The Annals of Statistics
1978, Vol. 6, No. 4, 802-812

ESTIMATION OF A MULTIVARIATE MODE

BY THOMAS W. SAGER

Stanford University

Consider a random sample from an absolutely continuous multivariate distribution. Let $\mathscr{S}$ be a class of sets which are not too long and thin. A point $\boldsymbol{\theta}_n$ chosen from a minimum volume set $S_n \in \mathscr{S}$ containing at least $r = r(n)$ of the data may be used as an estimate of the mode of the distribution. In this paper, it is shown that $\boldsymbol{\theta}_n$ converges almost surely to the true mode under very minor conditions on $\{r(n)\}$ and the distribution. Convergence rates are also obtained. Extensions to estimation of local and/or multiple modes are noted. Finally, computational simplifications resulting from choosing S_n from spheres or cubes centered at observations are discussed.

1. Introduction. Consider an absolutely continuous distribution function $F(\cdot)$ on Euclidean k-space E^k. Without being too precise at this point, we say that the probability distribution F has a mode at the point $\boldsymbol{\theta}$ if the greatest concentration of probability occurs around $\boldsymbol{\theta}$. Since sample characteristics tend to reflect characteristics of the parent population, we may expect that the plotted data points in a random sample of sufficiently large size will also tend to be most concentrated around $\boldsymbol{\theta}$. This simple observation suggests an estimate of $\boldsymbol{\theta}$ based on a set where the data are "most concentrated." The estimators of Venter [17] and Chernoff [2] for the case of univariate $F(k = 1)$ are both based on this idea: respectively, the shortest interval containing a given number of observations and the interval of given length containing the most observations. For multivariate $F(k > 1)$, the shape of the set used for the modal estimate becomes a more important consideration than in the univariate case. Possibilities include hyper-rectangles and hyper-spheres. In Sager [13] the set of greatest concentration of data was taken to be a convex set. At this time, the lack of a reasonable algorithm for locating the smallest-volume convex set containing a given number of observations lessens the utility of that method for higher dimensions. In this paper, we propose to define the set of greatest data concentration by means of a class of simple geometrical shapes, which will include rectangles or spheres. The greater tractability of these shapes increases the practical utility of our estimates.

Conceptually, estimators for the mode of a distribution may be classified as direct or indirect according to their paternity. When the estimator is generated as a by-product from estimating some other quantity—usually the density

Received January 1976; revised June 1977.

Research supported by National Science Foundation under Grant MPS75-09450-000 and by SIAM Institute for Mathematics and Society (SIMS).

AMS 1970 *subject classifications*. Primary 62G05; Secondary 62H99, 60F15.

Key words and phrases. Estimation, mode, multivariate, consistency, convergence rates.

function—we call it *indirect*. All of the standard density estimators (e.g., kernel [9], orthogonal series [6], nearest neighbor [7], spline [1], penalized likelihood [4], isotonic [10]) yield indirect modal estimates by the simple device of selecting the point at which the density estimate is maximized. On the other hand, when the estimator is specially designed for the sole purpose of estimating the mode as a statistical parameter in its own right, we call it *direct*. In this class are the estimators of Venter [17], Chernoff [2], Grenander [5]. These authors treat the univariate case only. Nonparametric multivariate modal estimation has not been extensively discussed in the statistical literature, except as occasional corollaries to multivariate generalizations of the above methods for density estimation.

Although the conceptual difference between direct and indirect estimators is pronounced, the theoretical and computational difference is often small. For example, in one dimension the mode of the density estimate with uniform kernel is just the Chernoff modal estimate and the mode of the nearest neighbor density estimate is the Venter estimate. Similarly, the multivariate estimates of this paper include as special cases the modes of spherical and cubical nearest neighbor density estimates (Loftsgaarden and Quesenberry [7] and Elkins [kernel method, 3]). It would seem, therefore, that many of the properties of modal estimates could be deduced from properties of density estimators. For example, an elementary and standard argument derives consistency of modal estimators from uniform consistency of the corresponding density estimators (e.g., [9]). Moreover, an interesting recent "duality" result (Moore and Yackel [8]) allows consistency results on nearest neighbor density estimators to be immediately converted into corresponding results for kernel estimators, and vice versa. Hence, one approach to consistency results for some of the multivariate nearest neighbor modal estimators of this paper lies in translating uniform consistency of uniform kernel density estimators (Theorems 2.1 and 2.2 of [8]) into uniform consistency of nearest neighbor density estimators. However, the baggage of attendant restrictions necessary to obtain uniform consistency must also be retained. More general results may be obtained by a direct proof from first principles. As noted by Schuster [14], uniform consistency of a density estimate is equivalent to uniform continuity of the density. In our proof, the density need not even be continuous; it may be nondifferentiable on a set of measure zero. Moreover, the number of nearest neighbors used in the estimate is freed somewhat from the restrictions of Van Ryzin [16, Theorem 2] and Moore and Yackel [8, Theorem 2.2]. To accomplish these gains we conceptualize our estimates as direct.

Consider a random sample of size n from the distribution F. Let $r = r(n)$ be a positive integer satisfying certain conditions to be specified later. From a class of specified k-dimensional sets $\mathscr{S}$ (see Definition 2.3) we choose a set S_n of smallest volume containing at least $r(n)$ of the data. A point $\boldsymbol{\theta}_n$ selected from S_n may be used as an estimate of the true mode $\boldsymbol{\theta}$. Possibilities for $\boldsymbol{\theta}_n$ include the centroid of S_n, the mean vector of the observations in S_n, etc. For the class

of candidate sets $\mathscr{S}$ we are principally interested in spheres and rectangles. But the theorems may be proved as easily for a much broader class of sets. Essentially, this class comprises unions and intersections of regions bounded by polynomial curves. In Section 2, we consider the almost sure convergence of $\boldsymbol{\theta}_n$ to $\boldsymbol{\theta}$. Certain assumptions will be necessary about the sequence $\{r(n)\}$ and the distribution F to obtain convergence. Stronger assumptions are necessary to obtain convergence rates in Section 3; these assumptions involve choice of an "optimal" sequence $\{r(n)\}$ for a given degree of "peakedness" of the density near $\boldsymbol{\theta}$ and are similar in spirit to those of [12] and [13]. Following the general results, we comment on some special cases and give suggestions for computational simplification.

One interesting application of these methods concerns geographically distributed variables. We may think of the distribution of illnesses attributed to or exacerbated by air pollution as a density function over an affected area. Identifying the mode of this distribution (perhaps after adjusting for population density) tells us where the effects of pollution are most severe and may suggest planning strategies for dealing with it. Local modes within subregions may also be identified. Data to estimate these modes could be collected from hospital records as occurrences of illness with the region and could be dated to correspond with time periods known to have experienced severe pollution. On the "causative" side of the dose-response relationship, the geographic mode(s) of the distribution of air pollutants is harder to ascertain partly because of the nonmobility and small number of air monitoring stations in most metropolitan areas. For this problem one might plausibly try to estimate the mode of a surrogate distribution. One thinks, for example, of substituting the geographic distribution of automobiles for that of carbon monoxide, which is known to be almost entirely attributable to mobile sources.

2. Almost sure convergence of $\boldsymbol{\theta}_n$. We begin by establishing some notation and definitions. Points in k-space will be denoted by boldface $\mathbf{x} = (x_1, x_2, \cdots, x_k)$. Let λ denote Lebesgue measure on the Borel sets in E^k with Euclidean metric $d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|$. An open ball centered at $\mathbf{x}$ with radius ε is the set $B(\mathbf{x}, \varepsilon) = \{\mathbf{y}; d(\mathbf{x}, \mathbf{y}) < \varepsilon\}$. The diameter of a set A is $\operatorname{diam} A = \sup\{d(\mathbf{x}, \mathbf{y}); \mathbf{x}, \mathbf{y} \in A\}$.

In the sequel we shall utilize the notion of differentiation of measures. The next two definitions and lemma contain the relevant ideas. A reference for this material is Rudin [11, Chapter 8].

DEFINITION 2.1. A collection $\mathscr{C} = \mathscr{C}(\beta)$ of open sets in E^k will be called a substantial family if both of the following hold:

(a) There is a finite constant β such that for each $A \in \mathscr{C}$ there is an open ball B (depending on A) containing A and satisfying $\lambda(B) < \beta \cdot \lambda(A)$.

(b) For each $\mathbf{x} \in E^k$ and for each $\delta > 0$, there is an $A \in \mathscr{C}$ satisfying $\operatorname{diam} A < \delta$ and $\mathbf{x} \in A$.

DEFINITION 2.2. If $\mathscr{C}$ is a substantial family in E^k and μ is a positive Borel measure on E^k, we define the upper derivative of μ with respect to $\mathscr{C}$ at $\mathbf{x}$ by

(a) $(\bar{D}\mu)(\mathbf{x}) = \lim_{\delta \to 0} \sup \{\mu(A)/\lambda(A); \mathbf{x} \in A, A \in \mathscr{A}, \operatorname{diam} A < \delta\}$ and the lower derivative of μ with respect to $\mathscr{A}$ at $\mathbf{x}$ by

(b) $(\underline{D}\mu)(\mathbf{x}) = \lim_{\delta \to 0} \inf \{\mu(A)/\lambda(A); \mathbf{x} \in A, A \in \mathscr{A}, \operatorname{diam} A < \delta\}$.

We say μ is differentiable with respect to $\mathscr{A}$ at $\mathbf{x}$ if the upper and lower derivatives of μ with respect to $\mathscr{A}$ at $\mathbf{x}$ are equal and finite. In that case we write $(\underline{D}\mu)(\mathbf{x}) = (\bar{D}\mu)(\mathbf{x}) = (\mathbf{D}\mu)(\mathbf{x})$.

LEMMA 2.1. *Let $\mathscr{A}$ be a substantial family in E^k. Let μ be a probability measure on E^k. If μ is absolutely continuous with respect to λ, then*

(a) *μ is differentiable almost everywhere $[\lambda]$ with respect to $\mathscr{A}$, and*

(b) *$\underline{D}\mu$, $\bar{D}\mu$, $\mathbf{D}\mu$ are all versions of the Radon–Nikodym derivative of μ with respect to λ.*

PROOF. See Rudin [11, Theorem 8.6].

Condition (a) of Definition 2.1 makes precise the idea that the sets in $\mathscr{A}$ may not be too long and thin; condition (b) ensures that $\mathscr{A}$ will be rich enough for the limits in Definition 2.2 to make sense for every $\mathbf{x}$. Two examples of substantial families are the collection of all open balls and the collection of all open hyperrectangles, each of whose longest edge is at most β times the length of its shortest edge. In view of Lemma 2.1, the computation of $D\mu$ is independent (a.e.) of the underlying substantial family.

At this point it may be helpful if we comment on the role of $D\mu$ and $\mathscr{A}$ in the sequel. By defining the density and mode of F in terms of the limiting process of Definition 2.2, we obtain a constructive representation of the density which will be more convenient for us than the version which springs (as if by magic!) from the Radon–Nikodym theorem. Since the interior of the set S_n from which $\boldsymbol{\theta}_n$ is selected will be a member of a substantial family, the sequence $\{S_n\}$ may be used in the limiting process, suggested by Definition 2.2, which defines $f(\boldsymbol{\theta})$. (The consistency argument does require a little more subtlety than this, since we do not know that $\boldsymbol{\theta} \in S_n$ for sufficiently large n; but $\{S_n\}$ will be compared with an auxiliary sequence $\{M_n\}$ which does contain $\boldsymbol{\theta}$ and may be used in the limiting process.)

For our substantial family $\mathscr{S}$ we shall use polynomial regions. With fixed a, let $P_{a,k}$ denote the class of all polynomials in k variables which have degree not greater than a. If $g \in P_{a,k}$ then $A_g = \{\mathbf{x}; g(\mathbf{x}) > 0\}$ is called a polynomial region of degree m. For example, $g(x_1, \cdots, x_k) = -(x_1^2 + \cdots + x_k^2) + r^2 > 0$ determines a polynomial region which is the interior of a sphere of radius r and centered at the origin. All polynomial regions are open sets. This facilitates the use of Definition 2.1, but it makes no real difference in the sequel if we close these sets, for their boundaries have zero probability content.

DEFINITION 2.3. Let b be a fixed positive integer. Define $\mathscr{S}$ to be the class of all sets which are polynomial regions, or a union or intersection of no more than b polynomial regions, and satisfying Definition 2.1(a).

$\mathscr{S}$ then constitutes a legitimate substantial family. Moreover, allowing the membership of $\mathscr{S}$ to include sets formed by a finite number of set operations on polynomial regions greatly enriches $\mathscr{S}$. For example, the rectangle $(c_1, d_1) \times \cdots \times (c_k, d_k)$ is the intersection of $2k$ polynomial regions of the form $c_i < x_i$ or $x_i < d_i$.

DEFINITION 2.4. Let F be an absolutely continuous distribution function (probability measure) on E^k. The probability density function $f(\mathbf{x})$ is defined by $f(\mathbf{x}) = (\bar{D}F)(\mathbf{x})$ for each $\mathbf{x}$, with respect to the substantial family $\mathscr{S}$.

We note that $f(\cdot)$ is a version of the Radon–Nikodym derivative $dF/d\lambda$. The above version of f is adopted for the sake of definiteness on the exceptional set of measure zero where the Radon–Nikodym derivative fails to be uniquely defined.

DEFINITION 2.5. A point $\boldsymbol{\theta}$ is said to be the mode of F if for each $\varepsilon > 0$, there exists $\delta > 0$ such that $d(\mathbf{x}, \boldsymbol{\theta}) > \varepsilon$ implies $f(\mathbf{x}) + \delta < f(\boldsymbol{\theta})$.

This definition marks a departure from the usual definition of mode: customarily, $\boldsymbol{\theta}$ is the mode if $f(\mathbf{x}) < f(\boldsymbol{\theta})$ for all $\mathbf{x} \neq \boldsymbol{\theta}$. Our definition is intended to eliminate the pathological case of a sequence $\{\mathbf{x}_i\}$ bounded away from $\boldsymbol{\theta}$ but with $f(\mathbf{x}_i) \uparrow f(\boldsymbol{\theta})$.

DEFINITION 2.6. Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be a random sample of size n from F. Let $r(n)$ be a positive integer. We define $S_n = S_n(r(n))$ to be a minimum-volume set among the class of closures of $\mathscr{S}$ which contain at least $r(n)$ observations.

Since each g function is a polynomial of degree at most a and no more than b such functions are used in determining S_n, then there are at most a fixed finite number of observations which can lie on the surface of any polynomial region. If $r(n)$ is chosen larger than this number but less than n, then a minimum (positive) volume S_n exists with probability one. Note that S_n need not be unique because of the necessity of satisfying Definition 2.1(a), but we suppose that a systematic procedure has been defined to select one of the possible sets if there are several of minimum volume. For the same reason S_n need not contain precisely $r(n)$ data.

We shall use the next lemma as a guide in our choice of $r(n)$.

LEMMA 2.2. *Let F_n be the empirical distribution function corresponding to F. Then*

$$P[\sup_{A \in \mathscr{S}} |F(A) - F_n(A)| > c(n^{-1} \log n)^{\frac{1}{2}} \text{ infinitely often}] = 0$$

where c is a constant which does not depend on n.

PROOF. An immediate consequence of Steele [15, page 20, corollary to Theorem 2.1].

THEOREM 2.1. *Let F be an absolutely continuous distribution on E^k with density f as given in Definition 2.3 and mode $\boldsymbol{\theta}$ as given in Definition 2.5. Let $\{r(n)\}$ be a*

sequence of integers such that $r(n)/n = o(1)$, $n^{\frac{1}{2}}/r(n) \cdot (\log n)^{\frac{1}{2}} = o(1)$. *Let* $S_n = S_n(r(n))$ *be a smallest-volume set among the closures of* $\mathscr{S}$ *which contain at least* $r(n)$ *observations, and suppose* $(DF)(\boldsymbol{\theta})$ *exists. If* $\boldsymbol{\theta}_n \in S_n$ *for each* n, *then* $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$ *almost surely.*

PROOF. The proof proceeds in a series of small steps. First we define an ancillary sequence $\{M_n\}$ which will be convenient for comparison with $\{S_n\}$. M_n is to be a minimum volume set among the closures of $\mathscr{S}$-sets which contain at least $r(n)$ observations and for which the true mode $\boldsymbol{\theta}$ is an interior point.

We claim that $F(M_n)/\lambda(M_n) \to f(\boldsymbol{\theta})$ almost surely. Since F is differentiable at $\boldsymbol{\theta}$, this will follow from Definition 2.2 provided diam $M_n \to 0$ almost surely. But because M_n cannot become too "thin" (Definition 2.1(a)) its diameter goes to zero if and only if its volume does. Suppose $\lambda(M_n) > \varepsilon$ for infinitely many n. It follows from Definition 2.1 that we can find $A \in \mathscr{S}$ with $\boldsymbol{\theta} \in A$ and $\lambda(A) < \varepsilon$, $F(A) > 0$. Since $r(n)/n = o(1)$, Lemma 2.2 shows that A becomes a smaller-volume set containing $\boldsymbol{\theta}$ and at least $r(n)$ observations than M_n. This contradiction establishes the claim.

Next we observe that $|nF_n(S_n) - r(n)| \leqq d$ and $|nF_n(M_n) - r(n)| \leqq d$ both hold almost surely for some integer d, free of n. Let d be the maximum number of observations in the boundary of a set in $\mathscr{S}$. Then the interiors of S_n, M_n must contain fewer than $r(n)$ observations or else smaller-volume sets (interior to S_n, M_n) could have been used for S_n and M_n.

Then we claim that $F(S_n)/\lambda(S_n) \to f(\boldsymbol{\theta})$ almost surely. Now by the definitions of S_n and the mode, we have $\lambda(S_n) \leqq \lambda(M_n)$ and $F(S_n) \leqq f(\boldsymbol{\theta}) \cdot \lambda(S_n)$. By these facts, and the above observations and Lemma 2.2, the claim is proved in the following string of inequalities:

$$
\begin{aligned}
f(\boldsymbol{\theta}) &= \liminf_n F(M_n)/\lambda(M_n) \leqq \liminf_n F(S_n)/\lambda(S_n) \cdot \limsup_n \lambda(S_n)/\lambda(M_n)\,. \\
&\limsup_n F(M_n)n(r(n))^{-1} \cdot [\liminf F(S_n)n(r(n))^{-1}]^{-1} \\
&\quad \leqq \liminf_n F(S_n)/\lambda(S_n) \cdot 1 \cdot \limsup_n [F_n(M_n) + O(n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}})]n(r(n))^{-1}\,. \\
&\quad \{\liminf_n [F_n(S_n) + O(n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}})]n(r(n))^{-1}\}^{-1} \\
&\qquad\qquad = \liminf_n F(S_n)/\lambda(S_n) \leqq \limsup_n F(S_n)/\lambda(S_n). \\
&\qquad\qquad \leqq f(\boldsymbol{\theta}) \quad \text{almost surely.}
\end{aligned}
$$

We now may collect the various pieces to complete the proof of Theorem 2.1. Let $\varepsilon > 0$. Suppose $\{A_n\}$ is a sequence of measurable sets such that $d(\boldsymbol{\theta}, A_n) > \varepsilon$ infinitely often. Then by Definition 2.5, there is a $\delta > 0$ such that $F(A_n) + \delta \cdot \lambda(A_n) < f(\boldsymbol{\theta}) \cdot \lambda(A_n)$ infinitely often. Hence, $\liminf_n F(A_n)/\lambda(A_n) + \delta \leqq f(\boldsymbol{\theta})$. Therefore, we must have $d(\boldsymbol{\theta}, S_n) \leqq \varepsilon$ for all but finitely many n, with probability one. Since diam $S_n \to 0$, we have $S_n \subset B(\boldsymbol{\theta}, 2\varepsilon)$ for all but finitely many n, with probability one. Hence, $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$ almost surely.

We note some extensions. Suppose we are interested in estimating the mode of f locally to some known set A having nonempty interior in E^k. Suppose further that the collection of local modes of f on A is not necessarily a singleton.

If we choose S_n as before, except that we suitably restrict it to A, will the estimator $\boldsymbol{\theta}_n$ converge to the collection of local modes of f on A? The answer, which we present now, is affirmative.

DEFINITION 2.7. Let A be a subset of E^k with nonempty interior. A nonempty set M is said to be the collection of A-modes of F if $M \subset A$, $f(M)$ is a positive singleton, for each $\varepsilon > 0$ there is a $\delta > 0$ such that $\mathbf{x} \in A$ and $d(\mathbf{x}, M) > \varepsilon$ imply $f(\mathbf{x}) + \delta < f(M)$.

DEFINITION 2.8. Let A be a fixed set with nonempty interior. Let $r(n)$ be a positive integer. We define $S_n(A) = S_n(A, r(n))$ to be a minimum-volume closure of sets in $\mathscr{S}$ which contain at least $r(n)$ of the observations and which are contained in the interior of A.

THEOREM 2.2. *Let F be an absolutely continuous distribution in E^k with density f as defined in Definition* 2.4 *and collection M_A of A-modes of F. Let $\{r(n)\}$ be a sequence of integers such that $r(n)/n = o(1)$, $n^{\frac{1}{2}}/r(n) \cdot (\log n)^{\frac{1}{2}} = o(1)$. Suppose there is a $\boldsymbol{\theta} \in M_A$ such that $\boldsymbol{\theta}$ lies in the interior of A and $(DF)(\boldsymbol{\theta})$ exists. Then if $\boldsymbol{\theta}_n \in S_n(A)$ for each n, we have $d(\boldsymbol{\theta}_n, M_A) \to 0$ almost surely.*

PROOF. The proof of Theorem 2.1 may be modified in appropriate places to prove Theorem 2.2.

3. Convergence rates for $\boldsymbol{\theta}_n$. To obtain convergence rates for $\boldsymbol{\theta}_n$, further distributional assumptions will be necessary. In general, the speed of convergence depends upon the steepness of ascent of the density near $\boldsymbol{\theta}$. The steeper the ascent, the faster the rate of convergence. The measure of steepness used here is given in the next definition.

DEFINITION 3.1. Let $f(\cdot)$ be the density of F as given in Definition 2.4. Let $\boldsymbol{\theta}$ be the unique mode of the distribution and let $D > 1$ be fixed. For each $\delta > 0$, define $\alpha(\delta, D) = \inf\{f(\mathbf{x}); d(\boldsymbol{\theta}, \mathbf{x}) \leqq \delta\}/\sup\{f(\mathbf{x}); d(\boldsymbol{\theta}, \mathbf{x}) \geqq D\delta\}$.

As an example, consider the k-variate normal distribution with independent components $X_i \sim N(\theta_i, \sigma_i^2)$, $i = 1, \cdots, k$. With $D = 2$, we compute $\alpha(\delta, 2)$ and expand the exponential in a Taylor's series to obtain $\alpha(\delta, 2) = 1 + \frac{3}{2}\delta^2(1/\sigma_1^2 + 1/\sigma_2^2 + \cdots + 1/\sigma_k^2) + O(\delta^4) > 1 + \text{cons. } \delta^2$ for all δ sufficiently small.

THEOREM 3.1. *With the same assumptions as in Theorem* 2.1 *but also that $\alpha(\delta, D) \geqq 1 + \rho\delta^m$ for some $D > 1$, $\rho > 0$, $m > 0$, for all δ sufficiently small and $r(n)$ of the form $Qn^{(2m+k)/(2m+2k)}$ for some $Q > 0$, then $d(\boldsymbol{\theta}_n, \boldsymbol{\theta}) = O(\delta_n)$ almost surely, where $\delta_n = n^{-1/(2m+2k)}(\log n)^{1/2m}$.*

PROOF. The first step in the proof is to establish that the ancillary set M_n (see proof of Theorem 2.1) is sufficiently close to $\boldsymbol{\theta}$.

LEMMA 3.1. *$M_n \subset B(\boldsymbol{\theta}, \delta_n)$ for all n sufficiently large, with probability one.*

PROOF. Since $\alpha(\delta, D) > 1$, there exist positive δ_0 and h such that $f(\mathbf{x}) > h$ for all $\mathbf{x} \in B(\boldsymbol{\theta}, \delta_0)$. Thus $h(\beta l_n)^k < F(M_n)$ for all n, for some $\beta < 1$, where $l_n = \sup\{d(\mathbf{x}, \boldsymbol{\theta}); \mathbf{x} \in M_n\}$. It suffices to show $l_n < \delta_n$ for all large n, with probability

one. From Lemma 2.2 we know that $F(M_n) = r(n)/n + O((n^{-1}\log n)^{\frac{1}{2}})$ almost surely. Hence $l_n < \beta^{-1}h^{-1/k}Q^{1/k}n^{-1/(2m+2k)} + O((n^{-1}\log n)^{1/2k}) < \delta_n$ almost surely.

Let Ω_0 be the event of probability one for which the conclusion of Lemma 2.2 holds. Let Ω_1 be the event of probability one for which the containment relations of Lemma 3.1 hold. Let Ω_2 be the event of probability one for which $F(S_n) > 0$ for all but finitely many n. We show that if $\Omega_0 \cap \Omega_1 \cap \Omega_2$ occurs, then $d(\boldsymbol{\theta}_n, \boldsymbol{\theta}) = O(\delta_n)$.

Suppose $\Omega_0 \cap \Omega_1 \cap \Omega_2$ occurs. Since $\lambda(S_n) \leqq \lambda(M_n)$, the diameter of S_n is of no higher order than that of M_n, which is $O(\delta_n)$. Since $d(\boldsymbol{\theta}_n, \boldsymbol{\theta}) \leqq d(\boldsymbol{\theta}, S_n) + \operatorname{diam} S_n$, it suffices to show that S_n is contained in $B(\boldsymbol{\theta}, D \in \delta_n)$ for some $\varepsilon > 0$ for all n sufficiently large. Suppose, on the contrary, that for each $\varepsilon > 0$ there is a subsequence of $\{S_n\}$, each member of which is entirely contained in the complement of the corresponding $B(\boldsymbol{\theta}, D \in \delta_n)$ (convergence at the proper rate automatically follows when S_n has nonempty intersection with $B(\boldsymbol{\theta}, D \in \delta_n)$ since $\operatorname{diam} S_n = O(\delta_n)$). For these n, we have

$$\begin{aligned} 1 + \rho\varepsilon^m n^{-m/(2m+2k)}(\log n)^{\frac{1}{2}} &< \alpha(\varepsilon\delta_n, D) \cdot \lambda(M_n)/\lambda(S_n) < F(M_n)/F(S_n) \\ &= [r(n)/n + O((n^{-1}\log n)^{\frac{1}{2}})]/[r(n)/n + O((n^{-1}\log n)^{\frac{1}{2}})] \\ &= 1 + O(n^{-m/(2m+2k)}(\log n)^{\frac{1}{2}}) \quad \text{almost surely.} \end{aligned}$$

But for ε sufficiently large, the left-hand side of this inequality is larger than the right-hand side for all but finitely many n. (Note that the order on the right does not depend on ε, by virtue of Lemma 2.2.) This contradiction completes the proof of the theorem.

Theorem 3.1 may be extended to provide a convergence rate for estimation of multiple local modes in the same sense that Theorem 2.2 extends Theorem 2.1.

THEOREM 3.2. *Suppose the conditions of Theorem* 2.2 *hold. Suppose also that for some* $\boldsymbol{\theta} \in M_A$, *we have* $\alpha_{\boldsymbol{\theta}}(\delta, D) \geqq 1 + \rho\delta^m$ *for some* $D > 1$, $\rho > 0$, $m > 0$, *for all* δ *sufficiently small, where* $\alpha_{\boldsymbol{\theta}}(\delta, D) = \inf\{f(\mathbf{x}); d(\boldsymbol{\theta}, \mathbf{x}) \leqq \delta\}/\sup\{f(\mathbf{x}); d(\boldsymbol{\theta}, \mathbf{x}) \geqq D\delta, \mathbf{x} \in A\}$, *and let* $r(n)$ *be of the form* $Qn^{(2m+k)/(2m+2k)}$ *for some* $Q > 0$. *If* $\boldsymbol{\theta}_n \in S_n(A)$ *for each* n, *then* $d(\boldsymbol{\theta}_n, M_A) = O(\delta_n)$ *almost surely, where* $\delta_n = n^{-1/(2m+2k)}(\log n)^{1/2m}$.

PROOF. The extension to a local mode follows easily by observing that the proof of Theorem 3.1 depends only on properties of f near $\boldsymbol{\theta}$. A closer inspection of the proof of Theorem 3.1 shows that it may be extended to multiple modes, provided at least one of the modes satisfies the $\alpha(\delta, D)$ condition. We cannot conclude that $\boldsymbol{\theta}_n$ will be within $O(\delta_n)$ of that particular mode satisfying the $\alpha(\delta, D)$ condition, only that it will be within $O(\delta_n)$ of some mode satisfying the condition.

REMARKS. For the multivariate normal distribution discussed at the beginning of this section, we have a convergence rate of $O(n^{-1/(4+2k)}(\log n)^{\frac{1}{2}})$ for the error term. For $k = 1$, the estimate based on an interval S_n is identical to that of Sager [12], in which it was shown that $\theta_n = \theta + o(n^{-\frac{1}{3}}(\log n)^{\frac{1}{2}})$ almost surely for estimating the mode of a univariate normal. Of course, if we really knew

that the distribution were normal, multivariate or univariate, we would not use the estimators proposed here or in [13]. Our observation is that the convergence rates given here probably are not optimal. In [12] convergence rates were obtained through calculations based on special properties of order statistics on the line. That technique may not be extended to distributions in E^k, $k \geqq 2$. Nevertheless, we feel that sharper rates are possible for the multivariate case. However, it is evident that negligible gain in this direction may be realized from sharpening the rate in Lemma 2.2, which is already within an order of $\log n$ of the best possible. As was the case in [12], simple consistency of $\boldsymbol{\theta}_n$ should hold for $r(n)$ of the form Qn^ν, $0 < \nu < 1$, and convergence rates may approach $O(1/n)$ for m sufficiently small.

Finally, we remark that the convergence rates of Theorem 3.1 are not vacuous. For each $m > 0$, there exists a k-variate density to which the given error rate applies. To see this, let $\mathbf{X}$ be the k-variate distribution with independent components X_j, each having marginal density proportional to $\exp|x_j - \theta_j|^m$ for $j = 1, \cdots, k$, respectively. A calculation of $\alpha(\delta, 2)$ and expansion in a Taylor's series shows that $\alpha(\delta, 2) > 1 + \text{cons}\, \delta^m$ for all small δ.

4. Some special classes $\mathscr{S}$. The results of Sections 2 and 3 covered a broader class of sets $\mathscr{S}$ than it may be feasible to search in practice. If we replace $\mathscr{S}$ with appropriate subclasses, the theorems remain true provided the subclasses meet the requirements of substantial families. Two such subclasses are the family of all open balls and the family of all open hypercubes with sides parallel to the coordinate axes. Yet another is the class of open k-cells (sets of the form $(a_1, b_1) \times \cdots \times (a_k, b_k)$), each of whose longest edge is no more than β times the length of its shortest edge. It is easy to see that there is a unique minimum-volume closed ball containing at least $r(n)$ observations, for its surface will contain $k + 1$ data points and the probability is zero that the volumes of the spheres determined by any two distinct sets of $k + 1$ observations are the same. However, it is also easy to see that a volume-minimizing hypercube or closed k-cell is not unique, in general. For estimating a density at a fixed point, Elkins [3] found some differences between cubical and spherical kernels, but in terms of consistency there is nothing to choose between them. Intuitively, the more flexible the class $\mathscr{S}$, the better modal estimate one should obtain, but the more difficult the computations to find it. The volume-minimizing closed ball is easiest to find: by searching the $\binom{n}{k+1}$ spheres determined by each $k + 1$ of the data. The volume-minimizing hypercube is more difficult to calculate, and the situation in the case of general k-cells seems very complex indeed.

The practical difficulties in finding the appropriate volume-minimizing set suggest a number of shortcuts. If the balls and cubes were required to be centered at observations, then only n searches need be performed. For k-cells, we could restrict attention to the family of k-cell covers of all pairs of data points having ratio of longest edge to shortest at most β (the k-cell cover of a set A is

the intersection of all closed k-cells containing A; in the case of a two-point set A, the k-cell cover is particularly easy to write down explicitly). Although these simplifications may rather strongly restrict the class of potential sets for our modal estimates, we expect that their consistency properties would not suffer. Indeed, a careful examination shows that the proofs of Theorems 2.1 and 3.1 hinge on being able to find the set $M_n \in \mathscr{S}$ containing $\boldsymbol{\theta}$ as an interior point. This guarantees that $\lambda(S_n) \leqq \lambda(M_n)$ and makes the proofs work. We therefore have the following result.

THEOREM 4.1. *Suppose $\mathscr{S}'$ is a (possibly finite, possibly data dependent) subclass of the substantial family $\mathscr{S}$ of Definition* 2.3. *Suppose, with probability one for all n sufficiently large, $\mathscr{S}'$ has an element M_n' containing at least $r(n)$ observations and containing $\boldsymbol{\theta}$ as an interior point. If S_n' is a minimum-volume set among the closures of $\mathscr{S}'$ which contain at least $r(n)$ observations and $\boldsymbol{\theta}_n \in S_n'$, then $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$ almost surely. In addition, if the assumptions of Theorem* 3.1 *are met, then $\boldsymbol{\theta}_n$ has the asymptotic convergence rates given there.*

To see how this theorem might apply to the possible simplifications mentioned above, let $h(S) = \mathbf{x} \in S$ assign a single point $\mathbf{x}$ to each $S \in \mathscr{S}$ such that $\mathbf{x}$ lies no less than $c \cdot \operatorname{diam}(S)$ from the boundary of S for some $0 < c < 1$. Let $\mathscr{S}'$ be the subclass of $\mathscr{S}$ whose $h(\cdot)$-points are observations. If we can show the existence of an M_n' satisfying Theorem 4.1, then we immediately have consistency and convergence rates for estimators $\boldsymbol{\theta}_n$ based on spheres and cubes centered at observations. To see the existence of M_n', let $\mathbf{x}_{i(n)}$ denote the closest observation to $\boldsymbol{\theta}$ and let M_n' be a minimum-volume set from among the closures of $\mathscr{S}'$ whose h-point is $\mathbf{x}_{i(n)}$. Now the ball $B = B(\boldsymbol{\theta}, c \cdot \operatorname{diam} M_n')$ contains roughly $f(\boldsymbol{\theta}) \cdot \lambda(B)$ probability, which is at least a fixed positive fraction of the probability content of M_n'. Thus by applying Lemma 2.2, we see that B ultimately contains a fraction of the observations in M_n' and therefore contains $\mathbf{x}_{i(n)}$. Hence $\boldsymbol{\theta} \in B(\mathbf{x}_{i(n)}, c \cdot \operatorname{diam} M_n') \subset M_n'$.

To handle k-cells, let $\mathscr{S}'$ be the class of k-cell covers (see above) of all pairs of data-points having ratio of longest edge to shortest at most β and containing at least $r(n)$ data-points. As noted, $\mathscr{S}'$ is simple to construct. To obtain the existence of the required M_n', it is necessary and sufficient that at least one of the 2^{k-1} pairs of sets $(\{\mathbf{x}; x_1 < \theta_1, x_2 > \theta_2, \cdots, x_k > \theta_k\}, \{\mathbf{x}; \mathbf{x}_1 > \theta_1, x_2 > \theta_2, \cdots, x_k > \theta_k\}), \cdots$ have both pair-sets receiving positive probability, where the pairs are formed by changing the direction of the inequalities and coupling those whose inequalities differ for only one index. This is an assumption about the distribution F as it cannot be proved from our assumptions. But most distributions in practice will satisfy it.

It is evident that Theorem 4.1 admits consistency for a variety of ingenious shortcut methods.

REFERENCES

[1] BONEVA, L., KENDALL, D. and STEFANOV, I. (1971). Spline transformations: three new diagnostic aids for the statistical data analyst. *J. Roy. Statist. Soc. Ser. B* **33** 1–70.

[2] CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** (Part 1) 31–41.

[3] ELKINS, T. (1968). Cubical and spherical estimation of multivariate probability density. *J. Amer. Statist. Assoc.* **63** 1495–1513.

[4] GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

[5] GRENANDER, U. (1965). Some direct estimates of the mode. *Ann. Math. Statist.* **36** 131–138.

[6] KRONMAL, R. and TARTAR, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

[7] LOFTSGAARDEN, D. O. and QUESENBERRY, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36** 1049–1051.

[8] MOORE, D. S. and YACKEL, J. W. (1977). Consistency properties of nearest neighbor density function estimates. *Ann. Statist.* **5** 143–154.

[9] PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

[10] ROBERTSON, T. (1967). On estimating a density which is measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

[11] RUDIN, W. (1966). *Real and Complex Analysis.* McGraw-Hill, New York.

[12] SAGER, T. W. (1975). Consistency in nonparametric estimation of the mode. *Ann. Statist.* **3** 698–706.

[13] SAGER, T. W. (1975). An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* To appear.

[14] SCHUSTER, E. F. (1970). Note on the uniform convergence of density estimates. *Ann. Math. Statist.* **41** 1347–1348.

[15] STEELE, J. M. (1975). Combinatorial entropy and uniform limit laws. Ph.D. Dissertation, Department of Mathematics, Stanford Univ.

[16] VAN RYZIN, J. (1969). On strong consistency of density estimates. *Ann. Math. Statist.* **40** 1765–1772.

[17] VENTER, J. H. (1967). On estimation of the mode. *Ann. Math. Statist.* **37** 1446–1455.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305