Sociedad de Estadística e Investigación Operativa
Test (1997) *Vol.* **6**, *No.* 2, pp. 223–320

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.

Key Words: Density estimation, kernel estimate, convergence, smoothing factor, minimum distance estimate, asymptotic optimality, simulation study.
AMS subject classification: 62G05.

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

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

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 (1996) 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 (1997) 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,

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

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 $h = a_n\widehat{\sigma}$ for a function a_n (depending upon the family), where $\widehat{\sigma}$ 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 (1996) 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, $T = 1$ 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.

$\diamondsuit$ 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., $|K(x) - K(y)| \leq C\|x - y\|$ 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 $C'(1 - \|x\|^d)_+$, where $(u)_+ = \max(u, 0)$. For more general classes, we will show how to take the parameters in remarks below.

$\diamondsuit$ 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

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

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.

$\Diamond$ 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 2.1. *(Devroye and Lugosi, 1996). 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 2.1 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 2.1, we refer to Devroye and Lugosi (1996a). However, we will state four lemmas, which each contribute key elements to the proof.

Lemma 2.1. *(Devroye and Lugosi, 1996). For each density f and elegant kernel K, $C' = C2^d\sqrt{d} + d$, the estimate $f_n (\in \mathcal{G}_n)$ satisfies*

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

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 $\log(b_n/a_n) \leq c' n^a$ 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 2.2. *(Devroye, 1983). Assume $K \geq 0$. 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 2.1 follows from the following two lemmas.

Lemma 2.3. *(Devroye and Penrod, 1984). Assume $K \geq 0$. Then*

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

We note that Lemma 2.3 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,

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

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

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

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

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

Theorem 2.1 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 2.1, 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 $K(x) = (3/4)(1-x^2)_+$. Our method (with bandwidth called h_{dl}) has the following parameter settings:

A. $m = \lfloor n^{0.85} \rfloor$. Note that $\frac{m}{n^{4/5}} \to \infty$ 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 $h \leq H$, 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 $0 \leq i \leq k$. 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 (1997) 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 1 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

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

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 4.1. *(Devroye and Lugosi, 1997). 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,*

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

Corollary 4.1. *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 4.2. *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, $K \geq 0$ and d (see Devroye and Györfi, 1985), we have

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

This universal asymptotic bound is shared with the first bandwidth. However, Theorem 4.1 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 4.1. *(Devroye and Lugosi, 1997). For each n, m, and for all f,*

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

The first term on the right-hand side of the inequality of Lemma 4.1 may be bounded by the following result:

Lemma 4.2. *(Devroye and Lugosi, 1996). Let K be a bounded kernel. If $m > 0$ is a positive integer such that $2m \leq n$, then*

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

Therefore,

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

A suitable upper bound for $\sup_{A\in\mathcal{A}} \left|\int_A f - \mu_m(A)\right|$ may be obtained via 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 $\epsilon > 0$ and $\psi(n, m, k, \epsilon) = 4e^8(m^2 + 1)(1 + 2km^2(n - m))^2 e^{-2m\epsilon^2}$, we have in fact

$$\mathbf{P}\left\{ \sup_{A\in\mathcal{A}} \left| \mu_m(A) - \int_A f \right| > \epsilon \middle| X_1, \ldots, X_{n-m} \right\} \leq \psi(n, m, k, \epsilon),$$

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

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

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.

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

$\Diamond$ ISOSCELES TRIANGULAR DENSITY. If $K(x) = (1-|x|)_+$, then elementary calculation shows that for all n, $\kappa_n \leq n+1$.

$\Diamond$ SYMMETRIC UNIMODAL KERNELS. As a first main example, consider symmetric unimodal densities (i.e., $K \geq 0$ and $\int K = 1$) on the real line. Let β be the last positive value for which $\int_\beta^\infty K \leq 1/(4n)$. 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 $\int K' = 1$, $\int |K - K'| \leq 1/n$, and that K' is Riemann with parameter $k = 2N + 2 \leq 8nK(0)\beta + 10$. Thus, $\kappa_n \leq 8nK(0)\beta + 10$.

Example 5.1. BOUNDED COMPACT SUPPORT DENSITIES. If K is symmetric, nonnegative, unimodal (such as the Epanechnikov-Bartlett kernel) and $K(x) \leq aI_{[-b,b]}(x)$, then $\kappa_n \leq 8nab + 10$.

Example 5.2. 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 $\beta \geq 1$,

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

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

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

Example 5.3. 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,

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

Example 5.4. DENSITIES WITH POLYNOMIAL TAILS. Note that if K is a symmetric unimodal density, and $|K(x)| \leq c/(1+|x|^{\gamma+1})$ 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.

$\Diamond$ 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 $|K(x)| \leq c/(1+|x|^{\gamma+1})$ 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.

$\Diamond$ 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,

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

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

$\diamond$ 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 $\int_\beta^\infty c_d u^{d-1} M(u)du \leq 1/(2n)$. 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 $k = N + 1 \leq 2nc_d K(0)\beta^d + 2$. Moreover,

$$\begin{aligned}
\int |K(x) - K'(x)|dx &= \int_0^\beta c_d u^{d-1} |M(u) - M'(u)|du \\
&+ \int_\beta^\infty c_d u^{d-1} |M(u) - M'(u)|du \\
&\leq \frac{1}{2n} + c_d \beta^{d-1} \int_0^\beta |M(u) - M'(u)|du \\
&\leq \frac{1}{2n} + c_d \beta^{d-1} \frac{M(0)\beta}{N} \leq \frac{1}{n}.
\end{aligned}$$

Thus,

$$\kappa_n \leq 2nc_d M(0)\beta^d + 2 \ .$$

$\diamond$ 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

$$Q(h) = \max \left(\left| \int_A f_{n-m,h} - \mu_m(A) \right| \ , \left| \int_B f_{n-m,h} - \mu_m(B) \right| \right) \ ,$$

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

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

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 $h_{\text{dl,it}}$ 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 $h_{\text{dl,rot}}$.

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,

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

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

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

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

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

For this existence result, see section 7.5 of Devroye (1987). It suffices to take a symmetric kernel L integrating to one, having compact support,

possessing a zero second moment. We cannot in general tell how to choose h'. This is frustrating, because nobody likes to work with the knowledge that there is something better out there. However, it is also a blessing, as we will use this property to our advantage to design an automatic smoothing factor selector.

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.

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

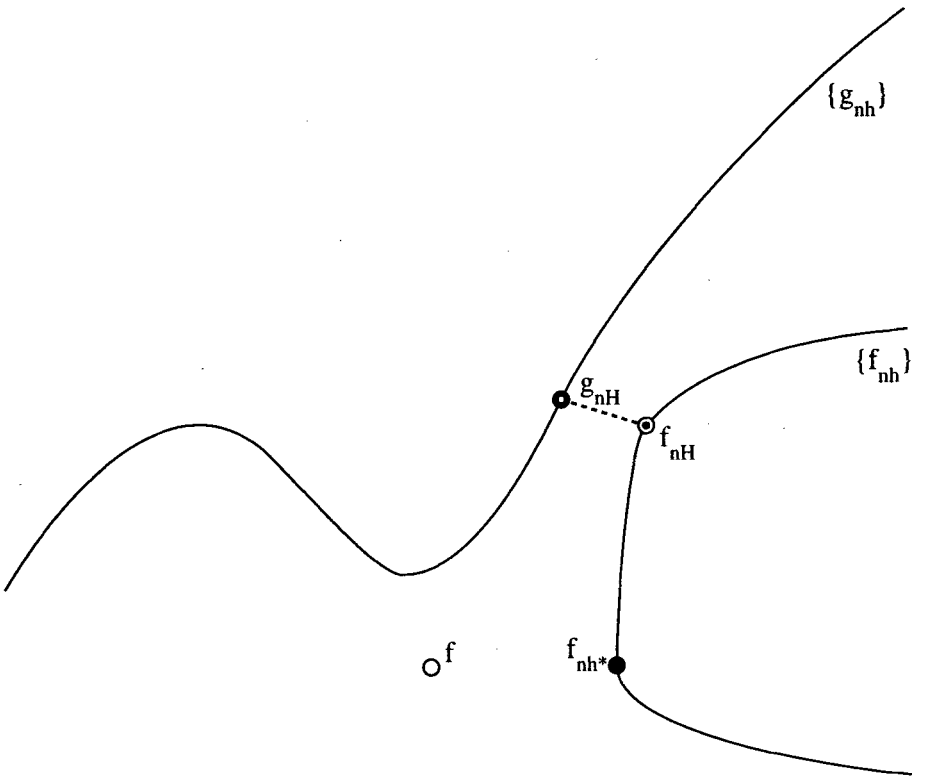

Figure 1: We show two families of density estimates in the set of all densities. The double kernel bandwidth minimizes the L_1 distance between f_{nh} and g_{nh}.

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 L_1 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

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

The choice of h for which we have

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

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

$$h = \left(\frac{\sqrt{15}\int\sqrt{f}}{\sqrt{2\pi}\int|f''|}\right)^{2/5} n^{-1/5}. \tag{8.1}$$

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 (8.1) is called an L_1 plug-in method.

8.2 L_1 reference density method

If (8.1) (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 (8.1) can be written as

$$h = \sigma\left(\frac{15e\sqrt{2\pi}}{8n}\right)^{1/5} = 1.6644\ldots\sigma\, n^{-1/5}. \tag{8.2}$$

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

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{8.3}$$

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

$$\widehat{\sigma} = \frac{X_{[3n/4]} - X_{[n/4]}}{F^{\text{inv}}(3/4) - F^{\text{inv}}(1/4)} = \frac{X_{[3n/4]} - X_{[n/4]}}{1.35\ldots}\ , \tag{8.4}$$

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 L_2 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.

- $h_{\text{ref},L_1} = 2.279\,\widehat{\sigma}\,n^{-1/5}$, where $\widehat{\sigma}$ is defined in (8.4).
- $h_{\text{DH},L_1} = 2.279\,\widehat{\sigma}\,n^{-1/5}$, where $\widehat{\sigma}$ is defined in (8.3). DH is a mnemonic for Deheuvels and Hominal.
- $h_{\text{ref},l_1} = 1.6644\,\widehat{\sigma}\,n^{-1/5}$, where $\widehat{\sigma}$ is defined by (8.4).

8.3 L_2 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:

$$h = \left(\frac{A(K)}{n\int f''^2}\right)^{1/5}, \tag{8.5}$$

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

$$h = \left(\frac{15}{n\int f''^2}\right)^{1/5} \tag{8.6}$$

(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

$$h_{\mathrm{pi,L_2}} = \min\left(h_{\mathrm{ms,L_2}}, \left(\frac{15}{n\rho}\right)^{1/5}\right) ,$$

where

$$h_{\mathrm{ms,L_2}} = 3(3/7n)^{1/5}\widehat{\sigma} = 2.532362\ldots\widehat{\sigma}n^{-1/5}$$

is a safe "maximal" bandwidth (Terrell, 1990; Scott and Terrell, 1985), and $\widehat{\sigma}$ is as in (8.3). It is easy to show that for all densities, $h_{\mathrm{pi,L_2}} \to 0$ and $nh_{\mathrm{pi,L_2}} \to \infty$ in probability whenever L'''' is uniformly bounded. This implies that $h_{\mathrm{pi,L_2}}$ is universally consistent.

Other L_2 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 L_2 reference density methods

The reference density method with a normal reference density used in (8.5) 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

- $h_{\mathrm{DH},L_2} = 2.345\, \hat{\sigma}\, n^{-1/5}$, where $\hat{\sigma}$ is (8.3).
- $h_{\mathrm{ref},L_2} = 2.345\, \hat{\sigma}\, n^{-1/5}$, where $\hat{\sigma}$ is (8.4).

8.5 L_2 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

$$\frac{\int K^2}{nh} + \frac{1}{n^2}\sum_{i \neq j} M_h(X_i - X_j) \tag{8.7}$$

for some function M. We conjecture that for any pair (M, K), the minimizer of (8.7) 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

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

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

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 $h_{\mathrm{dk},1}$, $h_{\mathrm{dk},2}$, $h_{\mathrm{dk},3}$ and $h_{\mathrm{dk},4}$ 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 $l \leq 1$.

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 $h_{\mathrm{ref},\mathrm{l}_1}/2^i$ with the property that $\chi(a) > J_{nh_{\mathrm{ref},\mathrm{l}_1}}$, and b as the smallest number of the form $h_{\mathrm{ref},\mathrm{l}_1} \times 2^i$ with the property that $\xi(a) > J_{nh_{\mathrm{ref},\mathrm{l}_1}}$. This procedure works with any starting point, not just $h_{\mathrm{ref},\mathrm{l}_1} J1865$. 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.

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

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

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

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

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

These bounds are used in all our computations of $\inf J_{nh}$ and $\inf J'_{nh}$.

$\Diamond$ 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.

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

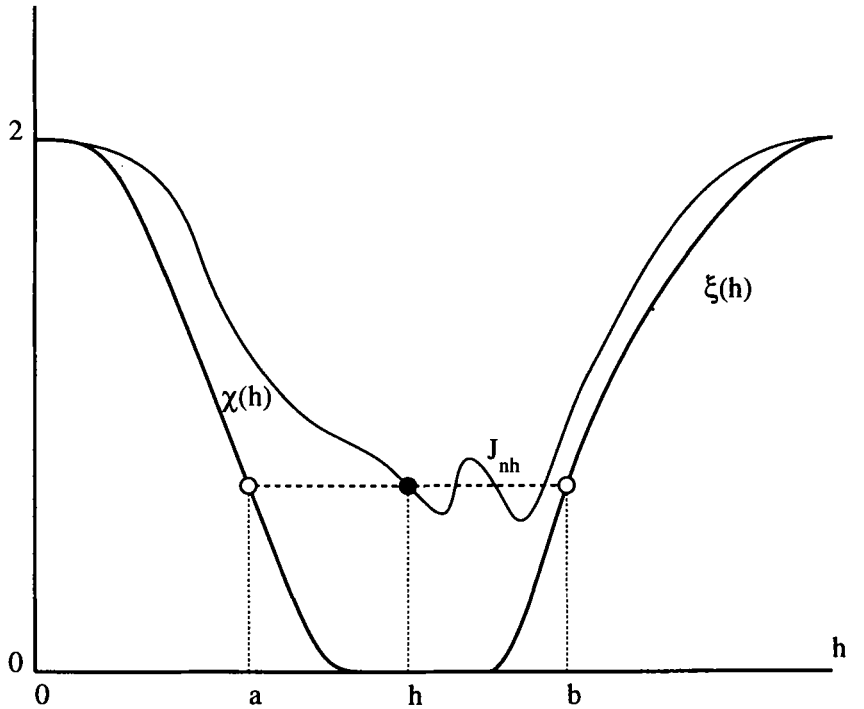

Figure 2: A hypothetical L_1 error J_{nh} is shown as a function of h. Also shown are the lower bounds $\chi(h)$ and $\xi(h)$. The figure illustrates how one computation of J_{nh} directly leads to an interval $[a, b]$ that contains the overall minimum, and yet stays bounded away from 0 and ∞.

upon $(K, L_{1.5})$. In algorithmic format, the bandwidths $h_{\mathrm{pi},\mathrm{l}_1}$ and $h_{\mathrm{pi},\mathrm{L}_1}$ 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.

$$
\begin{aligned}
h' &\leftarrow h_{\mathrm{ref},\mathrm{L}_1}(h_{\mathrm{dk}}) \\
A &= \int \sqrt{f_{nh'}} \\
R &= \frac{A\sqrt{\int (K-L)^2}}{\sqrt{nh'}\int |f_{nh'} - g_{nh'}|} \\
h'' &= h' \max\left(1, (10R)^{2/5}\right) \\
B &= \frac{2\int |f_{nh''} - g_{nh''}|}{h''^2 \int x^2 K}
\end{aligned}
$$

$\widehat{\sigma}$ is defined by (8.3)

$$h_{\mathrm{ms,L_1}} = 2.71042\ldots\widehat{\sigma}n^{-1/5}$$

$$h_{\mathrm{pi,l_1}}(h_{\mathrm{pi,L_1}}) = \min\left\{\left(\frac{\sqrt{15/(2\pi)}A}{B}\right)^{2/5} n^{-1/5}\ ,\ h_{\mathrm{ms,L_1}}\right\}$$

Remark 10.1. A is an estimate of $\int\sqrt{f}$, and B is an estimate of $\int|f''|$. $h_{\mathrm{ms,L_1}}$ is a safe "maximal" bandwidth derived on page 113 of Devroye and Györfi (1985). The coefficient 2.71042... 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 (1996, 1997) 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: $f(x) = \frac{1}{\sqrt{2\pi x}}e^{-x/2}$, $x > 0$.

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

$$\begin{aligned} f &= \frac{1}{2}\phi(0,1) + \frac{1}{10}\phi(-1,0.1) + \frac{1}{10}\phi(-0.5,0.1) \\ &+ \frac{1}{10}\phi(0,0.1) + \frac{1}{10}\phi(0.5,0.1) + \frac{1}{10}\phi(1,0.1) . \end{aligned}$$

24. The smooth comb: a normal mixture (density # 14 in Marron and Wand, 1992), with

$$\begin{aligned} f &= \frac{32}{63}\phi\left(-\frac{31}{21}, \frac{32}{63}\right) + \frac{16}{63}\phi\left(\frac{17}{21}, \frac{16}{63}\right) + \frac{8}{63}\phi\left(\frac{41}{21}, \frac{8}{63}\right) \\ &+ \frac{4}{63}\phi\left(\frac{53}{21}, \frac{4}{63}\right) + \frac{2}{63}\phi\left(\frac{59}{21}, \frac{2}{63}\right) + \frac{1}{63}\phi\left(\frac{62}{21}, \frac{1}{63}\right) . \end{aligned}$$

25. The caliper: The density of $S(X + 0.1)$, where S is a random sign, and X has density $f(x) = 4(1 - x^{1/3})$ on $[0, 1]$.

26. The trimodal uniform density:

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

where $f_{[a,b]}$ denotes the uniform density on $[a, b]$.

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

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

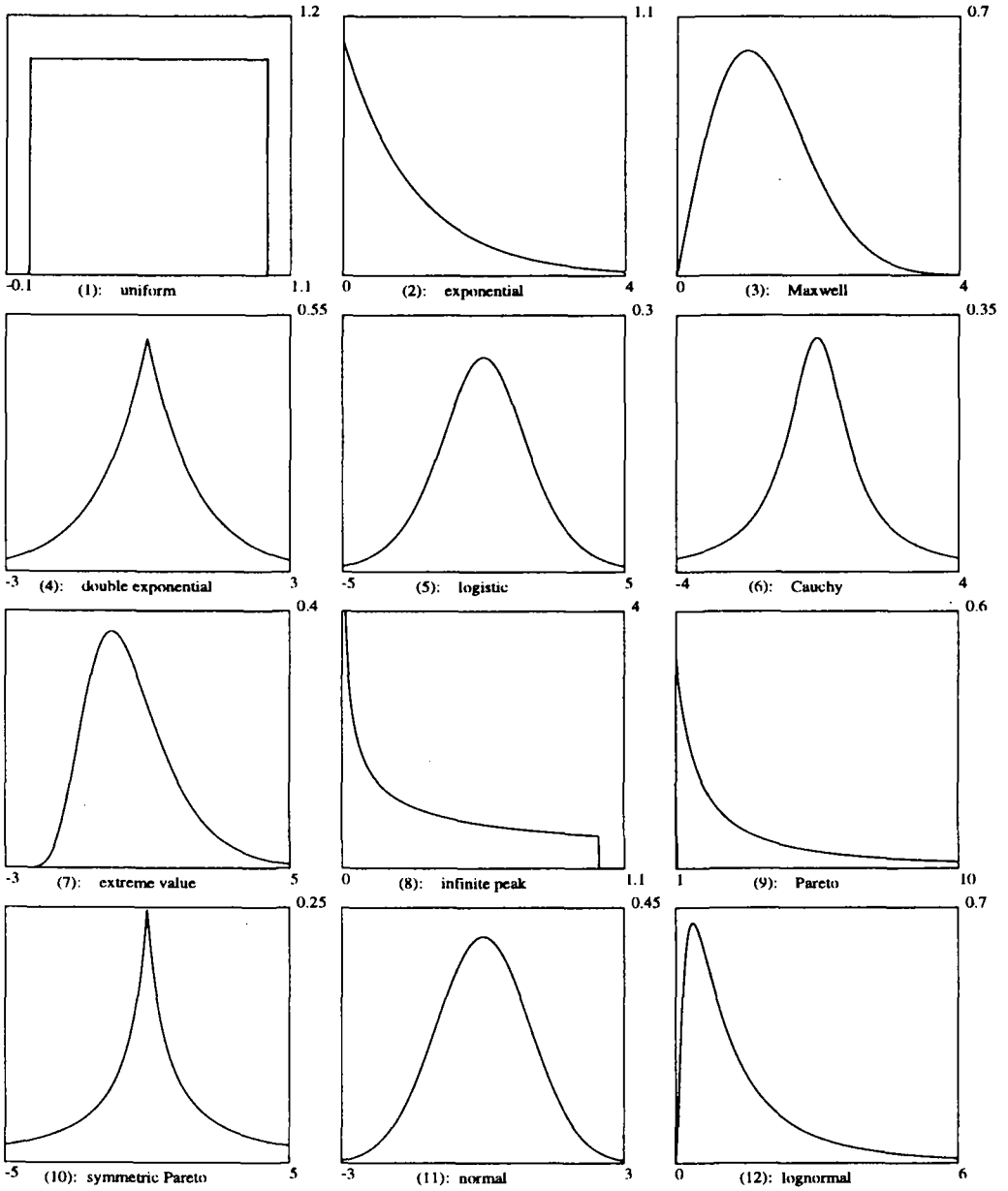

Figure 3: The unimodal densities in our collection (first part).

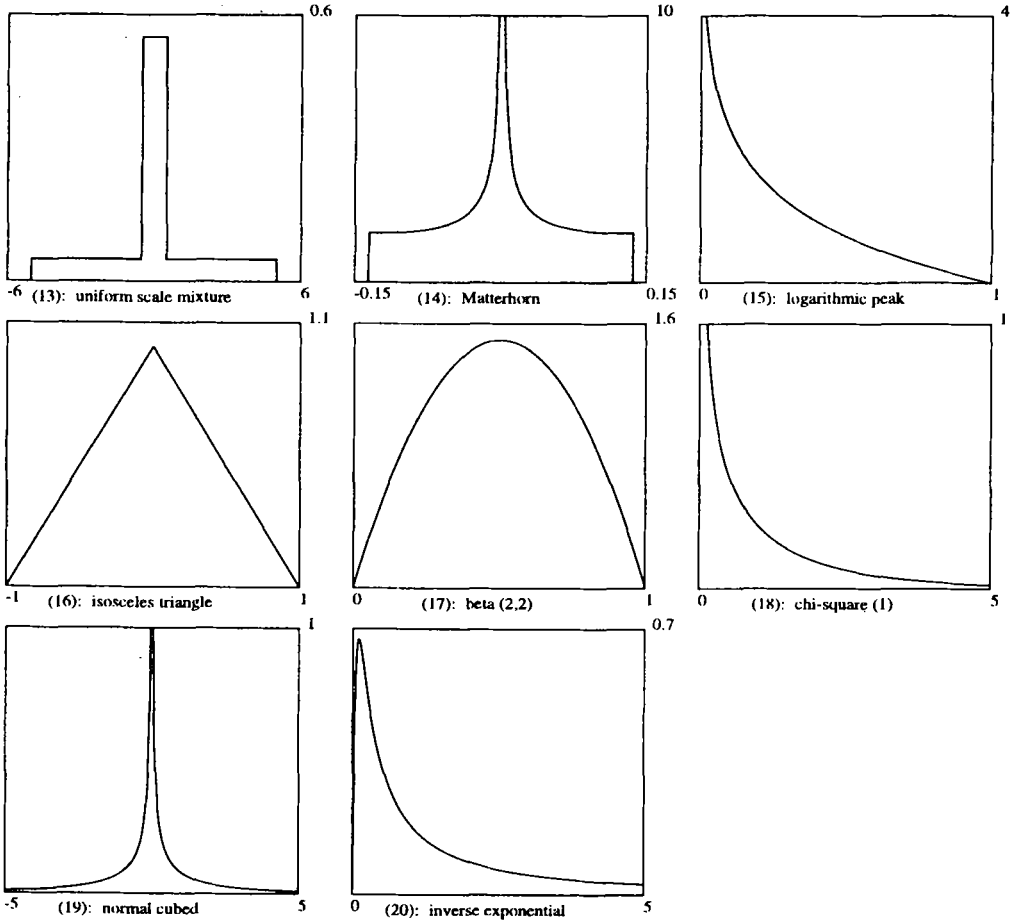

Figure 4: The unimodal densities in our collection (second part).

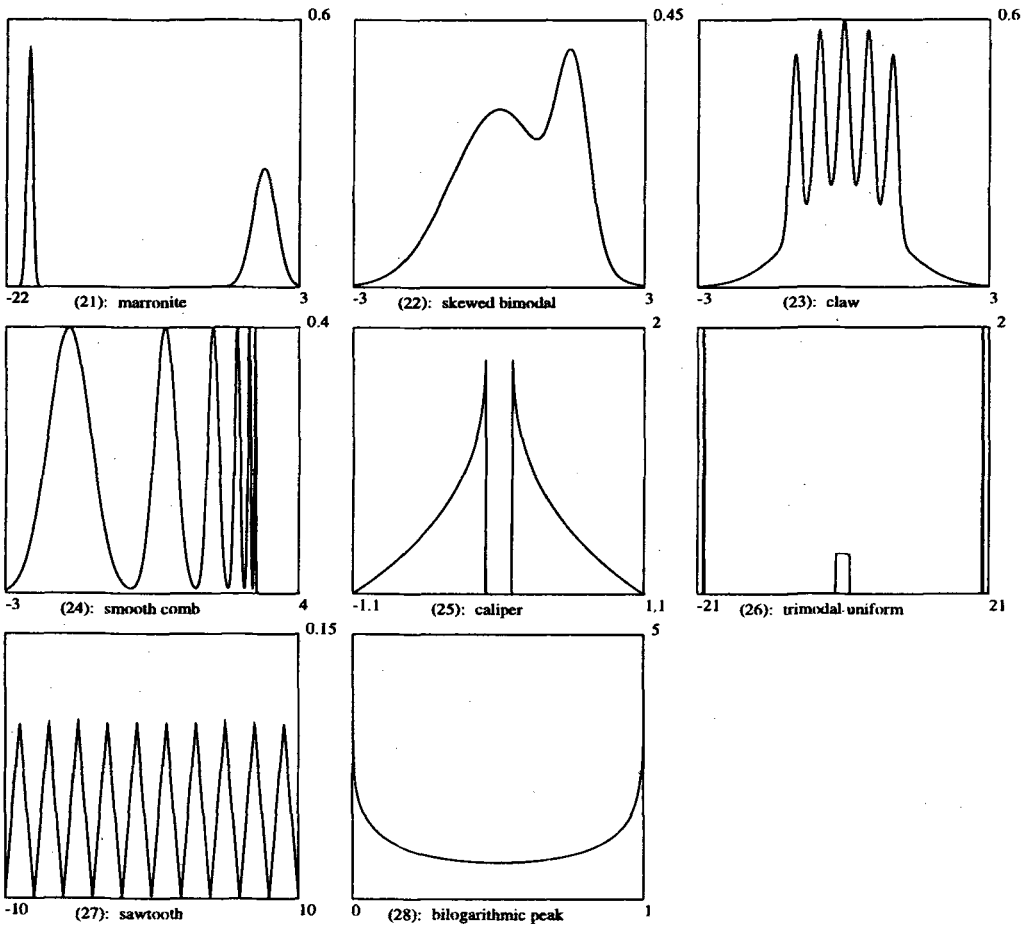

Figure 5: The multimodal densities in our collection.

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.

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}, $h_{\text{dl,it}}$, $h_{\text{dl,rot}}$), 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_{pi,L_1}, which has the best overall performance.

- The new methods h_{dl}, $h_{\text{dl,it}}$, $h_{\text{dl,rot}}$ 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_{dl}, $h_{\text{dl,it}}$, $h_{\text{dl,rot}}$ undersmooth and oversmooth about equally often. For this reason, rotating the sample as is done in $h_{\text{dl,rot}}$, 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.

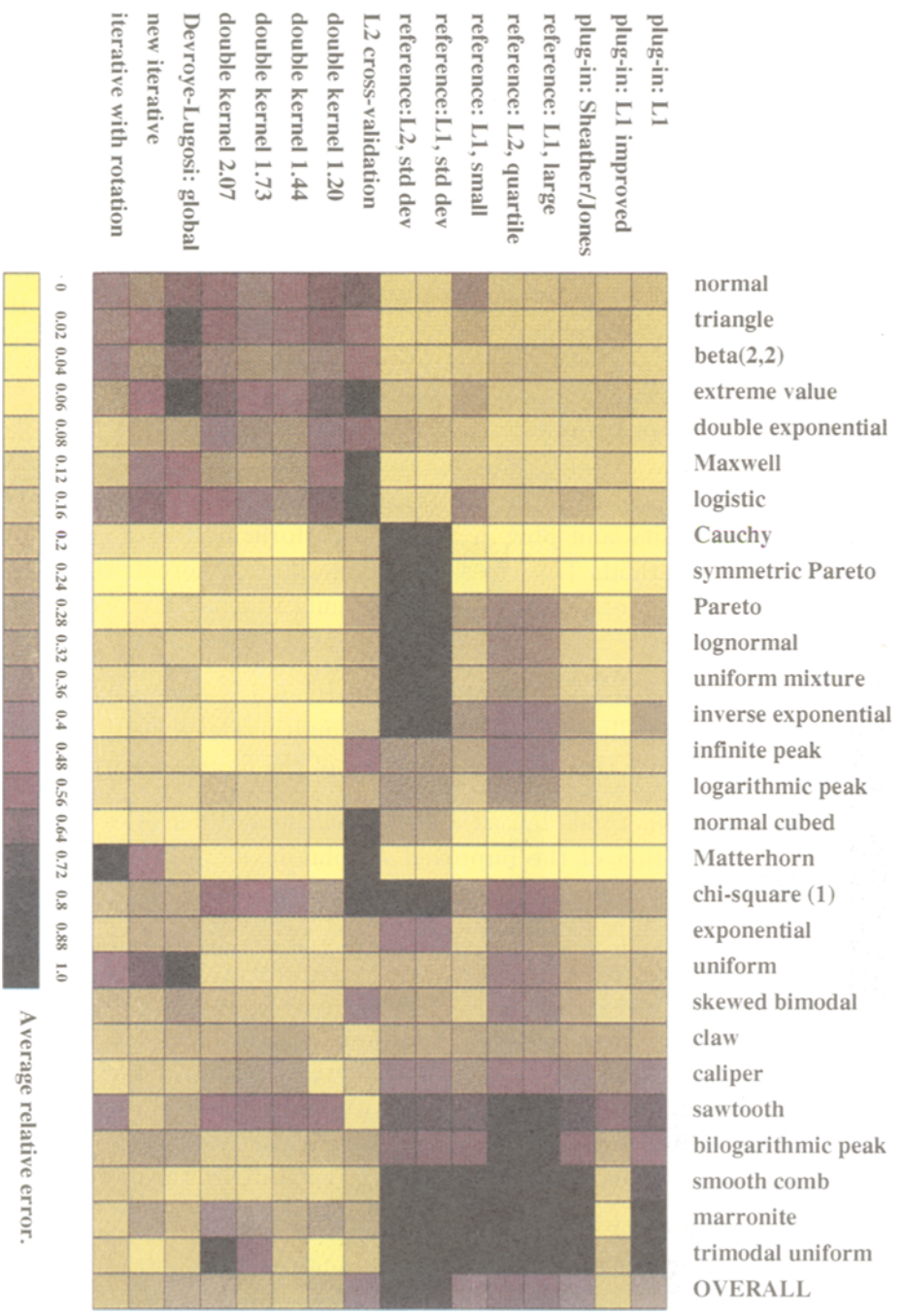

Figure 6: The average relative error (P_n) is shown for all densities.

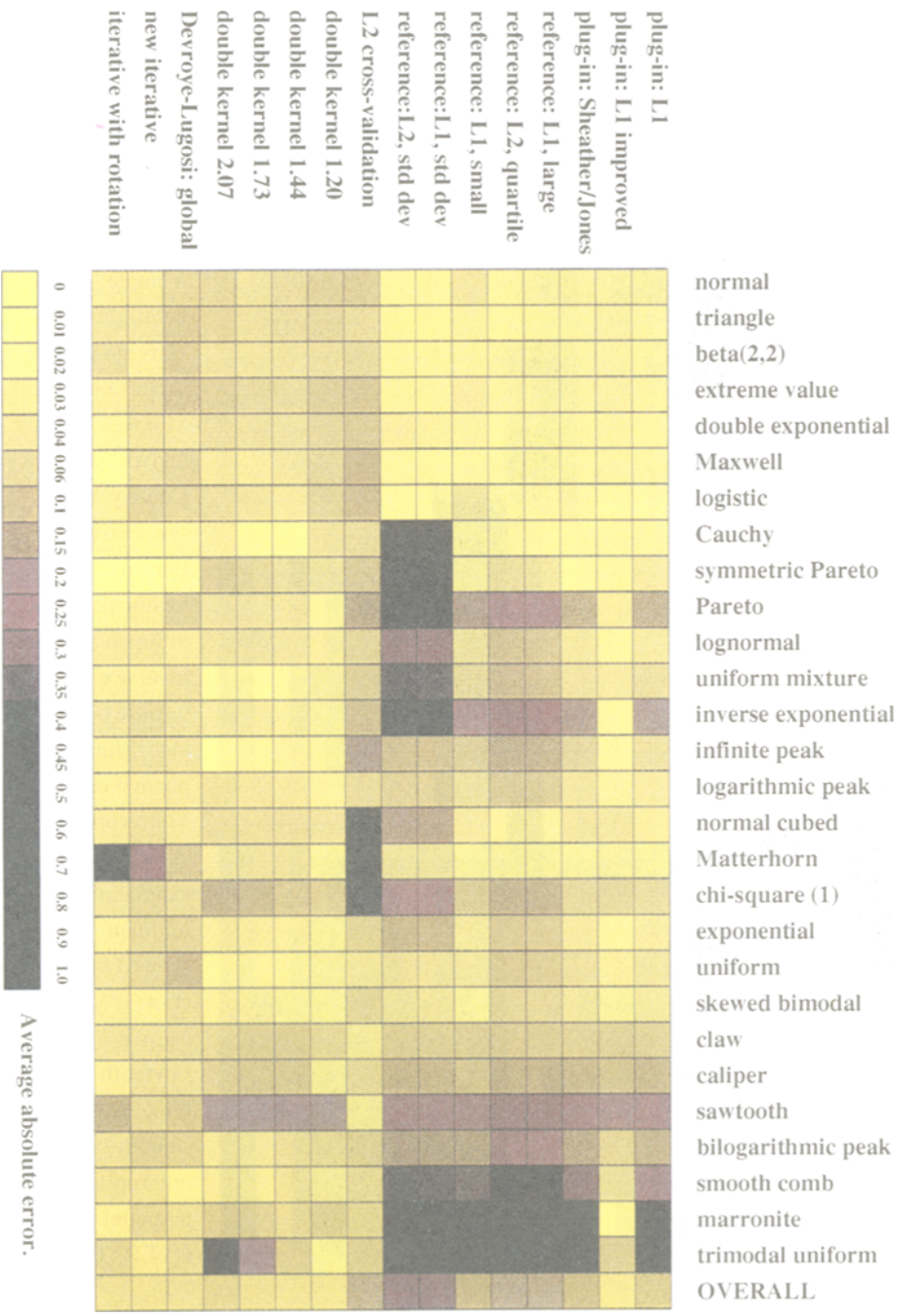

Figure 7: The average absolute error, defined as the average of $\int |f_{nH} - f|$ minus $\inf_h \int |f_{nh} - f|$, is shown for all densities.

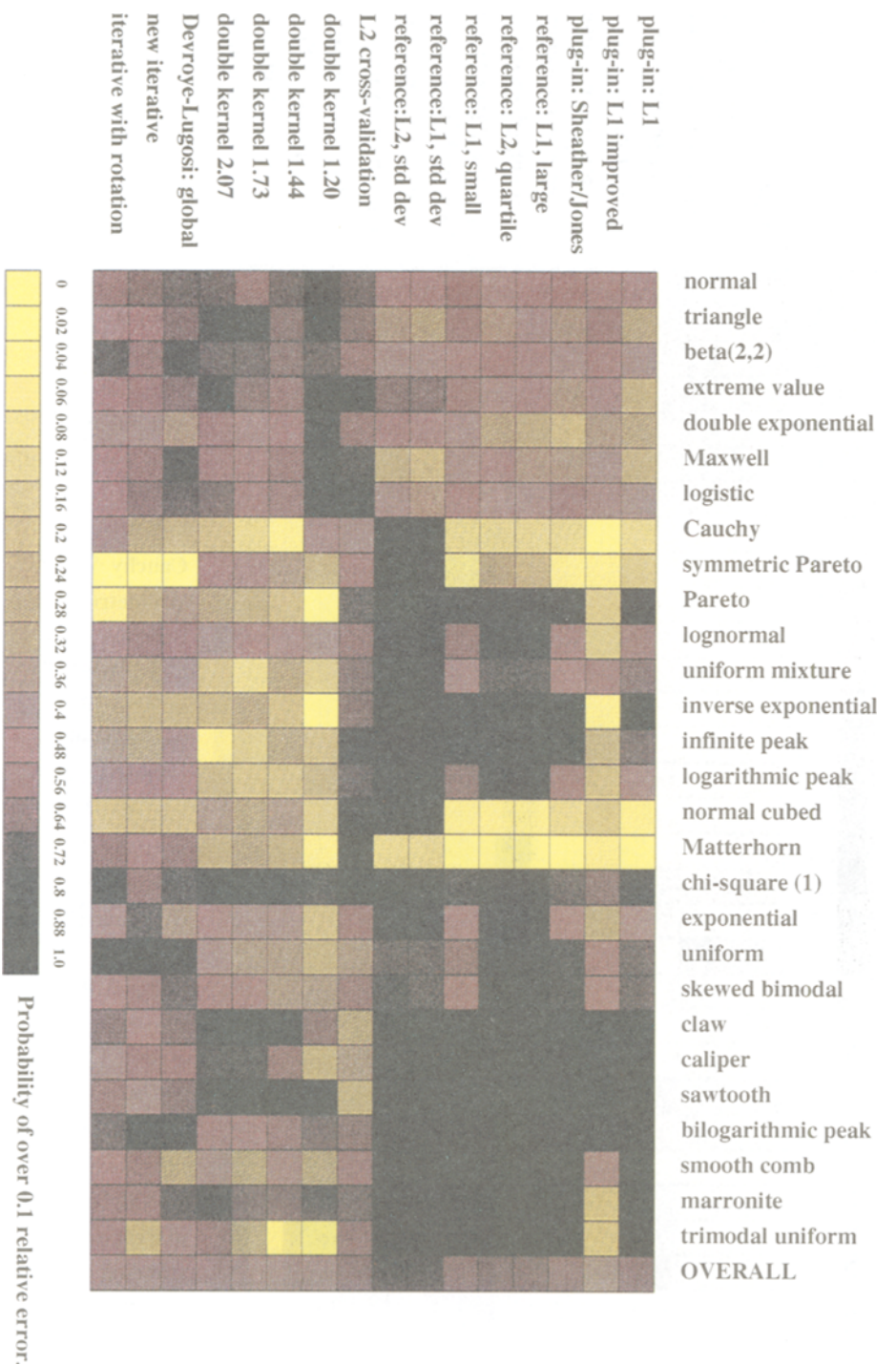

Figure 8: $\mathbf{P}\{P_n > 0.1\}$.

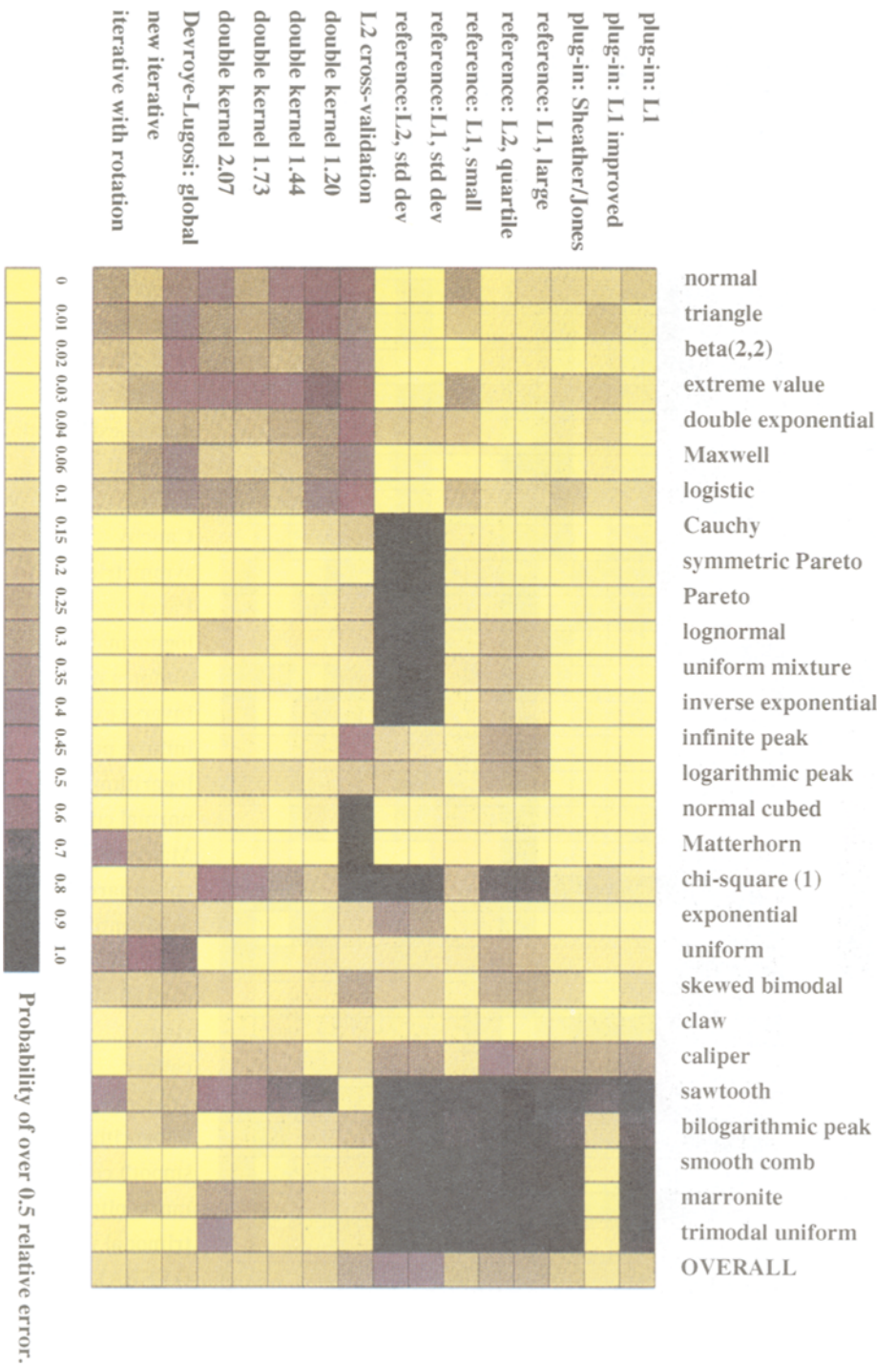

Figure 9: $\mathbf{P}\{P_n > 0.5\}$.

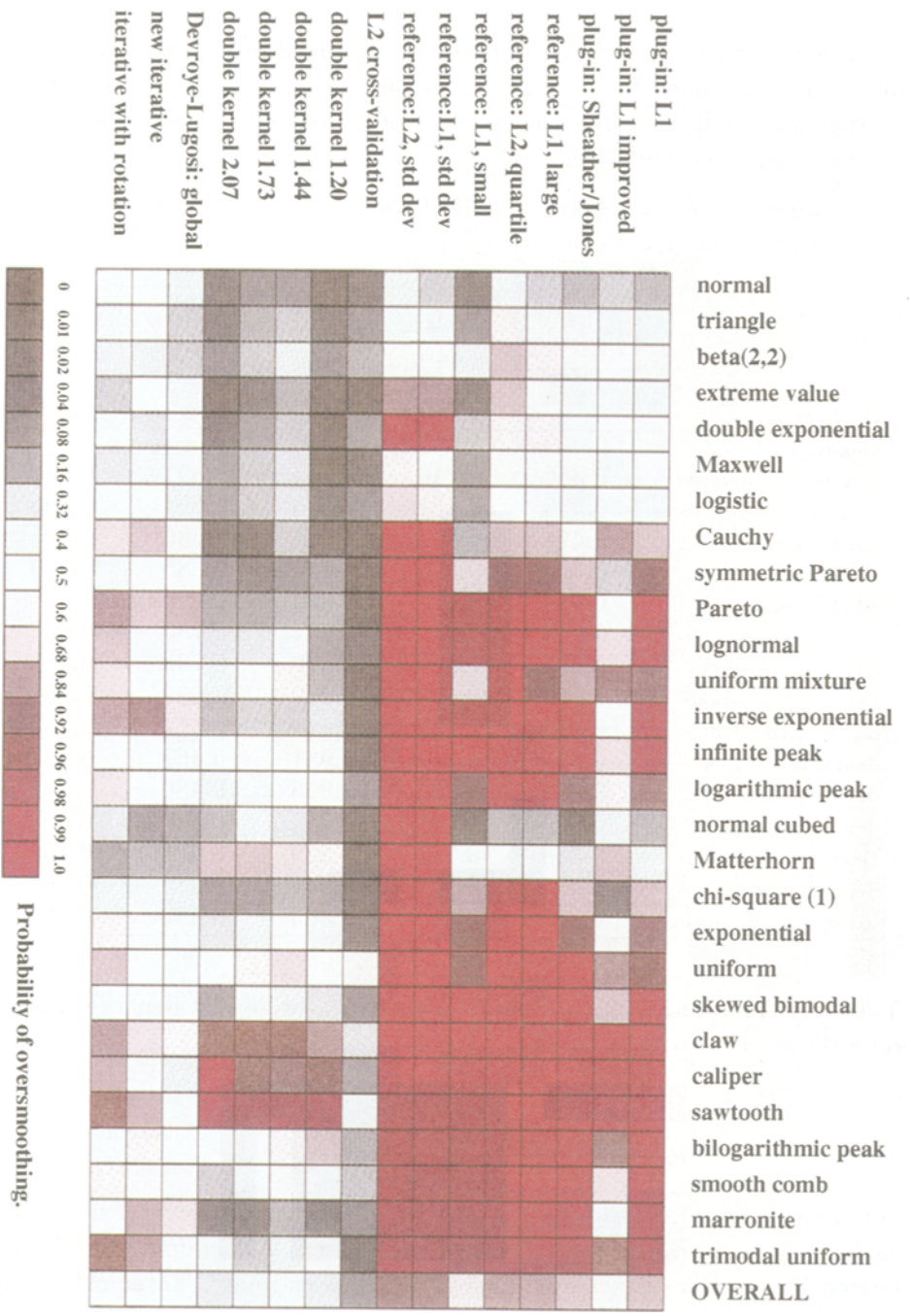

Figure 10: The probability of oversmoothing.

Table 1 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%.

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

	average error	average relative error	worst relative error	probability of relative error > 10%	pr bability of relative error > 50%
plug-in: L_1	0.429	0.345	0.784	0.583	0.185
plug-in: L_1 improved	0.366	0.146	0.613	0.391	0.075
plug-in: Sheather-Jones	0.448	0.421	0.935	0.594	0.196
ref: L_1, large	0.491	0.563	1.137	0.680	0.275
ref: L_2, quartile	0.495	0.576	1.157	0.685	0.280
ref: L_1, small	0.459	0.472	1.069	0.610	0.216
ref: L_1, std. dev.	0.643	0.843	1.501	0.789	0.446
ref: L_2, std. dev.	0.647	0.859	1.526	0.798	0.453
L_2, cross validation	0.480	0.472	1.496	0.639	0.298
double kernel 1.20	0.375	0.237	0.811	0.467	0.164
double kernel 1.44	0.379	0.220	0.931	0.467	0.137
double kernel 1.73	0.386	0.239	0.963	0.480	0.137
double kernel 2.07	0.404	0.301	1.055	0.525	0.166
Devroye-lugosi: global	0.388	0.303	1.229	0.541	0.166
new iterative	0.385	0.249	1.229	0.483	0.130
iterative with rotation	0.388	0.217	0.863	0.480	0.105

Table 1: 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 $h_{\mathrm{DH,L_1}}$, $h_{\mathrm{DH,L_2}}$) 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 $h_{\mathrm{ref,L_1}}$, $h_{\mathrm{ref,l_1}}$, $h_{\mathrm{ref,L_2}}$, $h_{\mathrm{DH,L_1}}$, $h_{\mathrm{DH,L_2}}$ 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. $h_{\mathrm{pi,l_1}}$ and $h_{\mathrm{pi,L_2}}$ are not immune. Among the tested bandwidths, only $h_{\mathrm{dk},1}$, $h_{\mathrm{dk},2}$, $h_{\mathrm{dk},3}$, $h_{\mathrm{dk},4}$, $h_{\mathrm{pi,L_1}}$, 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 $h_{\mathrm{DH,L_1}}$ or $h_{\mathrm{DH,L_2}}$). 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 $h_{\mathrm{pi,L_1}}$, 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 $h_{\mathrm{pi},\mathrm{l}_1}$, $h_{\mathrm{pi},\mathrm{L}_1}$, 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 (1997), 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 (1997). Then, using the inequalities of Theorem 4.1 and its second corollary, 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.

References

Akaike H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, **6**, 127-132.

Bartlett M.S. (1963). Statistical estimation of density functions. *Sankhya Series A*, **25**, 245-254.

Bean S.J. and C.P. Tsokos (1980). Developments in nonparametric density estimation. *International Statistical Review*, **48**, 267-287.

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

Bosq D. and J.P. Lecoutre (1987). *Théorie de l' estimation fonctionnelle*, Economica, Paris.

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

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

Bretagnolle J. and C. Huber (1979). Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **47**, 119-137.

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

Burman P. (1985). A data dependent approach to density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **69**, 609-628.

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

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

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

Chiu S.T. (1991). Bandwidth selection for kernel density estimation. *Annals of Statistics*, **19**, 1883-1905.

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

Cline D.B.H. (1988). Admissible kernel estimators of a multivariate density. *Annals of Statistics*, **16**, 1421-1427.

Cline D.B.H. (1990). Optimal kernel estimation of densities. *Annals of the Institute of Statistical Mathematics*, **42**, 287-303.

Deheuvels P. (1977a). Estimation non paramétrique de la densité par histogrammes generalisés. *Revue de Statistique Appliquée*, **25**, 5-42.

Deheuvels P. (1977b). Estimation nonparamétrique de la densité par histogrammes generalisés. *Publications de l'Institut de Statistique de l'Université de Paris*, **22**, 1-23.

Deheuvels P. and P. Hominal (1980). Estimation automatique de la densité. *Revue de Statistique Appliquée*, **28**, 25-55.

Devroye L. (1983). The equivalence of weak, strong and complete convergence in L1 for kernel density estimates. *Annals of Statistics*, **11**, 896-904.

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

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

Devroye L. (1988). Asymptotic performance bounds for the kernel estimate. *Annals of Statistics*, **16**, 1162-1179.

Devroye L. (1988a). The kernel estimate is relatively stable. *Probability Theory and Related Fields*, **77**, 521-536.

Devroye L. (1988b). Asymptotic performance bounds for the kernel estimate. *Annals of Statistics*, **16**, 1162-1179.

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

Devroye L. (1989b). The double kernel method in density estimation. *Annales de l'Institut Henri Poincaré*, **25**, 533-580.

Devroye L. (1989c). A universal lower bound for the kernel estimate. *Statistics and Probability Letters*, **8**, 419-423.

Devroye L. (1989d). On the non-consistency of the L2 cross-validated kernel density estimate. *Statistics and Probability Letters*, **8**, 425-433.

Devroye L. (1991). Exponential inequalities in nonparametric estimation. In: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, 31-44, NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Devroye L. (1992). A note on the usefulness of superkernels in density estimation. *Annals of Statistics*, **20**, 2037-2056.

Devroye L. (1994). On the non-consistency of an estimate of Chiu. *Statistics and Probability Letters*, **20**, 183-188.

Devroye L. (1994). On good deterministic smoothing sequences for kernel density estimates. *Annals of Statistics*, **22**, 886-889.

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

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

Devroye L. and G. Lugosi (1996). A universally acceptable smoothing factor for kernel density estimation. *Annals of Statistics*, **24**, 2499-2512.

Devroye L. and G. Lugosi (1997). Nonasymptotic universal smoothing factors, kernel complexity, and Yatracos classes. *Annals of Statistics*, **25**, to appear.

Devroye L. and M. P. Wand (1993). On the effect of density shape on the performance of its kernel estimate. *Statistics*, **24**, 215-233.

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

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

Eggermont P.P.B. and V.N. LaRiccia (1996). A simple and effective bandwidth selector for kernel density estimation. *Scandinavian Journal of Statistics*, **23**, 285-301.

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

Epanechnikov V.A. (1969). Nonparametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, **14**, 153-158.

Falk M. (1992). Bootstrap optimal bandwidth selection for kernel density estimation. *Journal of Statistical Planning and Inference*, **30**, 13-22.

Fan J. and J. S. Marron (1992). Best possible constant for bandwidth selection. *Annals of Statistics*, **20**, 2057-2070.

Faraway J. and M. Jhun (1990). Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association*, **85**, 1119-1122.

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

Habbema J.D.F., J. Hermans and K. Vandenbroek (1974). A stepwise discriminant analysis program using density estimation. In: COMPSTAT 1974, ed. G. Bruckmann, 101-110, Physica Verlag, Wien.

Hall P. (1982). Cross-validation in density estimation. *Biometrika*, **69**, 383-390.

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

Hall P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. in: *Multivariate Analysis VI*, ed. P. R. Krishnaiah,

289-309, North-Holland, Amsterdam.

Hall P. (1989). On convergence rates in nonparametric problems. *International Statistical Review*, **57**, 45-58.

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

Hall P., T.J. DiCiccio and J.P. Romano (1989). On smoothing and the bootstrap. *Annals of Statistics*, **17**, 692-704.

Hall P. and I.M. Johnstone (1992). Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society*, **B54**, 475-530.

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

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

Hall P. and J.S. Marron (1990). Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields*, **90**, 149-173.

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

Hall P., J.S. Marron and B. Park (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1-20.

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

Hall P. and M.P. Wand (1988). Minimizing L_1 distance in nonparametric density estimation. *Journal of Multivariate Analysis*, **26**, 59-88.

Hoeffding W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13-30.

Izenman A.J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, **86**, 205-225.

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

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

Jones M.C. (1991a). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters*, **12**, 51-56.

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

Jones M.C. and R.F. Kappenman (1992). On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, **19**, 337-349.

Jones M.C., J.S. Marron and B.U. Park (1991). A simple root n bandwidth selector. *Annals of Statistics*, **19**, 1919-1932.

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

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

Kim W.C., B.U. Park and J.S. Marron (1994). Asymptotically best bandwidth selectors in kernel density estimation. *Statistics and Probability Letters*, **19**, 119-127.

Mammen E. (1990). A short note on optimal bandwidth selection for kernel estimators. *Statistics and Probability Letters*, **9**, 23-25.

Marron J.S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Annals of Statistics*, **13**, 1011-1023.

Marron J.S. (1986). Will the art of smoothing ever become a science? *Contemporary Mathematics*, **59**, 169-178.

Marron J.S. (1987). A comparison of cross-validation techniques in density estimation. *Annals of Statistics*, **15**, 152-162.

Marron J.S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Economics*, **13**, 187-208.

Marron J.S. (1989). Comments on a data based bandwidth selector. *Computational Statistics and Data Analysis*, **8**, 155-170.

Marron J.S. (1991). Root n bandwidth selection. In: *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, 251-260, NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Marron J.S. (1992). Bootstrap bandwidth selection. In: *Exploring the Limits of the Bootstrap*, ed. R. LePage and L. Billard, 249-262, John Wiley, New York.

Marron J.S. and M.P. Wand (1992). Exact mean integrated square error. *Annals of Statistics*, **20**, 712-736.

McDiarmid C. (1989). On the method of bounded differences. In: *Surveys in Combinatorics 1989*, **141**, 148-188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge.

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

Müller H.G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics*, **12**, 766-774.

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

Park B.U. (1989). On the plug-in bandwidth selectors in kernel density estimation. *Journal of the Korean Statistical Society*, **18**, 107-117.

Park B.U. and J.S. Marron (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66-72.

Park B.U. and J.S. Marron (1992). On the use of pilot estimators in bandwidth selection. *Journal of Nonparametric Statistics*, **1**, 231-240.

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

Parzen E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

Ranneby B. (1984). The maximum spacings method: an estimation method. *Scandinavian Journal of Statistics*, **11**, 93-112.

Roeder K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617-624.

Rosenblatt M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.

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

Rudemo M. (1982) Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.

Schucany W.R. and J.P. Sommers (1977). Improvement of kernel type density estimators. *Journal of the American Statistical Association*, **72**, 420-423.

Schuster E.F. and G.G. Gregory (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. in: *Computer Science and Statistics:*

Proceedings of the 13th Symposium n the Interface, ed. W. F. Eddy, 295-298, Springer-Verlag, New York.

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

Scott D.W., Tapia R.A. and J.R. Thompson (1977). Kernel density estimation revisited. *Nonlinear Analysis*, **1**, 339-372.

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

Sheather S.J. (1993). The performance of six popular bandwidth selection methods on some real data sets (with discussion). *Computational Statistics*, **11**, 180-215.

Sheather S.J. and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density stimation. *Journal of the Royal Statistical Society*, **B53**, 683-60.

Stone C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, 1285-1297.

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

Stute W. (1992). Modified cross-validation in density estimation. *Journal of Statistical Planning and Inference*, **30**, 293-305.

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

Terrell G.R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, **85**, 470-477.

Terrell G.R. and D.W. Scott (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association*, **80**, 209-214.

Titterington D.M. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review*, **53**, 141-170.

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

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

van Es B. (1989). Likelihood cross-validation bandwidth selection for nonpara-

metric kernel density estimators. Technical Report 89-10, Faculty of Technical Mathematics and Informatics, University of Delft, The Netherlands.

Vapnik V.N. and A.Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their robabilities. *Theory of Probability and its Applications*, **16**, 264-280.

Wand M.P. and L. Devroye (1993). How easy is a given density to estimate? *Computational Statistics and Data Analysis*, **16**, 311-323.

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

Wand M.P. and W.R. Schucany (1990). Gaussian-based kernels. *Canadian Journal of Statistics*, **18**, 197-204.

Watson G.S. and M.R. Leadbetter (1963). On the estimation of the probability density. *Annals of Mathematical Statistics*, **34**, 480-491.

Woodroofe M. (1970). On choosing a delta sequence. *Annals of Mathematical Statistics*, **41**, 1665-1671.

Yatracos Y.G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, **13**, 768-774.

Young G.A. (1990). Alternative smoothed bootstraps. *Journal of the Royal Statistical Society*, **B52**, 477-484.

DISCUSSION

J. Beirlant
Katholieke Universiteit Leuven, Belgium

Luc Devroye is to be congratulated for his review of the state of the *art* on the illustrous problem of selecting the smoothing factor in nonparametric density estimation. I enjoyed the graphical procedure presenting the results from the simulation experiments. Figures 5-9 provide an efficient way to reduce the information and to show important trends.

The author has spent quite some of his research time focusing on density estimation in the true nonparametric sense: putting no restrictions on the densities. As such, from the theoretical point of view, the results presented

in Devroye and Lugosi (1996A and B) are really remarkable. The practical implementations of the corresponding bandwidths are quite involved and time-consuming, which constitutes an intrinsic drawback. However, the simulation results show that these new techniques have important potential in practical analysis, next to e.g. the popular plug-in methods, certainly in case of small and moderate sample sizes.

Next to the conclusions stated by the author as a consequence of the simulation study, I would like to add a few concerning the Devroye-Lugosi algorithms:

- they seem to have some trouble in case of densities with important discontinuities of the first kind at the border of the support. Their performance in case of the uniform and Matterhorn densities seems to indicate this.

- they perform especially well in case of smooth subexponential distributions, satisfying

$$\lim_{x \to \infty} \exp(\lambda x) f(x) = \infty$$

for any $\lambda > 0$.

This last observation is really intriguing to me. Any explanation is welcome here. These remarks make me conjecture that the choice of a bandwidth selection method could be combined with the estimation of the extreme value index of the underlying distribution at specific points, e.g. at infinity. For an expose on the latter see e.g. Beirlant *et al.* (1996). More study is needed here however.

Let me end with two comments. In view of my earlier comment on the performance of selectors for heavy tailed distributions, I wished the author had incorporated the transformation technique introduced by Wand *et al.* (1991). These authors proposed to combine any suitable bandwidth selection method for a normal distribution with the transformation to normality induced by the normal quantile plot. Finally, Devroye's paper makes it clear once more that enormous challenges lie ahead, not the least in gaining some theoretical insight in the global performance of the different selection methods.

References

Beirlant, J., P. Vynckier and J.L. Teugels (1996). Excess functions and estimation of the extreme value index. *Bernoulli*, **2**, 293-318.

Wand, M.P., J.S. Marron and D. Ruppert (1991). Transformations in density estimation. *Journal of the American Statistical Association*, **86**, 343-353.

R. Cao
Universidad de la Coruña, Spain

In this paper Luc Devroye gives an excellent survey of L_1-oriented methods for bandwidth selection, having in mind the universal suitability of the selectors. Two different versions of a new selector presented in Devroye and Lugosi (1996a) are developed and compared, via a simulation study, with many other existing competitors (even some L_2-oriented methods!).

The main idea behind the new selector is to split the sample into two parts and then select the smoothing parameter in order to minimize some total variation-like quantity:

$$\sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,h} - \mu_m(A) \right|, \tag{1}$$

where the class $\mathcal{A}$ is a strict subclass of the whole Borel class.

In my view, the richness of this class (that in the paper was fixed to be the Yatracos class) is a kind of new smoothing parameter, whose effect is also very important. In other terms, a kind of "pilot smoothing class". If $\mathcal{A}$ is too rich -let us assume, for instance, that it is the Borel class- the supremum would be equal to 2, since μ_m is atomic, and every possible bandwidth would be equally recommendable.. On the opposite side, for very poor choices of $\mathcal{A}$, such as a single orthant set, the method would be lead by the L_1 error of the cumulative distribution function estimation at a single point.

If we replace μ_m with μ, $f_{n-m,h}$ with $f_{n,h}$ and the Yatracos class with the Borel class in (1), we get one half of the L_1 distance between the kernel

estimator, $f_{n,h}$, and the underlying density, f. The relation between the L_1 distance and the total variation distance is a key idea in the definition of the new selector. I wonder if the idea behind this new selector can be still extended to an L_2 bandwidth selection setting.

When some preliminary information is assumed for f, one could incorporate it in the empirical measure, μ_m, in (1). For instance, if we happen to know that the underlying density is symmetric around the origin, one could reflect the last m data points and use the empirical measure pertaining to this reflected sample:

$$\mu'_m(A) = \frac{1}{2m} \sum_{i=n-m+1}^{n} (I_A(X_i) + I_A(-X_i))$$

instead of the μ_m in expression (1). When the preliminary information reduces the class of densities to a parametric family, μ_m may be replaced by the closest measure in the class (performing then a minimization in two "parameters": h and μ) and $f_{n-m,h}$ with $f_{n,h}$. This leads to L_1 minimum distance kernel estimation.

I also would like to comment a couple of things, concerning other aspects of the paper. The connection between the double kernel method and the bootstrap method is not very clear to me. As stated by Luc, the choice $L = 2K - K * K$ leads to the selection of the bandwidth that minimizes in h

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

The minimizer of this quantity is a degenerate bandwidth $H = 0$!

It should be mentioned that also the bootstrap bandwidth selectors are close related to expression (8.7) in the paper. In fact, most of these methods can be expressed as minimizers of

$$\frac{\int K^2}{nh} + \frac{1}{n^2} \sum_{i \neq j} M_h(X_i - X_j) + \frac{1}{n^3} \sum_{i \neq j} N_h(X_i - X_j),$$

for some functions M and N.

R. Fraiman
Universidad de la República, Uruguay

I would like to congratulate Professor Luc Devroye for this seminal paper. He analyzes two "universally suitable bandwidths", introduced in recent work by himself and Lugosi, solving one of the most compelling problems in density estimation. These are the first published smoothing factors that have been proved to have the property of been universally suitable, i.e. they found universal results for global smoothing factors selection for the first time. Moreover, they also obtained non asymptotic results for one of those proposals.

In this paper we found the result of several years of reflection about the problem of finding a method to select a bandwidth for kernel density estimation such that, for all densities in all dimensions, the L_1 error of the corresponding kernel estimate can be bounded in terms of the error of the estimate with the optimal smoothing factor. The problem of implementation of the proposed methods is also solved and a huge simulation to compare the performance of different bandwidth selection methods is included.

As usual in Devroye's work, I find here a lot of stimulating ideas. The way they define the bandwidths is somewhat related to the double kernel method (Devroye, 1989) but now looking at the distance to the empirical distribution over a "specially picked rich enough class".

Some nice new tools, like the use of Vapnik and Chervonenkis (1971) inequalities over Yatracos classes of sets, seem to be the key for these new results, and open the possibility to extend them to other nonparametric problems. Indeed, as an example, the extension of the results to nearest neighbor (and nearest neighbor with kernel) density estimates seems to be possible as follows.

Let us define

$$H_{n,k} = ||X^{(k)} - x||$$

be the distance from x to its k-nearest neighbor among $X_1, \ldots, X_{n-m}$, and

$$f_{n-m,k}(x) = \frac{k}{nH_{n,k}^d \lambda_1} = \frac{1}{\lambda_1 = nH_{n,k}^d} \sum_{i=1}^{n-m} I_{B(x,H_{n,k})}(X_i)$$

where λ_1 stands for the volume of the unit ball in R^k, I_A stands for the indicator function of the set A and $B(x,t)$ for the closed ball centered at x and radius t. The Yatracos class $\mathcal{A}$ for this problem would be the collection of sets

$$A_{k,k'} = \{x \in R^d : \frac{k}{H_{n,k}^d} \geq \frac{k'}{H_{n,k'}^d}\}$$

for k, k' nonnegative integers, while the smoothing factor K can be defined as the random value for which the supremum

$$sup_{A \in \mathcal{A}} | \int_A f_{n-m,k}(t)dt - \mu_m(A)|$$

is minimal, μ_m being the empirical measure based on the sample X_{n-m+1}, $\ldots, X_n$.

Once we know that it is possible to find universally suitable bandwidths for kernel based density estimates, one could try to solve the same problem for other nonparametric estimates. Devroye-Lugosi method seems to be a powerful tool in order to find universally suitable smoothing factors for histograms, Fourier-series density estimates and for nonparametric regression problems.

References

Devroye, L. (1989). The double kernel method in density estimation. *Annales de l'Institut Henri Poincaré*, **25**, 533-580.

Vapnik, V. and A. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**, 264-280.

P. Hall
The Australian National University, Australia

The literature contains quite a few comparative studies of bandwidth choice for density estimation, but none that approaches the breadth of this remarkable paper. Devroye's work blends deep theoretical insight with detailed

numerical analysis, for a range of densities and bandwidth selectors that is so vast that it takes us well beyond the sort of problems that most of us have in mind when we study methods for smoothing papameter choice.

The very breadth of Devroye's view, in terms of both target density and bandwidth choice method, argue in favour of a revised approach to bandwidth choice. As Devroye notes, no method works well across the spectrum of different targets. simpler, less variable bandwidth selectors are good performers for simple densities, but lack the adaptivity to deal with complex cases such as multimodal targets.

It is perhaps possible that good local bandwidth choice methods will produce respectable performance in complex cases. However, high complexity usually means that the target density changes fast in a local sense, and so local bandwidth selectors would often have only a small amount of information on which to base a local bandwidth function. This would inevitably lead to high variability, and most likely to poor performance. Perhaps more promising is a general method for global bandwidth choice which, before taking the plunge and computing the bandwidth, assessed the complexity of the target density.

Imagine that a software package first estimated a measure of the density's complexity, and then effectively "looked up a table" to find the bandwidth selector that was most appropriate for the level of complexity. In general, complexity could be at least a vector-valued quantity, but more siply, standar scalar measures suggest themselves.

Thus, while in Section 8.7 Devroye is rightly critical of so-called "entropy-based" methods for bandwidth choice, it is possible that a statistical estimate of entropy could be a valuable predictor of the relative performance of different bandwidth selectors. The densities in Devroye's tables are very roughly ordered in terms of increasing entropy, with the high-entropy densities generally doing better with double-kernel or Devroye-Lugosi methods than simpler bandwidth selectors. Therefore, Figures 6 to 10 tend to be lighter in top right-hand and bottom left-hand corners than elsewhere. Perhaps there would be more "light" in the table if entropy were estimated first, and bandwidth selectors assigned to data sets according to the estimated degree of complexity.

M.C. Jones
The Open University, United Kingdom

It is good to see the recent work of Luc Devroye on bandwidth selection for kernel density estimation gathered together in an interesting single source, and a pleasure to have the opportunity to comment upon it. The paper has three distinct phases. The first is the 'universal' theory which displays the author's superb mathematical virtuosity, and which I can only admire. The second is the suggestion of practical bandwidth selectors, those developed with Lugosi along with the double kernel methods. Most of these results and ideas have been published before, but it is very helpful to have them reviewed in one place. The third phase of the paper is a large, thorough and important simulation study.

To take the third phase first, the simulation results for individual densities are wide–ranging and very informative. As the author is well aware, any attempt to summarise the results in terms of single overall numbers (here by equally weighted averaging over the set of 28 chosen test densities) is open to criticism, and criticise it I shall. For example, in Figure 5, I suspect that I should be more concerned about the relatively poor performance of the lower block of methods for some of the easier–to–estimate densities than about the apparently even worse performance of the upper block of methods for the more eccentric of the multimodal densities. Can the author reassure me (by showing typical actual density estimates) that performance on the easier estimates is still acceptably good for the lower block of methods? The relatively good performance of the same methods for the eccentric multimodals may well be illusory; are the results of these methods acceptable or are they, as I suspect, just the best of a very bad lot?

In fact, Professor Devroye has rightly pointed out that for some of the latter densities, there is just not enough information in a sample of size 100 to be able to estimate them well nonparametrically at all. For many other densities, it would seem also that a basic kernel density estimate, however well the bandwidth is chosen, will provide quite poor density estimates, for example due to long tails, and more sophisticated methods might be considered (although I have my doubts about the practical usefulness of many of the 'improved bias' estimates, see Jones & Signorini, 1997). It seems that all 28 densities were considered as densities on the whole real

line, yet surely many of those with bounded support are usually used as models for situations where boundaries are known (e.g. values on a known interval, or nonnegative values). In such cases, it seems more realistic to treat them not as 'discontinuous densities' but as densities on a known finite support, assessing them only over that support and using appropriate forms of boundary correction in the estimates themselves.

I think some insight can be given into the interesting double kernel method of Professor Devroye by looking at the obvious L_2 modification of it. In Jones (1997), I show that when $h' = h$, the L_2 double kernel method provides (essentially) a raft of methods of the form (7) in the paper. Indeed, in the case $L = 2K - K * K$, the L_2 double kernel method is precisely Taylor's smoothed bootstrap approach. Intriguing things happen in the L_2 version when $h' >> h$. We do not get the improved rate of convergence of $\hat{h}$ to h_0, the minimiser of the *mean* integrated squared error (MISE), that we have come to expect from using the same trick in the plug–in literature. But we do seem, potentially, to get a particularly good estimate of $\hat{h}_0$, the minimiser of ISE (within the well known limitations of the problem).

The author follows Cao et al (1994) in utilising a modified, and inferior, form of Sheather/Jones bandwidth selector. But in Sheather & Jones (1991), we recommended the choice of two kernel estimation steps (not one) before employing a normal reference density, and of solving the defining equation for the optimal h and not directly plugging–in (admittedly losing the simplicity of an explicit formula). These were not arbitrary choices but the results of considerable practical work in which such choices were found to be much the better. (In fact, theory also dictates both that two stages are necessary, Sheather & Jones, 1991, and that solve–the–equation is better, Park, 1989.)

I was intrigued by the statement in Section 11 that "A density is only a tool for computing probabilities. Hence good bandwidth design should be based on probabilities." I suggest that the density also has major roles in envisaging and interpreting the properties of a distribution and in comparing different distributions. But if it really is probabilities that are wanted, why is it not the distribution function F itself, rather than the density f, that is the primary target? The L_2 theory clearly demonstrates that distribution function estimation is a much easier smoothing problem, that bandwidths need be of a considerably smaller size (often zero), and that — in my interpretation of recent work on bandwidth selection for estimation

of F — any reasonable method of bandwidth selection tailored to F estimation will perform just about as well. Questions truly about probabilities only do not require this paper's paraphernalia.

The Epanechnikov kernel is a reasonable choice for simulation studies because of computational ease, but I would not wish its use to be encouraged in practice. This is because the lack of continuous differentiability of the kernel is clearly visible as kinks in the estimated function. A good illustration of this can be found in Figure 7 of Keiding (1991). By the way, while the Epanechnikov kernel does appear in Bartlett (1963), I find it hard to interpret the latter paper as recognising the standard optimality property of the kernel. The paper referred to as Jones (1990) has since metamorphosed into Jones & Foster (1993).

References

Jones, M.C. (1997). On some kernel density estimation bandwidth selectors related to the double kernel method. Submitted.

Jones, M.C. and P.J. Foster (1993). Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, **3**, 81–94.

Jones, M.C. and D.F. Signorini (1997). A comparison of higher order bias kernel density estimators. *Journal of the American Statistical Association*, to appear.

Keiding, N. (1991). Age–specific incidence and prevalence: a statistical perspective (with discussion). *Journal of the Royal Statistiacl Society. Ser. A*, **154**, 371–412.

Park, B.U. (1989). On the plug–in bandwidth selectors in kernel density estimation. *Journal of the Korean Statistical Society*, **18**, 107–17.

Gábor Lugosi
Universitat Pompeu Fabra, Spain

It was a pleasure to read this great survey on kernel smoothing in density estimation.

For me the greatest surprise was that the new methods proposed in Devroye and Lugosi (1996a,b) and their variants introduced in this paper,

even in their raw form, perform comparably well in the simulations with the fine-tuned bandwidths matured over the years of the long history of the field. I suspected that—even though finite-sample performance guarantees exist for these estimates—their advantages start to show at significantly larger sample sizes than 100, for which the simulations were made.

Also, the simulation results provide empirical evidence that the factor of 3 appearing in the asymptotic performance bounds is an artifact of the analysis, and it is not an inherent property of the new methods, since for very small sample sizes even the best smoothing factor gives a large L_1 error, and a factor of 3 would push the error up to meaninglessly large levels. Another promising sign is that the new methods do not have a tendency for under or over smoothing, so their error may be credited to the variability resulting from the small sample size.

The paper concentrates on choosing the smoothing factor h once the kernel K has been fixed, and it is not concerned with the choice of K. However, the new methods may be adapted in a straightforward way to the data-dependent choice of a kernel. Consider the following situation: let $\mathcal{C} = \{K_\theta : \theta \in \Theta\}$ be a class of kernels on $\mathcal{R}^d$, where Θ is some set of parameters and $\int K_\theta = 1$ for each $\theta \in \Theta$. The class $\mathcal{C}$ may possibly vary with the sample size n. Each kernel in the class defines a kernel estimate

$$f_{n,\theta}(x) = \frac{1}{n}\sum_{i=1}^{n} K_\theta(x - X_i).$$

The goal is to select a parameter θ_n based on the available data such that the L_1 error $\int |f_{n,\theta_n} - f|$ is close to the optimal L_1 error in the class $\inf_{\theta\in\Theta} \int |f_{n,\theta} - f|$. In the special case when $\mathcal{C}$ is the "one-dimensional" class containing $K_h(x) = (1/h^d)K(x/h)$, $h > 0$ for some fixed kernel K, we are back in the situation of selecting a smoothing factor, as discussed in the paper in great detail. The estimate, called "the first bandwidth" in the paper, may be adapted to this more general situation in a straightforward way as follows. Choose an $n^{-1/2}$-covering of the class $\mathcal{C}$ of kernels, that is, let $\widehat{\mathcal{C}}$ be a finite set of functions of cardinality N such that for each $K_\theta \in \mathcal{C}$ there exists an $L_j \in \widehat{\mathcal{C}}$ (for some $j \le N$) such that $\int |L_j - K_\theta| \le 1/\sqrt{n}$. (We implicitly assume here that such a finite covering is possible, otherwise the estimate is undefined.) Then split the data into two parts, and based on the first (larger) part form N kernel estimates using the kernels $L_1, \ldots, L_N$

in the class $\widehat{\mathcal{C}}$:

$$f_{n-m,j}(x) = \frac{1}{n-m} \sum_{i=1}^{n-m} L_j(x - X_i), \qquad j = 1, \ldots, N.$$

Among these N estimates we choose the one which minimizes

$$\sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,j}(x) - \mu_m(A) \right|, \qquad j = 1, \ldots, N,$$

where μ_m is the standard empirical measure defined on the second (small) part of the data $X_{n-m+1}, \ldots, X_n$ and $\mathcal{A}$ is the class of sets containing all sets of the form

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

Let f_n denote the obtained kernel estimate.

Now it is easy to see that most of the analysis of Devroye and Lugosi (1996a) can be repeated for this more general situation. In particular with small straightforward modifications one can prove the following:

Theorem 1. *Assume that the covering numbers satisfy* $N = O(n^a)$ *for some* $a < \infty$ *and that for some* $b < 1/2$ *and* $c > 0$,

$$\inf_f \liminf_{n \to \infty} n^b \inf_{\theta \in \Theta} \mathbf{E} \int |f_{n,\theta} - f| \geq c.$$

Then if $m/n \to 0$ *and* $m/(n^{2b} \log n) \to \infty$, *then*

$$\sup_f \limsup_{n \to \infty} \frac{\mathbf{E} \int |f_n - f|}{\inf_{\theta \in \Theta} \mathbf{E} \int |f_{n,\theta} - f|} \leq 3. \tag{1}$$

Thus, the choice of the kernel is "universally suitable" within the class $\mathcal{C}$ of kernels if two properties are satisfied: (i) $\mathcal{C}$ should be "finite-dimensional" in the sense that the $1/\sqrt{n}$ covering numbers do not grow faster than a polynomial of n; (ii) the L_1 error of the best estimate in the class does not decrease too rapidly. This second condition assures that the denominator in (1) does not get smaller than the inevitable estimation error of order $n^{-1/2}$.

While the first condition is quite easy to check in many cases, checking the second condition may not be trivial for specific situations. Devroye (1988b) provides several such examples. Here we provide one simple case.

Example. Let K be a fixed elegant kernel on $\mathcal{R}$ (i.e., satisfying the Lipschitz condition $|K(x) - K(y)| \leq C|x - y|$ and vanishing outside $[-1, 1]$), and introduce the class $\mathcal{C}$ of kernels on $\mathcal{R}^d$ as the class of product kernels

$$K_\theta(x) = \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j}{h_j}\right), \qquad h_1, \ldots, h_d \in [e^{-n}, e^n],$$

where $\theta = (h_1, \ldots, h_d)$ is a vector of smoothing factors and $x_1, \ldots, x_d$ are the components of the vector $x \in \mathcal{R}^d$. Thus, here we are allowed to select a different smoothing factor in each coordinate direction. Just like in the single-smoothing-factor case, the restriction of the range of the h_j's is necessary to make sure that the class of kernels has finite covering numbers. Clearly, this restriction is insignificant in an asymptotic sense since the optimal smoothing factors eventually all fall in this interval.

To see why condition (i) is satisfied, consider the set of θ's whose coordinates are all of the form $e^{-n}(1+\delta_n)^k$ for some nonnegative integer k such that $e^{-n}(1 + \delta_n)^k \leq e^n$, where $\delta_n = 1/(2^d C^{d-1}(d + 1)\sqrt{n})$. The number of possible k's is at most

$$\frac{\log(b_n/a_n)}{\log(1 + \delta_n)} + 1 \leq \frac{4n}{\delta_n},$$

so the number N of different such θ's is at most $\left(\frac{4n}{\delta_n}\right)^d = O(n^{3d/2})$, a polynomial in n. To see that the kernels with such θ's indeed form an $n^{-1/2}$-covering of $\mathcal{C}$ note that if $\theta = (h_1, \ldots, h_d) \in \Theta$ is arbitrary and $\theta' = (h'_1, \ldots, h'_d)$ is the element in the finite subset such that h'_j is the smallest "discretized" value which is at least as large as h_j, $j = 1, \ldots, d$,

then

$$\begin{aligned}\int |K_\theta - K_{\theta'}| &= \int \cdots \int \left| \prod_{j=1}^{d} K(x_j) - \prod_{j=1}^{d} b_j K(b_j x_j) \right| dx_1 \ldots dx_d \\ &\quad \text{(where } b_j = h'_j/h_j\text{)} \\ &\leq \int \cdots \int \left| \prod_{j=1}^{d} K(x_j) - \prod_{j=1}^{d} K(b_j x_j) \right| dx_1 \ldots dx_d \\ &\quad + \int \cdots \int \left| \prod_{j=1}^{d} K(b_j x_j) - \prod_{j=1}^{d} b_j K(b_j x_j) \right| dx_1 \ldots dx_d.\end{aligned}$$

By applying the inequality $|ac - bd| \leq a|c - d| + c|a - b|$ and the fact that elegant kernels are bounded by C, it is easy to see that for each $x \in \mathcal{R}^d$, $\left|\prod_{j=1}^d K(x_j) - \prod_{j=1}^d K(b_j x_j)\right| \leq C^{d-1} \sum_{j=1}^d (b_j - 1)$, which implies that the first term on the right-hand side is at most $2^d C^{d-1} \sum_{j=1}^d (b_j - 1)$. On the other hand, the second term is easily seen to be bounded above by $\prod_{j=1}^d (b_j - 1)$, and therefore

$$\begin{aligned}\int |K_\theta - K_{\theta'}| &\leq 2^d C^{d-1} \sum_{j=1}^{d} (b_j - 1) + \prod_{j=1}^{d} (b_j - 1) \leq 2^d C^{d-1} d\delta_n + \delta_n^d \\ &\leq 2^d C^{d-1} (d+1)\delta_n = \frac{1}{\sqrt{n}},\end{aligned}$$

as desired.

To check condition (ii) just note that if g and $g_{n,\theta}$ denote the marginal densities of f and $f_{n,\theta}$ with respect to the first component of x, then by an inequality mentioned in the paper,

$$\mathbf{E} \int |f_{n,\theta} - f| \geq \mathbf{E} \int |g_{n,\theta} - g|.$$

Moreover, it is easy to see that $g_{n,\theta}$ is just the univariate kernel estimate of g with kernel K and smoothing factor h_1, so the cited result of Devroye and Penrod (1984) implies that

$$\inf_f \liminf_{n\to\infty} n^{2/5} \inf_{\theta\in\Theta} \mathbf{E} \int |f_{n,\theta} - f| \geq 0.86,$$

and therefore the conditions of the theorem are satisfied with $a = 3d/2$, $b = 2/5$, and $c = 0.86$.

It seems probable that this is just one of many interesting examples, and the full potential of the new new methods is yet to be explored.

E. Mammen
Universität Heidelberg, Germany

I enjoyed reading this paper. It is very impressive to have a density estimate $\hat{f}_n$ whose L_1-risk is universally bounded for all densities. At first sight the construction of this estimate suggested that this result was mainly of theoretical interest. Now, it is nice to see in the simulations of this paper that this was not true. In fact, the idea is very applied and it leads to an estimate with a very good practical performance.

I have some remarks on the practical use of the proposed kernel density estimate $\hat{f}_n$. In particular, I would like to highlight that it is very important to develop some asymptotic distribution theory for $\hat{f}_n$.

Kernel density estimation offers a tool box for statistical inference on the shape of the underlying density f. Inference on many practically relevant questions can be based on the inspection of kernel density estimates. This broad area of applications makes kernel density estimation so important and attractive. Examples of applications are the following questions: Is f a normal density or does it belong to another parametric class? Is f skewed? Does f have heavy tails? How many modes does f have? How do the densities of several samples compare?

A first mathematical check on the statistical performance of a density estimate is to study its asymptotic or finite sample risk. In this paper and in related work of L. Devroye and G. Lugosi this has been done for the L_1-risk of $\hat{f}_n$. However a good risk performance does not guarantee that a density estimate is suitable for a specific inference problem. For a rigorous treatment of a statistical inference problem, visual inspection of the plot of the density estimate does not suffice and for a more reliable approach one should consider test statistics (or other statistical procedures, e.g. confidence intervals or bands) based on the density estimate. Judgement on the

observed value of a test statistic requires some (asymptotic) distribution theory on $\hat{f}_n$.

There is another motivation to look at the distribution of $\hat{f}_n$. It is given by the simulations of this paper. Here, asymptotic knowledge would help to understand the results of the simulations. For instance, in the simulations the estimate $\hat{f}_n$ was outperformed by plug-in estimates for smooth unimodal densities. Intuitively, this would have been expected (as mentioned in the paper). However, more insight would be given by knowing the asymptotic distribution of $\hat{f}_n$ for smooth densities f.

For many kernel density estimates f_{nH_n} [with data adaptive bandwidth selector H_n] asymptotics is based on the following considerations. First it is shown that $H_n/h_n(f) = 1 + o_P(1)$ for a deterministic sequence $h_n(f)$ depending on the underlying density f. In general, this implies that the asymptotic behaviour of f_{nH_n} coincides with that of $f_{nh_n(f)}$. In particular, typically test statistics based on f_{nH_n} will have the same asymptotics as the test statistic where f_{nH_n} is replaced by $f_{nh_n(f)}$. So it is a very interesting research problem to study the following questions for the bandwidth selector H that has been proposed by L. Devroye and G. Lugosi

(1) Does it hold for the bandwidth selector H that $H/h_n(f) = 1 + o_P(1)$?

(2) How does $h_n(f)$ depend on f? [for densities f for which the first question can be answered positively].

(3) What is the speed of convergence of $H/h_n(f)$ to 1? [if it converges at all].

Study of the third question for other bandwidth selectors has been one of the major topics in the smoothing literature of the latest years. Clearly, fast convergence of $H/h_n(f)$ suggests that the approximation of f_{nH_n} by $f_{nh_n(f)}$ is accurate. On the other hand it has been argued that it implies more "stable" estimates f_{nH_n}.

Study of questions (1) - (3) for the bandwidth selector H of L. Devroye and G. Lugosi may be rather complicated. I conjecture that in particular the random nature of the family of sets $\mathcal{A}$ makes a mathematical analysis difficult. This may be the case even under restrictive assumptions on the density f. It is possible that modifications of H or related methods can be treated more easily. A promising related approach has been proposed

by Lepski (1990). He suggests to use the largest bandwidth H such that the absolute differences $|f_{nH} - f_{nh}|$ are not "significantly" large. For an implementation of this method for local bandwidth choice in a white noise model see Lepski, Mammen and Spokoiny (1997). There it has been shown that the relative pointwise risk of this estimate is uniformly bounded in $\mathcal{F}(L) = \{f : \sup_x f(x) - \inf_x f(x) \leq L\}$ for each $L > 0$. This result can be interpreted as a pointwise analogue of the results of L. Devroye and G. Lugosi.

References

Lepski, O.V. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, **35**, N. 3, 459-470.

Lepski, O.V., E. Mammen and V.G. Spokoiny (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics*, **25**, 929-947.

J. S. Marron
University of North Carolina, U.S.A.

Abstract

The author is to be congratulated on yet another deep and interesting paper on bandwidth selection. This discussion comes in two parts. The first is for "mass consumption", i.e. is mostly addressed to non-experts in bandwidth selection and tries to make clear my view of the big picture of this area, and how the approaches described in the present paper fit into that. The second is of a much more esoteric nature, and is intended for experts on bandwidth selection.

1 Big Picture Discussion

This is a nice paper with both the deep mathematics, and also the entertaining personal style, which characterize so many of this author's papers. It is also very good to see the author's continuing interest in how well his suggested methods actually work (beyond what is learned from the asymptotics), via a very detailed and informative simulation study.

Because I have not yet experienced religious conversion to the L^1 norm as a means of measuring error, the promising results indicated in the present paper, motivated me to test these methods myself, from other viewpoints.

A very important way to test is to actually look at density estimates, and I was pleased to see that the L^1 Improved Plug-In bandwidth, which performed very well in the present simulations, gave quite good performance. In my view, in this important respect, this method is in the same general range of effectiveness as a group of others that have been called "modern methods" in Jones, Marron and Sheather (1996). A number of other methods in this group have been discussed in the present paper, but note that this does not include cross-validation type methods, which I view as "unacceptably noisy".

However, as noted in the present paper, not all these methods are the same. My point is that in the "big picture sense for data analysis" I view these as close enough that differences between them are mostly esoteric in nature. In the next section I take a more microscopic look at a few of these methods, and show that these differences boil down to personal preference. In particular, I show that when the author's favorite L^1 norm is replaced by the "Visual Error Criterion" of Marron and Tsybakov (1995), then the Sheather Jones Plug In method becomes "superior". Hence the latter method remains my personal recommendation (and what I use first on a new data set). However I can understand how others would prefer to recommend other modern methods, and respect that choice.

While it is fun to debate exactly what is the "best bandwidth" (I heartily take this up in the next section), it is important to keep in mind the big picture: for practical use, most "modern" bandwidth selection methods, including the L^1 Improved Plug In, are quite useful.

2 Esoteric Details

This section discusses several fairly minor points, on some of which I take issue with the author.

2.1 Related kernel density estimates

In Section 1, the author gives a good overview of density estimation. Here are a few interesting new things it might be good to add to the list of things mentioned in the present paper.

1. Transformations combined with density estimation. This idea was made practical by Wand, Marron and Ruppert (1991), who developed a data based method for choosing among a parametric family of transformations. Yang (1997a) worked out the asymptotics of that approach, and also showed how it could be improved. Much larger families of transformations were shown to be useful in Yang and Marron (1997b). The Yang papers give several other references to work in this area.

2. Interesting unpublished work on location adaptive density estimation include the work of Farmen (1997), who shows that many of the common methods are often inferior to a well chosen constant bandwidth. The "zero bias" ideas of Sain and Scott (1996) are surprising and deep.

2.2 Choice of kernel function

The author continues his traditional use of flowery terminology, with a definition of "elegant" for a certain class of kernel functions in Section 2. I agree that "elegant" kernels should be nonnegative and at least Lipschitz continuous. However in my view a better use of that term would include much more smoothness, and allow infinite support, and in fact be a synonym for the Gaussian kernel.

In an upcoming paper by Chaudhuri and Marron, "elegance of kernels" is explored using some ideas from scale space theory in computer vision. In several senses, the Gaussian kernel is much more "natural and elegant", than any which satisfy the currently stated definition of "elegance". For example the Gaussian kernel is the only kernel for which "features", such as modes, in the density estimate monotonically diminish with increasing bandwidth. In addition, a set of "smoothing axioms" result in the family of smooths (indexed by the bandwidth) being a solution to the heat equation, which also essentially results in the Gaussian kernel. See Lindeberg (1994).

The Epanechnikov kernel has its advantages in some ways, but other kernels are quite competitive in other ways. A strong case can be made for the Gaussian kernel, especially on "elegance" grounds. Furthermore, the supposed computational advantages of compact support are negligible if a fast computational method such as binning is used. See, for example, Fan and Marron (1994).

2.3 Asymptotic approximations

Near the end of Section 8.1, the author points out that the asymptotically optimal bandwidth given in (1) is "often, but not always, close to the true optimal h". In fact, given any sample size n, it is straightforward to construct examples where these bandwidths can be arbitrarily far apart, as shown by Marron and Wand (1992). The idea discussed there (adapted here to the L^1 norm) is that there may not be enough information in the data so that $\int |f''|$ is practically relevant to the smoothing problem at hand. In particular, if f has features which cannot be discerned from the data (because the sample is too "small"), then $\int |f''|$ will be unrealistically large, resulting in (1) being far too small.

This point is illustrated in Figure 1, using the Double Claw target density from Marron and Wand (1992). The underlying density has small spikes that represent only two percent of the probability mass, and are thus indistinguishable from the $n = 100$ observations available. Yet the asymptotically optimal bandwidth formula (1) "feels" these spikes (because f has very strong curvature at those points), resulting in serious undersmoothing. The estimate resulting from the bandwidth (1) is interesting in that it seems to have spikes of the requested size, it is not reasonable in any other sense. The bandwidth that is optimal in the L^1 sense is much more reasonable for this pseudo data set. The poor recovery of the large left hand mode is quite common for any density estimate, because only $n = 100$ observations don't contain too much information about this density.

2.4 Why do density estimation?

In the second paragraph of Section 11, the author states "A density is only a tool for computing probabilities." I have a much different view on this point. First off, if one only cares about computing probabilities, then those

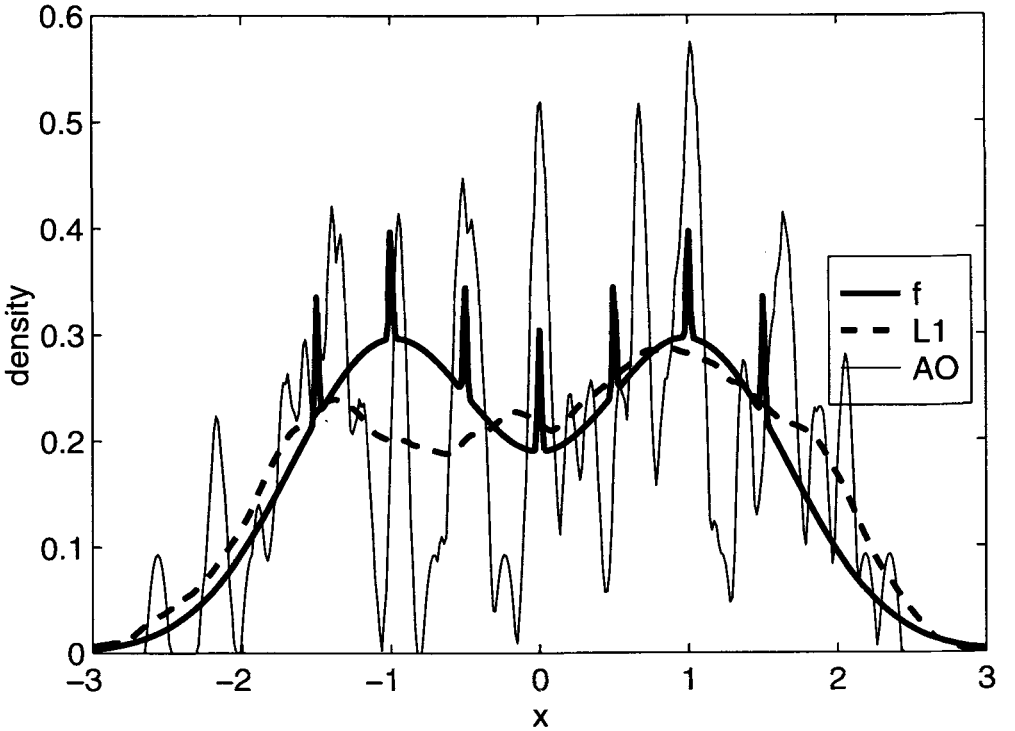

Figure 1: For the Double claw density, shown as the solid heavy curve, and a pseudo data set of size $n = 100$, this shows an Epanechnikov kernel estimator, with the L^1 optimal bandwidth as the heavy dashed curve, and the asymptotically optimal bandwidth as the thin solid curve. The asymptotically optimal bandwidth is grossly undersmoothed.

based on the empirical cumulative distribution function will typically be far more accurate than those optimally tuned for the L^1 norm. Even with a much different choice of bandwidth, there is no scope to do considerably better with a kernel estimator (following what has been called "deficiency theory").

But a much more important issue is: what is the motivation for studying density estimation? More mathematical researchers draw their motivation purely from the intellectual exercise, which is sensible because kernel density estimation really is fun. However, more statistically minded researchers draw motivation from the fact that kernel density estimation provides a powerful data analytic tool. In particular, it can allow the statistician to immediately discover features of the data that are much harder (if even possible) to find by other methods. A particularly compelling example of this type is the United Kingdom income data set, perhaps best analyzed in Schmitz and Marron (1992). See the monographs Silverman (1986), Scott (1992) and Wand and Jones (1995) for many more such examples.

2.5 Best bandwidth selector?

The author concludes that the L^1 Improved Plug In bandwidth is a very good choice, and I generally agree. However, I would not agree to suggestions that this is "best", for several reasons.

One place where different views can lead to different answers is in the design of the simulation study. Here the choice gets quite personal, and different motivations yield different preferences. Based on my personal motivation of data analysis experience, I find the present set of target densities to be "too unimodal". In particular, my personal interest is in finding features like bumps and modes, so I prefer more densities of this type, than the examples with poles and heavy tails that are predominantly studied here. The author may answer that my recommended bandwidth above fared the worst in such cases, but see below for another way of looking at this. However, this issue of target curves is not very important, unless one looks at summaries, as in Figure 10.

Another point which affects one's conclusions is the method that error is measured. This author and various co-authors have presented many interesting reasons as to why the L^1 norm is "most natural", especially in comparison to L^2. Many of these are quite interesting and compelling.

However, other methods of measuring errors have their relative advantages as well. For example, L^2 is enduring in the literature, doubtless because of its mathematical tractability (in fact many key ideas in density estimation are developed first in the simple L^2 context, and then after the idea is clear, the much harder L^1 version is developed, very often by this author). But recently several researchers have realized that none of the usual norms give the same bandwidth as "what one would choose by eye". See Marron and Tsybakov (1995) for further discussion, and development of an error measure which does work in this way. Marron (1997) applies this error measure to bandwidth selection.

This error measure is used in a few additional simulations, to investigate how much the conclusions can change, depending on the error measure. For simplicity, just the L^1 Improved Plug In, and the Sheather Jones Plug In, are compared, using average (over 500 pseudo data sets) L^1 norm, and the average of Marron and Tsybakov's VE, on the Smooth Comb target density, for $n = 100$. The results are summarized in Table 1 (these are not normalized as done in the major study, because only a single example is considered here). The results for the L^1 norm are consistent with those

	Avg. L^1 (CI rad.)	$10 *$ Avg VE (CI rad)
L^1 Improved Plug In	0.357 (0.003)	0.316(0.005)
Sheather Jones Plug In	0.371 (0.002)	0.266(0.003)

Table 1: Simulation results for $n = 100$, *Smooth Comb density. Average errors over pseudo data sets are reported with simple 95% Confidence Interval radii in parentheses to indicate Monte Carlo error.*

presented by the author, with the L^1 Improved Plug In appearing to be substantially better than Sheather Jones. However, when the error criterion shifts to VE, the ordering also shifts, and now Sheather Jones looks better. Insight into this apparent contradiction comes from Figure 2.

Figure 2a shows that these two bandwidth selection methods have rather different characteristics. For example the Sheather Jones Plug In is much more stable across different pseudo data sets, and there is a small proportion of data sets for which the L^1 Improved Plug In is surprisingly large. Figure 2c shows that the L^1 and VE performances are rather different as well. In particular, the Sheather Jones Plug In fares better with

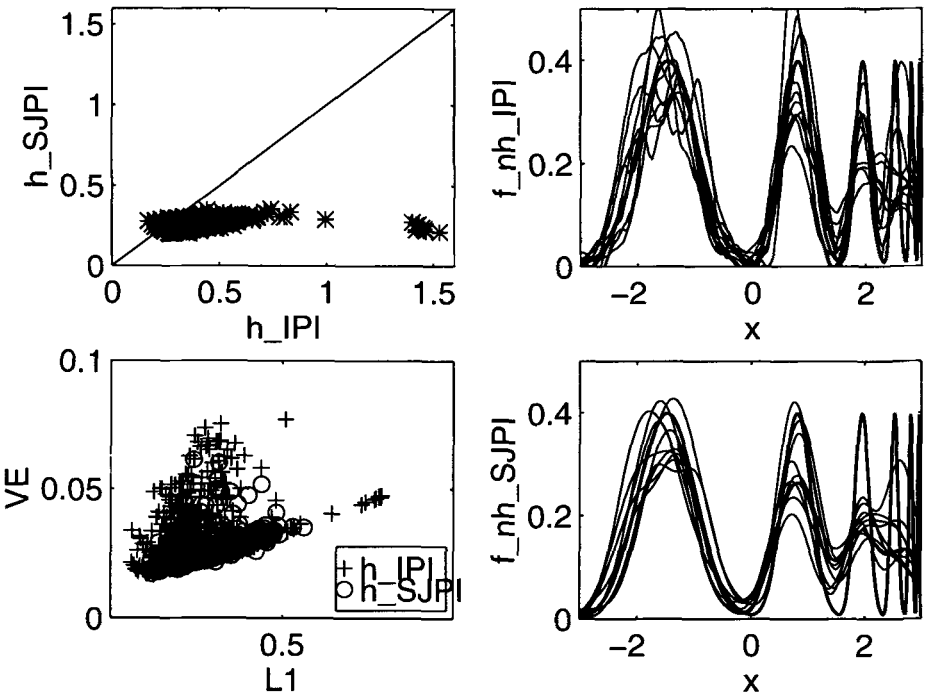

Figure 2: For 500 simulated data set of size $n = 100$, from the Smooth Comb density. Left hand panels: scatterplots of the L^1 Improved and Sheather Jones Plug In bandwidths in Figure 2a and of the error measures $L^1(h)$ and $VE(h)$ in Figure 2c. Right hand panels: 10 typical estimates for the L^1 Improved Plug In in figure 2b and the Sheather Jones Plug In in Figure 2d, together with the true underlying density, shown as the heavier curve.

respect to VE mostly by avoiding the proportion of quite large values that the L^1 Plug In suffers. On the other hand the L^1 Improved Plug In is better in the L^1 sense, because the L^1 error is often slightly smaller. Note that most of the time each method is quite acceptable with respect to both measures.

An important way of comparing the performance of different bandwidth selection methods is to look at the resulting density estimates, as done in the right hand panels, Figures 2b and 2d. Again personal opinions will vary, but I much prefer the Sheather Jones in this case, because I am most interested in finding features such as bumps, and prefer not to see "bumps that are not really there". The L^1 Improved Plug In has too many spurious bumps in the region of the first large peak, although there is improved performance in other regions. This density would be most effectively estimated by a location varying bandwidth, which is large on the left hand side, and smaller on the right, but it is still interesting to see how well a constant bandwidth method can perform, especially as location varying bandwidths are not so well understood yet.

2.6 Final remarks

It has been fun both reading this deep and informative paper, and also writing up this discussion. I am sure the author's response will also be interesting. Again, I would like to stress the point from Section 1 above, that I view all of these "modern" bandwidth selectors as being effective data analysis tools.

References

Fan, J. and J.S. Marron (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**, 35-56.

Farmen, M. (1997). An assessment of finite sample performance of adaptive methods in density estimation. Unpublished manuscript.

Jones, M.C., J.S. Marron and S.J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401-407.

Lindeberg, T. (1994). *Scale-space theory in computer vision*, Kluwer, Boston.

Marron, J.S. (1997). Assessing bandwidth selectors with visual error criteria, to appear in *Computational Statistics.*

Marron, J.S. and A.B. Tsybakov (1995). Visual error criteria for qualitative smoothing, *Journal of the American Statistical Association*, **90**, 499-507.

Sain, S.R. and D.W. Scott (1996). Zero-bias locally adaptive density estimators. Technical Report, Rice University, Available at: ftp.stat.rice.edu.

Schmitz, H.P. and J.S. Marron (1992). Simultaneous estimation of several size distributions of income. *Econometric Theory*, **8**, 476-488.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Wand, M. P., J.S. Marron and D. Ruppert (1991). Transformations in density estimation. *Journal of the American Statistical Association*, **86**, 343-361 (with discussion).

Yang, L. (1997). Root n convergent transformation kernel density estimation. Unpublished manuscript, available at: http://www.wiwi.hu-berlin.de/pub/papers/sfb373/sfb1997/dpsfb970006.ps.Z.

Yang, L. and J.S. Marron (1997). Iterated transformation kernel density estimation. To appear in *Journal of the American Statistical Association.* Avaliable at: http://www.wiwi.hu-berlin.de/pub/papers/sfb373/sfb1996/dpsfb960094.ps.Z.

C. Sánchez-Sellero
Universidad de Santiago de Compostela, Spain
J. de Uña
Universidad de Vigo, Spain

This paper provides a review of the new methodology developed by Luc Devroye and Gabor Lugosi for smoothing factor selection and a comparison with some of its best competitors. We would like to congratulate the author for the comprehensive study accomplished here and for the graphical presentation of the simulation results. We also admire the introduction of the minimum distance principle in the problem of density estimation and the appeal of establishing a relationship with the empirical measure. With respect to the theoretical results, we emphasize their application to all densities in all dimensions and their nonasymptotic character.

The author has developed this methodology in two steps, published in two different papers, and here defined as "first and second bandwidths". In this paper an implementation of the first bandwidth is provided, but not of the second one. We wonder whether the first bandwidth could in fact be considered as an implementation of the second one.

We would like to raise two points regarding the application of the method. First, the parameter b_n, that is, the upper bound of the candidate bandwidths is assumed to go to infinity in Theorem 1. We think that it is enough to require the convergence to a constant. In fact, this upper bound is taken as the interquartile distance in the implementation. Second, the possibility of nonuniqueness of the minimum is pointed out. We feel this as a drawback of the method. How often does this happen?

In the practical implementation of the double kernel method, the author proposes four rescaling values for the second kernel L and asserts: "The theory tells us that for large n, the scale factor of L should exceed that of K". This is nothing but a second bandwidth for the second kernel. We think that double kernel requires double bandwidth and we think that the author is of the same opinion.

According to the simulation study, the L1-plug-in improved selector seems to be the best one. The new methods, at the bottom of the figures, have an overall good performance. However, we are worried about their problems with the simpler densities, most of them at the left side of the figures. We also observe in the simulation results that the iteration and rotation do not supply significant improvements to the Devroye-Lugosi selector. Given the increase in the computational cost that these procedures imply, we are not sure about their usefulness.

This methodology can and should be adapted to obtain two main purposes: To take advantage of some known properties of the data (for instance, knowledge of the support), and to deal with special types of data. Regarding this latter case we have in mind the many types of incomplete data. In particular, we have designed a new bandwidth selector for random right censored data. This was done defining the total variation in this way:

$$T = \sup_A \left| \int_A \hat{f}_{n-m,h} - \int_A d\hat{F}_m \right|$$

where $\hat{f}_{n-m,h}$ is the convolution of the rescaled kernel with the product-limit estimate under random censorship constructed with the first $n - m$

data points, and $\hat{F}_m$ is the product-limit estimate with the test set of m data points. We stress the fact that suitable estimators of the distribution function could substitute the empirical estimate in order to obtain selectors adapted to many other contexts. In this sense, still under the model of random censorship, if we want to take advantage of the hypothesis of proportional censorship (submodel of the previous one), we could replace the product-limit estimator by the ACL estimator of the distribution function.

We carried out a small simulation study to assess the behavior of this selector and to compare it with a plug-in selector. We have chosen the plug-in selector adapted to censorship by Sánchez Sellero, González Manteiga and Cao (1997). Weibull distributions with scale parameter 1 and shape parameters 1, 2 and 3 were chosen for the variable of interest and a proportional censoring model was simulated with a probability of censoring of 25%. One thousand samples of size 100 were generated. The numbers represent the averages over the one thousand samples of the ISE and the IAE (the values were multiplied by 1000 for ease of presentation). We can observe that the Devroye-Lugosi bandwidth adapted to censoring is perfectly competitive. We also point out that the introduction of the proportional censoring information gives place to an improvement in the performance. This was expected. Finally, in our experience we noted that the kernel density estimate with DL bandwidth has a better performance when it is constructed with the whole sample than when it is computed using only the first $n - m$ data points.

	Means of the ISE			Means of the IAE		
Selector	$W(1,1)$	$W(2,1)$	$W(3,1)$	$W(1,1)$	$W(2,1)$	$W(3,1)$
DL with PLE	39.6	20.0	10.6	189	113	74.9
DL with ACL	34.3	16.9	8.98	176	103	68.6
Plug-in	44.1	21.0	8.85	181	109	65.3

Table 1: Results of the simulation study.

References

Sánchez-Sellero, C., W. González-Manteiga and R. Cao (1997). *Bandwidth selection in density estimation with truncated and censored data.* Preprint.

F. Udina
Universitat Pompeu Fabra, Spain

1 Introduction

I would like first to thank the editors for setting up this forum around the very interesting paper written by Luc Devroye and giving myself the opportunity to participate in it.

As I knew about the new methodology for bandwidth selection developed by Devroye and Lugosi (DL), I was attracted by both the elegance and simplicity of the involved ideas and, of course, by the universality they achieve. My second thought was about the practical performance of these selectors. Luc Devroye has made a really great work by comparing the new methods with the classical ones and showing that they are very competitive. It would be very interesting to extend it to include sample sizes other than $N = 100$.

In the following lines I want to discuss how Devroye-Lugosi bandwidth selectors behave in practice when facing real data sets. This concern two main issues: how to adapt the method to work in a binned data computational setting and how to protect the method against typical data manipulation such as rounding or sorting.

2 Binned version of Devroye-Lugosi selector

Using kernel estimation techniques in practice means to use some discretization technique like binning as described, for example, in Fan and Marron (1993) or, under the name of *warping* in Härdle and Scott (1992). It is the only practical way to deal with big or moderately big data sizes. For a given grid g_k, $\{j = 1 \ldots G\}$, bin counts c_k are computed for the first $n - m$ data points using linear binning. Then approximated estimates for density values are computed for every g_k needing only a few kernel function values. The finite convolution needed can be computed even faster using fast Fourier transform techniques. We use in this quick study the global version of the "First bandwidth" as described in Devroye's paper. The slight changes we make to the algorithm will be commented. In this binned context, the approximated Yatracos set $\tilde{A}_{ij}$ is simply a union of bins (the

ones where $f_{h_i}(g_k) > f_{h_j}(g_k)$) that can be coded as a bit vector of length G. Yatracos distance computation is really fast and simple in this situation.

A potential problem that would need some study appear: although binning gives very good approximated density values we have no guarantee of having good approximation to the true Yatracos sets (whatever this means) and this can distort the distances to be minimized. Our experience show that the choice of G and the binning width w do not affect the selected bandwidth, provided that G is big enough and the width is small. We usually take G in the hundreds. Once w is fixed, we use it to decide the minimum bandwidth to be considered, because it makes no sense to take bandwidth values that would include less than several grid values in the computation.

3 Dealing with *real* data sets

When we deal with real data sets we must be careful about how data has been transferred to the analyst. It's quite usual, for example, that data have been collected in rounded form or have been rounded a posteriori. Some bandwidth selectors are robust against any small change in data values, but the Devroye-Lugosi universal selector can have problems with repeated values. We will discuss some examples below.

Another source of problems can come from data being arranged in some way before arriving to the algorithm. Some of my time was spent investigating strange behavior in the algorithms proposed by the paper. The reason was that the data were sorted in increasing order in the file where it came to me. Obviously the method doesn't work at all with sorted data if split is done in the usual way, taking the first m values. Similar problems can appear if data are collected or transmitted in a non-random ordering, by different categories, for example. Even in some Monte-Carlo simulations I found that some of my algorithms were generating data from normal mixtures in a component-by-component way. This normally doesn't have any consequence but when splitting the data set to apply the Devroye-Lugosi procedures, problems arose.

The strategy to follow when data is suspected to have been sorted is not easy to devise. One possibility is to shuffle randomly the data set and then to take the first m data values to build the empirical measure. This

has the problem of producing a different selected bandwidth every time the algorithm is applied. As we will see in the examples, changing the m length subsample introduces a lot of variability in the selected bandwidth. In the paper, Devroye introduces the rotation idea to compensate it. The obvious drawback is that computation time grows, as the method itself is quite time consuming. Another possibility could be extracting the m data points in some systematic way. For example, a simple scheme could be, to take m elements out of n, sort the data set and take those with ordered index $\lfloor ni/m \rfloor$ for $i = 1 \ldots m$. But this needs some theoretical justification because violates the independence assumption and in case of repeated values would favor even more coincidences between the two subsamples.

4 Some real data sets

I was trying first to work with the Miguel Hidalgo stamps data set brought to the density estimation area by Izenmann and Sommer (1988). It's a data set that presents both a rich structure to study and a reliable sample size (n=485). Data comes from thickness measured from a collection of old stamps issued in 1872-74 in Mexico. But unfortunately, data is heavily rounded to thousandth of millimeters resulting in a high proportion of repeated data. This results in the impossibility to apply the described techniques to it. The Sheather-Jones plug-in bandwidth, for example, is 0.00272 for this data set, so it is in the same order of the resolution of the data. To select this bandwidth, DL would need to fix a minimum bandwidth so small that the estimates will result a series of single value peaks. When splitting the data, the m subsample results to be a subsample of the $n - m$ one. The result is that the minimum bandwidth is selected most of the times the algorithm is run. Probably Devroye-Lugosi theory might be modified to work with smoothed discrete distributions and then some work could be done with this data set.

Postman, Huchra and Geller (1986) studied the velocities of 82 galaxies moving away form our galaxy. The data set was deeply analyzed from the kernel smoothing point of view in Park and Turlach (1992). They gave bandwidth values selected by several automatic selectors. Translated to our kernel, they found from 1.37 for least squares cross-validation to 3.48 for biased cross-validation. For the Sheather-Jones plug-in they gave 1.44 and 2.72 for the normal-reference rule-of-thumb. The data set arrived to

me sorted in increasing order, so the re-shuffling strategy was applied.

	# of h values			
m/n	10	20	30	40
0.1	1.339	1.686	2.543	2.626
0.2	1.928	1.418	1.811	1.724
0.3	2.777	3.365	3.189	3.107
0.4	2.777	2.382	2.271	2.414
0.5	1.928	1.686	1.811	1.724

Figure 1: Values selected by the DL algorithm as described in the text, for several values of m/n and of the number of estimators in contest.

Applying the version of DL selector described above we obtained values around 2.0, depending on parameter choice. To show how the value selected depends on these parameters, we shuffled the data set once (it arrived to me sorted in increasing order again) and we run the algorithm with m being 10%, 20%, 30%, 40%, and 50% of the data size $n = 82$ and taking the number of different values for h in the range $[0.15, 4.0]$ as 10, 20, 30, and 40. The minimum bandwidth is fixed as 1.5 times the binning width (so ensuring that at least three bins are involved in any computation). The maximum bandwidth is chosen to be the interquartilic range. The resulting values are shown in the table in figure 1. The variability of the selected parameter is noticeable. As expected, higher values of the ratio m/n give less variability. But even using the same pair of parameters, shuffling data can give very different values for the selected bandwidth. For example, with $m/n = .3$ and 20 bandwidth values, we obtained 0.423, 1.418, 1.418, 1.686, and 1.193 in five runs of the algorithm with re shuffling. To test the rotation idea suggested by Devroye, we took $m = n/2$ and computed the geometric mean of the two selected bandwitdhs obtained using both halves alternately as m subsample. Number of bandwidths in contest were 20 in the range $[0.15, 4.0]$. The values obtained were $0.775, 1.004, 1.094, 1.546, 1.546$. We see that variability has not improved so much. As a matter of curiosity, we also show in figure 2 the shape that Yatracos sets appear to have. For 10 bandwidth values (from .15 to 4.0 in geometric steps) we generated 10 density estimates $f_1, \ldots, f_{10}$ and computed the Yatracos sets $A_{ij} = \{x | f_i(x) > f_j(x)\}$ Each horizontal line in figure 2 represent a set. The lower band composed by nine lines corresponds, as going up, to A_{1j}, $j = 2, \ldots, 10$ and so the other 10 bands.

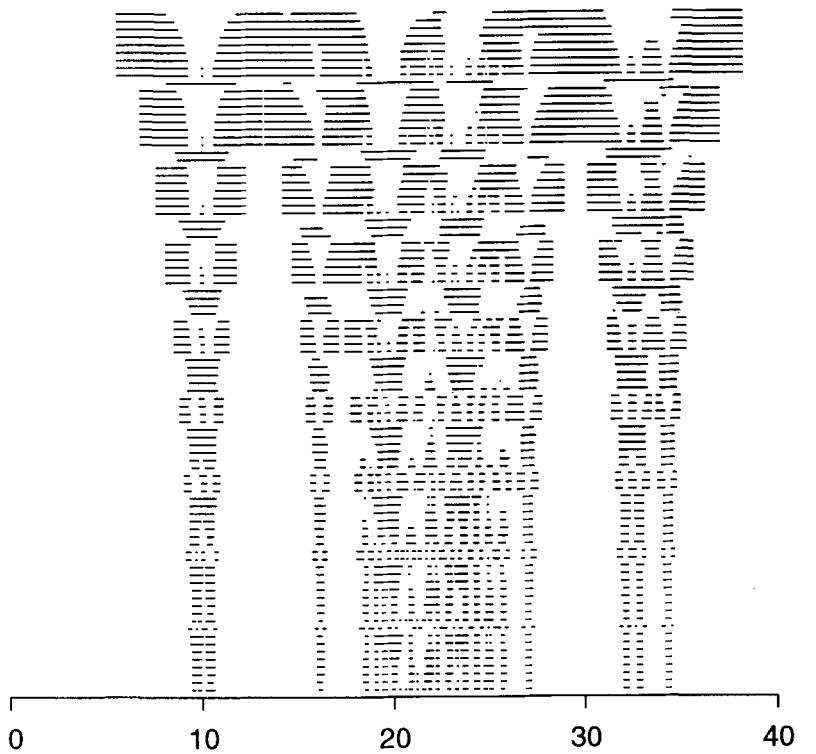

Figure 2: Yatracos sets for galaxies data

Most of the variability found in the selected bandwidths for the galaxies data can be explained by the sample size being too small. We decided to contrast it with the case of a very big data set, also well known in the smoothing literature. The net income of 7201 British households in the year 1975 was studied in Schmitz and Marron (1992) and also discussed in Wand, Marron and Ruppert (1991). Data range goes from 0.026 to 9.1225 the unit being the average income. Despite the data range and resolution, the set contains a quite high number of repeated values, namely 1357. We took a grid size of 800 in the range -1 to 10, so the minimum bandwidth to be considered is approx .022. The maximum was taken here as half the interquartilic range and 25 bandwidth values were taken in the resulting interval $[.022, .4]$. m was chosen to be 25% of the sample size. In figure 3 we plot Yatracos distance over (logarithm of) bandwidth values. 8 different runs are plotted resulting from re shuffling before split. The selected bandwidths were 0.051, 0.058 twice, 0.065 three times, 0.094 twice, 0.106 and 0.152. We see that variability is somewhat reduced but it's still quite high. Applying the systematic split discussed in section 2 selected a bandwidth of 0.025, clearly too small. The reason can be the relatively

high percentage of repeated values. Applying the rotation algorithm in this setting resulted to more reasonable and stable values, we got 0.10, 0.094, .081 in three different runs.

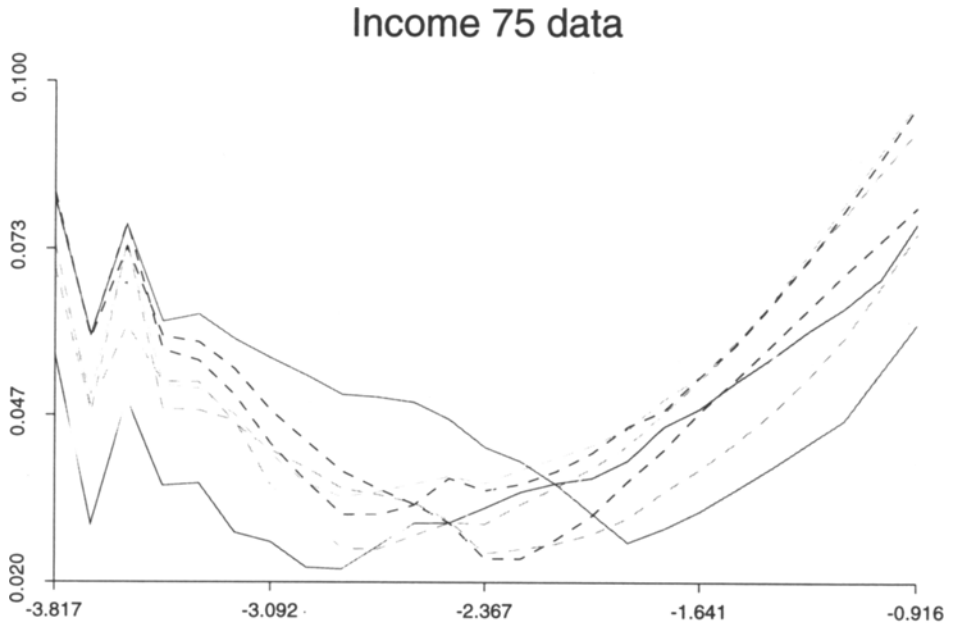

Figure 3: Values for the Yatracos distance in vertical axis. In horizontal axis, log of bandwidth. Each line is a diferent run of the algorithm with previous data shuffling.

5 Conclusion

We have seen that to use the DL selectors with real data, binning techniques can be used but some problems must be faced. The main one is how to deal with repeated values – this seems to need some theoretical adjustment. The other problem is to protect the algorithm against data that can be sorted or arranged in some non-random way. It looks like the shuffling-then-rotate approach is the safer one but further study is needed.

References

Fan, J. and J.S. Marron (1992). Best possible constant for bandwidth selection. *Annals of Statistics*, **20**, 2057–2070.

Härdle, W. K. and D.W. Scott (1992). Smoothing by Weighted Averaging of Rounded Points. *Computational Statistics*, **7**, 97–128.

Izenman, A. J. and C. Sommer (1988). Philatelic Mixtures and Multimodal densities. *Journal of the American Statistical Association*, **83**, 941–953.

Postman, M., J.P. Huchra and M.J. Geller (1986). Probes of large scale structures in the Corona Borealis region. *The Astronomical Journal*, **92**, 1238–1247.

Park, B. U. and B.A. Turlach (1992). Practical Performance of Several Data driven bandwidth selectors. *Computational Statistics*, **7**, 251–270.

Schmitz, H.P. and J.S. Marron (1989). Simultaneous estimation of several Size Distributions of Income. *Discussion paper A-186, SFB 303, University of Bonn, Dept. of Economics.*

Wand, M. P., J.S. Marron and D. Ruppert (1991). Transformations in Density Estimation. *Journal of the American Statistical Association*, **86**, 414, 343–361.

Rejoinder by L. Devroye

Sometimes, researchers charge ahead like young bulls in a bullring, hitting anything that moves in the arena, but oblivious to the sword that will eventually kill them. That is exactly how I feel after reading some of the comments. If the newly proposed methods are to be used in practice, as Frederic Udina points out, more work is needed to address the problems of the selection of a random subset and of real data with repeated values. On his transparencies, Frederic used the phrase "repeated values kill Devroye-Lugosi". César and Jacobo correctly point out that our work on the double kernel is unfinished: the unstudied double kernel/double h method is likely the one that will survive most matadorial attacks.

Many comments relate to generalizations of the new estimates or improvements of the kernel estimate. These include transformed kernel estimates (Steve and Jan), locally adaptive estimates (Steve again), support-

sensitive estimates (César and Jacobo), symmetry-sensitive estimates (Ricardo Cao), generalizations to censoring (César and Jacobo), and parameter selection in nearest neighbor estimates (Ricardo Fraiman). All these comments are pertinent and show real concern regarding the kernel estimate's performance. We are pleased that our study has uncovered new problems related to kernel density estimates. The existence of just a few patterns of behavior across the spectrum of densities (figures 6 through 10) prompted Peter Hall to suggest the automatic selection of a bandwidth selector based upon entropy estimates. This could also be done based on spacings. Indeed, if we had a further i.i.d. sample $Z_1, \ldots, Z_k$ with order statistics $Z_{(1)} < \cdots < Z_{(k)}$, then $\int_{Z_{i-1}}^{Z_i} f_n, 1 \leq i \leq k+1$, with $Z_{(0)} = -\infty$ and $Z_{(k+1)} = \infty$ should ideally be distributed as uniform spacings. A standard spacings test may be used to identify the best bandwidth selector. A related method was included in the simulations of Berlinet and Devroye (1994) to select the best h in the kernel method, but it yielded mediocre results. Nevertheless, this avenue of research should prove useful.

No simulation study can ever be conclusive. Both Chris Jones and Steve Marron point out some problems with the selection of our test densities and with the interpretation of the results. For example, it is clear that all bandwidths that optimize some criterion perform rather abysmally on near-normal densities relative to plug-in methods. To expose behavior of this sort was precisely one of our goals. One should not forget that we set out to develop bandwidths with new broad performance guarantees, and that the simulation study came later. The instability of the L_1-improved plug-in method discovered by Steve is indeed puzzling. To make matters worse, the second bandwidth of Devroye and Lugosi (1997) is simply unimplementable as it involves minimization of a function over an infinite space, and each function value requires a supremum over an uncountably infinite ensemble. At this stage, we should take the bull by the horns and place the selectors in the hands of agile bandwidth engineers to fine-tune, adjust, modify, robustize and tame.

Chris Jones mentions that we should have used a fuller version of the Sheather-Jones bandwidth. We picked the one used in the study by Cao, Cuevas and González-Manteiga (1994), in part to have a standard of comparison with that study. In his comment, Steve Marron confirms our simulation results with respect to the Sheather-Jones bandwidth, but it was not clear which version Steve used.

Chris Jones and to some extent Steve Marron wave the red banner for the choice of the Bartlett-Epanechnikov kernel in the simulation study. There were various reasons for our decision. Among positive kernels, it is optimal for L_1 and L_2. It is computationally convenient as the kernel estimate is piecewise parabolic. We also suspect that the relative performance of various bandwidth selectors will not change much among smooth unimodal positive kernels. However, more importantly, one might consider the joint data-based choice of K and h. This should improve the absolute performance, but perhaps not for small sample sizes. If the complexity of the family of kernels K can be controlled (in the computation of the Vapnik-Chervonenkis shatter coefficient), then methods similar to those of Devroye and Lugosi (1997) may be useful. Gábor's comments in fact read as a mini-paper in which he shows the way for parametric families of kernels. We thank him in particular for working out a d-dimensional example in which the kernel estimate has d scale parameters.

Steve sticks his banderilla in the right spot when he remarks that $n = 100$ is too small for some densities in our study. But that is precisely our point too. Those are the densities that can be classified as difficult. Thus, our simulation study covers many different virtual (or relative) sample sizes. We can get more information either by changing n or by adding more densities and keeping n fixed. Eventually, if we increase n, all densities will appear easy, so increasing n alone will not be sufficient in a study.

The choice of the L_1 criterion is based on arguments related to the estimation of probabilities, generating new approximate samples from the unknown density, invariance under monotone transformations, visual distance between curves, and the universality of the error scale. Our bandwidths are designed with the L_1 criterion in mind. Marron and Tsybakov's VE is different and requires appropriate and new bandwidth selectors. We unfairly threw sand in the readers' eyes by comparing the Sheather-Jones bandwidth with the L_1-improved plug-in method (as the former was designed for L_2). In the same vain, one should not compare our bandwidths with other ones based on VE, for example. It is also unclear whether in his comparison, Steve used VE(1), an L_1 version of VE, or VE(2), which by definition should favor L_2 bandwidth selectors such as Sheather-Jones. If f is Lipschitz, then VE(1) ($f_n \to f$) divided by $\int |f_n - f|$ is always between $1/\sqrt{1+C^2}$ and 1, so that we suspect VE(2) was used. Much blood can be spilled over the question of why we do density estimation. Chris Jones points out that if we are interested in probabilities, we should stick

to empirical distribution functions. These are only good for probabilities of intervals, not probabilities of Borel sets. Indeed, the finite set of the data receives mass one from the empirical distribution function but has in fact zero probability whenever a density exists. Density estimates with good L_1 properties provide good probability estimates uniformly over all (possibly data-dependent!) sets.

The paper builds on experiments first reported by Berlinet and Devroye (1994) and on theoretical results with Gábor Lugosi (1996, 1997). Alain's cooperation during the last few years is gratefully acknowledged, while Gábor should have been a coauthor of the present paper (I asked him but he refused to enter the ring). Various commenters refer to equation (7), which was changed to (8.7) in the final printed version. We also corrected typographical errors reported by César and Jacobo. We changed only one sentence in the original manuscript after a comment by Steve that (8.1) is not always close to the optimum, as we hastily and incorrectly claimed. We thank Domingo Morales for a thorough job as editor and we thank Test for allowing us to publish five color figures. Finally, we thank Antonio Cuevas and Wenceslao González-Manteiga for the warm reception in Santiago, where we read our paper on September 11, 1997, and hit all moving objects thrown at us by Ricardo Cao, Frederic Udina and César Sanchez-Sellero. They stopped just short of singing the Malagueña.

Let me conclude with a few open problems of my own. These show a strong personal bias and may be the last nervous jerks of an impetuous but wounded bull.

1. (1) Establish if the class of asymptotically optimal bandwidths is empty. Can the 3 in the bounds be replaced by 1 for some bandwidth selector?

2. (2) If the answer to (1) is negative, one should develop asymptotically optimal bandwidths for all densities in certain subclasses. We do not know for example how to pick an asymptotically optimal bandwidth for all monotone densities on the positive halfline. As there are infinitely many classes of densities, this is like letting the prize bull loose in the cow barn.

3. (3) Is the double kernel bandwidth suitable?

4. (4) Study the double kernel/double h method. Is it universally suitable?

5. (5) Study the properties of the iterative method of section 6.

6. (6) Study the sensitivity with respect to changes in subsets of the data of size k. That means studying the supremum of the L_1 error after the supremum is taken over all subsets of k out of n and all values on the real line for the k selected data points.