The Annals of Statistics
1997, Vol. 25, No. 6, 2273–2299

GRANULOMETRIC SMOOTHING

BY GUENTHER WALTHER

Stanford University

A new method for smoothing a multivariate data set is introduced that is based on a simple geometric operation. This method is applied to the problem of estimating level sets of a density and minimum volume sets with given probability content. Building on existing techniques, the resulting estimator combines excellent theoretical and computational properties: It converges with the minimax rates (up to log factors) in most cases where these rates are known and, at the same time, it can be computed, visualized, stored and manipulated by simple algorithms and tools. It is applicable to a wide class of sets that is characterized explicitly in terms of the underlying densities and includes nonconvex and disconnected sets, and it is argued that it should give reasonable results in completely general situations. Applications to the construction of multivariate confidence regions in frequentist and Bayesian contexts are briefly mentioned.

1. Introduction. An interesting and challenging problem facing the advance of computers in the field of statistics is the modeling of complex objects and the interest in the qualitative aspects, for example, the shape, of these objects. Some recent examples are as follows.

DasGupta, Ghosh and Zen (1995) consider the problem of constructing smallest volume multivariate confidence sets for the mode of a density as location parameter. For the special case where the underlying distribution has a density that is star unimodal and of a known form (up to the location parameter), they give an analytical formula for a star-shaped confidence set involving numerical integration and minimization.

In a related context, the so-called excess mass approach, introduced independently by Hartigan (1987) and Müller and Sawitzki (1987) and extended and investigated in detail by Polonik (1995), can be applied to various statistical problems, such as that of estimating level sets of a density (and thus the density itself): If $\mathscr{S}$ is a class of measurable subsets of $\mathbf{R}^d$, then for a distribution F on $\mathbf{R}^d$ the *excess mass over* $\mathscr{S}$ *at level* $\lambda \geq 0$ is defined as $E_{\mathscr{S}}(\lambda) = \sup\{F(S) - \lambda \operatorname{Leb}(S)\colon S \in \mathscr{S}\}$, where Leb denotes Lebesgue measure. If F has density f and the level set (or density contour cluster) at level λ, $L(\lambda) = \{x\colon f(x) \geq \lambda\}$, belongs to $\mathscr{S}$, then $F(L(\lambda)) - \lambda \operatorname{Leb}(L(\lambda))$ attains the excess mass $E_{\mathscr{S}}(\lambda)$. One can hence attempt to estimate the level set $L(\lambda)$ by a set $S \in \mathscr{S}$ that attains $\sup\{F_n(S) - \lambda \operatorname{Leb}(S)\colon S \in \mathscr{S}\}$, where F_n is the empirical measure of n i.i.d. observations $X_1, \dots, X_n$ drawn from F. A restriction of the form $S \in \mathscr{S}$ is necessary as otherwise the resulting estimator will be $S = \{X_1, \dots, X_n\}$; on the other hand, prescribing a certain class $\mathscr{S}$ allows us

Received March 1996; revised July 1997.

AMS 1991 *subject classifications*. Primary 62G05; secondary 65U05.

Key words and phrases. Granulometry, smoothing, level sets, minimum volume sets, excess mass, mode, multivariate confidence regions, highest posterior density regions.

to model qualitative aspects of the level sets. From a practical point of view, however, this approach will, in general, flounder on the maximization over the class $\mathscr{G}$ necessary to find the set that attains the sup above. Note that sets are infinite-dimensional objects. Algorithms have been devised only for simple classes $\mathscr{G}$, like the class of convex sets in $\mathbf{R}^2$ and the class of ellipsoids in $\mathbf{R}^d$ by Hartigan (1987) and Nolan (1991), respectively. The excess mass approach has the advantage that it outperforms traditional density estimates when the density has a jump.

A different problem arises when one intends to use an estimate of the density to estimate the minimum volume set with given probability content $1-\alpha$, that is, the set $L(\lambda)$ where λ is such that $F(L(\lambda)) = 1 - \alpha$. A trial-and-error procedure would be necessary to find the right level λ, with each trial involving contouring and numerical integration of the density estimate. This minimum volume set arises in a Bayesian context as a highest posterior density set if f given previously denotes the posterior density. In any case the challenging task arises to implement a data structure on the computer to store a possibly high-dimensional boundary of a complicated set and effectively work with that implementation.

These examples show that for such set-valued estimation problems the statistical theory should not be considered separately from computational issues. The aim of this paper is to show that in the preceding context the underlying geometry suggests methods and tools that give rise to estimators that offer both excellent computational and theoretical properties for a very wide and flexible class of sets. The idea for these methods derives from Blaschke's rolling theorem. Blaschke (1949) gave necessary and sufficient conditions for a ball to roll freely inside a convex body in $\mathbf{R}^2$ or $\mathbf{R}^3$. Section 2 will generalize this idea by characterizing the class of compact sets that allow a ball of fixed radius to roll freely inside and outside the set. The resulting class of sets is very flexible, allowing nonconvex and disconnected sets. It is shown that a large class of densities has level sets satisfying this requirement. Section 3 shows how this characterization can be used to "smooth" a multivariate data set and to devise estimators for the problem of estimating level sets and minimum volume sets with given probability content. The theoretical and computational properties of these estimators are analyzed in detail. The main theorem gives rates of convergence for these estimators that coincide with the minimax rates, up to log factors, in most cases where the minimax rates are known. Section 4 discusses the advantages of the new estimator over competing techniques and touches on some important applications. All proofs are deferred to Section 5.

For further literature on the problems of level set estimation and the vast area of confidence sets, see the references in Polonik (1995) and DasGupta, Ghosh and Zen (1995), respectively, and for minimax results in the context of set-valued estimation, see Mammen and Tsybakov (1995) and the references given therein.

2. The geometric approach. The setting throughout is $\mathbf{R}^d$ equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $|\cdot|$. $B_r(x)$ denotes the closed ball in $\mathbf{R}^d$ with radius r centered at x, and $B := B_1(0)$, $S^{d-1} := \partial B$. If A is a subset of $\mathbf{R}^d$, then A^c, $\overline{A}$, $\operatorname{int} A$ and ∂A denote the complement, closure, interior and boundary of A, respectively. Further $d(x, A) := \inf_{a\in A} |x-a|$ for $x \in \mathbf{R}^d$. For $A, C \subset \mathbf{R}^d$ and $\lambda \in \mathbf{R}$, write $\lambda A := \{\lambda a\colon a \in A\}$ and denote by

$$A \oplus C = \{a + c\colon a \in A,\ c \in C\}$$

the *Minkowski addition* of A and C, and by

$$A \ominus C = \{x\colon x + C \subset A\}$$

the *Minkowski difference*, where we write $x + C$ for $\{x\} + C$. One then checks that

$$A \ominus C = \big(A^c \oplus (-1)C\big)^c. \tag{1}$$

For $\varepsilon \in \mathbf{R}$ we use the abbreviation

$$A_\varepsilon = \begin{cases} A \oplus \varepsilon B, & \text{if } \varepsilon \geq 0,\\ A \ominus |\varepsilon| B, & \text{if } \varepsilon < 0.\end{cases}$$

As a measure of distance between sets we will either use the Hausdorff distance

$$d_H(A, C) := \inf\{\varepsilon > 0\colon A \subset C_\varepsilon \text{ and } C \subset A_\varepsilon\}$$

or the Lebesgue measure of the symmetric difference

$$d_{\mathrm{Leb}}(A, C) := \mathrm{Leb}(A \triangle C).$$

Minkowski addition and subtraction have become common tools in mathematical morphology and image processing [see Serra (1982)], and we will make use of two more definitions from that field: The class of compact sets A that satisfy $A = (A \oplus \lambda B) \ominus \lambda B = (A \ominus \lambda B) \oplus \lambda B$ for some $\lambda > 0$ is called *Serra's regular model*; see Serra (1982), page 144.

More generally, Matheron (1975), page 24, defines, for any $A \subset \mathbf{R}^d$ and $\lambda \geq 0$,

$$\Psi_\lambda(A) := (A \ominus \lambda B) \oplus \lambda B = \bigcup_{B_\lambda(x)\subset A} B_\lambda(x) \tag{2}$$

and calls the mapping $\lambda \to \Psi_\lambda(A)$ the *granulometry* of the set A with respect to the *structuring element* B. The granulometry represents the "size distribution" of the set A in the sense that the mapping $\lambda \to \mathrm{Leb}(\Psi_\lambda(A))$, where Leb denotes Lebesgue measure on $\mathbf{R}^d$, gives the volume occupied by the translates of λB that are included in A. Here we extend the definition of the granulometry to the whole line by setting

$$\Psi_{-\lambda}(A) := (A \oplus \lambda B) \ominus \lambda B, \qquad \lambda > 0. \tag{3}$$

For the generalization of Blaschke's rolling theorem, it is informative to introduce the following notion of generalized convexity, for which Mani-Levitska (1993) cites Perkal (1956) as a reference: The set $A \subset \mathbf{R}^d$ is called *r-convex* ($r > 0$) if $A = C_r(A)$, where $C_r(A) = \bigcap_{\operatorname{int} B_r(x)\cap A=\varnothing} (\operatorname{int} B_r(x))^c$ is called the *r-convex hull* of A. For an interpretation why this is a generalized notion of convexity and for further properties of granulometries, see Walther (1995).

Finally, if A is convex and compact, then a ball of radius r is said to *roll freely* in A if for each boundary point $b \in \partial A$ there exists $x \in \mathbf{R}^d$ such that $b \in B_r(x) \subset A$; see Schneider (1993). If A is only closed, then for rB to roll freely in A we require, in addition, that $A \ominus rB$ be path-connected in order to preserve the physical meaning of rolling freely.

The following generalization of Blaschke's rolling theorem links together all the preceding notions.

THEOREM 1. *Let $S \neq \varnothing$ be a compact subset of $\mathbf{R}^d$ and $r_0 > 0$. Then the following are equivalent*:

(i) $\Psi_\lambda(S) = S$ *for* $\lambda \in (-r_0, r_0]$;

(ii) *S and $\overline{S^c}$ are r_0-convex and* $\operatorname{int} S_i \neq \varnothing$ *for each path-connected component $S_i \subset S$*;

(iii) *a ball of radius r rolls freely inside each path-connected component of S and $\overline{S^c}$ for all* $0 \leq r \leq r_0$;

(iv) *∂S is a $(d-1)$-dimensional C^1 submanifold in $\mathbf{R}^d$ with the outward pointing unit normal vector $n(s)$ at $s \in \partial S$ satisfying the Lipschitz condition*

$$|n(s) - n(t)| \leq \frac{1}{r_0}|s - t| \quad \textit{for all } s, t \in \partial S.$$

Moreover, for some $r_0 > 0$ *the preceding is equivalent to*:

(v) *S belongs to Serra's regular model.*

The theorem shows that the smoothness of ∂S is linked to the behavior at the origin of the granulometry $\Psi_\lambda(S)$. The theorem is proved in Walther (1995) for path-connected S, but as remarked there, it extends to compact S in a straightforward way. Also, a further characterization is given there. In the following we denote by $\mathscr{S}(r_0)$ the class of all nonempty compact sets in $\mathbf{R}^d$ that satisfy the conditions of Theorem 1 with rolling parameter r_0. Further, for a set C let $\mathscr{S}_C(r_0)$ denote all sets in $\mathscr{S}(r_0)$ that also satisfy $S \subset C$. Examples of sets in $\mathbf{R}^2$ that satisfy the conditions of Theorem 1 are shown in Figure 1, where some level sets $L(\lambda) := \{x\colon f(x) \geq \lambda\}$ of the density f_M of the Gaussian mixture $\frac{1}{3}\sum_{i=1}^3 N(m_i, \frac{1}{10}I)$ are plotted with $m_1 = (0.8, 2.2)$, $m_2 = (2, 1.2)$ and $m_3 = (2, 2)$. It is, in fact, quite a general phenomenon that level sets of densities belong to $\mathscr{S}(r_0)$ for some r_0 as the next theorem shows.

THEOREM 2. *Let $f\colon \mathbf{R}^d \mapsto \mathbf{R}$ and $-\infty < l \leq u < \sup f$. Assume*:

(i) $f \in C^1(U)$, *where U is a bounded open set that contains $\overline{L(l-\eta)} \setminus \operatorname{int} L(u+\eta)$ for some $\eta > 0$*;

(ii) $\operatorname{grad} f$ *satisfies* $|\operatorname{grad} f| \geq m > 0$ *on U as well as a Lipschitz condition on U (or on $\partial L(\lambda)$ for all $\lambda \in [l, u]$)*:

$$|\operatorname{grad} f(x) - \operatorname{grad} f(y)| \leq k|x - y| \quad \textit{for } x, y \in U \; [\textit{or } \partial L(\lambda)].$$

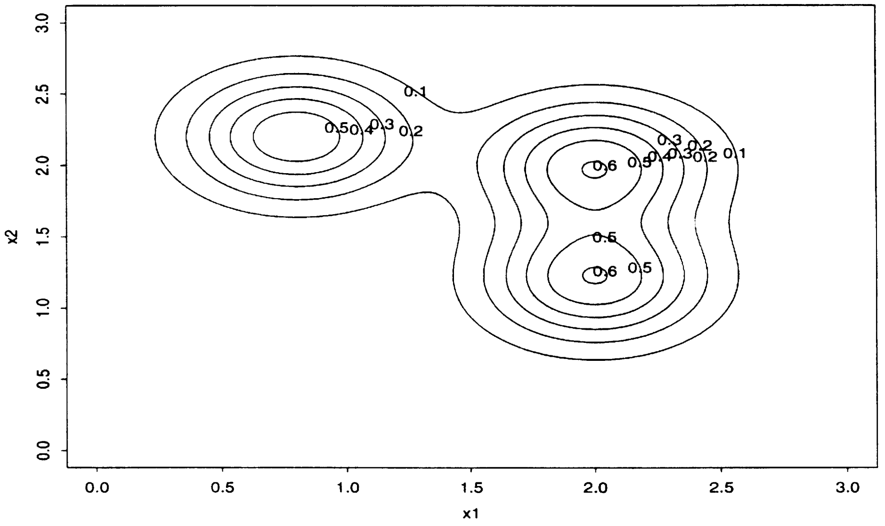

FIG. 1. *Some level sets of the Gaussian mixture density f_M.*

Then, for each $\lambda \in [l, u]$, $L(\lambda)$ satisfies the conditions of Theorem 1 *with $r_0 = m/k$.*

The theorem shows that in the example of the Gaussian mixture density f_M the level sets at almost all levels belong to $\mathscr{G}(r_0)$ for some r_0. The simple case of a Gaussian mixture is illustrative here because it shows how the generalized rolling theorem generalizes the well-known elliptical contours of a Gaussian distribution to the case of generalized convexity shown in Figure 1.

3. Estimating level sets and minimum volume sets. Note that first subtracting and then adding a ball B_λ to a set S corresponds to rolling a ball of radius λ inside S, as can be seen from the definition (2). Hence Ψ_λ "smooths" the set S in some sense and filters away small components of S. Similarly, $\Psi_{-\lambda}(S)$ is the smoothed set obtained by rolling a ball of radius λ along ∂S in S^c. This follows from (3) and (1), which imply $\Psi_{-\lambda}(A) = (\Psi_\lambda(A^c))^c$ for all $\lambda \in \mathbf{R}$ [the set $(\Psi_\lambda(A^c))^c =: \Psi_\lambda^*(A)$ is also called the *dual mapping* of Ψ_λ; see Matheron (1975), page 187]. The granulometric smoothing procedure $\Psi_{-\lambda}$ is especially interesting when applied to a point cloud. Adding B_λ fills the space between points so that the following subtraction of B_λ reveals the shape of the data set in a certain sense. A large λ will give a coarser summary of the shape with the convex hull in the limiting case $\lambda = \infty$, and a small λ will give more detailed information, with the limiting case $\lambda = 0$ recovering the original set $\mathscr{X}$. Figure 2 gives an illustration of the granulometric smoothing procedures Ψ_1 and Ψ_{-1} applied to the union S of two polytopes and a two-dimensional point cloud $\mathscr{X}$, respectively.

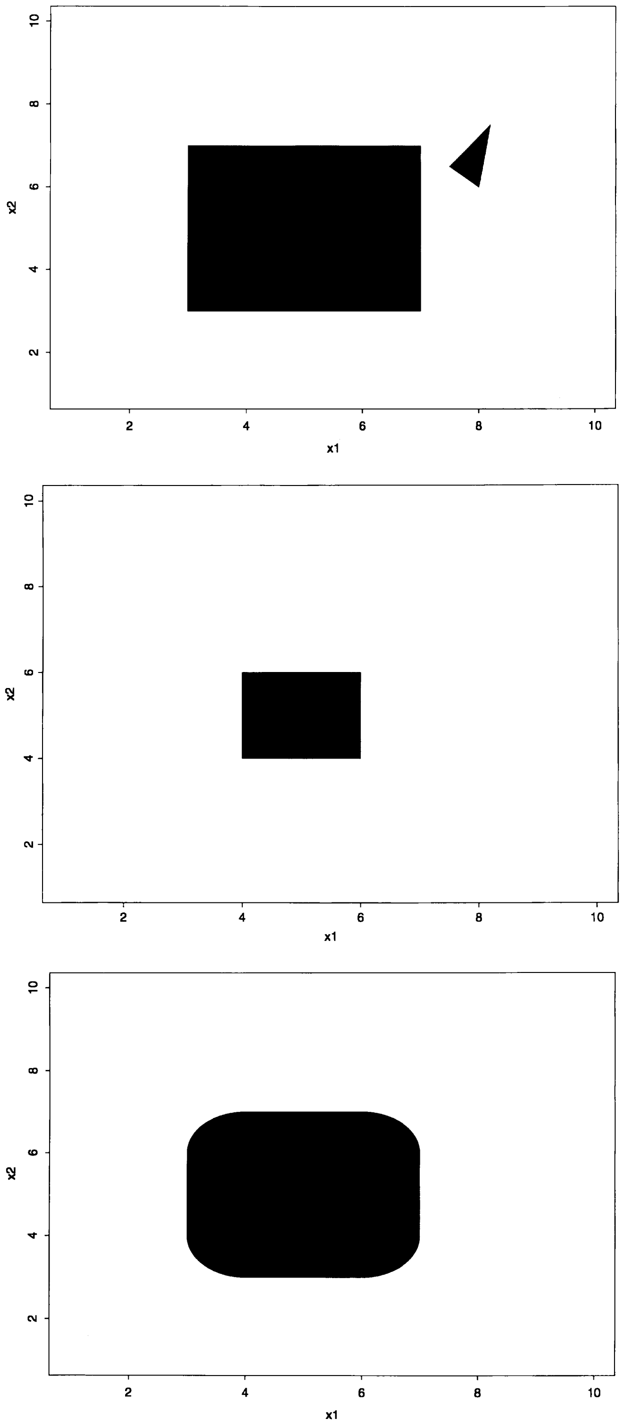

FIG. 2. *Left:* S, $S \ominus B_1$, $\Psi_1(S)$. *Right:* $\mathscr{X}$, $\mathscr{X} \oplus B_1$, $\Psi_{-1}(\mathscr{X})$.

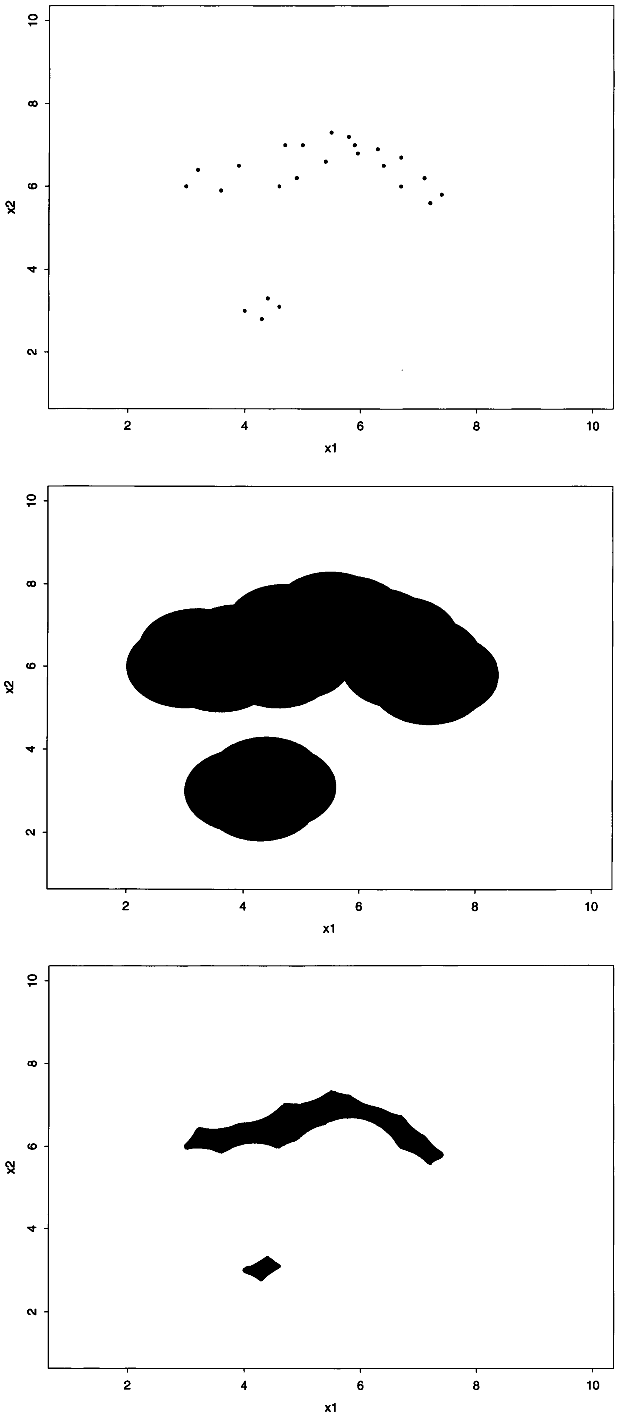

FIG. 2. *(Continued).*

In the following discussion an empirical version of Ψ_λ will be used to construct various estimators. A question that arises naturally in this context is whether applying the granulometric smoothing procedures Ψ_λ and $\Psi_{-\lambda}$, possibly repeatedly, to an arbitrary set S will always produce a smooth set in the sense of Theorem 1. The answer is negative, but it is shown in Walther (1995) that the smoothing works in a quite general context.

For the estimation of level sets and minimum volume sets, we distinguish the cases where the underlying density f is smooth and where it has a jump.

3.1. *The smooth case.* Let l, u be numbers such that $0 < l \le u < \sup f < \infty$. We will make the following assumption on f:

ASSUMPTION A. (i) $f \in C^p(U)$, where $p \ge 1$ and U is a bounded, open set that contains $\overline{L(l-\eta)} \setminus \operatorname{int} L(u+\eta)$ for some $\eta > 0$.

(ii) On U, grad f satisfies $|\operatorname{grad} f| \ge m > 0$ as well as a Lipschitz condition:

$$|\operatorname{grad} f(x) - \operatorname{grad} f(y)| \le k|x-y| \quad \text{for } x, y \in U.$$

Besides the level set $L(\lambda) := \{x\colon f(x) \ge \lambda\}$, we define, for $\gamma \in (0,1]$,

$$C(\gamma) := L(\lambda(\gamma)), \quad \text{where } \lambda(\gamma) := \sup\{\lambda\colon F(L(\lambda)) \ge \gamma\}.$$

Under Assumption A and if γ is such that $\lambda(\gamma) \in (l,u)$, then $F(C(\gamma)) = \gamma$. This follows from Theorem 2 and Lemma 2(b) (in Section 5) together with (6). Hence $C(\gamma)$ is then a minimum volume set with probability content γ by the Neyman–Pearson fundamental lemma.

It turns out that for smooth f the excess mass approach can be improved upon by using estimators that are based on density estimation procedures. For that reason we will investigate in the following the performance of a granulometric smoothing procedure that is based on a density estimator $\hat{f}$.

Let $\mathscr{X}_n := \{X_1, \dots, X_n\}$ where the X_i are i.i.d. f, and let $\hat{f}_n$ be any density estimate of f based on $\mathscr{X}_n$. In the following a kernel density estimate will be used, but any other density estimate can be employed analogously. Set $\mathscr{X}_n^+(\lambda) := \{X \in \mathscr{X}_n\colon \hat{f}_n(X) \ge \lambda\}$, $\mathscr{X}_n^-(\lambda) := \mathscr{X}_n \backslash \mathscr{X}_n^+(\lambda)$. Now apply an empirical version $\widehat{\Psi}$ of the previous smoothing procedure to obtain the estimators for $L(\lambda)$ and $C(\gamma)$:

$$L_n(\lambda) := \widehat{\Psi}_{r_n}(\mathscr{X}_n^+) := \big((\mathscr{X}_n^-(\lambda) \oplus r_n B)^c \cap \mathscr{X}_n^+(\lambda)\big) \oplus r_n B,$$

$$C_n(\gamma) := L_n(\lambda_n(\gamma)), \quad \text{where } \lambda_n(\gamma) := \max\{\lambda\colon F_n(L_n(\lambda)) \ge \gamma\}.$$

To see the analogy of $\widehat{\Psi}_r$ to Ψ_r, observe that, by (1), $\Psi_r(S) = (S^c \oplus rB)^c \oplus rB$. $\widehat{\Psi}$ has the advantage of possessing excellent computational properties; see the following discussion.

For the choice of r_n note that if m and k in Assumption A are known, then we know from Theorem 2 that $L(\lambda), C(\gamma) \in \mathscr{G}(m/k)$. This prior information can then be built into the estimator by setting $r_n = m/k$ (or slightly smaller). Observe that, by definition, a ball of radius r_n rolls inside L_n and C_n. But, in general, the boundaries of $L_n(\lambda)$ and $C_n(\gamma)$ will have vertices pointing inside, so the estimators will not belong to $\mathscr{G}(r)$ for any r by Theorem 1. However, Theorem 3 together with Theorem 2 of Walther (1995) shows that rolling a ball of radius r_n along the outside of these estimators will "project" them into $\mathscr{G}(\tilde{r}_n)$, where $\tilde{r}_n/r_n \to 1$, without affecting their rates of convergence. This is shown for $C_n(0.75)$ in Figure 3, using a sample of size $n = 100$ from the mixture density f_M given in Section 2, $r_n = 0.2$ and a kernel density estimate with Gaussian kernel.

If m and k are unknown, then one has to let r_n shrink to 0 slowly as the sample size increases; see Theorem 3.

Observe that there is a straightforward algorithm to compute these estimators: $L_n(\lambda)$ consists of the union of balls around those points in $\mathscr{X}_n^+$ that have a distance of at least r_n from each point in $\mathscr{X}_n^-$. Computing the centers of these balls takes $O(dn^2)$ steps, thus depending on the dimension d only in a linear way. The same complexity applies for the density estimation on $\mathscr{X}_n$, for example, in the case of a kernel density estimate. Further, using a bisection search, $C_n(\gamma)$ can be computed in $O(dn^2 \log n)$ steps, because $L_n(\lambda)$ is constant in λ except for at most n values $\lambda = \hat{f}_n(X_i)$, $i = 1, \dots, n$. The representation of the estimators can be further simplified by replacing several overlapping balls with a larger one; this could be done, for example, by a tree-structured algorithm. In any case the estimator can be stored as a list of centers and radii and is easily worked with.

The next theorem investigates the properties of these estimators in the case where $\hat{f}_n$ is a kernel density estimate, that is,

$$\hat{f}_n(x) := \frac{1}{n\sigma_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\sigma_n}\right),$$

where σ_n is a bandwidth sequence and K a kernel that will be assumed to satisfy

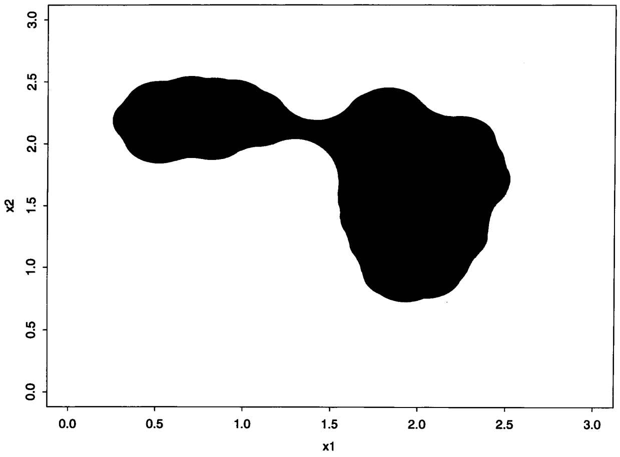

FIG. 3. *Estimate of $C(0.75)$ based on a sample of size 100 from f_M.*

ASSUMPTION K. K is a continuous kernel of order at least p with bounded support and finite variation.

In the following $a_n \gg b_n$ for sequences $\{a_n\}, \{b_n\}$ means $a_n/b_n \to \infty$.

THEOREM 3. *Let* $f\colon \mathbf{R}^d \mapsto \mathbf{R}$ *be a density satisfying Assumption* A *and let* K *be a kernel satisfying Assumption* K. *Let* σ_n *be of the order* $(\log n/n)^{1/(d+2p)}$ *and let* r_n *satisfy*

$$\max\left(\sigma_n^p, \left(\frac{\log n}{n}\right)^{1/(d+1)}\right) \ll r_n < r_0 - \varepsilon \quad \textit{for some } \varepsilon > 0,$$

where $r_0 := m/k$. *Then*

$$\begin{aligned}&\sup_{\lambda\in[l,u]} d_H(L(\lambda), L_n(\lambda))\\ &\quad = O\left(\max\left(\left(\frac{\log n}{n}\right)^{2/(d+1)} r_n^{-(d-1)/(d+1)}, \left(\frac{\log n}{n}\right)^{p/(d+2p)}\right)\right) \quad a.s.\end{aligned} \tag{4}$$

If further $\underline{\gamma} < \overline{\gamma}$ *are such that* $l < \lambda(\overline{\gamma}), \lambda(\underline{\gamma}) < u$, *then*

$$\begin{aligned}&\sup_{\gamma\in[\underline{\gamma},\overline{\gamma}]} d_H(C(\gamma), C_n(\gamma))\\ &\quad = O\left(\max\left(\left(\frac{\log n}{n}\right)^{2/(d+1)} r_n^{-(d-1)/(d+1)}, \left(\frac{\log n}{n}\right)^{p/(d+2p)}\right)\right) \quad a.s.\end{aligned} \tag{5}$$

The proof of the theorem proceeds by establishing an exponential inequality and is given in Section 5.

REMARK 1. The statements of the theorem also hold when d_{Leb} is used instead of d_H. This follows from (9), (10) and (6) in the case of $L_n(\lambda)$ and can be seen similarly in the case $C_n(\lambda)$.

REMARK 2. Theorem 3 shows that $L_n(\lambda)$ has a faster rate of convergence than the empirical λ-cluster studied in Polonik (1995), Proposition 3.7, when f is smooth enough [to apply said proposition note that (A) and (8) imply $\gamma = 1$ and $r = (d-1)/2$ there]. The reason is that the empirical λ-cluster does not make use of the smoothness of f.

REMARK 3. The proof of Theorem 3 shows that the term $(\log n/n)^{p/(d+2p)}$ stems from the estimation of f and the term $(\log n/n)^{2/(d+1)} r_n^{-(d-1)/(d+1)}$ from the specific method used to reconstruct the set $L(\lambda)$ based on $\mathscr{X}_n^+(\lambda)$, $\mathscr{X}_n^-(\lambda)$. This latter reconstruction can be improved by basing it on a larger set of points: Sample another m points from a density that is bounded away from zero on a set containing $L(\lambda) \oplus \varepsilon B$ (e.g., resample from $\hat{f}_n$) and evaluate $\hat{f}_n$, based on the original sample only, on the augmented sample $\mathscr{X}_{n,m}$ of size $n+m$

to obtain $\mathscr{X}^+_{n,m}(\lambda) := \{X \in \mathscr{X}_{n,m}\colon \hat{f}_n(X) \geq \lambda\}$ and analogously $\mathscr{X}^-_{n,m}(\lambda)$. The proof of Theorem 3 then shows that the resulting estimator

$$L_{n,m}(\lambda) := \big((\mathscr{X}^-_{n,m}(\lambda) \oplus r_n B)^c \cap \mathscr{X}^+_{n,m}(\lambda)\big) \oplus r_n B$$

converges at the rate

$$O\bigg(\max\bigg(\bigg(\frac{\log(n+m)}{n+m}\bigg)^{2/(d+1)} r_{n+m}^{-(d-1)/(d+1)}, \bigg(\frac{\log n}{n}\bigg)^{p/(d+2p)}\bigg)\bigg).$$

REMARK 4. In a general situation, if $L(\lambda)$ does not belong to $\mathscr{G}(r_0)$ for any r_0, for example, if $L(\lambda)$ has vertices, the estimator will still be consistent if f is not flat at $\partial L(\lambda)$ and $L(\lambda)$ can be approximated by sets in $\mathscr{G}(r_0)$ as $r_0 \downarrow 0$. This will, in general, be the case for sets encountered in practice.

REMARK 5. A simple estimator can be built using the regularization concept of Grenander (1981), page 373, and setting $\widetilde{L}_n(\lambda) = \bigcup_{X \in \mathscr{X}_n^+} X \oplus r_n B$; see also Cuevas (1990). Using techniques similar to the proof of Theorem 3, one can show that the rate of the reconstruction step is not faster than $(\log n/n)^{1/d}$ (and is obtained when r_n is of the same order). The corresponding rate $(\log n/n)^{2/(d+1)} r_n^{-(d-1)/(d+1)}$ for $L_n(\lambda)$ and fixed or slowly decreasing r_n shows that $L_n(\lambda)$, although of the same structure (a union of balls), greatly outperforms $\widetilde{L}_n(\lambda)$.

3.2. *The nonsmooth case.* In the case where the density has a jump along $\partial L(\lambda)$, the excess mass approach turns out to be superior to procedures based on kernel density estimates. It will be shown how the preceding granulometric smoothing technique can be applied to the excess mass approach to produce an estimator that converges to $L(\lambda)$ with the minimax rates, up to log factors, if $d > 2$. The estimate employs a kernel density estimator as a pilot estimate and uses an approximation scheme based on Monte Carlo sampling. The estimate is hence computer-intensive to obtain, but once computed, it is of the same simple form as in the smooth case.

Specifically, we consider densities satisfying:

ASSUMPTION B. For fixed numbers $0 < \underline{l} < l < \lambda < u, r_0 > 0$ and some compact set $C \subset \mathbf{R}^d$, f satisfies $L(\lambda) \in \mathscr{G}_C(r_0)$, $L(\lambda) \oplus r_0 B \subset C$ and $\underline{l} \leq f \leq l$ on $(L(\lambda))^c \cap C$, $f \geq u$ on $L(\lambda)$.

The assumptions involving the set C are imposed to facilitate the exposition. The following estimator can be modified to handle the case $f \leq l$ on $(L(\lambda))^c$, $f \geq u$ on $L(\lambda)$, by putting down auxiliary points as described previously in Remark 3.

The estimator is constructed as follows. Start again by computing on $\mathscr{X}_n$ a kernel density estimate $\hat{f}_n$ with bandwidth $\sigma_n \to 0$, $\sigma_n \gg (\log n/n)^{1/d}$, and set $\mathscr{X}_n^+ := \{X \in \mathscr{X}_n\colon \hat{f}_n(X) \geq \lambda\}$, $\mathscr{X}_n^- := \mathscr{X}_n \setminus \mathscr{X}_n^+$. Setting $\widetilde{\mathscr{X}_n^+} := \mathscr{X}_n^+ \cap (\mathscr{X}_n^- \oplus$

$2\sigma_n B)^c$, $\widetilde{\mathscr{X}_n^-} := \mathscr{X}_n^- \cap (\mathscr{X}_n^+ \oplus 2\sigma_n B)^c$ and $R := \mathscr{X}_n \setminus (\widetilde{\mathscr{X}_n^+} \cup \widetilde{\mathscr{X}_n^-})$ will make sure $\widetilde{\mathscr{X}_n^+} \subset L(\lambda)$ and $\widetilde{\mathscr{X}_n^-} \subset (L(\lambda))^c$ for large n. Next we will successively smooth the sets $\widetilde{\mathscr{X}_n^+} \cup R_i$ for various $R_i \subset R$ by employing the granulometric smoothing procedure twice. This will ensure that the following candidate sets $Z_{n,i}$ are regular enough, that is, close enough to elements of $\mathscr{G}_C(r)$, for the excess mass approach to work. The sets R_i, $i = 1, \dots, I$, are obtained by a randomization procedure, for example, accepting each $X \in R$ into R_i with probability proportional to $\hat{f}_n(X)$, or by considering all possible subsets $R_i \subset R$ if $|R|$ is small. One obtains

$$Z_{n,i} := \left\{ X \in \mathscr{X}_n \colon |X - Y| > 2r_n \text{ for all } Y \in \left(\left(\widetilde{\mathscr{X}_n^+} \cup R_i \right) \oplus r_n B \right)^c \cap \widetilde{\mathscr{X}_n^-} \right\} \oplus r_n B.$$

This estimator is again a union of balls, the centers of which are easily computed by a straightforward algorithm comparing various distances between the elements of $\mathscr{X}_n$. Computing all these centers takes $O(dn^2)$ time. Finally, choose as an estimator $L_n^I(\lambda)$ a set $Z_{n,j}$ that attains the excess mass over the competing estimators $Z_{n,i}$:

$$(F_n - \lambda\,\mathrm{Leb})(L_n^I(\lambda)) \ge (F_n - \lambda\,\mathrm{Leb})(Z_{n,i}) \quad \text{for all } i = 1, \dots, I.$$

An appealing choice for I in practice is to monitor the increase in the largest excess mass found so far and to terminate the algorithm once this increase tends to level off. For the computation of $\mathrm{Leb}(Z_{n,i})$, one can use the formula of Naiman and Wynn (1992) that reduces the volume of the union of a finite number of balls with equal radius in $\mathbf{R}^d$ to an expression involving volumes of intersections of at most $d + 1$ balls. Another possibility is to approximate $\mathrm{Leb}(Z_{n,i})$ by random methods, for example, sampling from a uniform distribution on a cube or ball containing $Z_{n,i}$ and using the fraction of the sample falling into $Z_{n,i}$ to estimate its volume. Using such a random sample of size J for each $Z_{n,i}$ gives a total cost of $O((n^2 + J)I)$ to compute the final estimate $L_n^I(\lambda)$.

The estimator $L_n^I(\lambda)$ serves as an approximation to the estimator $L_n(\lambda)$ that employs Naiman and Wynn's exact formula and considers all possible subsets $R_i \subset R$. For this estimator $L_n(\lambda)$ one obtains the following result.

THEOREM 4. *Under Assumption* B, *if* $0 < r_n < r_0$ *is fixed*, *then*

$$d_{\mathrm{Leb}}(L(\lambda), L_n(\lambda)) = \begin{cases} O\left(\left(\dfrac{\log n}{n}\right)^{2/(d+1)}\right) & a.s., \quad \text{if } d > 3, \\ O\left(\dfrac{\log n}{\sqrt{n}}\right) & a.s., \quad \text{if } d \le 3. \end{cases}$$

If r_0 is unknown, one has to let r_n shrink to 0 so that $r_n \gg (\log n/n)^{1/(d+1)}$ and incurs a penalty in the convergence rate as in Theorem 3. Of course, this penalty can be made arbitrarily small if r_n shrinks slowly enough.

REMARK 6. Up to log factors, the rates for $d > 2$ in Theorem 4 agree with those established by Polonik (1995), Proposition 3.7, for the empirical generalized λ-cluster [observe that (8) and Assumption B give $r = (d-1)/2$ and $\gamma = \infty$ for Polonik's Proposition 3.7].

REMARK 7. Following Proposition 3.8, Polonik remarks that for the problem of estimating the support of a uniform distribution, the empirical generalized λ-cluster converges with the minimax rate. Similarly, for $d > 2$ the rates of Theorem 4 are up to log factors minimax rates for Assumption B. This follows from Mammen and Tsybakov (1995), Theorem 5.1, Model 2. The class $\mathscr{G}_{2,L}$ used there is larger than $\mathscr{G}(r_0)$, but the proof of their Theorem 5.1 shows that the minimax rate is determined by the smoothness of a parametrization of the boundary in terms of a Lipschitz condition on the derivatives, and that condition (iv) of Theorem 1 implies the same minimax rate for $\mathscr{G}(r_0)$ as for $\mathscr{G}_{2,L}$. Mammen and Tsybakov (1995) also state that optimal rates for the problem of level set estimation for general smoothness classes are not yet known and comment on the difficulty of finding practical estimators.

REMARK 8. As in the smooth case, one can obtain a version of the estimator that belongs to the class $\mathscr{G}(r_0)$ by rolling a ball along the outside.

4. Comparison with other approaches and applications. This section puts the results of the previous sections in the context of competing techniques, which consist of two fundamentally different approaches: contouring a density estimate and the excess mass approach, which specifically incorporates shape information. Both of these techniques have several disadvantages when evaluated against the two main criteria, statistical performance and computational feasibility. The excess mass approach produces an estimator that satisfies the prescribed shape restrictions and has a better statistical performance if the density is not smooth. In the case where the density is smooth, however, contouring a density estimate based, for example, on an appropriate kernel allows one to exploit the smoothness and results in better rates of convergence. But this method does not allow one to incorporate prior shape information on the level set. The computational aspects are unsatisfactory for both methods: The excess mass approach, in general, does not seem amenable to a computational realization, and algorithms exist only for the classes of convex sets in $\mathbf{R}^2$ and ellipsoids in $\mathbf{R}^d$. Contouring a density estimate, based on a kernel or other method, provides the challenging task of implementing a data structure to store several multidimensional components of the resulting set. In addition, when estimating a minimum-volume set with given probability content, the right level for contouring has to be found by a search procedure that involves numerical integration of the density estimate over a candidate level set in each step.

It was shown in Theorem 2 that the class $\mathscr{G}(r)$ described in Theorem 1 plays a very important role in density estimation: Under some mild smoothness condition on the density, the level sets will belong to $\mathscr{G}(r)$ for some r. This explains the significance of the shape-restricted inference tailored to this class that is provided by the granulometric smoothing procedure. This procedure allows one to incorporate prior information on the rolling radius r. Furthermore, Theorem 3 and the following Remark 4 show how this procedure can be employed without this prior information, leading to a penalty term in the rate of convergence, and that this procedure is still consistent should the level set not belong to the widely applicable class $\mathscr{I}(r)$. It was shown both how the excess mass approach can be applied to the granulometric smoothing procedure and also how that procedure can be used in conjunction with a density estimator to exploit smoothness of the density, yielding excellent statistical performance in the smooth case as well as in the nonsmooth case. The analysis in Section 3 makes clear the computational advantages of the procedure: The computational complexity of this estimator depends on the dimension only in a linear way, and there exist simple algorithms and a simple data structure that allow easy computation, storage and manipulation of the estimator.

To put this technique into a concrete context, two important applications shall briefly be delineated. In a Bayesian context, highest posterior density regions are a common tool to summarize the structure of the posterior distribution. Chapter 5.3 of Tanner (1993) describes some of the difficulties involved in computing the contents and boundaries of highest posterior density regions even in low-dimensional situations. Note that a highest posterior density region with content $1-\alpha$ is a minimal volume set of the posterior density with that content. Thus, if one can sample from the posterior, for example, with the Gibbs sampler, the Metropolis algorithm or the Monte Carlo EM algorithm, then the set $C_n(1-\alpha)$ introduced in Section 3 gives a readily implemented estimator for the $(1-\alpha)$ highest posterior density region in any dimension.

In a frequentist context, Hall [(1992), page 160] describes the construction of multivariate bootstrap confidence regions whose shape is determined by the data so that the regions are (approximately) likelihood based in the sense described in Hall (1992), page 17. See also Cox and Hinkley [(1974), pages 236, 238] for arguments supporting likelihood-based confidence regions. Let $\hat{\theta}_n$ be an estimate of a d-dimensional parameter θ_0 based on n i.i.d. observations, $(1/n)\widehat{\Sigma}$ be an estimate of the covariance matrix of $\hat{\theta}_n$, and consider the root $T = n^{1/2}\widehat{\Sigma}^{-1/2}(\hat{\theta}_n - \theta_0)$. The goal is to find a $(1-\alpha)$-confidence region $\mathscr{R}$ for θ_0 by constructing (an estimate of) a level set S of the density f of T such that $\mathbb{P}(T \in S) = 1-\alpha$. So S is a minimum volume set for f with probability content $1-\alpha$. Following Hall [(1992), page 160] draw b resamples of size n to obtain $T_1^*, \dots, T_b^*$. Apply a density estimator to these b values of T^* to obtain a density estimate $\hat{f}_b$ of f. The resulting $(1-\alpha)$-confidence region as given on page 160 of Hall (1992) is then $\mathscr{R} = \{\hat{\theta} - n^{-1/2}\widehat{\Sigma}^{1/2}x \colon x \in S\}$, where S is the level set of $\hat{f}_b$ that contains a proportion $1-\alpha$ of the b values $T_1^*, \dots, T_b^*$. The difficulties arising in the construction of such a set were described in Section 1. A solution is given by the estimator $S = C_b(1-\alpha)$ introduced in Section 3.

Finally, it is possible to build on the ideas of Romano (1988a, b) and DasGupta, Ghosh and Zen (1995) to construct multivariate bootstrap confidence regions for the mode as a location parameter in a quite general setting. This idea will be developed elsewhere.

5. Proofs.

First, some auxiliary results will be given that will be used to prove Theorems 3 and 4. Theorem 2 will conveniently be proved together with Lemma 2.

For any compact set C in $\mathbf{R}^d$,

$$\operatorname{Leb}(S_{\pm\varepsilon}) = \operatorname{Leb}(S) + O(\varepsilon) \quad \text{as } \varepsilon \downarrow 0, \text{ uniformly in } S \in \mathscr{G}_C(r_0). \tag{6}$$

This can be shown using the property that if $S \in \mathscr{G}_C(r_0)$, then a ball of radius r_0 rolls freely in S and in $\overline{S^c}$.

LEMMA 1. *If* $0 < \varepsilon < r_1 \le r_2$ *and* $x_1, x_2 \in \mathbf{R}^d$ *with* $|x_1 - x_2| \le r_1 + r_2 - \varepsilon$, *then*

$$\operatorname{Leb}\big(B_{r_1}(x_1) \cap B_{r_2}(x_2)\big) \ge \frac{4c_{d-1}}{(d+1)2^{(d+1)/2}} r_1^{(d-1)/2} \varepsilon^{(d+1)/2},$$

where $c_d = \pi^{d/2}/\Gamma(1 + d/2)$.

PROOF. Let e be any unit vector. Elementary considerations show

$$\begin{aligned}\operatorname{Leb}\big(B_{r_1}(x_1) \cap B_{r_2}(x_2)\big) &\ge \operatorname{Leb}\big(B_{r_1}(0) \cap B_{r_2}((r_1 + r_2 - \varepsilon)e)\big)\\ &\ge \operatorname{Leb}\big(B_{r_1}(0) \cap B_{r_1}((2r_1 - \varepsilon)e)\big)\\ &= 2\operatorname{Leb}\left(B_{r_1}(0) \cap \left\{y\colon \langle y, e\rangle \ge r_1 - \frac{\varepsilon}{2}\right\}\right).\end{aligned}$$

For $0 \le x \le r_1$ the hyperplane $\{y\colon \langle y, e\rangle = r_1 - x\}$ intersects $B_{r_1}(0)$ in a $(d-1)$-dimensional ball of radius $\sqrt{(2r_1 - x)x}$, which has volume $c_{d-1}(\sqrt{(2r_1 - x)x})^{d-1}$. So by Fubini's theorem the last line in the preceding equation equals

$$\begin{aligned}2\int_0^{\varepsilon/2} c_{d-1}\Big(\sqrt{(2r_1 - x)x}\Big)^{d-1}\,dx &\ge 2c_{d-1}\int_0^{\varepsilon/2} (r_1 x)^{(d-1)/2}\,dx \quad \text{as } \frac{\varepsilon}{2} \le r_1\\ &= \frac{4c_{d-1}}{(d+1)2^{(d+1)/2}} r_1^{(d-1)/2} \varepsilon^{(d+1)/2}. \qquad \square\end{aligned}$$

LEMMA 2. *Under the assumptions of Theorem* 2 *there exists* $h > 0$ *such that*:

(a) *For all* $\lambda \in [l, u]$ *and all* t *with* $0 < |t| \le h$,

$$\inf_{x \in L(\lambda)_t} f(x) > \lambda - \left(\frac{m}{2}t \vee 2mt\right)$$

and

$$\sup_{x \in (L(\lambda)^c)_t} f(x) < \lambda + \left(\frac{m}{2}t \vee 2mt\right).$$

(b) *If* λ_1, λ_2 *are such that* $l \le \lambda_1 < \lambda_2 \le u$ *and* $\lambda_2 - \lambda_1 \le (m/2)h$, *then*

$$L(\lambda_1) \ominus \frac{2}{m}(\lambda_2 - \lambda_1)B \subset L(\lambda_2) \subset L(\lambda_1) \ominus \frac{1}{2m}(\lambda_2 - \lambda_1)B,$$

$$L(\lambda_2) \oplus \frac{1}{2m}(\lambda_2 - \lambda_1)B \subset L(\lambda_1) \subset L(\lambda_2) \oplus \frac{2}{m}(\lambda_2 - \lambda_1)B.$$

(c) *For* λ_1, λ_2 *as in* (*b*),

$$\mathrm{Leb}(L(\lambda_1)) \ge \mathrm{Leb}(L(\lambda_2)) + d\big(\mathrm{Leb}(L(\lambda_2))\big)^{(d-1)/d} \frac{c_d^{1/d}}{2m}(\lambda_2 - \lambda_1),$$

where $c_d = \pi^{d/2}/\Gamma(1 + d/2)$.

PROOF OF THEOREM 2 AND LEMMA 2. Recall that a nonempty set $S \subset \mathbf{R}^d$ is called a $(d-1)$-dimensional surface if $S = f^{-1}(0) := \{x\colon f(x) = 0\}$ for some smooth function $f\colon U \to \mathbf{R}$, $U \subset \mathbf{R}^d$ open and $\operatorname{grad} f(x) \ne \mathbf{0}$ for all $x \in S$ [see Thorpe (1979), page 16]. S is a $(d-1)$-dimensional C^1 submanifold iff it is locally a $(d-1)$-dimensional surface with $f \in C^1$; see Theorem 2.1.2 in Berger and Gostiaux (1988).

If $x \in \overline{L(l-\eta)} \setminus \operatorname{int} L(u+\eta) \subset U$, then

$$f\left(x \pm \varepsilon \frac{\operatorname{grad} f(x)}{|\operatorname{grad} f(x)|}\right) - f(x) \ \substack{> \\ <}\ \pm \frac{m}{2}\varepsilon \quad \text{for } 0 < \varepsilon \le \eta_1 \tag{7}$$

and some $\eta_1 > 0$. This is so because $|\operatorname{grad} f(x)| \ge m$ on U gives (7) for some $\eta_1 = \eta_1(x)$, but as $\overline{L(l-\eta)} \setminus \operatorname{int} L(u+\eta)$ is compact and $f \in C^1(U)$, one can pick a universal $\eta_1 > 0$.

Now let $\lambda \in [l, u]$ and $x \in f^{-1}(\lambda) := \{y\colon f(y) = \lambda\}$. Then $x \in \overline{L(l)} \setminus \operatorname{int} L(u+\eta)$, so (7) shows $f^{-1}(\lambda) \subset \partial L(\lambda)$. Conversely, $\partial L(\lambda) \subset f^{-1}(\lambda)$ as f is continuous in a neighborhood of $\partial L(\lambda)$ as $\partial L(\lambda) \subset \overline{L(\lambda)} \setminus \operatorname{int} L(u+\eta) \subset U$. So $\partial L(\lambda) = f^{-1}(\lambda) = (f|_U)^{-1}(\lambda)$ and thus $\partial L(\lambda)$ is a $(d-1)$-dimensional C^1 submanifold. By elementary differential geometry [see, e.g., Thorpe (1979), page 14], $\operatorname{grad} f(x)$ is a normal vector to $f^{-1}(\lambda)$ at $x \in f^{-1}(\lambda)$. This shows that

$$n(x) := -\frac{\operatorname{grad} f(x)}{|\operatorname{grad} f(x)|}$$

is an outward-pointing unit normal vector at $x \in \partial L(\lambda)$. Then one obtains, for $x, y \in \partial L(\lambda)$,

$$\begin{aligned}|n(x) - n(y)|^2 &\le \frac{|\operatorname{grad} f(x)||\operatorname{grad} f(y)|}{m^2}\left(2 - 2\frac{\langle \operatorname{grad} f(x), \operatorname{grad} f(y)\rangle}{|\operatorname{grad} f(x)||\operatorname{grad} f(y)|}\right) \\ &= \frac{2|\operatorname{grad} f(x)||\operatorname{grad} f(y)|}{m^2} \\ &\quad + \frac{1}{m^2}\big(|\operatorname{grad} f(x) - \operatorname{grad} f(y)|^2 - |\operatorname{grad} f(x)|^2 - |\operatorname{grad} f(y)|^2\big)\end{aligned}$$

$$= \frac{1}{m^2}\big(|\operatorname{grad} f(x) - \operatorname{grad} f(y)|^2 - \big(|\operatorname{grad} f(x)| - |\operatorname{grad} f(y)|\big)^2\big)$$
$$\le \frac{k^2}{m^2}|x-y|^2.$$

Hence $\partial L(\lambda)$ satisfies condition (iv) of Theorem 1 with $r_0 = m/k$. Further, $L(\lambda)$ is compact as U is bounded and as $\partial L(\lambda) \subset f^{-1}(\lambda)$ was shown previously. This proves Theorem 2.

The assertion concerning the inf in Lemma 2(a) is obtained by showing that for some $h > 0$ the following three inequalities hold for all $\lambda \in [l, u]$ and all $t > 0$: $f(x) > \lambda + (m/2)t$ if $x \in L(\lambda) \ominus tB$ and $d(x, \partial L(\lambda)) \le \eta_1$; $f(x) > \lambda + (m/2)h$ if $x \in L(\lambda) \ominus tB$ and $d(x, \partial L(\lambda)) > \eta_1$; and $f(x) > \lambda - 2mt$ if $x \in L(\lambda) \oplus tB$ and $0 < t \le h$. The routine proof of these inequalities is omitted. The assertion concerning the sup in Lemma 2(a) is proved analogously.

As for part (b), setting $t = -(2/m)(\lambda_2 - \lambda_1)$ in part (a) gives

$$\inf_{x \in L(\lambda_1) \ominus (2/m)(\lambda_2 - \lambda_1)B} f(x) > \lambda_1 + (\lambda_2 - \lambda_1) = \lambda_2,$$

proving the first inclusion. Further, setting $t = (1/2m)(\lambda_2 - \lambda_1)$ and using (1), one obtains $\sup_{x \in (L(\lambda_1) \ominus (1/2m)(\lambda_2 - \lambda_1)B)^c} f(x) < \lambda_1 + (\lambda_2 - \lambda_1) = \lambda_2$, whence $(L(\lambda_1) \ominus (1/2m)(\lambda_2 - \lambda_1)B)^c \subset L(\lambda_2)^c$, which yields the second inclusion of the first assertion. The second assertion follows analogously.

Part (c) follows from (b) and the Brunn-Minkowski inequality [see, e.g., Burago and Zalgaller (1988)]: As $L(\lambda_2)$ is nonempty and compact by Theorem 2, said inequality gives

$$\operatorname{Leb}\bigg(L(\lambda_2) \oplus \frac{1}{2m}(\lambda_2 - \lambda_1)B\bigg) \ge \bigg[\big(\operatorname{Leb}(L(\lambda_2))\big)^{1/d} + \bigg(\operatorname{Leb}\bigg(\frac{1}{2m}(\lambda_2 - \lambda_1)B\bigg)\bigg)^{1/d}\bigg]^d.$$

Recall

$$\operatorname{Leb}\bigg(\frac{\lambda_2 - \lambda_1}{2m}B\bigg) = c_d\bigg(\frac{\lambda_2 - \lambda_1}{2m}\bigg)^d. \qquad \square$$

LEMMA 3. *Let $C \subset \mathbf{R}^d$ be compact, $r_0 > 0$ and X_i, $1 \le i \le n$, be i.i.d. from some density f on $\mathbf{R}^d$.*

(a) *If $f \ge b > 0$ on $S \in \mathscr{S}_C(r_0)$ and $0 < \varepsilon < r/2 \wedge r_0$, then*

$$\mathbb{P}\big(S_{r-2\varepsilon} \not\subset \big(S \cap \{X_i,\ 1 \le i \le n\}\big) \oplus rB\big)$$
$$\le D(\varepsilon, S_r)\exp\big(-nab\min(r-\varepsilon, r_0)^{(d-1)/2}\varepsilon^{(d+1)/2}\big),$$

where $D(\varepsilon, S_r) := \max\{\operatorname{card} M\colon M \subset S_r, |x-y| > \varepsilon$ for different $x, y \in M\}$ is the packing number for S_r, and a is a dimensional constant.

(b) *If further $f \ge b > 0$ on C, $0 < \varepsilon < r/3 \wedge 1$ and $r_0 \ge r - 2\varepsilon$, then*

$$\mathbb{P}\big(S_{r-3\varepsilon} \not\subset \big(S \cap \big\{X_i,\ 1 \le i \le n\big\}\big) \oplus rB \text{ for some } S \in \mathscr{S}_C(r_0)\big)$$
$$\le D(\varepsilon, C_r)D\bigg(\frac{\varepsilon}{10r}, S^{d-1}\bigg)\exp\big(-nab(r-2\varepsilon)^{(d-1)/2}(\varepsilon/2)^{(d+1)/2}\big).$$

PROOF. We will employ a variation of the packing argument used in the proof of Theorem 1[a] in Dümbgen and Walther (1996). Write $\mathscr{X}_n^S$ for $S \cap \{X_i,\ 1 \le i \le n\}$. Let M be a maximal subset of $S_{r-2\varepsilon}$ such that $|x-y| > \varepsilon$ for different $x, y \in M$. So, for any set $T \subset \mathbf{R}^d$, $M \subset T$ implies $S_{r-2\varepsilon} \subset T \oplus \varepsilon B$. Hence

$$\begin{aligned}\mathbb{P}\big(S_{r-2\varepsilon} \not\subset \mathscr{X}_n^S \oplus rB\big) &\le \mathbb{P}\big(M \not\subset \mathscr{X}_n^S \oplus (r-\varepsilon)B\big)\\ &\le \sum_{m\in M} \mathbb{P}\big(\mathscr{X}_n^S \cap B_{r-\varepsilon}(m) = \varnothing\big)\\ &\le D(\varepsilon, S_r) \sup\nolimits_{m\in S_{r-2\varepsilon}} \mathbb{P}\big(\mathscr{X}_n^S \cap B_{r-\varepsilon}(m) = \varnothing\big).\end{aligned}$$

Let $m \in S_{r-2\varepsilon}$. Then there exists $s \in S$ with $|m-s| \le r-2\varepsilon$. As S satisfies the conditions of Theorem 1 with $r_0 > 0$, there exists $x \in S$ with $s \in B_{r_0}(x) \subset S$, so $|m-x| \le r_0 + r - 2\varepsilon$. Lemma 1 shows that

$$\mathrm{Leb}(B_{r-\varepsilon}(m) \cap B_{r_0}(x)) \ge a \min(r-\varepsilon, r_0)^{(d-1)/2} \varepsilon^{(d+1)/2},$$

where a is a dimensional constant. Hence

$$\begin{aligned}\mathbb{P}\big(\mathscr{X}_n^S \cap B_{r-\varepsilon}(m) = \varnothing\big) &= \mathbb{P}\big(X_i \notin S \cap B_{r-\varepsilon}(m) \text{ for all } 1 \le i \le n\big)\\ &= \prod_{i=1}^n \big(1 - \mathbb{P}\big(X_i \in S \cap B_{r-\varepsilon}(m)\big)\big)\\ &\le \big(1 - \mathbb{P}\big(X_1 \in B_{r_0}(x) \cap B_{r-\varepsilon}(m)\big)\big)^n\\ &\le \big(1 - ab \min(r-\varepsilon, r_0)^{(d-1)/2} \varepsilon^{(d+1)/2}\big)^n\\ &\le \exp\big(-nab \min(r-\varepsilon, r_0)^{(d-1)/2} \varepsilon^{(d+1)/2}\big),\end{aligned}$$

which proves part (a). To prove (b), let M be a maximal subset of $C_{r-3\varepsilon}$ such that $|x-y| > \varepsilon$ for different $x, y \in M$ and set $M^S := M \cap S$. If $S \in \mathscr{G}_C(r_0)$ and $s \in S_{r-3\varepsilon}$, then it follows from Theorem 1(iii) that $s \in B_{r_0}(x) \subset S_{r-3\varepsilon}$ for some x. This implies $S_{r-3\varepsilon} \subset (M^{S_{r-3\varepsilon}}) \oplus 2\varepsilon B$. Hence

$$\begin{aligned}&\mathbb{P}\big(\text{there exists } S \in \mathscr{G}_C(r_0)\text{: } S_{r-3\varepsilon} \not\subset \mathscr{X}_n^S \oplus rB\big)\\ &\quad\le \mathbb{P}\big(\text{there exists } S \in \mathscr{G}_C(r_0)\text{: } M^{S_{r-3\varepsilon}} \not\subset \mathscr{X}_n^S \oplus (r-2\varepsilon)B\big)\\ &\quad= \mathbb{P}\big(\text{there exists } m \in M\text{: } m \in S_{r-3\varepsilon} \text{ and } m \notin \mathscr{X}_n^S \oplus (r-2\varepsilon)B\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for some } S \in \mathscr{G}_C(r_0)\big)\\ &\quad\le D(\varepsilon, C_{r-3\varepsilon}) \sup_{m\in C_{r-3\varepsilon}} \mathbb{P}\big(m \in S_{r-3\varepsilon} \text{ and } \mathscr{X}_n^S \cap B_{r-2\varepsilon}(m) = \varnothing\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for some } S \in \mathscr{G}_C(r_0)\big).\end{aligned}$$

If $m \in S_{r-3\varepsilon}$, then as in the proof of part (a) one sees that there exists x with $|m-x| \le r_0 + r - 3\varepsilon$ and $B_{r_0}(x) \subset S$. Hence the last probability in the preceding inequality is not larger than

$$\mathbb{P}\Big(\mathscr{X}_n \cap B_{r_0}(x) \cap B_{r-2\varepsilon}(m) = \varnothing \\ \text{for some } x \text{ with } B_{r_0}(x) \subset C \text{ and } |m-x| \le r_0 + r - 3\varepsilon\Big).$$

One verifies that $r_0 \ge r - 2\varepsilon$ implies for $e := (x-m)/|x-m|$ (and any unit vector e if $|x-m| = 0$):

$$\begin{aligned} B_{r_0}(x) \cap B_{r-2\varepsilon}(m) &\supset B_{r_0}(m + (r_0 + r - 3\varepsilon)e) \cap B_{r-2\varepsilon}(m) \\ &\supset B_{r_0}\big(m + \big(r_0 + r - \tfrac{5}{2}\varepsilon\big)\tilde{e}\big) \cap B_{r-2\varepsilon}(m), \end{aligned}$$

whenever $\tilde{e}$ is a unit vector with $|e - \tilde{e}| \le \varepsilon/10r$ and $\varepsilon \le 1$. So if $\widetilde{M}$ is a maximal subset of the unit sphere so that $|x-y| > \varepsilon/10r$ for different $x, y \in \widetilde{M}$, then the previous probability is not larger than

$$\begin{aligned} &\mathbb{P}\bigg(\mathscr{X}_n \cap B_{r_0}\bigg(m + \bigg(r_0 + r - \frac{5}{2}\varepsilon\bigg)\tilde{e}\bigg) \cap B_{r-2\varepsilon}(m) = \varnothing \text{ for some } \tilde{e} \in \widetilde{M} \text{ with} \\ &\qquad\qquad B_{r_0}\bigg(m + \bigg(r_0 + r - \frac{5}{2}\varepsilon\bigg)\tilde{e}\bigg) \cap B_{r-2\varepsilon}(m) \subset C\bigg) \\ &\quad\le \operatorname{card} \widetilde{M}\bigg(1 - \mathbb{P}\bigg(X_1 \in B_{r_0}\bigg(m + \bigg(r_0 + r - \frac{5}{2}\varepsilon\bigg)\tilde{e}\bigg) \cap B_{r-2\varepsilon}(m)\bigg)\bigg)^n \\ &\qquad\qquad \text{where } B_{r_0}\bigg(m + \bigg(r_0 + r - \frac{5}{2}\varepsilon\bigg)\tilde{e}\bigg) \cap B_{r-2\varepsilon}(m) \subset C \\ &\quad\le \operatorname{card} \widetilde{M}\big(1 - ab\min(r - 2\varepsilon, r_0)^{(d-1)/2}(\varepsilon/2)^{(d+1)/2}\big)^n \\ &\quad\le D\bigg(\frac{\varepsilon}{10r}, S^{d-1}\bigg)\exp\big(-nab(r-2\varepsilon)^{(d-1)/2}(\varepsilon/2)^{(d+1)/2}\big). \qquad \square \end{aligned}$$

PROPOSITION 1. *Let F be a probability measure on $\mathbf{R}^d$ with bounded density, $C \subset \mathbf{R}^d$ be compact and $r_0 > 0$. Then*

$$\sup_{S \in \mathscr{G}_C(r_0)} |F_n(S) - F(S)| = \begin{cases} O\big(n^{-2/(d+1)}\big) & a.s., \quad \text{if } d > 3, \\ O\bigg(\dfrac{\log n}{\sqrt{n}}\bigg) & a.s., \quad \text{if } d = 1, 2, 3. \end{cases}$$

PROOF. Let

$$N_I(\varepsilon, \mathscr{G}_C(r_0), F) := \inf\{m\colon \text{there exist measurable } G_1, \dots, G_m \subset \mathbf{R}^d \text{ such that for every } G \in \mathscr{G}_C(r_0) \text{ there exist } i, j \in \{1, \dots, m\} \text{ with } G_i \subset G \subset G_j \text{ and } F(G_j \setminus G_i) < \varepsilon\}$$

be the metric entropy with inclusion of $\mathscr{G}_C(r_0)$ with respect to F. We will show, for $d \geq 2$,

$$\log N_I(\varepsilon, \mathscr{G}_C(r_0), F) \leq M\varepsilon^{-(d-1)/2} \quad \text{for } 0 < \varepsilon \leq 1 \text{ and some } M. \tag{8}$$

Then Corollary 2.5 in Alexander (1984) gives the assertion for the case $d \geq 3$. [In his definition of metric entropy N_2^B, Alexander requires that the bracketing sets G_i, G_j be in $\mathscr{G}_C(r_0)$. However, the proof of his Theorem 2.3 (on which Corollary 2.5 is based) goes through without this requirement. So we may substitute $N_I(\varepsilon^2, \mathscr{G}_C(r_0), F)$ for $N_2^B(\varepsilon, \mathscr{G}_C(r_0), F)$.]

In the case $d = 1$ one can readily find an explicit bound for N_I, but for the following it is enough to note that N_I for $d = 1$ is not larger than N_I for $d = 2$. So if $d = 1, 2$, then $\int_0^1 (\log N_I(\varepsilon^2, \mathscr{G}_C(r_0), F))^{1/2}\, d\varepsilon < \infty$, and Theorem 6.2.1 in Dudley (1984) gives the assertion for this case.

It remains to prove (8). Let $G \in \mathscr{G}_C(r_0)$. Part (iv) of Theorem 1 and Theorem 2.1.2(iv) in Berger and Gostiaux (1988) show that ∂G is locally the graph of a real-valued function f defined on $\mathbf{R}^{d-1}$ (possibly after a permutation of the coordinate axes) and that grad f satisfies a uniform Lipschitz condition. Hence we can find a fixed number $N = N(r_0)$ of overlapping rectangles R_i, $i = 1, \dots, N$, covering C such that for each $G \in \mathscr{G}_C(r_0)$, $G = \bigcup_{i \in J}(G \cap R_i)$, $J \subset \{1, \dots, N\}$, and $G \cap R_i \in \mathscr{C}(2, k, d)$ for some k, where $\mathscr{C}(2, k, d)$ is defined in Dudley (1984), page 51. Theorem 7.1.1 in Dudley (1984) shows $\log N_I(\varepsilon, \mathscr{C}(2, k, d), F) \leq M_1 \varepsilon^{-(d-1)/2}$, $0 < \varepsilon \leq 1$. But $G = \bigcup_{i \in J}(G \cap R_i)$ and $|J| \leq N$ imply $N_I(\varepsilon, \mathscr{G}_C(r_0), F) \leq (N_I(\varepsilon/N, \mathscr{C}(2, k, d), F))^N$. Equation (8) follows. □

PROOF OF THEOREM 3. We will first prove (4) by showing that there exist constants $c, M > 0$ such that, for $\varepsilon_n := c(\log n/n)^{2/(d+1)} r_n^{-(d-1)/(d+1)}$

$$\mathbb{P}\big(L(\lambda)_{-2M\sigma_n^p - \varepsilon_n} \subset L_n(\lambda) \text{ for all } \lambda \in [l, u] \text{ eventually}\big) = 1, \tag{9}$$

$$\mathbb{P}\big(L_n(\lambda) \subset L(\lambda)_{\varepsilon_n + 2M\sigma_n^p} \text{ for all } \lambda \in [l, u] \text{ eventually}\big) = 1. \tag{10}$$

Then (4) follows as $L(\lambda)_{-2M\sigma_n^p - \varepsilon_n} \subset L_n(\lambda)$ implies $L(\lambda) \subset L_n(\lambda)_{2M\sigma_n^p + \varepsilon_n}$ for n large enough by Lemma B.1(c) of Walther (1995), because $L(\lambda) \in \mathscr{G}(m/k)$ for all $\lambda \in [l, u]$ by Theorem 2. Equations (9) and (10) will be proved in two steps. Define the events

$$A_n := \big\{\omega \colon \mathscr{X}_n^+(\lambda)(\omega) \subset L(\lambda)_{(M/2)\sigma_n^p} \text{ for all } \lambda \in [l, u]\big\},$$
$$B_n := \big\{\omega \colon \mathscr{X}_n^-(\lambda)(\omega) \subset \big(L(\lambda)_{-(M/2)\sigma_n^p}\big)^c \text{ for all } \lambda \in [l, u]\big\}.$$

First, it will be shown that

$$\mathbb{P}(A_n \text{ eventually}) = \mathbb{P}(B_n \text{ eventually}) = 1, \tag{11}$$

$$\begin{aligned} &\mathbb{P}\Big(L(\lambda)_{M\sigma_n^p} \setminus L(\lambda)_{-(r_n - \varepsilon_n) + 2M\sigma_n^p} \\ &\qquad \subset \mathscr{X}_n^-(\lambda) \oplus r_n B \text{ for all } \lambda \in [l, u] \text{ eventually}\Big) = 1. \end{aligned} \tag{12}$$

Then these auxiliary results will be used to prove (9) and (10).

To prove (11), observe that one can find $r > 0$ such that the compact set $U(r) := (L(l) \setminus \operatorname{int} L(u)) \oplus rB$ satisfies $U(r) \subset U$. The kernel K satisfies the assumptions of Theorem 3.1 in Stute (1984). That theorem and Taylor's theorem, respectively, show that there is a compact set $C \supset U(r/2)$ such that

$$\sup_C |\hat{f}_n - f_{\sigma_n}| = O\big(n^{-p/(d+2p)}\big) \quad \text{a.s.}, \qquad n \to \infty, \tag{13}$$

$$\sup_C |f_{\sigma_n} - f| = O\big(\sigma_n^p\big), \qquad n \to \infty, \tag{14}$$

where we write f_{σ_n} for $\int \sigma_n^{-d} K((\cdot - x)/\sigma_n) f(x)\,dx$. Using (MON) of Walther (1995), one verifies

$$(L(u) \cap C^c) \oplus \frac{r}{4} B \subset L(u) \ominus \frac{r}{4} B.$$

Together with the fact that K has bounded support, $\sigma_n \to 0$ and Lemma 2(a), this shows that, for some $v > 0$ and n large enough,

$$\inf_{L(u)\cap C^c} f_{\sigma_n} \geq \inf_{(L(u)\cap C^c)\oplus(r/4)B} f > u + v \quad \text{and} \quad \sup_{(L(l))^c \cap C^c} f_{\sigma_n} < l - v,$$

where the last inequality follows in an analogous way. $\hat{f}_n$ converges to f_{σ_n} uniformly on C^c, as $\sigma_n^d \gg \log n/n$, $\sup f < \infty$ and the assumptions on K guarantee that the graphs of $K(\cdot - x/\sigma)$ have polynomial discrimination; see Pollard (1984), page 36. Hence we obtain

$$\begin{aligned} &\mathbb{P}\Big(\inf_{L(u)\cap C^c} \hat{f}_n(x) > u + \frac{v}{2} \text{ eventually}\Big) = 1, \\ &\mathbb{P}\Big(\sup_{(L(l))^c \cap C^c} \hat{f}_n(x) < l - \frac{v}{2} \text{ eventually}\Big) = 1. \end{aligned} \tag{15}$$

Equations (13)–(15) together with Lemma 2(a) imply that there exists $M > 0$ such that

$$\mathbb{P}\Big(\sup_{x \in (L(\lambda)\oplus(M/2)\sigma_n^p B)^c} \hat{f}_n(x) < \lambda \text{ for all } \lambda \in [l, u] \text{ eventually}\Big) = 1,$$

$$\mathbb{P}\Big(\inf_{x \in L(\lambda)\ominus(M/2)\sigma_n^p B} \hat{f}_n(x) > \lambda \text{ for all } \lambda \in [l, u] \text{ eventually}\Big) = 1.$$

Equation (11) follows.

To prove (12), let N_n be the integer part of $2(u-l)((m/2)M\sigma_n^p)^{-1}$ and let $l = \lambda_1 < \cdots < \lambda_{N_n} = u$ be such that $\lambda_{k+1} - \lambda_k \leq (m/2)M\sigma_n^p$ for all $k < N_n$. Then

$$L(\lambda_k)_{-(r_n-\varepsilon_n)+M\sigma_n^p} \subset L(\lambda_{k+1})_{-(r_n-\varepsilon_n)+2M\sigma_n^p} \tag{16}$$

for all $k < N_n$ by Lemma B.1(c) of Walther (1995) and Lemma 2(b), provided n is large enough. Now suppose

$$L(\lambda_k)_{M\sigma_n^p} \setminus L(\lambda_k)_{-(r_n-\varepsilon_n)+M\sigma_n^p} \subset \mathscr{X}_n^-(\lambda_k) \oplus r_n B \quad \text{for all } k = 1, \dots, N_n. \tag{17}$$

Let $\lambda \in [l, u]$ and let k be such that $\lambda \in [\lambda_k, \lambda_{k+1}]$. Then $L(\lambda_{k+1}) \subset L(\lambda) \subset L(\lambda_k)$ and (16) give

$$\begin{aligned} L(\lambda)_{M\sigma_n^p} \setminus L(\lambda)_{-(r_n-\varepsilon_n)+2M\sigma_n^p} &\subset L(\lambda_k)_{M\sigma_n^p} \setminus L(\lambda_k)_{-(r_n-\varepsilon_n)+M\sigma_n^p} \\ &\subset \mathscr{X}_n^{-}(\lambda_k) \oplus r_n B \\ &\subset \mathscr{X}_n^{-}(\lambda) \oplus r_n B \quad \text{by the definition of } \mathscr{X}_n^{-}. \end{aligned}$$

So if the set on the left-hand side of the preceding inclusion is not contained in $\mathscr{X}_n^{-}(\lambda) \oplus r_n B$ for some λ, then the inclusion in (17) must fail for some k. This shows

$$\begin{aligned} &\mathbb{P}\big(\big\{L(\lambda)_{M\sigma_n^p} \setminus L(\lambda)_{-(r_n-\varepsilon_n)+2M\sigma_n^p} \not\subset \mathscr{X}_n^{-}(\lambda) \oplus r_n B \text{ for some } \lambda \in [l, u]\big\} \cap A_n\big) \\ &\qquad \le \sum_{k=1}^{N_n} \mathbb{P}\big(\big\{L(\lambda_k)_{M\sigma_n^p} \setminus L(\lambda_k)_{-(r_n-\varepsilon_n)+M\sigma_n^p} \not\subset \mathscr{X}_n^{-}(\lambda_k) \oplus r_n B\big\} \cap A_n\big) \end{aligned} \tag{18}$$

(A_n is included for later use of this inequality.)

Next, by Assumption A one can find $0 < r < r_0/2$ such that $f \ge l/2$ on $L(l)_r$. Pick any $\lambda \in [l, u]$. One verifies that for large enough n the (nonrandom) set $S(n) := L(\lambda)_r \setminus \operatorname{int} L(\lambda)_{M\sigma_n^p}$ satisfies the conditions of Theorem 1 with rolling radius at least $r/4$. On the event A_n we have $S(n) \cap \{X_i, i = 1, \ldots, n\} \subset \mathscr{X}_n^{-}(\lambda)$. Hence Lemma 3(a) gives

$$\begin{aligned} &\mathbb{P}\big(\big\{S(n)_{r_n-\varepsilon_n} \not\subset \mathscr{X}_n^{-}(\lambda) \oplus r_n B\big\} \cap A_n\big) \\ &\qquad \le D(\varepsilon_n/2, S(n)_{r_n}) \\ &\qquad\qquad \times \exp\big(-2nal \min(r_n - \varepsilon_n/2, r/4)^{(d-1)/2} (\varepsilon_n/2)^{(d+1)/2}\big). \end{aligned} \tag{19}$$

It is easy to check that if $S, T \subset \mathbf{R}^d$, $0 \le b$, $0 < c$ and $S \oplus cB \subset T$, then $(T \setminus S)_b = T_b \setminus S_{-b}$. Together with $r + r_n - \varepsilon_n \ge M\sigma_n^p$ and Lemma B.1(c) of Walther (1995), this gives

$$\begin{aligned} S(n)_{r_n-\varepsilon_n} &\supset L(\lambda)_{r+r_n-\varepsilon_n} \setminus (L(\lambda)_{M\sigma_n^p})_{-r_n+\varepsilon_n} \\ &\supset L(\lambda)_{M\sigma_n^p} \setminus L(\lambda)_{-(r_n-\varepsilon_n)+M\sigma_n^p}. \end{aligned} \tag{20}$$

Using $\operatorname{diameter}(S(n)_{r_n}) \le \operatorname{diameter}(L(l)) + 2r_0$, the bound in Pollard (1990), page 14, yields

$$D(\varepsilon_n/2, S(n)_{r_n}) \le \big(4(\operatorname{diameter}(L(l)) + 2r_0)\varepsilon_n^{-1} + 1\big)^d. \tag{21}$$

Equations (19)–(21) and $\varepsilon_n < r_n$ show that for n large enough the right-hand side of (18) is bounded by

$$N_n C_1 \varepsilon_n^{-d} \exp\big(-C_2 n r_n^{(d-1)/2} \varepsilon_n^{(d+1)/2}\big) \le n^q \exp\big(-C_2 c^{(d+1)/2} \log n\big),$$

with some constants $C_1, C_2, q > 0$, as $N_n \le \text{const}\, \sigma_n^{-p} = \text{const}(n/\log n)^{p/(d+2p)}$. Choosing the factor c in ε_n large enough yields $\sum_n n^q \exp(C_2 c^{(d+1)/2} \log n) < \infty$, so (12) follows from (18) via the Borel–Cantelli lemma and because $\mathbb{P}(A_n \text{ eventually}) = 1$.

Equation (9) can be proved in a similar way: One checks that $L(\lambda_k)_{-M\sigma_n^p-\varepsilon_n} \subset L_n(\lambda_k)$ for all $k = 1, \dots, N_n$ implies $L(\lambda)_{-2M\sigma_n^p-\varepsilon_n} \subset L_n(\lambda)$ for all $\lambda \in [l, u]$. So once we find a constant $r > 0$ such that for each $\lambda \in [l, u]$ and n large enough (and not depending on λ)

$$\begin{aligned} &\mathbb{P}\big(\{S(n)_{r_n-\varepsilon_n} \not\subset L_n(\lambda)\} \cap B_n\big) \\ &\qquad \le D(\varepsilon_n/2, S(n)_{r_n}) \exp\big(-nal \min(r_n - \varepsilon_n/2, r)^{(d-1)/2}(\varepsilon_n/2)^{(d+1)/2}\big), \end{aligned} \tag{22}$$

where $S(n) = L(\lambda)_{-M\sigma_n^p-r_n}$, then (9) will follow from the Borel–Cantelli lemma as before. This is so because $\mathbb{P}(B_n$ eventually$) = 1$ and $S(n)_{r_n-\varepsilon_n} = L(\lambda)_{-M\sigma_n^p-\varepsilon_n}$ for n large enough; see Lemma B.1(c) of Walther (1995). Lemma B.2 of Walther (1995) shows that $S(n) \in \mathscr{G}(r)$ for large enough n. Using (1), one sees that on the event B_n we have $S(n) \cap \{X_i,\ i = 1, \dots, n\} \subset (\mathscr{X}_n^-(\lambda) \oplus r_n B)^c \cap \mathscr{X}_n^+(\lambda)$. So (22) follows from Lemma 3(a).

Finally, (11) and (12) show that $\mathbb{P}((\mathscr{X}_n^-(\lambda) \oplus r_n B)^c \cap \mathscr{X}_n^+(\lambda) \subset L(\lambda)_{-(r_n-\varepsilon_n)+2M\sigma_n^p}$ for all $\lambda \in [l, u]$ eventually$) = 1$. So (10) follows from Lemma B.1(c) of Walther (1995).

To prove the claim (5) concerning $C_n(\gamma)$, set $a_n := \varepsilon_n + 2M\sigma_n^p$. Assume

$$\begin{aligned} &L(\lambda)_{-a_n} \subset L_n(\lambda) \subset L(\lambda)_{a_n} \quad \text{for all } \lambda \in [l, u], \\ &\sup_{\lambda \in [l, u]} |F_n(L(\lambda)) - F(L(\lambda))| \le a_n. \end{aligned} \tag{23}$$

We will show that (23) implies

$$\sup_{\gamma \in [\underline{\gamma}, \overline{\gamma}]} |\lambda_n(\gamma) - \lambda(\gamma)| \le D a_n \tag{24}$$

for some constant D specified later, provided n is large enough.

But (24) implies, for every $\gamma \in [\underline{\gamma}, \overline{\gamma}]$, $L(\lambda(\gamma))_{-(2/m)Da_n} \subset L(\lambda_n(\gamma)) \subset L(\lambda(\gamma))_{(2/m)Da_n}$ if $Da_n \le (m/2)h$ by Lemma 2(b). Together with (23), this gives $L(\lambda(\gamma))_{-(2/m)Da_n-a_n} \subset L_n(\lambda_n(\gamma))$ and also $L_n(\lambda_n(\gamma)) \subset L(\lambda(\gamma))_{(2/m)Da_n+a_n}$. This shows

$$C(\gamma) \subset C_n(\gamma)_{(2/m)Da_n+a_n} \quad \text{and} \quad C_n(\gamma) \subset C(\gamma)_{(2/m)Da_n+a_n} \tag{25}$$

for n large enough, because then

$$L(\lambda(\gamma)) = \big(L(\lambda(\gamma))_{-(2/m)Da_n-a_n}\big)_{(2/m)Da_n+a_n} \quad \text{as } L(\lambda(\gamma)) \in \mathscr{G}(m/k).$$

Now the assertion of the theorem follows because the statements in (23) hold eventually a.s. by (9), (10) and Proposition 1 together with Theorem 2.

It remains to prove that (23) implies (24). Set $D := \widetilde{C} + 2m$ with

$$\widetilde{C} = \frac{4mc_d^{-1/d}}{ld(\mathrm{Leb}(L(u)))^{(d-1)/d}},$$

where $c_d = \pi^{d/2}/\Gamma(1 + d/2)$, and suppose $\lambda_n(\gamma) > \lambda(\gamma) + Da_n$ for some $\gamma \in [\underline{\gamma}, \overline{\gamma}]$. Then

$$C_n(\gamma) \subset L_n(\lambda(\gamma) + Da_n) \subset L(\lambda(\gamma) + Da_n)_{a_n} \subset L(\lambda(\gamma) + Da_n - 2ma_n),$$

where the last two inclusions follow from (23) and Lemma 2(b), respectively. Hence

$$\begin{aligned}
F_n(C_n(\gamma)) &\le F_n\big(L(\lambda(\gamma)+\widetilde{C}a_n)\big)\\
&\le F\big(L(\lambda(\gamma)+\widetilde{C}a_n)\big)+a_n \quad \text{by (23)}\\
&\le F(L(\lambda(\gamma)))-\lambda(\gamma)\cdot \mathrm{Leb}\big(L(\lambda(\gamma))\setminus L\big(\lambda(\gamma)+\widetilde{C}a_n\big)\big)+a_n\\
&= \gamma-\lambda(\gamma)\big[\mathrm{Leb}(L(\lambda(\gamma)))-\mathrm{Leb}\big(L(\lambda(\gamma)+\widetilde{C}a_n)\big)\big]+a_n\\
&\le \gamma - ld(\mathrm{Leb}(L(u)))^{(d-1)/d}\,\frac{c_d^{1/d}}{2m}\,\widetilde{C}a_n+a_n \quad \text{by Lemma 2(c)}\\
&= \gamma-a_n.
\end{aligned}$$

This contradiction to the definition of $C_n(\gamma)$ shows that, in fact, $\lambda_n(\gamma)\le\lambda(\gamma)+Da_n$. Note that $\lambda_n(\gamma)\ge\lambda(\gamma)-Da_n$ can be shown analogously, proving (24). □

PROOF OF THEOREM 4. The technical arguments of the proof are very similar to those in the proof of Theorem 3, so only the main steps of the proof will be given.

One checks $\inf_{L(\lambda)\ominus\sigma_n B}\hat{f}_n \ge \lambda$ and $\sup_{(L(\lambda)\oplus\sigma_n B)^c}\hat{f}_n < \lambda$ eventually a.s., so the argument of Lemma 3(a) together with the Borel–Cantelli lemma shows that

$$\widetilde{\mathscr{X}_n^{+}}\subset L(\lambda),\qquad \widetilde{\mathscr{X}_n^{-}}\subset (L(\lambda))^c,\qquad R\subset L(\lambda)_{3\sigma_n}\setminus L(\lambda)_{-3\sigma_n}\quad \text{eventually a.s.}$$

and hence, also by Lemma 3(a) and using $\widehat{\Psi}_{-r_n}(\widetilde{\mathscr{X}_n^{+}})\subset\widehat{\Psi}_{-r_n}(\widetilde{\mathscr{X}_n^{+}}\cup R_i)\subset\widehat{\Psi}_{-r_n}(\widetilde{\mathscr{X}_n^{+}}\cup R)$,

$$L(\lambda)_{-4\sigma_n}\subset\widehat{\Psi}_{-r_n}\Big(\widetilde{\mathscr{X}_n^{+}}\cup R_i\Big)\subset L(\lambda)_{4\sigma_n}\quad\text{eventually a.s. uniformly in } i. \tag{26}$$

Now

$$\Psi_{-r_n}\Big(\widehat{\Psi}_{-r_n}\Big(\widetilde{\mathscr{X}_n^{+}}\cup R_i\Big)\Big)=\widehat{\Psi}_{-r_n}\Big(\widetilde{\mathscr{X}_n^{+}}\cup R_i\Big) \tag{27}$$

by (1) and as $\Psi_{-r_n}(A_{-r_n})=A_{-r_n}$ for any set A [use (MON) in Walther (1995)]. Further, $L(\lambda)=\Psi_{r_0}(L(\lambda))=\Psi_{-r_0}(L(\lambda))$. This together with (26), (27) and Theorem 2 of Walther (1995) shows that

(28) the sets $S_{n,i}:=\Psi_{r_n}(\Psi_{-r_n}(\widehat{\Psi}_{-r_n}(\widetilde{\mathscr{X}_n^{+}}\cup R_i)))=\Psi_{r_n}(\widehat{\Psi}_{-r_n}(\widetilde{\mathscr{X}_n^{+}}\cup R_i))$ satisfy the conditions of Theorem 1 with rolling radius at least $(r_0+r_n)/2$ as well as $L(\lambda)_{-4\sigma_n}\subset S_{n,i}\subset L(\lambda)_{4\sigma_n}$ eventually a.s. uniformly in i.

One checks that $Z_{n,i}=\bigcup(\{X\in\mathscr{X}_n\colon X\oplus r_nB\subset S_{n,i}\}\oplus r_nB)$, and as $S_{n,i}\subset C$ for all i eventually a.s., Lemma 3(b) together with the Borel–Cantelli lemma yields the uniform convergence of $Z_{n,i}$ to $S_{n,i}$:

$$(S_{n,i})_{-c(\log n/n)^{2/(d+1)}}\subset Z_{n,i}\subset S_{n,i}\quad\text{eventually a.s. uniformly in } i \tag{29}$$

for some constant c. Setting $R_j := R \cap L(\lambda)$ and using Lemma 3(a) and Theorem 2 of Walther (1995), one also finds that $L(\lambda)_{-c(\log n/n)^{2/(d+1)}} \subset S_{n,j} \subset L(\lambda)_{c(\log n/n)^{2/(d+1)}}$ eventually a.s., so together with (29) one obtains

$$\begin{aligned}\mathbb{P}\Big(&L(\lambda)_{-2c(\log n/n)^{2/(d+1)}} \subset Z_{n,j} \\ &\subset L(\lambda)_{2c(\log n/n)^{2/(d+1)}} \text{ for some } j \text{ eventually}\Big) = 1.\end{aligned} \tag{30}$$

Now set $H_\lambda(\cdot) := F(\cdot) - \lambda\,\mathrm{Leb}(\cdot)$, $H_{n,\lambda}(\cdot) := F_n(\cdot) - \lambda\,\mathrm{Leb}(\cdot)$, $p_n := c_1(\log n/n)^{2/(d+1)}$ if $d > 3$ and $p_n := c_1(\log n/\sqrt{n})$ if $d \le 3$, where $c_1 > 2c$ is a constant that derives from Proposition 1. It will be shown that, for measurable $Z \subset C$:

(31) If $d_{\mathrm{Leb}}(L(\lambda), Z) > a$, then $H_\lambda(Z) < H_\lambda(L(\lambda)) - \min(|l-\lambda|, |u-\lambda|)a$.

(32) If $d_{\mathrm{Leb}}(L(\lambda), Z) > kp_n$ and $S_{-p_n} \subset Z \subset S_{p_n}$ for some $S \in \mathscr{G}_C(\tilde{r})$ with $\tilde{r} > 0$, then $H_{n,\lambda}(Z) < H_\lambda(L(\lambda)) - (\min(|l-\lambda|, |u-\lambda|)k - (2\tilde{c}\sup f + 1))p_n$ eventually a.s. for some $\tilde{c} > 0$.

(33) If $L(\lambda)_{-p_n} \subset Z \subset L(\lambda)_{p_n}$, then $H_{n,\lambda}(Z) \ge H_\lambda(L(\lambda)) - (2u\tilde{c} + 1)p_n$ eventually a.s.

Equations (6), (30) and (33) show that eventually a.s. there exists a candidate set $Z_{n,j}$ with $d_{\mathrm{Leb}}(L(\lambda), Z_{n,j}) \le c_2(\log n/n)^{2/(d+1)}$ for some c_2 and $H_{n,\lambda}(Z_{n,j}) \ge H_\lambda(L(\lambda)) - (2u\tilde{c}+1)p_n$. If we choose k in (32) large enough so that $\min(|l-\lambda|, |u-\lambda|)k - (2\tilde{c}\sup f + 1) > 2u\tilde{c} + 1$, then (32) and (29) show that all $Z_{n,i}$ with $d_{\mathrm{Leb}}(L(\lambda), Z_{n,i}) \ge kp_n$ obey $H_{n,\lambda}(Z_{n,i}) < H_{n,\lambda}(Z_{n,j})$ eventually a.s. uniformly in i. It follows that $d_{\mathrm{Leb}}(L(\lambda), L_n(\lambda)) \le \max(c_2, (\log n/n)^{2/(d+1)}, kp_n)$ eventually a.s., proving the theorem.

It remains to prove (31)–(33). For (31), observe $F(Z \setminus L(\lambda)) \le l\mathrm{Leb}(Z \setminus L(\lambda))$ and $F(L(\lambda) \setminus Z) \ge u\mathrm{Leb}(L(\lambda) \setminus Z)$. Hence

$$\begin{aligned}H_\lambda(Z) &= F(L(\lambda)) + F(Z \setminus L(\lambda)) - F(L(\lambda) \setminus Z) \\ &\quad - \lambda\big(\mathrm{Leb}(L(\lambda)) + \mathrm{Leb}(Z \setminus L(\lambda)) - \mathrm{Leb}(L(\lambda) \setminus Z)\big) \\ &\le H_\lambda(L(\lambda)) - \min\big(|l-\lambda|, |\lambda-u|\big)d_{\mathrm{Leb}}(L(\lambda), Z).\end{aligned}$$

As for (32),

$$\begin{aligned}H_{n,\lambda}(Z) &\le F_n(S_{p_n}) - \lambda\,\mathrm{Leb}(Z) \\ &\le F(S_{p_n}) + p_n - \lambda\,\mathrm{Leb}(Z) \quad \text{eventually a.s.} \\ &\qquad\qquad \text{[by Proposition 1 and Lemma B.2 of Walther (1995)]} \\ &\le F(S_{-p_n}) + 2\tilde{c}p_n \sup f - \lambda\,\mathrm{Leb}(Z) + p_n \quad \text{for some } \tilde{c} \text{ by (6)} \\ &< H_\lambda(L(\lambda)) - \min\big(|l-\lambda|, |u-\lambda|\big)kp_n + (2\tilde{c}\sup f + 1)p_n \quad \text{by (31).}\end{aligned}$$

Finally, (33) follows from

$$\begin{aligned} F_n(Z) - \lambda\,\mathrm{Leb}(Z) &\geq F(L(\lambda)_{-p_n}) - p_n - \lambda\,\mathrm{Leb}(L(\lambda)_{p_n}) \quad \text{eventually a.s.} \\ &\qquad \text{[by Proposition 1 and Lemma B.2 of Walther (1995)]} \\ &\geq F(L(\lambda)) - u\tilde{c}p_n - p_n - \lambda\big(\mathrm{Leb}(L(\lambda)) + \tilde{c}p_n\big) \quad \text{by (6)} \\ &\geq H_\lambda(L(\lambda)) - \big(2u\tilde{c}+1\big)p_n. \qquad \square \end{aligned}$$

Acknowledgments. I would like to thank an Associate Editor and two referees for their many helpful comments and suggestions.

REFERENCES

ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.

BERGER, M. and GOSTIAUX, B. (1988). *Differential Geometry: Manifolds, Curves and Surfaces.* Springer, New York.

BLASCHKE, W. (1949). *Kreis und Kugel.* Chelsea, New York.

BURAGO, Y. D. and ZALGALLER, V. A. (1988). *Geometric Inequalities.* Springer, New York.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

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

DASGUPTA, A., GHOSH, J. K. and ZEN, M. M. (1995). A new general method for constructing confidence sets in arbitrary dimensions: with applications. *Ann. Statist.* **23** 1408–1432.

DUDLEY, R. M. (1984). A course on empirical processes. *École d'Été de Probabilités de Saint Flour XII. Lecture Notes in Math.* **1097** 1–142. Springer, New York.

DÜMBGEN, L. and WALTHER, G. (1996). Rates of convergence for random approximations of convex sets. *Adv. in Appl. Probab.* **28** 384–393.

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

HALL, P. (1992). *The Bootstrap and Edgeworth Expansions.* Springer, New York.

HARTIGAN, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.

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

MANI-LEVITSKA, P. (1993). Characterizations of convex sets. In *Handbook of Convex Geometry* (P. M. Gruber and J. M. Wills, eds.) **A** 19–41. North-Holland, Amsterdam.

MATHERON, G. (1975). *Random Sets and Integral Geometry.* Wiley, New York.

MÜLLER, D. W. and SAWITZKI, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint 398, SFB 123, Univ. Heidelberg.

NAIMAN, D. Q. and WYNN, H. P. (1992). Inclusion–exclusion–Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Statist.* **20** 43–76.

NOLAN, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348–371.

PERKAL, J. (1956). Sur les ensembles ε-convexes. *Colloq. Math.* **4** 1–10.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications.* IMS, Hayward, CA.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23** 855-881.

ROMANO, J. P. (1988a). Bootstrapping the mode. *Ann. Inst. Statist. Math.* **40** 565–586.

ROMANO, J. P. (1988b). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647.

SCHNEIDER, R. (1993). *Convex Bodies: The Brunn–Minkowski Theory.* Cambridge Univ. Press.

SERRA, J. (1982). *Image Analysis and Mathematical Morphology* **1**. Academic Press, San Diego.

STUTE, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab.* **12** 361–379.

TANNER, M. A. (1993). *Tools for Statistical Inference.* Springer, New York.

THORPE, J. A. (1979). *Elementary Topics in Differential Geometry.* Springer, New York.

WALTHER, G. (1995). On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Math. Methods Appl. Sci.* To appear.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: walther@stat.stanford.edu