THE CONSISTENCY OF A SMOOTHED MINIMUM DISTANCE ESTIMATE

Ricardo Cao
Department of Mathematics
University of La Coruña
La Coruña, Spain

and

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

Abstract. We give sufficient conditions for the consistency of a minimum distance density estimate, obtained by minimizing the L_1 distance between a kernel estimate with smoothing factor h and a density in a collection $\mathcal{G}$ of densities. The optimization is performed over all $h > 0$ and all of $\mathcal{G}$. Various parametric and nonparametric target classes $\mathcal{G}$ are considered. It turns out that the choice of the kernel is crucial for the consistency of this minimum distance method.

Keywords and phrases. Density estimation, kernel estimate, convergence, smoothing factor, minimum distance estimate.

1991 Mathematics Subject Classifications: Primary 62G05.

Authors' address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2A7. The first author's research was sponsored by DGICYT Grants PB91-0794 and PB94-0494. The second author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. Part of this research carried out while the second author was visiting the Department of Mathematics, University of La Coruña, Spain.

1. Introduction.

Sometimes, when we estimate a density, we have no idea whether or not this density falls in a given class of densities such as the class of Pearson densities, or the class of all log-concave densities. So, we would like to bet on two horses and get two density estimates: one is obtained under the assumption that the density is in the target class, and the other one is nonparametric—it assumes nothing. This may be achieved by two independently constructed density estimates. We believe that these estimates should be designed together, and propose in this paper a method—the smoothed minimum distance method— that was first suggested by Cao, Cuevas and Fraiman (1993). The estimate provides a general and useful tool for estimating densities in the presence of a priori information. It yields two density estimates. This paper shows which one is useful under what circumstances. Interestingly, the smoothing factor in the nonparametric estimate is automatically picked by the data; absolutely no fine-tuning or manual adjusting is required. Universal conditions of consistency are provided.

More formally, we are given an i.i.d. sample $X_1, \ldots, X_n$ drawn from an unknown density f on $\mathbb{R}^d$. The target class of densities (to which our density f may or may not belong) is denoted by $\mathcal{G}$. Our second ingredient is the Akaike-Parzen-Rosenblatt density estimate

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

where K is an almost everywhere continuous density (the kernel), $K_h(x) = (1/h)K(x/h)$, and $h > 0$ is the smoothing factor (Akaike, 1954; Parzen, 1962; Rosenblatt, 1956).

Define the pair (g_n, H) by

$$(g_n, H) = \arg \min_{g \in \mathcal{G}, h>0} \int |g - f_{nh}| \ .$$

In other words, we find the pair of densities $g_n \in \mathcal{G}$ and f_{nH} (a kernel estimate) that is closest to each other. The existence of at least one minimum follows for most classes considered by us, by compactness. It is tacitly assumed throughout this paper. Uniqueness is of no concern to us as all results given below remain valid if we select any pair at which a minimum is attained. This paper considers conditions that guarantee the convergence of f_{nH} to f, and if $f \in \mathcal{G}$, we also study the convergence of g_n to f. It turns out that the choice of K is critical: it must be adapted to $\mathcal{G}$ to make the family of kernel estimates indexed by h sufficiently different from $\mathcal{G}$.

Minimum distance estimation under various criteria have been applied in many contexts. For work that is loosely related to ours in ideology, we refer to Beran (1977). However, we bring two new key ingredients to the foreground: the L_1 criterion, and the idea of minimizing with respect to the smoothing factor in the kernel estimate. In fact, both estimates g_n and f_{nH} solve the omnipresent problem of the data-based selection of a smoothing factor. At the same time, the conceptual simplicity of the estimates may make their use widespread. It is, however, necessary to better understand the properties of such estimates. The main property that we need to verify is the consistency—this paper addresses this issue. In future work, one should study the goodness of the estimates as a function of n.

PARAMETRIC CLASSES. If we consider $\mathcal{G}$ as a parametric class indexed by a parameter vector θ, then any consistency study must work with the equivalence classes $E(f)$ consisting of all θ that lead to the same density f. In particular, to study convergence in the space of the θ's, it is necessary to introduce uniqueness and identifiability notions. For this reason, we will not deal with the convergence of parameters, but rather with convergence of densities in the L_1 sense.

PARAMETRIC MEGACLASSES. The class of all Pearson densities could be called a megaclass as it houses several different subclasses of densities. Our method furnishes a density estimate g_n from this class; note in particular that the data automatically picks a particular Pearson subfamily and a collection of parameters for a density from this subfamily.

OTHER SELECTION PRINCIPLES. We opted for an L_1-based minimum distance estimate. In this sense, the kernel estimate became a natural choice in the definition of the criterion to be minimized, as the total variation distance between any density and the standard empirical measure is one; hence the need for smoothing the empirical measure. Furthermore, methods such as moment matching and maximum likelihood do not share the generality of application that is exhibited in this paper. Early experimental results in Cao, Cuevas and Fraiman (1993) indicate that their minimum distance method is indeed competitive with the maximum likelihood method for mixtures of normal densities.

RELATED ESTIMATES. The estimate of Devroye (1989b) chooses between an estimator in $\mathcal{G}$ and a kernel estimate f_{nH}, basing the decision on the size of the L_1 distance between the estimates. Other references to similar work may be found there. Our focus here is different: we plainly want our estimate to be in or near $\mathcal{G}$.

Automatic selection. Further work might include the following: consider a kernel estimate f_{nh} with kernel K, a kernel estimate f^*_{nh} with another kernel L, and a class of densities $\mathcal{G}$. We find smoothing factors (H', H^*) defined by

$$(H', H^*) \stackrel{\text{def}}{=} \arg\min_{h,h^*>0} \int |f_{nh} - f^*_{nh^*}| \ ,$$

following ideas taken from the double kernel method (Devroye, 1989a). Define g_n and H as before. Now decide to estimate the unknown f by $f_{nH'}$ if the L_1 distance between $f_{nH'}$ and $f^*_{nH^*}$ is less than the L_1 distance between g_n and f_{nH}, and to estimate f by g_n otherwise. This would be an alternative to other parametric/nonparametric selection methods such as those of Olkin and Spiegelman (1987) and Devroye (1989b).

2. The main result.

Theorem 1. *Let $f \in \mathcal{G}$. Assume the following conditions:*

(i) For every $\delta > 0$,

$$\inf_{g\in\mathcal{G},h\geq\delta} \int |g - K_h * f| > 0 \ .$$

(ii) For every $M < \infty$ and $\epsilon > 0$, there exist positive constants a, n_0 such that

$$\mathbf{P}\left\{ \inf_{g\in\mathcal{G},h\leq M/n} \int |f_{nh} - g| \geq a \right\} \geq 1 - \epsilon \ , \ \forall n \geq n_0 \ .$$

Then

$$\int |f_{nH} - f| \to 0 \text{ in probability} \ ,$$

and

$$\int |g_n - f| \to 0 \text{ in probability} \ .$$

Before we embark on the proof of this result, we note that there are no restrictions imposed on h; in other words, the method is fully automatic. Also, no a priori restrictions are put on f or $\mathcal{G}$, other than those implicit in (i)-(ii).

Condition (i) states that the kernel K should be sufficiently different from the densities in $\mathcal{G}$. Condition (ii) states that $\mathcal{G}$ should not be too rich: if it contains too many densities, then the information in the data is lost because just about any density can be matched with one from $\mathcal{G}$. Simple sufficient conditions for (ii) are provided further on.

3. Proof of Theorem 1.

If we set $h = 1/\sqrt{n}$, then by the consistency of the kernel estimate for all densities (see page 148 of Devroye and Györfi, 1985),

$$\inf_{g \in \mathcal{G}, h>0} \int |g - f_{nh}| \le \int |f - f_{n,1/\sqrt{n}}| \to 0 \text{ almost surely .} \tag{1}$$

We will show that (1) entails the almost sure convergence to 0 of $\int |f - f_{nH}|$. Therefore, by the triangle inequality, $\int |g_n - f| \to 0$ almost surely. As a first intermediate result, we prove the following Lemma. The conditions of the Theorem are of course assumed.

LEMMA 1. *Let f be any density. For any $\delta > 0$, we have*

$$\sup_{h \ge \delta} \int |K_h * f - f_{nh}| \to 0 \textit{ almost surely .}$$

In particular, if H is any random variable, possibly dependent upon the data,

$$\int |K_H * f - f_{nH}| I_{[H \ge \delta]} \to 0 \textit{ almost surely .}$$

PROOF OF LEMMA 1. Let M be a large integer and $\beta > 1$ be a constant to be picked later. For arbitrary $\eta > 0$, we will show that almost surely, for all n large enough,

$$\sup_{h \ge \delta} \int |K_h * f - f_{nh}| \le 3\eta .$$

Split the interval $[\delta, \infty)$ into $[\delta\beta^M, \infty)$ and $[\delta\beta^{j-1}, \delta\beta^j)$, $1 \le j \le M$. Note the following:

$$\begin{aligned} &\sup_{h \ge \delta\beta^M} \int |K_h * f - f_{nh}| \\ &\quad \le \sup_{h \ge \delta\beta^M} \frac{1}{n} \sum_{i=1}^n \left(\int |K_h * f - K_h| + \int |K_h(x - X_i) - K_h(x)|\, dx \right) . \end{aligned}$$

We have $\int |K_h - K_h * f| = \int |K - K * f_{1/h}| \to 0$ as $h \to \infty$ where $f_{1/h}(x) = hf(hx)$ (see Theorem 3.3, page 23, of Devroye, 1987). If $\delta\beta^M$ is large enough, we see that one term in the last chain of inequalities may be made as small as desired. Consider the other term. It is useful to recall that as $\epsilon \to 0$, $\int |K(x+\epsilon) - K(x)|\, dx \to 0$ (to see this, use Scheffé's theorem, the integrability of K, the almost everywhere continuity of K, and the Lebesgue dominated convergence theorem). Thus,

$$\begin{aligned} \sup_{h \ge \delta\beta^M} \frac{1}{n} \sum_{i=1}^n \int |K_h(x - X_i) - K_h(x)|\, dx &= \sup_{h \ge \delta\beta^M} \frac{1}{n} \sum_{i=1}^n \int |K(x - X_i/h) - K(x)|\, dx \\ &\le \sup_{h \ge \delta\beta^M} \left\{ \eta + \frac{2}{n} \sum_{i=1}^n I_{[|X_i/h| > A]} \right\} \end{aligned}$$

$$\begin{aligned}&\text{(where } A \text{ depends upon } \eta)\\ &\leq \eta + \frac{2}{n}\sum_{i=1}^{n} I_{[|X_i|>A\delta\beta^M]}\\ &\leq 2\eta\end{aligned}$$

almost surely for all n large enough, by the law of large numbers and our choice of the product $A\delta\beta^M$. Finally, we look at M terms, the j-th one of which is

$$\sup_{\delta\beta^{j-1}\leq h<\delta\beta^j} \int |K_h * f - f_{nh}| \ .$$

Define $a_j = \delta\beta^j$, and bound the supremum above by

$$\sup_{a_{j-1}\leq h,h'\leq a_j} \left(\int |f_{nh} - f_{nh'}| + \int |K_h * f - K_{h'} * f|\right)$$
$$+ \int |K_{a_j} * f - f_{na_j}| + \int |K_{a_{j-1}} * f - f_{na_{j-1}}| \ .$$

It is a simple exercise to show that the last two terms in the above expression tend to zero (this is like the variation term in the L_1 error when the bandwidth remains constant). The second term in the supremum is bounded by

$$\sup_{a_{j-1}\leq h,h'\leq a_j} \int |K_h - K_{h'}| = \sup_{1\leq u\leq\beta} \int |K - K_u|$$

and this may be made as small as desired by picking β close enough to one. To see this, note that as $u \downarrow 1$, $\int |K - K_u| \to 0$ by Scheffé's theorem, the almost everywhere continuity of K and the Lebesgue dominated convergence theorem. Similarly, the first term in the supremum is bounded by

$$\sup_{a_{j-1}\leq h,h'\leq a_j} \frac{1}{n}\sum_{i=1}^{n} \int |K_h(x - X_i) - K_{h'}(x - X_i)|\, dx \leq \sup_{1\leq u\leq\beta} \int |K - K_u| \ .$$

This concludes the proof of Lemma 1. □

From (1) and Lemma 1, we conclude that

$$\int |K_H * f - g_n| I_{[H\geq\delta]} \to 0 \text{ almost surely .}$$

However, by condition (i), the left-hand-side of the last expression is bounded from below by

$$cI_{[H\geq\delta]}$$

for some $c > 0$ possibly depending upon δ. Therefore, as δ is arbitrary, we conclude that necessarily, $H \to 0$ almost surely. It remains then to show that $nH \to \infty$ in probability, so that we may conclude, by a general theorem on the consistency of data-based kernel estimates (see page 148 of Devroye and Györfi, 1985) that

$$\int |f - f_{nH}| \to 0 \text{ in probability} .$$

Let ϵ, M, a and n_0 be as in condition (ii). Then,

$$\begin{aligned}
&\mathbf{P}\{H \leq M/n\} \\
&= \mathbf{P}\left\{H \leq M/n\ ,\ \int |g_n - f_{nH}| < a\right\} + \mathbf{P}\left\{H \leq M/n\ ,\ \int |g_n - f_{nH}| \geq a\right\} \\
&\leq \mathbf{P}\left\{\inf_{h \leq M/n, g \in \mathcal{G}} \int |g - f_{nh}| < a\right\} + \mathbf{P}\left\{\int |g_n - f_{nH}| \geq a\right\} \\
&\leq \epsilon + \mathbf{P}\left\{\int |g_n - f_{nH}| \geq a\right\} \quad (n \geq n_0) \\
&= \epsilon + o(1)\ ,
\end{aligned}$$

where we used (1) in the last step. Because M is arbitrary, it follows that $nH \to \infty$ in probability. This concludes the proof of the Theorem.

4. Condition (ii).

In this section, we provide easy-to-verify sufficient conditions for (ii). Lemma 2 below states that it suffices to have a continuity of norms between the Kolmogorov-Smirnov distance and the L_1 distance within $\mathcal{G}$.

LEMMA 2. *Assume that the characteristic function of K is not identically one on any open neighborhood of the origin and that K is absolutely integrable with integral one (both conditions hold when K is a density). Condition (ii) of Theorem 1 holds when for fixed $f \in \mathcal{G}$,*

$$\lim_{a \downarrow 0} \sup_{g \in \mathcal{G}_a} \int |f - g| = 0\ ,$$

where

$$\mathcal{G}_a = \mathcal{G} \cap \{g : \sup_x |G(x) - F(x)| < a\}\ ,$$

and F and G are the distribution functions for f and g respectively. (We will refer to this as the continuity-of-norms condition.)

PROOF. We define $\mathcal{G}_a^* = \mathcal{G} - \mathcal{G}_a$. Note that

$$\inf_{g\in\mathcal{G},h\le M/n} \int |f_{nh} - g| \ge \min\left(\inf_{g\in\mathcal{G}_a,h\le M/n} \int |f_{nh} - g| \ , \ \inf_{g\in\mathcal{G}_a^*,h\le M/n} \int |f_{nh} - g| \right) \ .$$

We need two auxiliary results. The first one is Lemma S1 of Devroye (1989a) which states that under the condition on K given in Lemma 2, for every $M > 0$, there exists a constant $c(M) > 0$ such that

$$\liminf_{n\to\infty} \inf_{h\le M/n} \mathbf{E} \int |f_{nh} - f| \ge c(M) > 0 \ .$$

If $J_{nh} \stackrel{\text{def}}{=} \int |f_{nh} - f|$, then Lemma S2 of Devroye (1989a) shows that

$$\sup_{h>0} |J_{nh} - \mathbf{E}J_{nh}| \to 0 \quad \text{almost surely}$$

when K is absolutely integrable. Both results taken together imply that for every $\delta \in (0,1)$, there exists an n_0 such that for $n \ge n_0$,

$$\mathbf{P}\left\{ \inf_{h\le M/n} \int |f_{nh} - f| \ge c(M)/2 \right\} > \delta \ .$$

But

$$\inf_{g\in\mathcal{G}_a,h\le M/n} \int |f_{nh} - g| \ge \inf_{h\le M/n} \int |f_{nh} - f| - \sup_{g\in\mathcal{G}_a} \int |f - g| \stackrel{\text{def}}{=} A - B \ .$$

We recall that $\mathbf{P}\{A \ge c(M)/2\} > \delta$ for $n \ge n_0$. Also, we may pick a so small that $B \le c(M)/4$. Therefore, for $n \ge n_0$,

$$\mathbf{P}\left\{ \inf_{g\in\mathcal{G}_a,h\le M/n} \int |f_{nh} - g| \ge c(M)/4 \right\} \ge \delta \ .$$

Next, by Scheffé's identity (Scheffé, 1947), if F_n denotes the standard empirical distribution function, and if f has distribution function F,

$$\begin{aligned}
&\inf_{g\in\mathcal{G}_a^*,h\le M/n} \int |f_{nh} - g| \\
&\quad\ge \inf_{g\in\mathcal{G}_a^*,h\le M/n} \sup_x \left| \int_{-\infty}^x f_{nh} - \int_{-\infty}^x g \right| \\
&\quad\ge \inf_{g\in\mathcal{G}_a^*,h\le M/n} \left(\sup_x \left| \int_{-\infty}^x f - \int_{-\infty}^x g \right| - \sup_x \left| \int_{-\infty}^x f_{nh} - F_n(x) \right| - \sup_x |F_n(x) - F(x)| \right) \\
&\quad\ge a - \sup_{h\le M/n} \sup_x \left| \int_{-\infty}^x f_{nh} - F_n(x) \right| - \sup_x |F_n(x) - F(x)| \ .
\end{aligned}$$

By the Glivenko-Cantelli theorem, the last term tends to zero almost surely with n. Note that if the kernel K has distribution function L, then

$$\left| \int_{-\infty}^x f_{nh} - F_n(x) \right| \le \frac{1}{n} \sum_{i=1}^n |L((x - X_i)/h) - S(x - X_i)| \ ,$$

where $S(u)$ is one if $u \geq 0$ and zero otherwise. For fixed $X_1, \ldots, X_n$, the right-hand-side decreases monotonically to zero as $h \downarrow 0$. Hence, it reaches its supremum over $h \leq M/n$ at $h = M/n$. Thus,

$$\sup_{h \leq M/n} \left| \int_{-\infty}^{x} f_{nh} - F_n(x) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |L((x - X_i)/(M/n)) - S(x - X_i)| \stackrel{\text{def}}{=} C(x) \ .$$

Let $1 - L(u) < \theta$ for $u > \eta$, and let $L(u) < \theta$ for $u < -\eta$, where $\theta > 0$ is arbitrary, and $\eta > 0$ depends upon θ. Then,

$$\begin{aligned}
\sup_x C(x) &\leq \theta + \sup_x \left(F_n(x + M\eta/n) - F_n(x - M\eta/n)\right) \\
&\leq \theta + \sup_x |F_n(x + M\eta/n) - F(x + M\eta/n)| \\
&\quad + \sup_x |F_n(x - M\eta/n) - F(x - M\eta/n)| \\
&\quad + \sup_x \left(F(x + M\eta/n) - F(x - M\eta/n)\right) \\
&= \theta + 2 \sup_x |F_n(x) - F(x)| + \sup_x \left(F(x + 2M\eta/n) - F(x)\right) \ .
\end{aligned}$$

The last term tends to zero with n by the uniform continuity of F. The middle term tends to zero almost surely by the Glivenko-Cantelli lemma. As θ was arbitrary, we conclude that $\sup_x C(x) \to 0$ almost surely. Therefore,

$$\sup_{h \leq M/n} \sup_x \left| \int_{-\infty}^{x} f_{nh} - F_n(x) \right| \to 0 \text{ almost surely } ,$$

and thus, with probability one,

$$\inf_{g \in \mathcal{G}_a^*, h \leq M/n} \int |f_{nh} - g| \geq a$$

for all n large enough.

We combine the two previous results to conclude that for n large enough, and a as picked earlier,

$$\mathbf{P} \left\{ \inf_{g \in \mathcal{G}, h \leq M/n} \int |f_{nh} - g| \geq \min(a, c(M)/4) \right\} \geq \delta \ .$$

This conclude the proof of Lemma 2. □

The condition of Lemma 2 in turn is satisfied for many interesting classes. Assume for example that for each pair of densities $f, g \in \mathcal{G}$, the function $f - g$ changes sign at most k times. That is, there are numbers $a_1 < a_2 < \cdots < a_k$, such that on each interval (a_i, a_{i+1}), with $0 \leq i \leq k$ and $a_0 = -\infty$ and $a_{k+1} = \infty$, $f - g$ is of one sign. For such classes, we note that

$$\int |f - g| \leq (k+1) \sup_x |F(x) - G(x)| \ .$$

Clearly, the condition of Lemma 2 then holds. Examples of such classes include:

A. The class of all normal densities.

B. The class of all mixtures of N normal densities, with N fixed.

C. The class of all gamma densities.

D. The class of all Pearson densities (beta, gamma, Pearson IV, beta of the second kind, Student t, and normal).

E. The class of all Burr densities.

F. The exponential family defined by densities of the form

$$f(x) = e^{\sum_{i=0}^{N} a_i x^i}$$

for some constants a_i and N fixed.

We may also take finite unions of such classes and even finite mixtures of densities from such classes. Our estimate allows us to make a useful link with mixture distributions—for more on this, see for example Titterington, Smith and Makov (1985). We present two further lemmas offering sufficient conditions. These are geared towards nonparametric classes, to which the sign change argument does not apply.

LEMMA 3. *The continuity-of-norms condition holds when $\mathcal{G}$ is sequentially relatively compact in the L_1 topology, i.e., when for every sequence $(g_n) \subset \mathcal{G}$, there exists a subsequence (g_{n_k}) and a nonnegative measurable function g, such that $\int |g_{n_k} - g| \to 0$. (By Lemma 2, this in turn implies condition (ii) of Theorem 1.)*

PROOF. We take the notation of the proof of Lemma 2. We will use reductio ad absurdum. Assume that the continuity-of-norms condition does not hold. Then there exists a positive ε and a sequence $(g_n) \subset \mathcal{G}$ such that

$$\sup_x |G_n(x) - F(x)| \to 0 \ , \ \int |f - g_n| > \varepsilon.$$

Since $\sup_x |G(x) - G_{n_k}(x)| \leq \int |g - g_{n_k}|$, our condition in this Lemma implies that G_{n_k} converges to G with the supremum norm. But we already noted that G_{n_k} converges to F in this norm. Hence $F = G$ and thus $f = g$ almost everywhere. As a consequence,

$$\int |f - g_{n_k}| \to 0$$

which contradicts

$$\int |f - g_n| > \varepsilon \ . \ \square$$

Lemma 3 is useful to prove that some Lipschitz classes fulfill condition (ii). This will be shown in a later section. Lemma 3 cannot be applied to any class $\mathcal{G}$ that contains a density $g(.)$ and all its scale transforms $ag(a.)$ or a density g and all its translates $g(. - a)$. For this reason, we also offer another sufficient condition and a theorem by Kolmogorov.

LEMMA 4. *Assume $f \in \mathcal{G}$. The continuity-of-norms condition holds if for every sequence $(g_n) \subset \mathcal{G}$ with $\sup_x |G_n(x) - F(x)| \to 0$ (where G_n and F are the cumulative distribution functions for g_n and f), there exists a subsequence (g_{n_k}) and a nonnegative measurable function, g, such that*

$$\int |g_{n_k} - g| \longrightarrow 0 \ .$$

PROOF. Mimic the proof of Lemma 3. $\square$

THEOREM 2 (Kolmogorov; see Yosida, 1980). *A class of functions S in L_1 is relatively compact if*

(a) *There exist a constant C such that $\int |g| \leq C$ for all $g \in S$.*

(b) $\lim_{x \to 0} \int |g(x+y) - g(y)| dy = 0$, *uniformly in $g \in S$.*

(c) $\lim_{B \to \infty} \int_{|y|>B} |g(y)| dy = 0$, *uniformly in $g \in S$.*

5. Sufficient conditions for condition (i).

Condition (i) makes our estimate theoretically interesting, as we must choose K different enough from any density in $\mathcal{G}$. K should also not occur as a limit of densities in $\mathcal{G}$. To make things more precise, we define closure($\mathcal{G}$) as the class of densities g that may be approximated in L_1 by a sequence of densities from $\mathcal{G}$. Thus, classes $\mathcal{G}$ that are dense in the space of all densities in the L_1 sense have closure$(\mathcal{G}) = \{$all densities$\}$. These will be useless for our method, and should be avoided. Examples include the class of all analytic densities, the class of all Lipschitz densities, or the class of all indecomposable densities. We define $\mathcal{G}_0$ as the class of densities g that are L_1 limits of sequences of densities of the form $(1/a_n)g_n(x/a_n)$, where $g_n \in \mathcal{G}$ and $a_n \to 0$ is an arbitrary sequence.

LEMMA 5. *Condition (i) of Theorem 1 holds if for every $h > 0$, $K_h * f \notin$ closure$(\mathcal{G})$, and if $K \notin \mathcal{G}_0$.*

PROOF. Take $\delta > 0$ as in condition (i). We argue again by creating an absurdum. Assume that there exist sequences $(g_n) \subset \mathcal{G}$ and $h_n \geq \delta$ such that $\int |g_n - K_{h_n} * f| \to 0$. Assume first that (h_n) has a finite limit $h \geq \delta$ along a subsequence. As $\int |K_{h_n} * f - K_h * f| \leq \int |K_{h_n} - K_h| \to 0$ for all integrable K along this subsequence, we note that $\int |g_n - K_h * f| \to 0$. This is impossible, as $K_h * f \notin$ closure$(\mathcal{G})$. Thus, we must have $h_n \to \infty$. But then $\int |K_{h_n} * f - K_{h_n}| \to 0$ as in the proof of Lemma 1. Therefore, $\int |h_n g_n(xh_n) - K(x)|\, dx \to 0$. This implies $K \in \mathcal{G}_0$, which is impossible. □

6. When f is not in $\mathcal{G}$.

The applications we have in mind when we wrote this paper all involve the sure knowledge that $f \in \mathcal{G}$. In this section, we briefly discuss strategies to follow when f is possibly not in $\mathcal{G}$. A bit of thought shows that the present method can only work when none of the densities $f * K_h$, $h \geq \delta$, is closer to $\mathcal{G}$ than f, for any $\delta > 0$. This seems to be difficult to enforce. However, we may be able to detect that f is not in $\mathcal{G}$ by a simple device, following the ideas of Devroye (1989b). We consider three density estimates, g_n (as above), f_n' and f_n'', where f_n' and f_n'' are ordinary kernel estimates based upon kernels K and L that have Fourier transforms that do not coincide in any neighborhood of the origin—this condition is identical to that imposed in the double kernel method (Devroye, 1989a). The latter two estimates have an identical bandwidth H' found by minimizing $\int |f_n' - f_n''|$ with respect to the smoothing parameter over $(0, \infty)$, where ties are broken arbitrarily. It is known from that reference that $\int |f_n' - f_n''|$, $\int |f_n' - f|$ and $\int |f_n'' - f|$ all tend to zero almost surely if the smoothing factor is thus selected. There are no conditions on f. We might select the most suitable estimate as follows. If f_{nH} is the kernel estimate discussed in this paper, then define

$$f_n^* = \begin{cases} g_n & \text{if } \int |f_{nH} - g_n| < \int |f_n' - f_n''| \\ f_n' & \text{otherwise} . \end{cases}$$

To have a satisfactory statistical procedure or to accept or reject the hypothesis that $f \in \mathcal{G}$, we need something like

$$\mathbf{P}\{f_n^* \equiv g_n\} \to \begin{cases} 1 & \text{if } f \in \mathcal{G} \\ 0 & \text{if } f \notin \mathcal{G} . \end{cases}$$

This is nearly impossible unless we impose the condition

$$\text{closure}(\mathcal{G}) = \mathcal{G} . \tag{2}$$

Assume for example that $\mathcal{G}$ is the class of all unit variance normal densities with rational mean. The property above states that we should be able to asymptotically decide whether the mean of a normal is rational or irrational! Such classes must therefore be excluded, and this is why (2) is so natural.

THEOREM 3. *Let f be an arbitrary unknown density. Assume that $\mathcal{G}$ satisfies (2) and that conditions (i) and (ii) of Theorem 1 hold. Then $\int |f_n^* - f| \to 0$ in probability. Furthermore, if $f \notin \mathcal{G}$, $\mathbf{P}\{f_n^* \equiv g_n\} \to 0$ as $n \to \infty$.*

Theorem 3 states that the selection method suggested above protects us from disasters by rejecting g_n with high probability when $f \notin \mathcal{G}$. When $f \in \mathcal{G}$, we can construct small parametric classes $\mathcal{G}$ and obtain $\mathbf{P}\{f_n^* \equiv g_n\} \to 1$ as $n \to \infty$. However, for larger classes $\mathcal{G}$, the picture is not as clear. One would have to study the rate of convergence of g_n to $f \in \mathcal{G}$, and there is no room in the present article for this. If $\int |g_n - f|$ tends to zero faster than $\int f_n' - f_n''|$, then we would have the sought guarantee.

PROOF OF THEOREM 3. We only provide a rough outline. For $f \in \mathcal{G}$, Theorem 3 follows from Theorem 1 and the results on double kernel estimates given by Devroye (1989a) since both g_n and f_n' are L_1 consistent. So, we assume $f \notin \mathcal{G}$. It suffices to exhibit an $\epsilon > 0$ such that

$$\lim_{n\to\infty} \mathbf{P}\left\{ \inf_{h>0, g\in\mathcal{G}} \int |f_{nh} - g| > \epsilon \right\} = 1 ,$$

as this implies

$$\lim_{n\to\infty} \mathbf{P}\{f_n^* \equiv g_n\} = 0$$

and thus

$$\int |f_n^* - f| \to 0 \text{ in probability.}$$

By the closure property (2), we know that

$$\xi \stackrel{\text{def}}{=} \inf_{g\in\mathcal{G}} \int |g - f| > 0 .$$

Fix $\delta > 0$. First we find a constant c such that

$$\mathbf{P}\left\{ \sup_{1/(cn)\le h\le c} \int |f_{nh} - f| > \frac{\xi}{2} \right\} < \delta$$

for all n large enough. The existence of such a constant follows again from results of Devroye (1989a). By definition of ξ and the triangle inequality, we note that

$$\mathbf{P}\left\{ \inf_{1/(cn)\le h\le c, g\in\mathcal{G}} \int |f_{nh} - g| > \frac{\xi}{2} \right\} > 1 - \delta$$

for all n large enough. By condition (ii), there exists $a > 0$ such that

$$\mathbf{P}\left\{ \inf_{h\le 1/(cn), g\in\mathcal{G}} \int |f_{nh} - g| > a \right\} > 1 - \delta$$

for all n large enough. Define

$$\mu = \inf_{h\ge c, g\in\mathcal{G}} \int |g - f * K_h| .$$

By condition (i), $\mu > 0$. Next, we show that

$$\mathbf{P}\left\{\inf_{h\geq c, g\in\mathcal{G}} \int |f_{nh} - g| > \frac{\mu}{2}\right\} > 1-\delta$$

for all n large enough. Indeed.

$$\begin{aligned}
&\mathbf{P}\left\{\inf_{h\geq c, g\in\mathcal{G}} \int |f_{nh} - g| > \frac{\mu}{2}\right\} \\
&\geq \mathbf{P}\left\{\inf_{h\geq c, g\in\mathcal{G}} \int |f*K_h - g| > \mu\right\} - \mathbf{P}\left\{\inf_{h\geq c} \int |f*K_h - f_{nh}| > \frac{\mu}{2}\right\} \\
&\stackrel{\text{def}}{=} I - II\ .
\end{aligned}$$

Clearly, $II = o(1)$ by Lemma 1, and $I = 0$ by definition of μ.

Collecting the three bounds, we see that for all n large enough,

$$\mathbf{P}\left\{\inf_{h, g\in\mathcal{G}} \int |f_{nh} - g| > \min\left(\frac{\xi}{2}, a, \frac{\mu}{2}\right)\right\} > 1-3\delta\ .$$

This is what we set out to prove. □

Condition (i) in Theorems 1 and 3 will be illustrated below. In Theorem 3, it must hold for all f, while in Theorem 1, it needs only hold for $f \notin \mathcal{G}$. Condition (i) is satisfied if K is chosen as a function of the structure of $\mathcal{G}$. It should be sufficiently different from the members of $\mathcal{G}$ in a manner to be made precise below. A helpful tool for checking (i) is the fact that if f and g are densities with characteristic functions ϕ and ψ respectively, then

$$\int |f-g| \geq \sup_t |\phi(t) - \psi(t)|\ . \tag{3}$$

Two simple examples follow.

- Assume that $\mathcal{G}$ is any family of densities with support on $[0,\infty)$. If we take K symmetric and unimodal, for any f, $f*K_h$ puts some weight on the negative axis, and increasingly so as $h\uparrow$. Therefore, condition (iii) is satisfied.
- Assume that $\mathcal{G}$ is the family of normal densities on the real line. Using (3), it is easy to see that (iii) holds if we take for K the Cauchy density, as its characteristic function $e^{-|t|}$ behaves sufficiently differently from the normal characteristic functions in a neighborhood of the origin.

7. Examples: parametric classes.

We return to Theorem 1 and the situation $f \in \mathcal{G}$.

When $\mathcal{G}$ is the class of normal densities with mean zero, to satisfy condition (i), it suffices to take any kernel K that is not centered normal (check Lemma 5). Condition (ii) is trivially satisfied.

When $\mathcal{G}$ is the class of gamma densities with shape parameter a, and free translation and scale parameters, condition (ii) holds. Condition (i) tells us that we should exclude for K any gamma density, as well as any density in $\mathcal{G}_0$. The latter class is just the class of all normal densities. Therefore, K may be any density except gamma or normal.

Assume next that $\mathcal{G}$ is the class of mixtures of three normal densities. Used in many models and approximations, finite mixtures of normals are an important family. Condition (i) holds if K is non-normal. Finally, condition (ii) holds. This example has been studied from a theoretical viewpoint and via simulation in Cao, Cuevas and Fraiman (1993).

8. $\mathcal{G}$ is the class of Lipschitz densities on $[0, 1]$.

If we take K with infinite support, then condition (i) necessarily holds. Condition (ii) holds when $\mathcal{G}$ is the subclass of Lipschitz densities on $[0, 1]$ with fixed constant λ. To prove this, observe that this class is closed, bounded and equicontinuous and, by the Ascoli-Arzelà theorem, a compact set of $C[0, 1]$. Hence, Lemma 3 applies.

9. $\mathcal{G}$ is the class of Lipschitz densities on the real line.

This class is too big. In fact, observe that the class $\mathcal{G}$ of all Lipschitz densities on the line is dense in the L_1 sense for the class of all densities, so that there is no hope to satisfy the condition of Lemma 5 or to satisfy condition (i) in Theorem 1. Hence the necessity to use proper restrictions. It is easy to verify that when f is Lipschitz with constant λ, then $f * K_h$ is Lipschitz with the same constant regardless of the form of K_h. Condition (i) is always violated. A little thought shows that our method would in fact not be consistent.

10. $\mathcal{G}$ is the class of symmetric unimodal densities with mode at 0.

Condition (i) is violated when K too is symmetric unimodal, as the convolution of two symmetric unimodal densities is again symmetric unimodal (see, e.g., Ibragimov, 1956). An acceptable choice for condition (i) is a shifted negative exponential density, for example. To prove condition (ii) in this context (as well as for the class in the next example), we apply Lemma 4. It suffices to verify the conditions in Theorem 2 for $S = (g_n)$, where g_n and f are density functions with unique modes and satisfy that $\sup_x |G_n(x) - F(x)| \longrightarrow 0$.

Condition (a) in Theorem 2 is evident. To prove condition (c), fix $\varepsilon > 0$ and let $B > 0$. Then

$$\int_{|y|>B} g_n(y)dy = G_n(-B) + 1 - G_n(B) < F(-B) + 1 - F(B) + \varepsilon < 2\varepsilon$$

for n and B large enough. Hence condition (c) follows.

It only remains to prove condition (b). Let $x > 0$, and let δ_n denote the unique mode of g_n. Define the density $\gamma_n(x) = g_n(x + \delta_n)$, with unique mode at 0 . Now, the unimodality of γ_n implies

$$\begin{aligned}
\int |g_n(x+y) - g_n(x)|dy &= \int |\gamma_n(x+y) - \gamma_n(y)|dy \\
&= \Gamma_n(0) - \Gamma_n(-x) + \int_{-x}^{0} |\gamma_n(x+y) - \gamma_n(y)|dy + \Gamma_n(x) - \Gamma_n(0) \\
&\leq \Gamma_n(x) - \Gamma_n(-x) + \int_{-x}^{0} (\gamma_n(x+y) + \gamma_n(y))dy \\
&= 2(\Gamma_n(x) - \Gamma_n(-x)) \\
&\leq 2(F(x+\delta_n) - F(-x+\delta_n) + \varepsilon),
\end{aligned}$$

for n large enough. Condition (b) follows immediately.

11. $\mathcal{G}$ is the class of log-concave densities with a unique mode.

To prove condition (ii) we may use the same argument as in the previous example. As kernel, we may take any density K with tails decreasing at a polynomial rate or slower (such as the Cauchy density). This insures that for all $h > 0$, $K_h * f \notin \mathcal{G}$ and that $K \notin \text{closure}(\mathcal{G})$, because for any log-concave density g, there exist finite constants a and b such that $g(x) \leq ae^{-b|x|}$. By Lemma 5, condition (i) then holds. Again, we are picking K very different from any density in $\mathcal{G}$.

12. Acknowledgment.

Discussions with Ricardo Fraiman, Antonio Cuevas and Wenceslao Gonzalez-Manteiga are gratefully acknowledged. The reports of both referees were extremely helpful as well. One referee suggested considerable improvements in sections 8, 9 and 10. We thank them.

13. References.

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

R. Beran, "Minimum Hellinger distance estimates for parametric models," *Annals of Statistics*, vol. 5, pp. 445–463, 1977.

R. Cao, A. Cuevas and R. Fraiman, "Smoothed minimum distance estimation of mixture parameters," *Computational Statistics and Data Analysis*, vol. 20, pp. 611–6311995.

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

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

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

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

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

L. Devroye and A. Berlinet, “The double kernel estimate revisited,” Technical Report, School of Computer Science, McGill University, Montreal, 1993.

I.A. Ibragimov, “On the composition of unimodal distributions,” *Theory of Probability and its Applications*, vol. 1, pp. 255–260, 1956.

I. Olkin and C.H. Spiegelman, “A semiparametric approach to density estimation,” *Journal of the American Statistical Association*, vol. 82, pp. 858–865, 1987.

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

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

H. Scheffé, “A useful convergence theorem for probability distributions,” *Annals of Mathematical Statistics*, vol. 18, pp. 434–458, 1947.

D.M. Titterington, A.F.M. Smith and U.E. Makov, “Statistical Analysis of Finite Mixture Distributions,” John Wiley, New York, 1985.

K. Yosida, “Functional Analysis,” Springer-Verlag, Berlin, 1980.