The Annals of Statistics
1997, Vol. 25, No. 6, 2300–2312

A PLUG-IN APPROACH TO SUPPORT ESTIMATION

BY ANTONIO CUEVAS[1] AND RICARDO FRAIMAN[2]

Universidad Autónoma de Madrid and Universidad de la República, Montevideo

We suggest a new approach, based on the use of density estimators, for the problem of estimating the (compact) support of a multivariate density. This subject (motivated in terms of *pattern analysis* by Grenander) has interesting connections with detection and clustering.

A natural class of density-based estimators is defined. Universal consistency results and convergence rates are established for these estimators, with respect to the usual measure-based metric d_μ between sets. Further convergence rates (with respect to both d_μ and the Hausdorff metric d_H) are also obtained under some, fairly intuitive, shape restrictions.

1. Introduction.

1.1. *The problem*: *background and motivation.* This paper is concerned with a problem of nonparametric set estimation: let f be a (Lebesgue) probability density on $\Re^d$. Define, as usual, the support S of f as the minimal closed set having f-probability 1. Assume that S is compact. We want to estimate S from a random sample $X_1, \dots, X_n$ of f. Some references are Geffroy (1964), Chevalier (1976), Devroye and Wise (1980), Grenander (1981), Cuevas (1990), Korostelev and Tsybakov (1993), Mammen and Tsybakov (1995), Korostelev, Simar and Tsybakov (1995), Härdle, Park and Tsybakov (1995), Polonik (1995) and Tsybakov (1997).

A very simple and intuitive estimator is defined by

$$\hat{S}_n = \bigcup_{i=1}^{n} B(X_i, \varepsilon_n), \tag{1}$$

where $B(x, a)$ denotes the closed ball centered at x with radius a and ε_n is a sequence of smoothing parameters.

Some suitable criterion of proximity between sets is required in order to analyze the performance of the estimates. A standard choice is the *measure-based distance* defined by

$$d_\mu(T, S) = \mu(T \Delta S), \tag{2}$$

Received March 1996; revised November 1996.

[1]Research partially supported by DGICYT Spanish Grants PB94-0179 and PB95-0826.

[2]Research partially supported by Grant 37 from the CONICYT at Montevideo.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Support estimation, kernel density estimators, Hausdorff metric, L_1-approach, multivariate spacings.

where Δ denotes the symmetric difference, μ is a measure on $\Re^d$ (very often $\mu = \mu_L$, the Lebesgue measure) and P_X is the common underlying distribution of the observations X_i (we identify T and S if they differ by a null set).

Devroye and Wise (1980) have proved d_μ-consistency results for the *naïve estimator* (1) under some conditions on the sequence $\{\varepsilon_n\}$ which are analogous to those imposed on the bandwidth parameters in kernel density estimation [see, e.g., Devroye and Györfi (1985)]. These results are universal: they hold for any S and any probability μ such that $\mu \ll P_X$ (on S).

However, some assumptions on both the density f and the shape of the support S are usually needed to get d_μ-convergence rates or optimality results [see Korostelev and Tsybakov (1993), Chapter 7, and Härdle, Park and Tsybakov (1995)].

Another natural criterion of proximity between sets is given by the *Hausdorff metric*,

$$d_H(T, S) = \inf\{\varepsilon > 0\colon T \subset S^\varepsilon \text{ and } S \subset T^\varepsilon\}, \tag{3}$$

where S^ε denotes the union of all open balls with radius ε around points of S. This distance corresponds to an intuitive notion of "physical proximity" between sets. It has been used in different settings, including fractal theory and random sets. Some results on d_H-based support estimation can be found in Cuevas (1990), Korostelev and Tsybakov (1993) and Korostelev, Simar and Tsybakov (1995).

As for the practical applications of support estimation let us mention cluster analysis [Hartigan (1975)] and detection of abnormal behavior in a system [Devroye and Wise (1980)]. A more detailed account, including references to other related problems, can be seen in Korostelev and Tsybakov [(1993), pages 182–183].

1.2. *Connections with density estimation: our proposal.* An explicit connection between the support problem and the theory of density estimation was suggested to us by Dobrow (1992) [for related ideas see also Sager (1979)], who, basically, proposed a *plug-in* idea to address the estimation of the support as a by-product of the usual nonparametric kernel estimation. In a way, the situation would be similar to that arising in the estimation of some functions or quantities of interest (f', $\int f^2$, mode of $f, \ldots$) which are estimated by replacing the unknown density by an estimator f_n in the corresponding functional [for a related approach in the nonparametric regression setting see, e.g., Boularan, Ferré and Vieu (1995)].

Dobrow's proposal was to estimate S by $\tilde{S}_n = \{\hat{f}_n > 0\}$, where $\hat{f}_n$ is a kernel density estimator,

$$\hat{f}_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \equiv \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \tag{4}$$

where $h = h_n$ is the sequence of smoothing parameters (bandwidths), K is the kernel function and $K_h(t) = (1/h^d)K(t/h)$.

An estimator of type $\tilde{S}_n$ is a very simple and natural choice but it presents two appreciable limitations. First, we are restricted to using compact-supported K in order to avoid the useless estimator $\tilde{S}_n = \Re^d$. Second, if $S(K) = \{K > 0\}$ is bounded, the estimator $\tilde{S}_n = \{\hat{f}_n > 0\}$ is again a finite union of type (1), where the balls $B(X_i, h)$ are replaced by $X_i + hS(K)$.

We consider here a modified version of the above idea which overcomes these problems by introducing a new *tuning parameter*, in addition to the smoothing parameter of $\hat{f}_n$. To be concrete, our proposal is to estimate S by

$$S_n = \{f_n > \alpha_n\}, \tag{5}$$

where f_n is a nonparametric density estimator (usually, but not necessarily, of kernel type: in this case we will use the notation $\hat{f}_n$ instead of f_n) and α_n is a sequence converging to zero. A related idea was also suggested by Dobrow (1992) under convexity restrictions on S. We assume throughout that f_n is a bona fide estimator (i.e., $f_n \geq 0$, $\int f_n = 1$). As we will show below, asymptotic results (with respect to both d_μ and d_H) for the estimator (5) can be obtained under very general conditions. The additional parameter α_n provides more flexibility in the shape of S_n which typically will have [unlike the estimator (1)] a differentiable boundary. Hence, (5) can be considered as a smoothed version of (1) in the same spirit as the kernel density estimator compares with the simpler (but rougher) histogram.

This paper is organized as follows. Consistency and convergence rates for S_n with respect to d_μ are given in Section 2. Section 3 is devoted to analyzing the convergence rates with respect to the Hausdorff metric d_H. Some final remarks are given in Section 4.

2. Consistency and d_μ-convergence rates. Throughout this section we consider the distance d_μ defined in (2) by taking $\mu = \mu_L$, the Lebesgue measure on $\Re^d$. Unless otherwise stated, the arrow $\longrightarrow$ denotes convergence as n tends to infinity.

2.1. *Universal results.* We first prove a theorem which provides three results on strong consistency and convergence rates (in probability) for the estimator (5) where f_n is a general density estimate. These results are *universal* in the sense that they impose *no restriction* on the support S, except a very mild one [in part (a)] which only excludes pathological cases. As for the density f, we will impose [in parts (b) and (c)] two conditions related to the way in which f "decreases to the ground."

To be concrete, we will use the following assumptions:

(S1) The Lebesgue measure $\mu_L(E_0) = 0$, where $E_0 = \{x \in S\colon f(x) = 0\}$.

This condition excludes only those pathological cases where the set $\{f > 0\}$ is far away from the support S; for instance, let $A \subset [0, 1]$ be an open set dense in $[0, 1]$ such that $0 < \mu_L(A) < 1$ (A could be, e.g., the complement in $[0, 1]$ of a Cantor-type set of positive measure). Let f be the uniform density constant on A and null on A^c. The support of f is $[0, 1]$ and $\mu_L(E_0) = 1 - \mu_L(A) > 0$.

(R1) $\alpha_n^{-1} \int |f_n - f| \longrightarrow 0$, a.s. (resp., in probability).

(R2) $\rho_n \int |f_n - f| \longrightarrow 0$ in probability, where ρ_n is a sequence such that $\rho_n \to \infty$ and $\rho_n \alpha_n a_n \to 0$, where $a_n := \mu_L(\{f \leq 2\alpha_n\} \cap S)$.

We will also need the following definition, which has to do with the *sharpness* in the decrease of f to zero: sharper cases correspond to faster decreases. Related concepts can be found in Härdle, Park and Tsybakov (1995) and Hall (1982).

DEFINITION 1. Let f be a density on $\Re^d$ with compact support S; define $f^*(t) = \mu_L(\{f < t\} \cap S)$. We will say that $\gamma > 0$ is the *sharpness order* of f if $f^*(t)$ has the same order (when $t \to 0^+$) as t^γ, that is,

$$0 < \liminf_{t \to 0^+} \frac{f^*(t)}{t^\gamma} \leq \limsup_{t \to 0^+} \frac{f^*(t)}{t^\gamma} = c$$

for some finite constant c.

We will denote by $\mathscr{S}_\gamma(S)$ the space of densities with support S and sharpness order γ. Finally, denote

$$\mathscr{S}_\infty(S) = \{f\colon f^*(t) = o(t^\gamma) \text{ when } t \to 0^+, \text{ for all } \gamma > 0\}.$$

Let us observe that some densities do not belong to any space $\mathscr{S}_\gamma$; this is the case of $f(x) = c_1 \exp(-1/x)$, $x \in (0, 1]$. However, a density of type $f(x) = c_2\, x^{1/p}$, $x \in [0, 1]$, $p > 0$, satisfies $f \in \mathscr{S}_p([0, 1])$. If f is bounded away from zero on S, then $f \in \mathscr{S}_\infty(S)$.

THEOREM 1. *Let f be a density on $\Re^d$ with a compact support S. Given a sequence $\{f_n\}_{n \geq 1}$ of density estimators, define an associated sequence of support estimators $S_n = \{f_n > \alpha_n\}$, where $\alpha_n \downarrow 0$.*

(a) *If* (S1) *and* (R1) *hold, then $d_\mu(S_n, S) \longrightarrow 0$, a.s.* (*resp., in probability*).
(b) *If* (R2) *holds then $\beta_n d_\mu(S_n, S) \longrightarrow 0$, in probability, where*

$$\beta_n = \frac{1}{a_n + (\rho_n \alpha_n)^{-1}}. \tag{6}$$

(c) *Let us suppose that $f \in \mathscr{S}_\gamma(S)$. Assume that* (R2) *holds with $\rho_n = n^\rho$ ($\rho > 0$) and take $\alpha_n = n^{-\alpha}$, where $0 < \alpha < \rho$ and $\rho - \alpha < \alpha\gamma$. Then $n^{\rho-\alpha} d_\mu(S_n, S) \longrightarrow 0$, in probability. If $f \in \mathscr{S}_\infty(S)$, the estimation of S can be performed at any convergence rate of type n^β with $\beta < \rho$.*

PROOF. Define $A_n = \{x\colon |f_n(x) - f(x)| \geq \alpha_n\}$. By considering a suitable partition of $S_n \,\Delta\, S$ and taking into account $\mu_L(S_n \cap S^c \cap A_n^c) = 0$ and $S_n^c \cap S \cap A_n^c \subset \{f \leq 2\alpha_n\} \cap S$, we get

$$d_\mu(S_n, S) \leq \mu_L(A_n) + \mu_L(S_n^c \cap S \cap A_n^c) \leq \mu_L(A_n) + a_n.$$

From (S1) $a_n \downarrow 0$, since $\{f \le 2\alpha_n\} \cap S \downarrow E_0$. We also have $\mu_L(A_n) \to 0$, a.s.: this follows directly from (R1) since

$$\alpha_n^{-1} \int |f_n - f| \ge \mu_L(A_n) + \alpha_n^{-1} \int_{A_n^c} |f_n - f|.$$

This concludes the proof of part (a). To prove (b) take β_n as given in (6). Then, for any $\delta > 0$ and n large enough,

$$\begin{aligned} P\{\beta_n d_\mu(S_n, S) > \delta\} &\le P\left\{\mu_L(A_n) + a_n > \frac{\delta}{\beta_n}\right\} \\ &\le P\left\{\frac{1}{\alpha_n(\delta/\beta_n - a_n)} \int |f_n - f| > 1\right\}, \end{aligned}$$

where we have used that $\delta/\beta_n - a_n$ is eventually positive (which follows from the assumption $\rho_n a_n \alpha_n \to 0$). Now, from (R2) we conclude the convergence to zero of the right-hand side of the last inequality.

Finally, (c) follows directly from expression (6): since $\alpha_n = n^{-\alpha}$, the assumption $f \in \mathscr{S}_\gamma(S)$ implies that the sequence a_n is of (exact) order $n^{-\alpha\gamma}$ and, therefore, β_n is of exact order $n^{\rho-\alpha}$. □

REMARKS. (a) In the case where $\{f_n\}$ is a sequence of d-variate kernel estimators, assumption (R1) would be typically fulfilled (in probability) by a sequence $\{\alpha_n\}$ of type $\alpha_n^{-1} = o(n^{2/(d+4)})$ [see Holmströn and Klemelä (1992)].

(b) According to (6), we need $\rho_n \alpha_n \to \infty$ in order to ensure $\beta_n \to \infty$. So α_n must go to zero slowly enough, depending on the convergence rate ρ_n of the density estimator.

(c) The sequence a_n in (R2) depends directly on the way in which f "decreases to the ground." In the *sharp* cases where f is bounded away from zero we have $a_n = 0$ eventually. This is the most favorable situation. In general, the slower the decreasing to zero of a_n, the worse the convergence rate β_n one can get in (6). This is fairly intuitive, since a slow decrease of a_n is associated with the existence of wide "empty" areas of low probability (where f is very small) which will be underrepresented in the sample.

The purpose of Theorem 1(c) is to quantify these ideas in terms of the value of three real parameters ρ, α and γ associated, respectively, with the convergence rate of f_n, the tuning parameter α_n and the sharpness in the decay to zero of f. The inequality $\rho - \alpha < \alpha\gamma$ can be seen as the "support version" of the typical trade-off arising in all problems of nonparametric smoothing: whereas the expression of the convergence rate $\beta_n = n^{\rho-\alpha}$ suggests that α should be chosen as small as possible, the bound $\alpha\gamma$ goes in the opposite sense. Observe that this bound is not operative when the association between f and S is *sharp* ($\gamma = \infty$). In this case we can estimate the support with a convergence rate arbitrarily close to n^ρ by taking α small enough.

2.2. *Convergence rates under shape restrictions.* We will establish here a result about rates for d_μ-convergence in mean for the support estimator (5).

It holds in the case where the auxiliary density estimate f_n is of kernel type, under some shape restrictions on the support S.

In particular, we will need the following notion of standardness, which has been considered by Cuevas (1990) in the support estimation setting. The intuitive idea is to exclude some pathological sets (for instance, those having infinitely many sharper and sharper peaks). This notion is related to the *cone condition* and the $\mathscr{G}_N$ classes introduced in Korostelev and Tsybakov [(1993), page 137].

DEFINITION 2. A bounded set $S \subset \Re^d$ is said to be *standard* if for every $\lambda > 0$ there exists $\delta \in (0, 1)$ such that

$$\mu_L(B(x, \varepsilon) \cap S) \geq \delta \mu_L(B(x, \varepsilon)) \qquad \forall\, x \in S,\ 0 < \varepsilon \leq \lambda.$$

Another geometrical condition which will appear in a natural way has to do with the volume increase from S to S^h, as measured by the *blowing-up function* $\Delta(S; h) := \mu_L(S^h) - \mu_L(S)$. Clearly, this function provides some information about the complexity of S: the simpler the structure of S, the smaller $\Delta(S, h)$. A typical behavior is $\Delta(S; h) = O(h)$; this is the case when S is a finite union of convex sets: this follows from the isoperimetric inequality [see, e.g., Bhattacharya and Ranga Rao (1976), Theorem 3.1, page 24].

The following condition on the kernel K also will be used.

(K1) $c_1 I_{B(0, r_1)}(t) \leq K(t) \leq c_2 I_{B(0, r_2)}(t)$, for some constants c_1, $c_2 > 0$ and $0 < r_1 < r_2$, where I_A denotes the indicator function of the set A.

Finally, for every $\varepsilon > 0$ let us denote by $R(S; \varepsilon)$ the minimum number of balls, centered at points of S with radius ε, required to cover S. We have the following result:

THEOREM 2. *Let $S_n = \{\hat{f}_n > \alpha_n\}$, where $\hat{f}_n$ is a kernel density estimator whose kernel K fulfils* (K1).

Assume that the density f is bounded away from zero on its support S which is supposed to be standard. Then

$$Ed_\mu(S_n, S) \leq c_3 h_n^d R(S; r_1 h_n/2) \exp(-c_4 n h_n^d) + \Delta(S; r_2 h_n), \tag{7}$$

for n large enough, where c_3 and c_4 are positive constants. As a direct consequence, if we assume in addition $\Delta(S; h_n) = O(h_n)$ as h_n tends to zero, we have

$$Ed_\mu(S_n, S) = O\big(\exp(-c_4 n h_n^d) + h_n\big). \tag{8}$$

Hence, by taking a suitable sequence h_n, one can obtain any rate of type $Ed_\mu(S_n, S) = O(n^{-s})$ with $0 < s < 1/d$.

PROOF. We have

$$\mu_L(S_n \Delta S) = \mu_L\{\hat{f}_n > \alpha_n,\ f = 0\} + \mu_L\{f > 0,\ \hat{f}_n \leq \alpha_n\}. \tag{9}$$

From assumption (K1), $x \in (S^{r_2 h})^c$ implies $\hat{f}_n(x) = 0$ (a.s.). Therefore the first term in the right-hand side of (9) is easily bounded,

$$\mu_L\{\hat{f}_n > \alpha_n,\ f = 0\} \leq \mu_L(S^{r_2 h_n}) - \mu_L(S) = \Delta(S; r_2 h_n) \quad \text{a.s.} \tag{10}$$

To handle the second term of (9) let us consider a minimal covering of S with balls $B(Z_j, r_1 h_n/2)$, $Z_j \in S$, $j = 1, \dots, R(S; r_1 h_n/2)$. Write $B_j := B(Z_j, r_1 h_n/2)$ and $R := R(S; r_1 h_n/2)$. Then

$$\mu_L(S \cap S_n^c) \leq \mu_L\left\{\left(\bigcup_{j=1}^{R} B_j\right) \cap S_n^c\right\} \leq \sum_{j=1}^{R} \mu_L(B_j \cap S_n^c). \tag{11}$$

Let

$$A_{n,j} = \left\{\frac{1}{nh_n^d} \sum_{i=1}^{n} I_{B_j}(X_i) > \frac{\alpha_n}{c_1}\right\}.$$

Observe that if $\omega \in A_{n,j}$, then $B_j \cap S_n^c(\omega) = \varnothing$; to see this take $t \in B_j$. Then

$$\begin{aligned} \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) &\geq \frac{1}{nh_n^d} \sum_{i=1}^{n} c_1 I_{B(t, r_1 h_n)}(X_i) \\ &= \frac{c_1}{nh_n^d} \#\{i\colon X_i \in B(t, r_1 h_n)\} \\ &\geq \frac{c_1}{nh_n^d} \#\{i\colon X_i \in B_j\} > \alpha_n. \end{aligned} \tag{12}$$

Hence, denoting $\lambda_1 = \mu_L(B(0,1))$, we have

$$\begin{aligned} E\left(\sum_{j=1}^{R} \mu_L(B_j \cap S_n^c)\right) &\leq E\left(\sum_{j=1}^{R} I_{A_{n,j}^c} \lambda_1 \left(\frac{r_1 h_n}{2}\right)^d\right) \\ &= \sum_{j=1}^{R} h_n^d \lambda_1 \left(\frac{r_1}{2}\right)^d P(A_{n,j}^c). \end{aligned} \tag{13}$$

We next find an upper bound for $P(A_{n,j}^c)$. If δ is the standardness constant of S (for a given $\lambda \geq \sup_n r_1 h_n/2$; see Definition 2) and $f > a > 0$ on S, we have

$$p_{j,n} := p_j = E\big(I_{B_j}(X_i)\big) = P(X_i \in B_j) > a\delta\left(\frac{r_1 h_n}{2}\right)^d \lambda_1 := a_1 h_n^d, \tag{14}$$

which entails $p_j/2 - \alpha_n h_n^d/c_1 > a_1 h_n^d/2 - \alpha_n h_n^d/c_1$; thus, since $\alpha_n \to 0$ and $a_1 > 0$, there exists n_0 such that $p_j/2 - \alpha_n h_n^d/c_1 > 0$, for all $n \geq n_0$ and

$$\begin{aligned} P(A_{n,j}^c) &= P\left\{\sum_{i=1}^{n} (p_j - I_{B_j}(X_i)) \geq np_j - \frac{nh_n^d \alpha_n}{c_1}\right\} \\ &\leq P\left\{\sum_{i=1}^{n} (p_j - I_{B_j}(X_i)) \geq \frac{np_j}{2}\right\} \quad \text{for } n \text{ large enough.} \end{aligned}$$

Now, using Bernstein's inequality [see, e.g. Shorack and Wellner (1986), page 855], we get

$$P(A^c_{n,j}) \le 2\exp\left(-\frac{3np_j}{28}\right). \tag{15}$$

Then, since $p_j > a_1 h^d$, from (15) and (10), denoting $c_3 = 2\lambda_1 r_1^d 2^{-d}$, $c_4 = 3a_1/28$, we get (7) and (8). $\square$

Inequality (7) has a direct intuitive interpretation: the simpler the set S, the faster it can be estimated. The covering function R (which, of course, is essentially the classical entropy) and the blowing-up function Δ are the relevant features in order to quantify the complexity of S.

A related result, with a different approach, has been established by Korostelev and Tsybakov [(1993), Theorem 7.2.2, page 184]: these authors obtain a result of type $\sup_{S\in\mathscr{G}} Ed_\mu(\hat{S}_n, S) = O((\log n/n)^{1/d})$ for the classical estimator $\hat{S}_n$, defined in (1), in the uniform case ($f \equiv c$), where $\mathscr{G}$ is a class of domains having piecewise Lipschitz boundaries with the number of pieces and the Lipschitz constants uniformly bounded.

3. Convergence rates with respect to d_H. In this section we obtain results of type

$$\beta_n d_H(S_n, S) \longrightarrow 0 \quad \text{a.s.}, \tag{16}$$

where $\beta_n \uparrow \infty$. The auxiliary density estimator is assumed to be of kernel type throughout.

Again the results here are very general: only mild assumptions (concerning boundedness and standardness) will be imposed on f and S.

We will also use the following assumptions on the kernel K and the bandwidths h_n of the kernel estimator $\hat{f}_n$:

(K2) The kernel K is a bounded density, uniformly Lipschitz on $\Re^d$, such that $\|t\|^d K(t)$ is bounded and there exist positive constants c_1, r_1 satisfying

$$c_1 I_{B(0,r_1)}(u) \le K(u), \tag{17}$$

where I_A stands for the indicator function of the set A.

(K3) $K(u)$ is a decreasing function of $\|u\|$ such that $\|u\|^{d+1}K(u) \to 0$ as $\|u\| \to \infty$.

(H1) $h_n \longrightarrow 0$ and $nh_n^d/\log n \longrightarrow \infty$, as $n \to \infty$.

We will also use the assumption that K is compact-supported, which must be understood in the "functional" sense that K is zero outside a compact set.

THEOREM 3. *Let us assume that the support S is a compact standard set and that f is bounded and there exists $a > 0$ such that $f > a$ on S.*

(a) *If* (K2), (K3) *and* (H1) *hold, then*

$$\beta_n d_H(S_n, S) \longrightarrow 0 \quad a.s., \tag{18}$$

for every sequence $\beta_n \uparrow \infty$ such that $\beta_n h_n \to 0$ and $\{\beta_n^{d+1} h_n/\alpha_n\}$ is bounded.

This conclusion also holds if (K3) *is replaced by the assumption that K is compact-supported. In this case the boundedness of the sequence $\{\beta_n^{d+1}h_n/\alpha_n\}$ is no longer required and one can achieve any rate β_n of type $o((n/\log n)^{1/d})$ by taking a suitable sequence h_n.*

(b) *This result cannot be improved by using the classical estimator* (1). *More generally, if we consider the estimator S_n with $\alpha_n = 0$ and a compact-supported K, then any sequence β_n such that $(n/\log n)^{1/d} = O(\beta_n)$ does not satisfy* (16), *not even in probability.*

PROOF. (a) Assume (K2), (K3) and (H1). Let us first prove that

$$\text{there exists } a_0 \text{ such that } \inf_{x\in S} \hat{f}_n(x) > a_0 \text{ eventually, a.s.} \tag{19}$$

By using the standardness of S, the fact that f is bounded away from zero on S and inequality (17), we have, for all $x \in S$ and $h_n < r_1$,

$$\begin{aligned} E\big(\hat{f}_n(x)\big) &= \int K_h(x-t)f(t)\,dt \geq a\int_S K_h(x-t)\,dt \\ &\geq c_1 a\int_S \frac{1}{h_n^d} I_{B(0,\,r_1)}\left(\frac{x-t}{h_n}\right)dt \geq \frac{c_1 a}{h_n^d}\delta\mu_L(B(x, r_1h_n)) \\ &= \frac{c_1 a}{h_n^d}\delta\mu_L(B(0,1))r_1^d h_n^d := 2a_0 > 0. \end{aligned}$$

The proof of (19) will be complete if

$$\sup_x |\hat{f}_n(x) - E(\hat{f}_n(x))| \longrightarrow 0 \quad \text{a.s., as } n\to\infty,$$

holds. A proof of this result [under hypothesis (H1) and the boundedness and Lipschitz conditions established in (K2)] can be found, for example, in Prakasa Rao [(1983), pages 185–187].

Now, in order to prove (18), let us take $\varepsilon > 0$; we will show that $S_n \subset S^{\varepsilon/\beta_n}$ a.s. eventually. To see this note that, for all $x \neq 0$,

$$\frac{1}{h_n^d}K\left(\frac{x}{\|x\|}\frac{\varepsilon}{\beta_n h_n}\right) = \frac{\beta_n^{d+1}h_n}{\varepsilon^{d+1}}K\left(\frac{x}{\|x\|}\frac{\varepsilon}{\beta_n h_n}\right)\left(\frac{\varepsilon}{\beta_n h_n}\right)^{d+1}.$$

Since $\beta_n h_n \to 0$, (K3) and the boundedness of $\{\beta_n^{d+1}h_n/\alpha_n\}$ imply

$$\forall\ \varepsilon > 0\ \exists\ n_1\colon K_h\left(\frac{x}{\|x\|}\frac{\varepsilon}{\beta_n}\right) < \alpha_n \qquad \forall\ n > n_1,\ \forall\ x \neq 0. \tag{20}$$

Note also that, if K is compact-supported, (20) follows without assuming the boundedness of $\{\beta_n^{d+1}h_n/\alpha_n\}$.

If $x \in S_n$ for $n > n_1$, then (20) and the assumption that $K(x)$ is a decreasing function of $\|x\|$ imply that there exists X_j ($\in S$ a.s.) such that

$$K_h(x - X_j) > \alpha_n \quad \text{and} \quad \|x - X_j\| < \frac{\varepsilon}{\beta_n},$$

which implies $S_n \subset S^{\varepsilon/\beta_n}$ a.s. for $n > n_1$.

On the other hand, since $\alpha_n \downarrow 0$, (19) implies that $S \subset S_n$, and, of course, $S \subset S_n^{\varepsilon/\beta_n}$ eventually, a.s.

We have thus obtained

$$S \subset S_n^{\varepsilon/\beta_n} \quad \text{and} \quad S_n \subset S^{\varepsilon/\beta_n} \quad \text{eventually, a.s., for any } \varepsilon > 0,$$

which is equivalent to (18).

Note that any rate β_n of type $o(n/\log n)^{1/d}$ can be achieved by using any h_n with $h_n = o(\beta_n^{-1})$ and $(\log n/n)^{1/d} = o(h_n)$; for instance, we could take $h_n = \beta_n^{-1}/\log[\beta_n^{-1}(n/\log n)^{1/d}]$.

(b) It suffices to consider the case where f is uniform on $[0, 1]^d$ and the set $\{K > 0\}$ is $[-1/2, 1/2]^d$. Let us first prove that the condition $\beta_n h_n \to 0$ is necessary as well; that is, if $\beta_n h_n \not\to 0$, then β_n cannot be a convergence rate in probability. Suppose that there is some subsequence (denoted also by $\beta_n h_n$) and some $b_0 > 0$ such that $\beta_n h_n > b_0$ eventually. Take $0 < b < b_0/2$ and define $y_n = (b/\beta_n) + 1$ and $\gamma_n = h_n/2 - b/\beta_n$ (thus $\gamma_n > 0$ eventually). Denote $\mathbf{1} = (1, \dots, 1)^t \in \Re^d$ and $Y_n = y_n\mathbf{1}$.

We have, for n large enough,

$$\begin{aligned} P\big\{\beta_n d_H(S, S_n) > b/2\big\} &\geq P\big\{\hat{f}_n(Y_n) > 0\big\} \\ &= 1 - P\left\{\bigcap_{i=1}^{n}\{I_{(Y_n - \mathbf{1}h_n/2,\, Y_n + \mathbf{1}h_n/2)}(X_i) = 0\}\right\} \\ &= 1 - \big(1 - \gamma_n^d\big)^n \\ &\geq 1 - \exp\big(-n\gamma_n^d\big) \\ &= 1 - \exp\big(-nh_n^d((\beta_n h_n - 2b)/2\beta_n h_n)^d\big) \\ &\geq 1 - \exp\big(-nh_n^d(1/2 - b/b_0)^d\big). \end{aligned} \tag{21}$$

Then, if $nh_n^d \not\to 0$ on the same subsequence for which $\beta_n h_n > b_0$, we conclude from (21) that $P\{\beta_n d_H(S, S_n) > b/2\} \not\to 0$, and so β_n cannot be a convergence rate. There remains only the case where $nh_n^d \to 0$ and $\beta_n h_n > b_0$; then $h_n = o(n^{-1/d})$, $\beta_n > b_0 n^{1/d}$ (for large n), and it suffices to show that $P\{n^{1/d} d_H(S, S_n) \geq 1\} \not\to 0$, which follows from

$$\begin{aligned} P\{n^{1/d} d_H(S, S_n) \geq 1\} &\geq P\big\{\hat{f}_n(x) = 0, \text{ for all } x \in [0, 1/n^{1/d}]^d\big\} \\ &\geq P\left(\bigcap_{i=1}^{n}\{X_i \notin [0, 2/n^{1/d}]^d\}\right) \\ &= \big(1 - 2^d/n\big)^n \to \exp(-2^d). \end{aligned}$$

We have thus proved that if β_n satisfies (16), then we must have $\beta_n h_n \to 0$. Thus, to prove part (b) of the theorem take β_n satisfying $(n/\log n)^{1/d} = O(\beta_n)$ and $h_n = o((\log n/n)^{1/d})$. Then $(n/\log n)^{1/d} \leq c\beta_n$ for some constant $c > 0$. We may assume $c = 1$ since $\beta_n^* = c\beta_n$ is a convergence rate if and only if β_n also is one.

Now the result will follow as a consequence of a theorem [due to Janson (1987)] on the asymptotic distribution of the maximal uniform spacing in the multivariate case. This theorem is a multivariate extension of the classical Lévy results on univariate spacings [see, e.g., Shorack and Wellner (1986), Chapter 21]. In precise terms, let $X_1, \dots, X_n$ be a uniform sample on $S = [0,1]^d$. Define

$$\Delta_n = \sup\{r\colon \exists\ x,\ \text{ with } x + rA \subset S \setminus \{X_1, \dots X_n\}\},$$

where $A = [-1/2, 1/2]^d$. Deheuvels (1983) defined the maximal spacing by $V_n = \Delta_n^d$, which is the volume of the largest cubical gap (parallel to the unit cube). Janson (1987) proved the weak convergence

$$nV_n - \log n - (d-1)\log\log n \to_w U,$$

where U has the extreme value distribution $P\{U \le u\} = \exp(-e^{-u})$.

Coming back to our proof, we have, for n large enough,

$$\begin{aligned}
&P\left\{\beta_n d_H(S, S_n) \ge \frac{1}{4}\right\}\\
&\quad \ge P\left\{\left(\frac{n}{\log n}\right)^{1/d} d_H(S, S_n) \ge \frac{1}{4}\right\}\\
&\quad \ge P\left\{\hat{f}_n(x) = 0 \text{ on some cube of side length } \frac{1}{2}\left(\frac{\log n}{n}\right)^{1/d} \text{ in } [0,1]^d\right\}\\
&\quad \ge 1 - P\left\{\Delta_n^d \le \frac{\log n}{n}\right\}\\
&\quad = 1 - P\{nV_n - \log n - (d-1)\log\log n \le -(d-1)\log\log n\}\\
&\quad \ge 1 - P\{nV_n - \log n - (d-1)\log\log n \le 0\} \to 1 - \exp(-1),
\end{aligned}$$

which proves that β_n is not a convergence rate. □

COMMENTS. (a) It is worth noting that, under the conditions of Theorem 3(a), the estimator S_n is "shape-preserving" in the sense that if S is a connected set, then S_n is also connected (eventually, a.s.). This follows easily from the fact that $S \subset S_n$ eventually a.s.

(b) A sharp result on optimal (in the minimax sense) d_H-rates for convergence in mean has been given by Korostelev, Simar and Tsybakov (1995). They provide not only the optimal estimator but also the exact asymptotic bound for the minimax risk. Further convergence rates (in mean) and optimality results can be found in Korostelev and Tsybakov [(1993), Chapter 7].

4. Final remarks.

4.1. *On the tuning parameter α_n.* Theorem 1 provides some guide about the appropriate asymptotic order for α_n. Also, Theorems 2 and 3 show that, in many cases of practical interest, this parameter is asymptotically irrelevant.

As for the choice of α_n for a given sample, let us note that this parameter is more tractable than the bandwidth h_n in (4), in the sense that every choice of α_n is directly interpretable in population terms: $S_n = \{f_n > \alpha\}$ can be seen as an estimator of the α-support $\{f > \alpha\}$ which plays an important role in cluster theory [see, e.g., Hartigan (1975)]. A reasonable approach would be to determine α_n, for a given sample, in an indirect way by specifying the "outside probability" $p_n = P\{f_n \le \alpha_n\}$ which has a more direct intuitive interpretation.

Note also that the estimator $S_n = \{f_n > \alpha_n\}$ can be considered as a robust alternative to the estimator (1) suitable for rejecting isolated outliers. Such a property could be particularly useful when S_n is used in a detection problem [as outlined in Devroye and Wise (1980)].

4.2. *A smoothness property.* Whereas the border of the rough estimator (1) is not smooth, that of S_n, ∂S_n (which, for most usual choices of K, coincides a.s. with $\{f_n = \alpha_n\}$), will be typically a differentiable manifold of dimension $d-1$. It is known that a sufficient condition for this is $\nabla f_n(x) \ne 0$ for all x such that $f_n(x) = \alpha_n$ (∇f denotes the gradient of f). In turn, this can be guaranteed by imposing a suitable condition of boundedness away from zero for $\|\nabla f\|$ on sets $\{f > a\}$ together with the uniform a.s. convergences $f_n \to f$ and $\nabla f_n \to \nabla f$ [see Sarda and Vieu (1988)].

Acknowledgments. We are very grateful to Robert Dobrow for sending us his interesting manuscript. We also thank Luc Devroye for helpful suggestions. The constructive criticisms of an Associate Editor and two referees led to substantial improvements in the paper.

REFERENCES

BHATTACHARYA, R. N. and RANGA RAO, R. (1976). *Normal Approximations and Asymptotic Expansions*. Wiley, New York.

BOULARAN, J., FERRÉ, L. and VIEU, P. (1995). Location of particular points in nonparametric regression analysis. *Austral. J. Statist.* **37** 161–168.

CHEVALIER, J. (1976). Estimation du support et du contenu de support d'une loi de probabilité. *Ann. Inst. H. Poincaré Sec. B* **12** 339–364.

CUEVAS, A. (1990). On pattern analysis in the non-convex case. *Kybernetes* **19** 26–33.

DEHEUVELS, P. (1983). Strong bounds for multidimensional spacings. *Z. Wahrsch. Verw. Gebiete* **64** 411–424.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1-View*. Wiley, New York.

DEVROYE, L. and WISE, G. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.* **38** 480–488.

DOBROW, R. (1992). Estimating level sets of densities. Unpublished manuscript.

GEFFROY, J. (1964). Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique des Universités de Paris* **13** 191–210.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

HALL, P. (1982). On estimating the endpoint of a distribution. *Ann. Statist.* **10** 556–568.

HÄRDLE, W., PARK, B. U. and TSYBAKOV, A. B. (1995). Estimation of non-sharp support boundaries. *J. Multivariate Anal.* **55** 205–218.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York.

HOLMSTRÖN, L. and KLEMELÄ, J. (1992). Asymptotic bounds for the expected L_1-error of a multivariate kernel density estimator. *J. Multivariate Anal.* **42** 245–266.

JANSON, S. (1987). Maximal spacings in several dimensions. *Ann. Probab.* **15** 274–280.

KOROSTELEV, A. P., SIMAR, L. and TSYBAKOV, A. B. (1995). Efficient estimation of monotone boundaries. *Ann. Statist.* **23** 476–489.

KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. Springer, New York.

MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical mimimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23** 855–882.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation.* Academic Press, New York.

SAGER, T. W. (1979). An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* **74** 329–339.

SARDA, P. and VIEU, P. (1988). Vitesses de convergence d'estimateurs non-parametriques d'une régression et de ses dérivées. *C. R. Acad. Sci. Paris Ser. I Math.* **306** 83–86.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969.

DEPARTAMENTO DE MATEMÁTICAS
FACULTAD DE CIENCIAS
UNIVERSIDAD AUTÓNOMA DE MADRID
28049-MADRID
SPAIN
E-MAIL: antonio.cuevas@uam.es

CENTRO DE MATEMÁTICA
UNIVERSIDAD DE LA REPÚBLICA
EDUARDO ACEVEDO, 1139
MONTEVIDEO
URUGUAY
E-MAIL: rfraiman@cmat.edu.uy