shown for all densities.

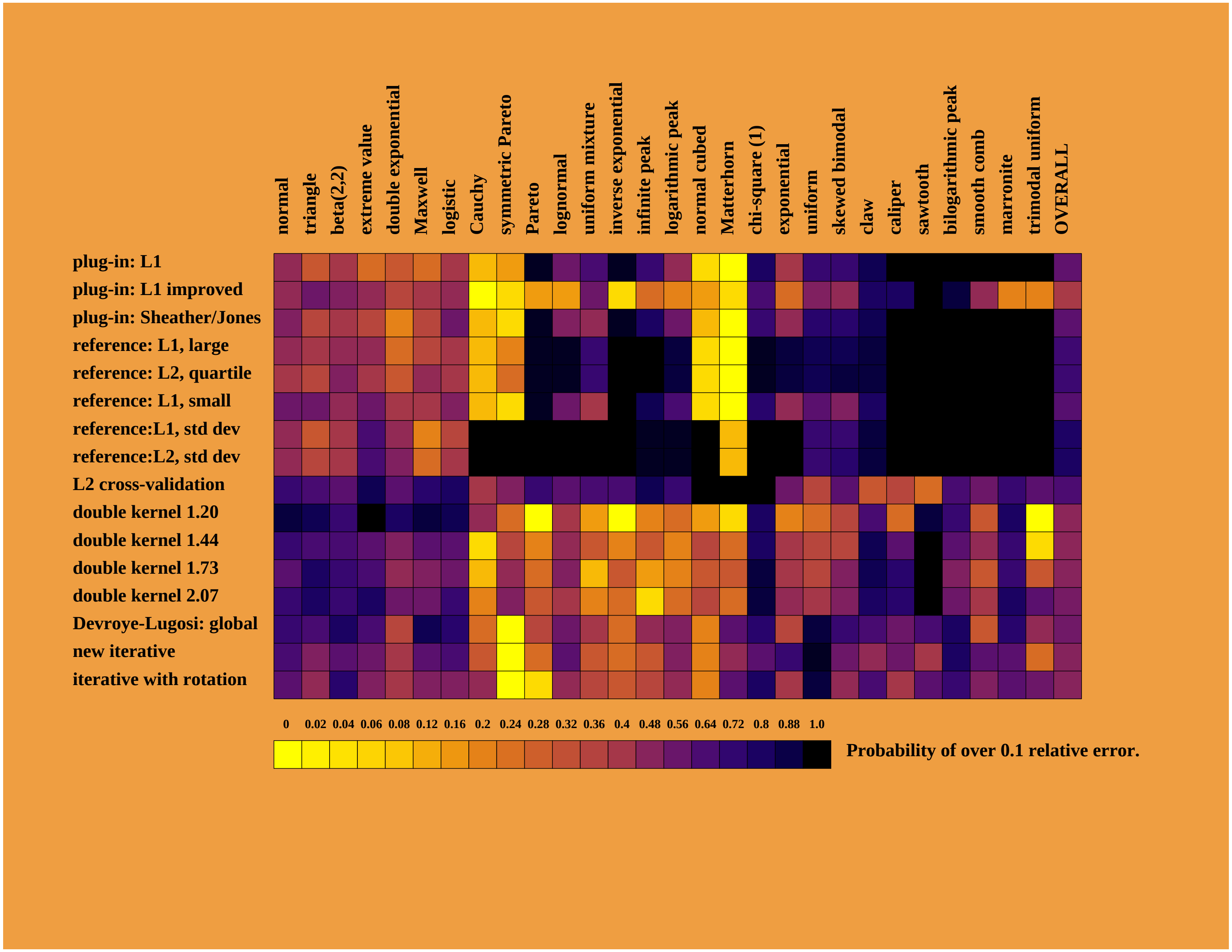

FIGURE 7. $\mathrm{P}\{P_n > 0.1\}$.

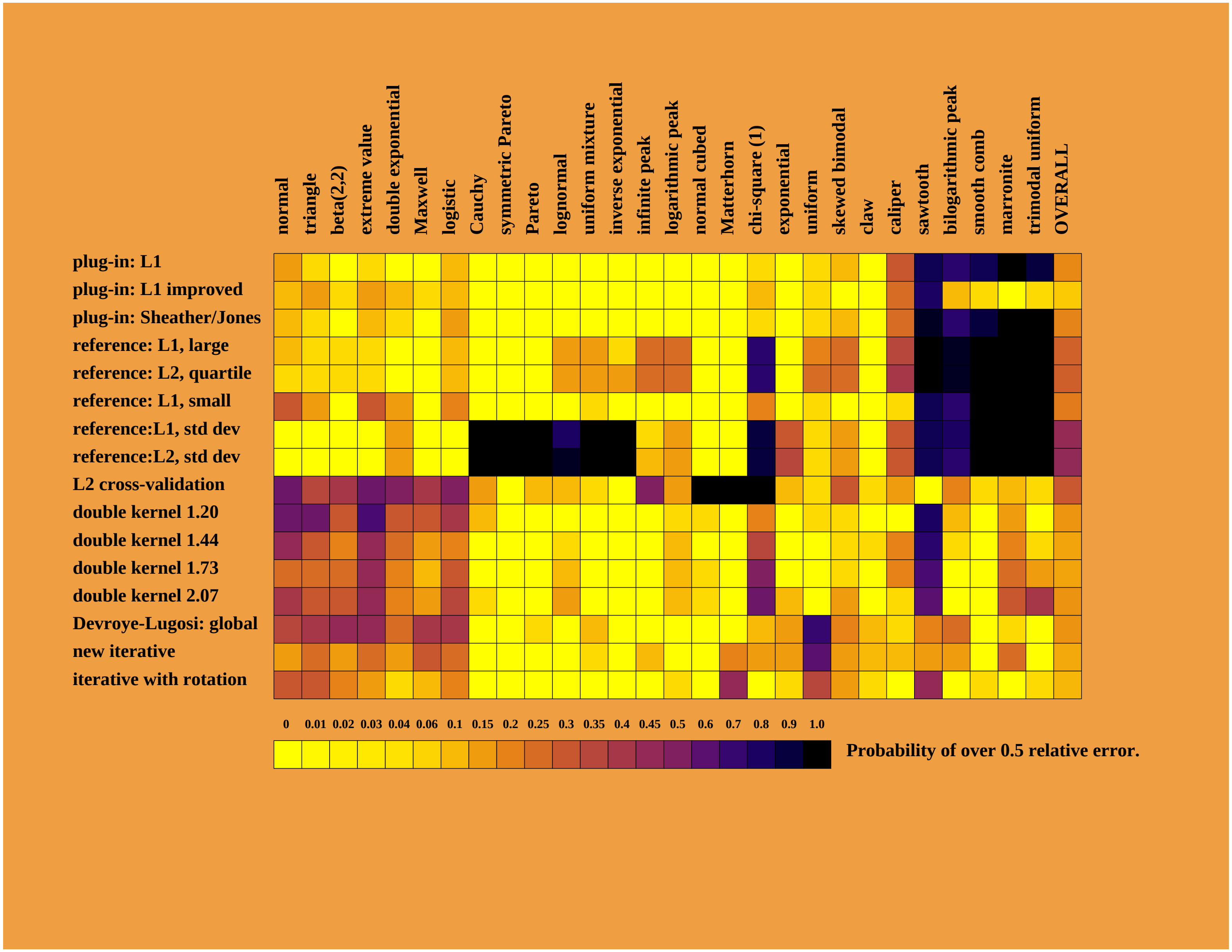

FIGURE 8. $\mathrm{P}\{P_n > 0.5\}$.

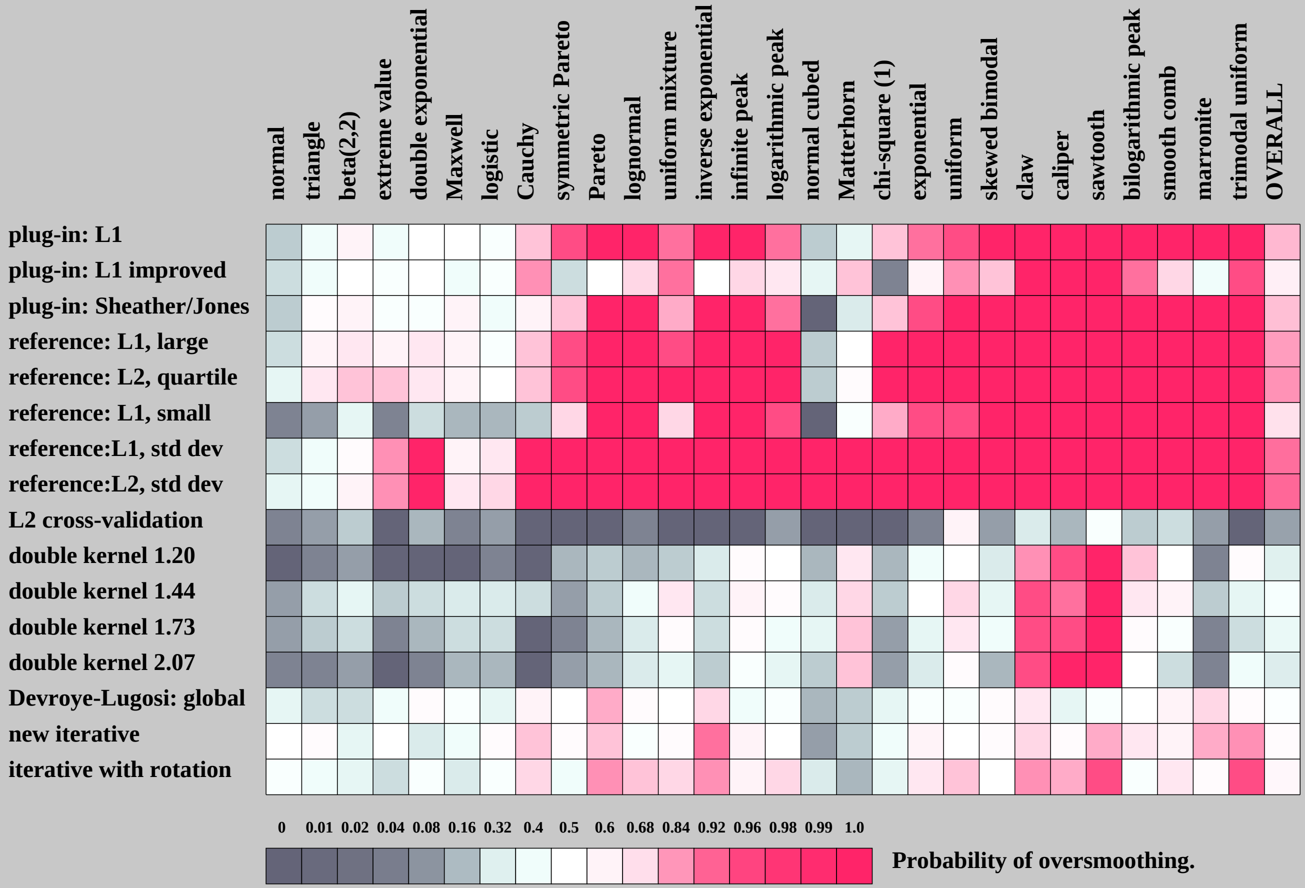

FIGURE 9. The probability of oversmoothing.

Our experiments shows why density estimation is fascinating—every method seems to "like" certain types of densities. The L_1-based plug-in methods are admissible with respect to the basket of criteria given above 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.

The conclusions of the study may be summarized as follows:

- For smooth unimodal densities (grouped on top in the figures), the reference density and plug-in methods perform better than the optimization methods (L_2 cross-validation, double kernel, h_{dl}, $h_{\mathrm{dl,it}}$, $h_{\mathrm{dl,rot}}$), simply because the plug-in formula is relatively accurate in such situations.
- The reference density methods are clearly not useful in general as they fail abysmally for long-tailed and multimodal densities (which are grouped near the midle and bottom of the figures respectively).
- Most plug-in methods fail as well for multimodal densities with the notable exception of the improved L_1 plug-in method $h_{\mathrm{pi,L1}}$, which has the best overall performance.
- The new methods $h_{\mathrm{dl}}, h_{\mathrm{dl,it}}, h_{\mathrm{dl,rot}}$ are robust across the spectrum and seem at par or slightly better than the double kernel methods.
- The plug-in, reference density and double kernel methods typically oversmooth, the L_2 cross-validation method usually undersmooths, while the new methods $h_{\mathrm{dl}}, h_{\mathrm{dl,it}}, h_{\mathrm{dl,rot}}$ undersmooth and oversmooth about equally often. For this reason, rotating the sample as is done in $h_{\mathrm{dl,rot}}$, should reduce the variation in H and stabilize the performance.
- 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.

The table below 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%.

average error	average relative error	worst relative error	probability {relative error > 10 pct}	probability {relative error > 50 pct}	
0.429	0.345	0.784	0.583	0.185	plug-in: L1
0.366	0.146	0.613	0.391	0.075	plug-in: L1 improved
0.448	0.421	0.935	0.594	0.196	plug-in: Sheather/Jones
0.491	0.563	1.137	0.680	0.275	reference: L1, large
0.495	0.576	1.157	0.685	0.280	reference: L2, quartile
0.459	0.472	1.069	0.610	0.216	reference: L1, small
0.643	0.843	1.501	0.789	0.446	reference:L1, std dev
0.647	0.858	1.526	0.798	0.453	reference:L2, std dev
0.480	0.472	1.496	0.639	0.298	L2 cross-validation
0.375	0.237	0.811	0.467	0.164	double kernel 1.20
0.379	0.220	0.931	0.467	0.137	double kernel 1.44
0.386	0.239	0.963	0.480	0.137	double kernel 1.73
0.404	0.301	1.055	0.525	0.166	double kernel 2.07
0.388	0.303	1.229	0.541	0.166	Devroye-Lugosi: global
0.385	0.249	1.229	0.483	0.130	new iterative
0.388	0.217	0.863	0.480	0.105	iterative with rotation

FIGURE 10. A summary of the results, averaged over 28 test densities and 20 repetitions each, with $n = 100$.

11.1. 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 $h_{\text{DH,L1}}, h_{\text{DH,L2}}$) 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_{cv} is inconsistent when the densities have too large infinite peaks. The double kernel and plug-in bandwidths as well as h_{dl} 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 $h_{\mathrm{ref,L1}}$, $h_{\mathrm{ref,l1}}$, $h_{\mathrm{ref,L2}}$, $h_{\mathrm{DH,L1}}$, $h_{\mathrm{DH,L2}}$ are inadmissible. 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. $h_{\mathrm{pi,l1}}$ and $h_{\mathrm{pi,L2}}$ are not immune. Among the tested bandwidths, only $h_{\mathrm{dk,1}}$, $h_{\mathrm{dk,2}}$, $h_{\mathrm{dk,3}}$, $h_{\mathrm{dk,4}}$, $h_{\mathrm{pi,L1}}$, h_{dl} and h_{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_{cv}) or $\sup_{x_1} H(x_1, X_2, X_3, \ldots, X_n) = \infty$ (as in the case of $h_{\mathrm{DH,L1}}$ or $h_{\mathrm{DH,L2}}$). This idea may be generalized to insensitivity with respect to an ϵ-fraction of the sample.

11.2. Expediency. The previous sections describe scenarios for catastrophic behavior that must be avoided at all costs. So, to narrow the scope, let us look at the behavior of the bandwidth selectors on 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 $h_{\mathrm{pi,L2}}$, 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. The bandwidths $h_{\mathrm{pi,l1}}$, $h_{\mathrm{pi,L1}}$, and the double kernel choices $h_{\mathrm{dk,1}}, h_{\mathrm{dk,2}}, h_{\mathrm{dk,3}}, h_{\mathrm{dk,4}}$ are expedient. And of course, h_{dl} is expedient because it is universally suitable (so, $\mathcal{N}$ in the supremum may be replaced by the class of all densities).

12. Adaptation for minimax criteria

In a minimax setting, a subclass $\mathcal{F}$ of densities of interest is given, and the minimax risk is commonly defined by

$$R_n(\mathcal{F}) \stackrel{\mathrm{def}}{=} \inf_{f_n} \sup_{f\in\mathcal{F}} \mathbf{E} \int |f_n - f| \ ,$$

where the infimum is over all density estimates. For many smoothness classes it is known that if f_{nh} is the kernel estimate with an appropriate kernel K, then

$$\sup_{f\in\mathcal{F}} \inf_h \mathbf{E} \int |f_{nh}-f| \le C R_n(\mathcal{F})$$

for some universal constant $C > 1$ (see, e.g., Devroye, 1987). In fact, the proof of such a result usually reveals a formula for h as a function of $f \in \mathcal{F}$. However, we do not know f, and so we are stuck. If we use the data-dependent H_n of Devroye and Lugosi (1996b), then with $m = o(n)$ and $\kappa_n = O(n^a)$ for some finite a, we have

$$\sup_{f\in\mathcal{F}} \mathbf{E} \int |f_{nH}-f| \le (3C + o(1)) R_n(\mathcal{F}) + O(\sqrt{\log n/m}) \ .$$

In many cases, the last term is negligible. Thus, our results may be used for existence proofs of minimax optimal estimators; if one can find a formula $h = h(f,n)$ for the bandwidth that gives a certain rate, then that same rate will be achieved with H.

A more interesting problem occurs when we define $\mathcal{F}$ up to a parameter, such as the class of all Lipschitz densities on $[0,1]$ with unknown Lipschitz constant α. For fixed α, the class is denoted by $\mathcal{F}_\alpha$. Assume that we know that for each α,

$$\sup_{f\in\mathcal{F}_\alpha} \inf_h \mathbf{E} \int |f_{nh}-f| \le C_\alpha R_n(\mathcal{F}_\alpha) \ . \tag{8}$$

When α is not given beforehand, the challenge is to find a data-dependent H_n such that

$$\sup_\alpha \frac{\sup_{f\in\mathcal{F}_\alpha} \mathbf{E}\int |f_{nH_n}-f|}{R_n(\mathcal{F}_\alpha)} \le C'$$

for some suitable constant C'. In that case, we may say that H_n adapts itself nicely to the union of the classes $\mathcal{F}_\alpha$. Such a point of view is not without merit. Assume that H_n is picked by the second method of Devroye and Lugosi (1996b). Then, using the inequalities of Theorem 2 and Corollary 2, assuming $\kappa_n = O(n^a)$ for some finite $a > 0$, we see that there exist universal constants D and E such that

$$\begin{aligned}
\sup_\alpha \frac{\sup_{f\in\mathcal{F}_\alpha} \mathbf{E}\int |f_{nH_n}-f|}{R_n(\mathcal{F}_\alpha)} &\le \sup_\alpha \frac{\sup_{f\in\mathcal{F}_\alpha} D\inf_h \mathbf{E}\int |f_{nh}-f| + E\sqrt{\frac{\log n}{n}}}{R_n(\mathcal{F}_\alpha)} \\
&\le \sup_\alpha \frac{DC_\alpha R_n(\mathcal{F}_\alpha) + E\sqrt{\frac{\log n}{n}}}{R_n(\mathcal{F}_\alpha)} \\
&= D\sup_\alpha C_\alpha + \frac{E\sqrt{\frac{\log n}{n}}}{\inf_\alpha R_n(\mathcal{F}_\alpha)} \ .
\end{aligned}$$

In the majority of the interesting cases, this is $D\sup_\alpha C_\alpha + o(1)$. Indeed, then, one may use H_n and be assured of good adaptive capabilities whenever (8) holds and the constants C_α are uniformly bounded. Typically, (8) is easy to verify, so that one need not be concerned with the details of the random bandwidth H_n. Furthermore, the universal nature of the above result says something very powerful about the kernel estimate and about the bandwidths described in the first part of the paper.

13. References

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 and L. Devroye, "A comparison of kernel density estimates," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 38(3), pp. 3–59, 1994.

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.

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

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, "A universal lower bound for the kernel estimate," *Statistics and Probability Letters*, vol. 8, pp. 419–423, 1989.

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

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, "Exponential inequalities in nonparametric estimation," in: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 31–44, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

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 an estimate of Chiu," *Statistics and Probability Letters*, vol. 20, pp. 183–188, 1994.

L. Devroye, "On good deterministic smoothing sequences for kernel density estimates," *Annals of Statistics*, vol. 22, pp. 886–889, 1994.

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

L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

L. Devroye and G. Lugosi, "A universally acceptable smoothing factor for kernel density estimation," *Annals of Statistics*, vol. 0, pp. 0–0, 1996a.

L. Devroye and G. Lugosi, "Nonasymptotic universal smoothing factors, kernel complexity, and Yatracos classes," Technical Report, McGill University, Montreal, 1996b.

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.

P. P. B. Eggermont and V. N. LaRiccia, "Maximum smoothed likelihood density estimation for inverse problems," *Annals of Statistics*, vol. 23(1), pp. 199–220, 1995.

P. P. B. Eggermont and V. N. LaRiccia, "A simple and effective bandwidth selector for kernel density estimation," *Scandinavian Journal of Statistics*, vol. 23, pp. 285–301, 1996.

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 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.

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.

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, "On convergence rates in nonparametric problems," *International Statistical Review*, vol. 57, pp. 45–58, 1989.

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.

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, "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.

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

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

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, 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. 11, pp. 180–215, 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.

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.

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. Ya. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities,” *Theory of Probability and its Applications*, vol. 16, pp. 264–280, 1971.

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 M. C. Jones, *An Introduction to Kernel Smoothing*, Chapman and Hall, London, 1995.

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.

Y. G. Yatracos, “Rates of convergence of minimum distance estimators and Kolmogorov’s entropy,” *Annals of Statistics*, vol. 13, pp. 768–774, 1985.

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