# The Excess Mass Ellipsoid

By

D. Nolan
Department of Statistics
University of California at Berkeley
Berkeley, California 94720

Technical Report No. 216
December 1989
revised December 1990

Department of Statistics
University of California
Berkeley, California

# The Excess-Mass Ellipsoid

D. Nolan*
University of California, Berkeley

Abstract: The excess-mass ellipsoid is the ellipsoid that maximizes the difference between its probability content and a constant multiple of its volume, over all ellipsoids. When an empirical distribution determines the probability content, the sample excess-mass ellipsoid is a random set that can be used in contour estimation and tests for multimodality. Algorithms for computing the ellipsoid are provided, as well as comparative simulations. The asymptotic distribution of the parameters for the sample excess-mass ellipsoid are derived. It is found that a $n^{1/3}$ normalization of the center of the ellipsoid and lengths of its axes converge in distribution to the maximizer of a Gaussian process with quadratic drift. The generalization of ellipsoids to convex sets is discussed.

keywords: test for bimodality, contour estimation, minimum volume ellipsoid, cube-root asymptotics, empirical process, Gaussian process, quadratic drift, nearest-neighbor distance, rates of convergence.

*Research partly supported by NSF Grant no. DMS-8701426

## 1 Introduction

Hartigan (1987) and Muller and Sawitzki (1989) independently proposed a set statistic to estimate the contours of a density and to test for bimodality. Hartigan notes that for $P$ the unknown distribution of interest with density $p$, the $\alpha$-level contour $\{\boldsymbol{x} : p(\boldsymbol{x}) \geq \alpha, \boldsymbol{x} \in \mathbb{R}^d\}$ can be defined as the set that maximizes

$$P(S) - \alpha V(S)$$

over all sets $S$, where $V(S)$ is the volume of $S$. The $\alpha$-level contour of $P$ can be estimated from a sample if the empirical distribution $P_n$ is substituted for the unknown distribution. In particular, if $\mathcal{C}$ is the collection of convex sets in $\mathbb{R}^d$ then

$$\text{(1)} \qquad \arg\sup_{\mathcal{C}} P_n(C) - \alpha V(C)$$

approximates $\arg\sup_{\mathcal{C}} P(C) - \alpha V(C)$, which we call $C_\alpha$. Note: $C_\alpha$ coincides with the $\alpha$-level contour of $P$ when the density has nested convex contours.

Additionally, Hartigan (1987) and Muller and Sawitzki (1989) base a test for bimodality on the search for a second convex set, exterior to $C_\alpha$, where the density also exceeds the level $\alpha$:

$$\text{(2)} \qquad \sup_{\alpha} \sup_{C \subset C_\alpha^c} P(C) - \alpha V(C).$$

If the density is bimodal then the test statistic formed by substituting $P_n$ in (2) should be quite large.

In this paper, the case is treated when the supremum in (1) is restricted to the collection of ellipsoids. The limit distribution of the ellipsoid that maximizes (1) is found; we call this ellipsoid the empirical $\alpha$-level excess-mass ellipsoid. The simplification from convex sets to ellipsoids does not seem unduly restrictive. Practically speaking, many distributions are nearly elliptical or a transformation makes them so, and the density itself need not have elliptical contours in order for an ellipse to find a primary or secondary mode, or in order for (1) to be uniquely maximized over ellipsoids. The class of ellipsoids allows a parametrization of the problem that enables us to find the limit distribution of the empirical $\alpha$-level excess-mass ellipsoid, which is also key to determining the rate of convergence of the test-statistic for bimodality. We show that the center of the ellipsoid of interest and the lengths of the axes of the ellipsoid converge at a $n^{1/3}$ rate to a Gaussian process with quadratic drift.

The asymptotic results presented here are closely related to those that arise from other set statistics that are contour estimates, density estimates, and tests for multimodality. We describe a few of them briefly now.

Chernoff (1964) and Venter (1967) estimate the mode of a density function in one dimension by the center of the interval of fixed length to contain the greatest number of observations and by the center of the shortest interval to contain at least half of the observations, respectively. Sager (1979) generalized these univariate set statistics to the multidimensional case. He estimates the contours of a unimodal density by a sequence of nested convex sets. The first and largest set is the smallest convex set to contain a fixed proportion $q$ of the observations; the second set is the smallest convex set that contains proportion $q$ of the observations within the first set, and so on. Eddy and Hartigan (1977) proposed a similar multidimensional estimator.

These set statistics also arise in density estimation: the center of the fixed-length interval of Chernoff coincides with the mode of a fixed-bandwidth uniform-kernel density estimate; and the center of the shortest interval with a fixed proportion of the observations locates the mode of the $k^{\text{th}}$ nearest-neighbor density estimate, for $k = \frac{1}{2}n$. However, they are not completely comparable, because in density estimation the bandwidth and the number of nearest neighbors shrink with $n$. More recently, tests for bimodality constructed from density estimates have been suggested. Silverman (1986, p.139) proposes a test based on the size of the 'critical bandwidth' that provides a kernel density estimate which borders on bimodality. That is, a smaller bandwidth gives a density estimate with two or more modes, and a larger bandwidth yields a unimodal estimate. Wong and Schaack (1985) assess multimodality with $k^{\text{th}}$ nearest-neighbor density estimates. For values of $k$ from 1 to $n$, they count those $k^{\text{th}}$ nearest neighbor estimates that are bimodal. The number of $k$'s that produce a bimodal estimate represents the size of the smallest modal cluster among density estimates with two modes. A large count indicates the presence of a second mode.

Asymptotic results for these set statistics include those of Chernoff (1964), Andrews et. al (1972), Grübel (1988) and Kim and Pollard (1990). Chernoff shows the center of the fixed-length interval converges at a $n^{1/3}$ rate to the maximum of a Gaussian process with quadratic drift. Grübel finds the length of the shortest interval to contain half of the observations (the shorth) has a $\sqrt{n}$ asymptotic normal distribution. Kim and Pollard find general conditions for which cube-root rates of convergence are obtained in arbitrary dimension.

We make use of their results here to show that not only does the center of the empirical excess-mass ellipsoid converge at a $n^{1/3}$ rate to a Gaussian process with quadratic drift, but additionally, unlike the case of the shorth, the length of the axes of the ellipsoid also have cube-root rates of convergence.

The following comparison of the shorth and the $\alpha$-level excess-mass interval points out the difference between the convergence rates of these set statistics. The $\alpha$-level contour estimate is a maximization over $\mathcal{I}$, the collection of intervals in $\mathbb{R}$, i.e.

$$I_\alpha = \arg\sup_{\mathcal{I}} P_n(I) - \alpha V(I),$$

whereas the shorth is a constrained maximization over $\{I : P_n I \geq 1/2\}$. This constraint is responsible for the different convergence rates. Another factor that plays a role in both set statistics' distributions is the nondifferentiability of indicator functions.

To see how these two factors determine the rates of convergence, parametrize each interval $I$ by $\mu$ its center and $r$ half its length, so $I_{\mu,r} = [\mu - r, \mu + r]$. Then

$$P_n(I_{\mu,r}) - \alpha V(I_{\mu,r}) = [P(I_{\mu,r}) - \alpha V(I_{\mu,r})] + [P_n(I_{\mu,r}) - P(I_{\mu,r})].$$

Typically, we would take a Taylor-series expansion of the deterministic term about $P(I_0) - \alpha V(I_0)$, where $I_0 = I_{\mu_0,r_0}$ is the unique maximizer of $P(I) - \alpha V(I)$, and at the same time we would expand $I_{\mu,r}$ in the stochastic part about $I_0$. Nondifferentiability of indicators does not allow the latter expansion. Therefore, assuming $P$ has a differentiable density $p$,

$$\begin{aligned} 0 &\leq \sup_{\mu,r} P_n(I_{\mu,r}) - P_n(I_0) - \alpha V(I_{\mu,r}) + \alpha V(I_0) \\ &= [(r - r_0)^2 + \mu^2]p'(\mu_0 + r_0) + [P_n(I_{\mu,r} - I_0) - P(I_{\mu,r} - I_0)]. \end{aligned}$$

The coefficient of the first term is negative, as expected in a maximization. The variance of the second term is $O((|r-r_0|+|\mu-\mu_0|)/n)$. Kim and Pollard (1990) point out that the maximization occurs for values of the quadratic trend that balance those of the noise in the second term, which implies $|r_n - r_0| + |\mu_n - \mu_0|$ is of order $n^{-1/3}$, for the optimal interval $I_{\mu_n,r_n}$.

Both $\mu_n - \mu_0$ and $r_n - r_0$ of the $\alpha$-level excess-mass interval have nondegenerate limits when normalized by $n^{1/3}$. Not so for the length of the shorth.

Here is where the constraint that $P_n(I_n) = 1/2$ enters the picture; a faster $n^{1/2}$ normalization is needed for a nondegenerate limit for $r_n - r_0$. See Grübel (1988) and Kim and Pollard (1990) for a more thorough explanation. They show that for $\Delta_n = \sup_{\mathcal{I}} |P_n(I) - P(I)|$ and some constant $c$ that depends on $P$,

$$P_n(I_{\mu_0, r_0 + c\Delta_n}) \geq \Delta_n + P(I_{\mu_0, r_0 + c\Delta_n}) \geq 1/2$$

and

$$P_n(I_{\mu_0, r_0 - c\Delta_n}) \leq -\Delta_n + P(I_{\mu_0, r_0 - c\Delta_n}) < 1/2.$$

These inequalities imply $r_n - r_0$ is of order $n^{-1/2}$. However, this constraint does not change $\mu_n$'s rate of convergence.

In addition to providing a consistency result for the convex set in (1) and a heuristic argument for a bound on its rate of convergence, Hartigan provides an algorithm for finding the empirical $\alpha$-level excess-mass convex set in two dimensions. The algorithm builds up polygons from triangles with vertices at the observations. It requires $O(n^3)$ computations. Hartigan also proposes a faster method that approximates (1) by dividing the plane into $N^2$ cells using only the centers of cells as potential vertices for the convex set. In Section 3, we show that in two dimensions, determination of the $\alpha$-level excess-mass ellipsoid also requires many computations, about $O(n^6)$. We too present a faster algorithm, based on that of Rousseeuw and Leroy (1987), to approximate the ellipsoid.

Muller and Sawitzki provide algorithms for finding multiple modes in one dimension. They build a density contour cluster tree which they call a 'silhoutte' by varying $\alpha$ and finding the $M$ disjoint intervals $I_{n,1}(\alpha) \ldots I_{n,M}(\alpha)$ that maximize, over the collection of all possible $M$ disjoint intervals,

$$\sum_{j=1}^{M} P_n I_j - \alpha V(I_j).$$

They do not extend the silhoutte to higher dimensions.

The remainder of the paper is organized as follows: Section 2 contains the formal definition of the excess-mass ellipsoid as well as a proof of consistency. Section 3 compares the ellipsoid that estimates the $\alpha$-level contour to the minimum-volume ellipsoid, which is a robust estimator of location and scale in the multivariate setting. This comparison leads to algorithms for computing the ellipsoid of interest here. Section 4 contains weak convergence results for the sample excess-mass ellipsoid and Section 5 generalizes the main result of Kim and Pollard (1990) to include other rates of convergence. This result is then applied to set statistics of interest to Hartigan (1987) and Muller and Sawitzki (1989). Proofs are in Section 6.

## 2 The Set-up

Let $X_1, \ldots, X_n$ be $n$ independent observations from the distribution $P$ with density $p$ on $\mathbb{R}^d$, and let $P_n$ represent the empirical distribution constructed from the observations. The constant $\alpha$ is assumed positive. Let $\mathcal{E}$ denote the collection of ellipsoids $\{E\}$ in $\mathbb{R}^d$ and $\mathcal{S}$ the collection of spheres $\{S\}$ in $\mathbb{R}^d$.

**Definition:** *Define the $\alpha$-level excess-mass ellipsoid to be the ellipsoid $E_0$ that maximizes, over $\mathcal{E}$,*

$$P(E) - \alpha V(E). \tag{3a}$$

*Similarly, the $\alpha$-level empirical-excess-mass ellipsoid $E_n$ maximizes, over $\mathcal{E}$,*

$$P_n(E) - \alpha V(E). \quad \square \tag{3b}$$

**Definition:** *The $\alpha$-level excess-mass sphere $S_0$ is defined to be the sphere that maximizes*

$$P(S) - \alpha V(S), \tag{4a}$$

*over all spheres $S$ in the collection $\mathcal{S}$. The empirical version $S_n$ maximizes,*

$$P_n(S) - \alpha V(S). \quad \square \tag{4b}$$

If the excess-mass ellipsoid or sphere is not uniquely defined, use an arbitrary fixed rule to choose a candidate ellipsoid or sphere from among the possibilities. We ignore the slight complication in the above definitions that this rule entails. Only those distributions where $E_0$ and $S_0$ are uniquely determined are considered below.

If $P$ belongs to a family of elliptical distributions then the density $p$ can be expressed as

$$g((\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})),$$

for some function $g : \mathbb{R}^+ \to \mathbb{R}^+$ and some symmetric positive definite matrix $\Sigma$. The density need not be of this form in order for $E_0$ to be uniquely determined, but this restriction is placed on $p$ to simplify the central limit theory for $E_n$.

The advantage of restricting the maximization to ellipsoids, or even spheres, is clear from the parametrization of these sets. Each element $S$ of $\mathcal{S}$ can be parametrized by $\boldsymbol{\mu} \in \boldsymbol{R}^d$ its center and $r \in \boldsymbol{R}^+$ its radius. Denote such a sphere $S(\boldsymbol{\mu}, r)$. This parametrization is insufficient for ellipses. An ellipse can be represented via the quadratic form:

$$(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq 1,$$

where $\boldsymbol{\mu} \in \boldsymbol{R}^d$ and $\Sigma$ is a positive-definite symmetric matrix.

Parametrize the ellipse by $(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})$ where $\boldsymbol{\mu}$ is the center of the ellipse and $\tilde{\boldsymbol{\lambda}}$ is the $d + \binom{d}{2}$ vector of elements of $\Lambda = \Sigma^{1/2}$, the symmetric square root of $\Sigma$. That is, $\tilde{\boldsymbol{\lambda}} = (\lambda_{11}, \lambda_{22}, \ldots, \lambda_{dd}, \lambda_{12}, \lambda_{13}, , \lambda_{23}, \ldots, \lambda_{d-1,d})$ where $\Lambda = [\lambda_{ij}] = \mathrm{mat}(\tilde{\boldsymbol{\lambda}})$.

The collection of spheres is the subset of $\mathcal{E}$ such that $\Lambda$ can be written as $\lambda I$, for some positive scalar $\lambda$. In this case, the quadratic form above reduces to $(\boldsymbol{x} - \boldsymbol{\mu})'(\boldsymbol{x} - \boldsymbol{\mu}) \leq \lambda^2$, the first $d$ elements of $\tilde{\boldsymbol{\lambda}}$ are $\lambda$ and the last $\binom{d}{2}$ elements are 0.

Express $S_0$, $E_0$, $S_n$ and $E_n$ as $S(\boldsymbol{\mu}_0, r_0)$, $E(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)$, $S(\boldsymbol{\mu}_n, r_n)$ and $E(\boldsymbol{\mu}_n, \tilde{\boldsymbol{\lambda}}_n)$, respectively. We determine the limiting distribution of the parameters $(\boldsymbol{\mu}_n, r_n)$ and $(\boldsymbol{\mu}_n, \tilde{\boldsymbol{\lambda}}_n)$ in Section 4. As for consistency, the lemma below provides an almost-sure result for the parameters of $E_n$. The result for $S_n$ is implied by that of $E_n$.

**Consistency:** *Suppose the ellipse $E_0 = E(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)$ uniquely maximizes (3a) and $P$ has bounded density $p$. Then $E(\boldsymbol{\mu}_n, \tilde{\boldsymbol{\lambda}}_n)$ the ellipse that maximizes (3b) is such that*

$$\boldsymbol{\mu}_n \to \boldsymbol{\mu}_0$$

*and*

$$\tilde{\boldsymbol{\lambda}}_n \to \tilde{\boldsymbol{\lambda}}_0 \quad \textit{almost surely.}$$

**Proof:** The proof follows from an application of the Glivenko-Cantelli result for ellipsoids in $\mathbb{R}^d$. That is,

$$\sup_{\mathcal{E}} |P_n(E) - P(E)| \to \quad 0 \;\; a.s.$$

(See Pollard 1984, Theorem II.24). In particular, $P_n(E_n) - P(E_n) \to 0$ a.s. Denote the supremum above by $\Delta_n$.

The definition of $E_n$ implies

$$P_n(E_n) - \alpha V(E_n) \geq P_n(E_0) - \alpha V(E_0)$$

or

$$P(E_n) - \alpha V(E_n) + \Delta_n \geq P(E_0) - \alpha V(E_0) - \Delta_n.$$

Likewise, the definition of $E_0$ implies

$$P(E_0) - \alpha V(E_0) \geq P(E_n) - \alpha V(E_n).$$

Let $J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}) = P(E(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})) - \alpha V(E(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}))$. Combine the above two inequalities to show

$$|J(\boldsymbol{\mu}_n, \tilde{\boldsymbol{\lambda}}_n) - J(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)| \leq 2\Delta_n.$$

The function $J$ is a continuous function of $(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})$, because $P$ has a bounded density. Therefore, the uniqueness of the maximum yields the desired consistency, provided $J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})$ is bounded away from $J(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)$ outside some compact region about $(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)$. This provision is implied by boundedness of $p$. $\square$

A similar argument works for the Hausdorf distance between $E_n$ and $E_0$ (see Hartigan (1987)). Notice that if $\alpha \geq \sup_{\boldsymbol{x}} p(\boldsymbol{x})$ then the ellipse $E_0$ is degenerate, and so, is not unique.

## 3 Minimum-Volume Ellipsoids and Computational Algorithms

The excess-mass ellipsoid (EME) can be recast as a function of the minimum-volume ellipsoid (MVE). This representation is useful because it suggests algorithms for computing $E_n$. See for example Titterington (1978), Silverman and Titterington (1980), Preparata and Shamos (1985), Devroye (1983), Rousseeuw and Leroy (1987).

Observe that $P_n(E) = k/n$ for some $k = 1, 2, \ldots, n$. So, if $MVE_k$ is the ellipsoid of minimum volume that contains at least $k$ observations then

$$\sup_{\mathcal{E}}[P_n(E) - \alpha V(E)] = \sup_k [\frac{k}{n} - \alpha V(MVE_k)]$$

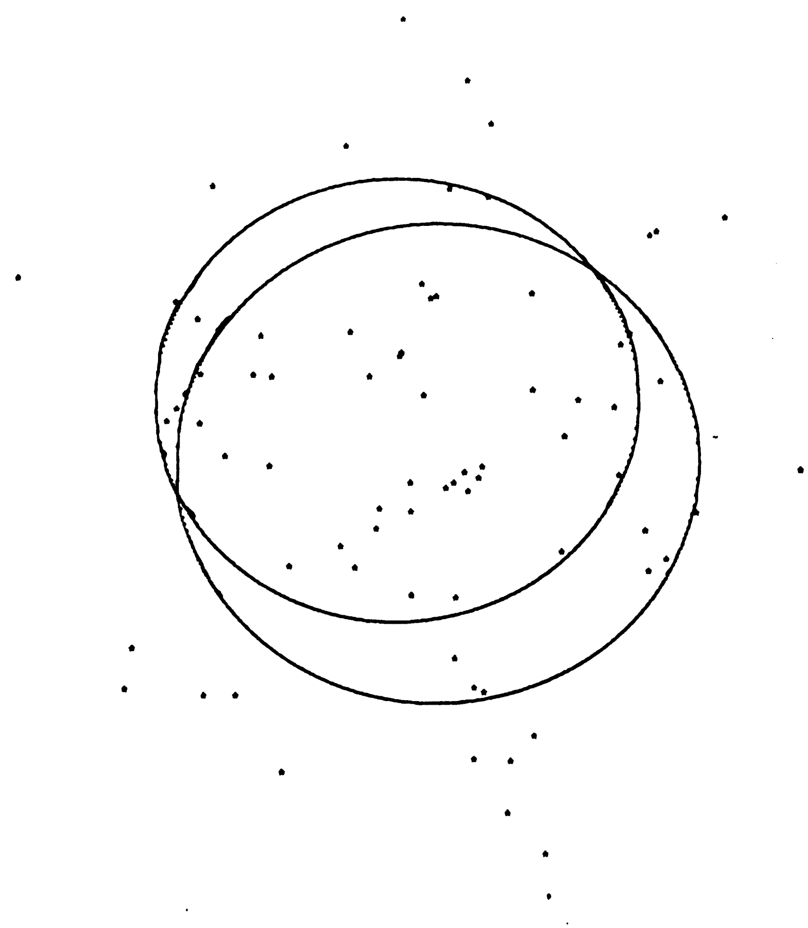

Figure 1

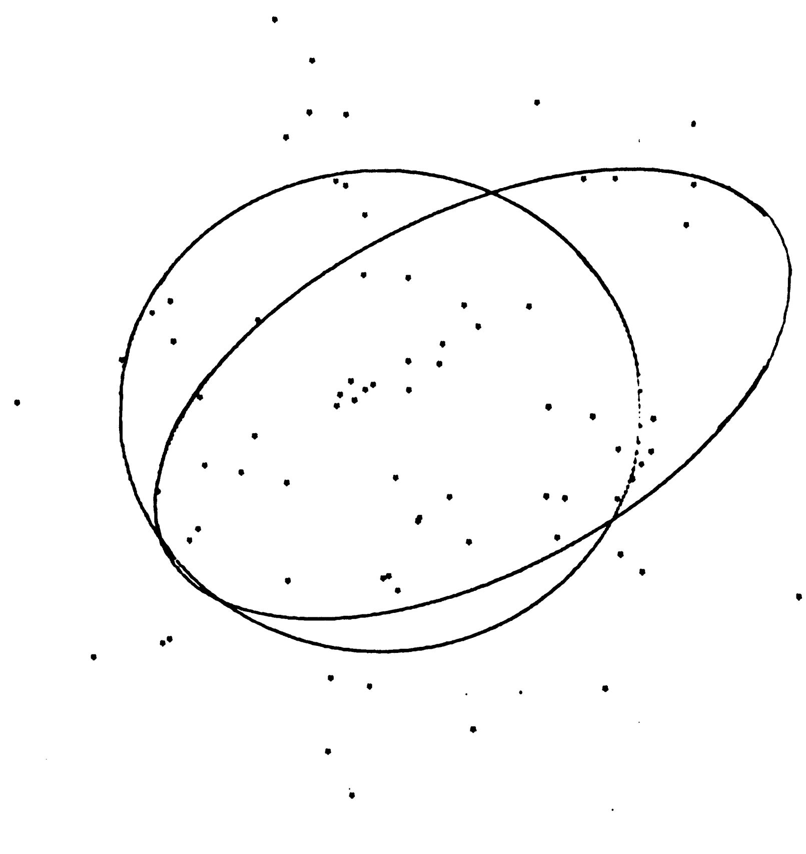

Figure 2

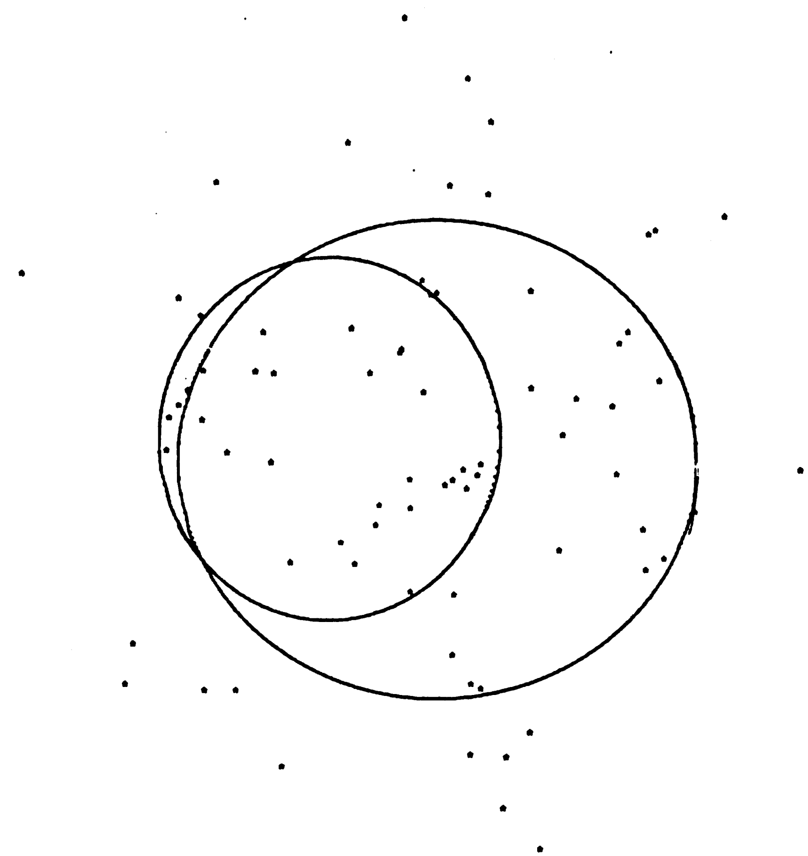

Figure 3

That is, the $EME$ is $MVE_k$ for some $k$.

This representation provides techniques for computing the excess-mass ellipsoid from a given set of observations. If one additionally restricts the search to spheres then the problem is equivalent to that of the smallest enclosing circle (Preparata and Shamos (1985)), or the problem of finding the center and radius of the smallest $k^{th}$-nearest-neighbor ball for each $k$. In two dimensions, all the $k^{th}$-nearest-neighbor balls can be found in $O(n^3)$ time. The restriction to spheres reduces computational complexity, and for many distributions, it can still be effective in mode hunting.

In the elliptical setting in two dimensions, the minimum-volume ellipse has either 3, 4 or 5 of the $n$ observations on its boundary (Titterington, 1978). Therefore, the computations required to find the EME are of $O(n^5)$, or possibly $O(n^6)$. To reduce the computational burden, Rousseeuw and Leroy (1987) propose an algorithm for estimating the MVE. This algorithm can be adapted to estimate the EME. We describe one iteration of the algorithm. Sample $d+1$ points $x_{\pi_1}, \ldots, x_{\pi_{d+1}}$ from $x_1, \ldots, x_n$, without replacement. Determine the average $\overline{x}_\pi = (x_{\pi_1} + \ldots + x_{\pi_{d+1}})/(d+1)$ and the covariance matrix $\sum_\pi = [(x_{\pi_1} - \overline{x}_\pi)\ (x_{\pi_1} - \overline{x}_\pi)' + \ldots + (x_{\pi_{d+1}} - \bar{x}_\pi)(x_{\pi_d} - \bar{x}_\pi)']/(d+1)$. Next calculate the order statistics

$$c_{\pi(k)} = k^{th}\text{smallest value of } \{(x_i - \overline{x}_\pi)'\Sigma_\pi^{-1}(x_i - \overline{x}_\pi)\}.$$

Then $c_{\pi(k)}$ can be used to magnify $\sum_\pi$ to contain $k$ observations. The volume of the resulting ellipse is proportional to $\sqrt{c^d_{\pi(k)} \det(\Sigma_\pi)}$. Repeat the above procedure $m$ times. For each $k$, find the smallest of the $m$ ellipses that contains $k$ observations. This is an estimate of the MVE(k), call it $MVEE_m(k)$. Finally, minimize $k/n - \alpha MVEE_m(k)$ over $k$. Here the number of operations to estimate the MVE are $O(nm)$. For comparison, Figures 1 and 2 show the EMS and an estimate of the EME based on $m = 60$, respectively, for a sample of size 75 from a standard bivariate normal. The $\alpha$-level contour is also displayed in these figures.

A third possible algorithm uses the sample covariance matrix to transform the data points to spherical symmetry and then proceeds with the first approach based on nearest-neighbor balls. Figure 3 shows this estimate for the same sample and $\alpha$ of figures 1 and 2.

## 4 Rates of Convergence

In this section, limit distributions of the maximal sets are obtained. Three cases are considered in turn. First the density is assumed to be spherically symmetric and the maximization of $P - \alpha V$ is restricted to the class of spheres in $\boldsymbol{R}^d$. Then the restriction that the density be spherically symmetric is relaxed to elliptical symmetry. Finally, the collection of sets is enlarged to include all ellipses. The proofs appear in the last section.

In the lemmas below $\|\cdot\|$ denotes Euclidean distance; $g : \boldsymbol{R} \to \boldsymbol{R}$ is a bounded nonnegative function; $v_d$ stands for the volume of the $d$-dimensional unit sphere; and $w_d = \int_{\{\boldsymbol{y}'\boldsymbol{y} \leq 1\}} y_1^2 dy$. Also $\boldsymbol{t}$ is a $d$-dimensional vector, $\boldsymbol{u}$ is a $d + \binom{d}{2}$-dimensional vector, $s$ is a scalar, and $Z$ is a mean-zero Gaussian process.

For ellipsoids, the limit process $Z$ is indexed by $(\boldsymbol{t}, \boldsymbol{u})$ and the covariance kernel is:

(5)
$$\begin{aligned} C((\boldsymbol{t}, \boldsymbol{u}), (\boldsymbol{t}^*, \boldsymbol{u}^*)) = \\ \lim_{\delta \downarrow 0} \delta^{-1} P(E(\boldsymbol{\mu}_0 + \boldsymbol{t}\delta, \tilde{\boldsymbol{\lambda}}_0 + \boldsymbol{u}\delta) - E(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)) \\ (E(\boldsymbol{\mu}_0 + \boldsymbol{t}^*\delta, \tilde{\boldsymbol{\lambda}}_0 + \boldsymbol{u}^*\delta) - E(\boldsymbol{\mu}_0, \tilde{\boldsymbol{\lambda}}_0)) \\ = \lim_{\delta \downarrow 0} \delta^{-1} [PA(\boldsymbol{t}\delta, \boldsymbol{u}\delta, 0, 0) + PA(0, 0, \boldsymbol{t}^*\delta, \boldsymbol{u}^*\delta) \\ - PA(\boldsymbol{t}\delta, \boldsymbol{u}\delta, \boldsymbol{t}^*\delta, \boldsymbol{u}^*\delta)] \end{aligned}$$

where $A(\boldsymbol{x}_1, \boldsymbol{y}_1, \boldsymbol{x}_2, \boldsymbol{y}_2) = E(\boldsymbol{\mu}_0 + \boldsymbol{x}_1, \tilde{\boldsymbol{\lambda}}_0 + \boldsymbol{y}_1) \cap E(\boldsymbol{\mu}_0 + \boldsymbol{x}_2, \tilde{\boldsymbol{\lambda}}_0 + \boldsymbol{y}_2)^c$.

In the case of spheres, take $Z$ to be a centered Gaussian process with covariance kernel:

(6)
$$C((\boldsymbol{t}, s), (\boldsymbol{t}^*, s^*)) = \lim_{\delta \downarrow 0} \delta^{-1} P(S(\boldsymbol{\mu}_0 + \boldsymbol{t}\delta, r_0 + s\delta) - S(\boldsymbol{\mu}_0, r_0))(S(\boldsymbol{\mu}_0 + \boldsymbol{t}^*\delta, r_0 + s^*\delta) - S(\boldsymbol{\mu}_0, r_0)).$$

If $P$ is spherically symmetric about $\boldsymbol{\mu}_0$ with differentiable density then the limit in (6) becomes

(7)
$$\begin{aligned} C((\boldsymbol{t}, s), (\boldsymbol{t}^*, s^*)) &= \alpha(d-1) v_{d-1} r_0^{d-1} [s L(\tfrac{s}{\|\boldsymbol{t}\|}) + \|\boldsymbol{t}\| M(|s|, \|\boldsymbol{t}\|) \\ &- s^* L(-\tfrac{s^*}{\|\boldsymbol{t}^*\|}) + \|\boldsymbol{t}^*\| M(|s^*|, \|\boldsymbol{t}^*\|) \\ &+ (s - s^*) L(\tfrac{s^* - s}{\|\boldsymbol{t} - \boldsymbol{t}^*\|}) + \|\boldsymbol{t} - \boldsymbol{t}^*\| M(|s - s^*|, \|\boldsymbol{t} - \boldsymbol{t}^*\|)] \end{aligned}$$

where

$$
\begin{aligned}
L(x) &= \int_{-1}^{x^*}(1-y^2)^{\frac{d-3}{2}}dy, \quad x^* = \max(-1, \min(1,x)) \\
M(x,y) &= (1-\min(1, x^2/y^2))^{\frac{d-1}{2}}.
\end{aligned}
$$

At first glance, this representation of the covariance kernel does not appear symmetric in $(\boldsymbol{t}, s)$ and $(\boldsymbol{t}^*, s^*)$. But, the equality

$$
L(x) = L(1) - L(-x)
$$

can symmetrize (7) when $L(x)$ is replaced with $\frac{1}{2}[L(x) + L(1) - L(-x)]$ in the first 2 terms on the right hand side.

The lemma below finds the limit distribution of $(\boldsymbol{\mu}_n, r_n)$ when $P$ has spherical contours.

**4.1. Lemma:** *Suppose*

*(i)* $S(0, r_0)$ *is the unique maximizer of (4a) and* $S(\boldsymbol{\mu}_n, r_n)$ *maximizes (4b) over* $\mathcal{S}$*;*

*(ii)* $p(\boldsymbol{x}) = g(\boldsymbol{x}'\boldsymbol{x})$*;*

*(iii)* $g$ *has two derivatives* $g^{(1)}$ *and* $g^{(2)}$*, and* $g^{(1)}(r_0^2) < 0$.

*Then* $n^{1/3}(\boldsymbol{\mu}_n, r_n - r_0)$ *converges in distribution to* $(\boldsymbol{t}^*, s^*)$ *the almost-surely unique maximizer of*

$$
r_0^d g^{(1)}(r_0^2)[v_d \boldsymbol{t}'\boldsymbol{t} + w_d(d^2 - 2d)s^2] + Z(\boldsymbol{t}, s),
$$

where $Z$ is a centered Gaussian process with covariance kernel (7). □

The density need not be spherically symmetric in order for a sphere to uniquely maximize (4a) over all spheres. For example, if $p$ is elliptically symmetric then some sphere, say $S(0, r_0)$, uniquely maximizes (4a) over $\mathcal{S}$ (Tong, 1980). In this case, provided $g$ is strictly decreasing, we get a limit process not unlike that of the previous lemma.

**4.2. Lemma:** *Suppose*

*(i)* $S(0, r_0)$ *uniquely maximizes (4a) and* $S(\boldsymbol{\mu}_n, r_n)$ *maximizes (4b), both over* $\mathcal{S}$*;*

*(ii)* $p(\boldsymbol{x}) = g(\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x})$ *where* $\Sigma$ *is a positive definite diagonal matrix with* $\Sigma = diag(\boldsymbol{\sigma})$, $\boldsymbol{\sigma} \in \boldsymbol{R}^d$*;*

*(iii)* $g$ *is decreasing on* $\boldsymbol{R}^+$ *and* $g$ *has two derivatives* $g^{(1)}$ *and* $g^{(2)}$*.*

*Then* $n^{1/3}(\boldsymbol{\mu}_n, r_n - r_0)$ *converges in distribution to the almost-surely-unique maximizer of*

$$\frac{1}{2}(\boldsymbol{t}, s)'W(\boldsymbol{t}, s) + Z(\boldsymbol{t}, s),$$

*where* $Z$ *is a centered Gaussian process with covariance kernel (6) and* $W$ *is a* $(d+1) \times (d+1)$ *diagonal matrix with diagonal elements:*

$$w_{ii} = \int_{\{\boldsymbol{y}'\boldsymbol{y} \le r_0^2\}} \frac{2}{\sigma_i} g^{(1)}(\boldsymbol{y}'\Sigma^{-1}\boldsymbol{y}) + \frac{4y_i^2}{\sigma_i^2} g^{(2)}(\boldsymbol{y}'\Sigma^{-1}\boldsymbol{y}) d\boldsymbol{y} \quad i \le d$$

$$w_{d+1,d+1} = \int_{\{\boldsymbol{y}'\boldsymbol{y} \le r_0^2\}} \frac{4}{r_0^2} (\boldsymbol{y}'\Sigma^{-1}\boldsymbol{y})^2 g^{(2)}(\boldsymbol{y}'\Sigma^{-1}\boldsymbol{y}) d\boldsymbol{y} - \frac{d(d+2)}{r_0^2} J(0, r_0). \quad \square$$

Notice that when $\sum = I$ the expectation and covariance kernel of the limit process reduce to the special case of Lemma 4.1. This is made clear in the proof of Lemma 4.1.

The next lemma finds the asymptotic behavior of the maximizing ellipse $E_n$ when $p$ is elliptically symmetric. Recall $\Lambda_0 = \Sigma_0^{1/2}$ and $\tilde{\boldsymbol{\lambda}}_0$ is the vector of elements in $\Lambda_0$.

**4.3. Lemma:** *Suppose*

*(i)* $p(\boldsymbol{x}) = g(\boldsymbol{x}'\Sigma_0^{-1}\boldsymbol{x})$ *for* $\Sigma_0 = diag(\boldsymbol{\sigma})$ *a positive definite diagonal matrix and* $\alpha = g(1)$*;*

*(ii)* $E(0, \tilde{\boldsymbol{\lambda}}_0)$ *uniquely maximizes (3a) over* $\mathcal{E}$ *and* $E_n = E(\boldsymbol{\mu}_n, \tilde{\boldsymbol{\lambda}}_n)$ *maximizes (3b) over* $\mathcal{E}$*.*

*(iii)* $g$ *has two derivatives* $g^{(1)}$ *and* $g^{(2)}$*, and* $g^{(1)}(1) < 0$*.*

*Then* $n^{1/3}(\boldsymbol{\mu}_n, \tilde{\boldsymbol{\lambda}}_n - \tilde{\boldsymbol{\lambda}}_0)$ *converges in distribution to the almost surely unique maximizer for*

$$(\boldsymbol{t}, \boldsymbol{u})'W(\boldsymbol{t}, \boldsymbol{u}) + Z(\boldsymbol{t}, \boldsymbol{u})$$

*where $Z$ is a centered Gaussian process with covariance kernel (5) and $W$ is a $2d + \binom{d}{2}$ diagonal matrix with*

$$\begin{aligned} w_{ii} &= g^{(1)}(1) v_d \det(\Sigma_0^{1/2})/\sigma_i \quad 1 \le i \le d \\ & \quad 3 g^{(1)}(1) w_d \det(\Sigma_0^{1/2})/\sigma_{i-d} \quad d+1 \le i \le 2d \end{aligned}$$

*and for $i > 2d$, if the $i^{th}$ element of $\tilde{\boldsymbol{\lambda}}$ is $\lambda_{jk}$ then*

$$w_{ii} = g^{(1)}(1) w_d \det(\Sigma_0^{1/2})\left(\frac{1}{\sqrt{\sigma_j}} + \frac{1}{\sqrt{\sigma_k}}\right)^2 . \square$$

## 5 Other Rates of Convergence

Muller and Sawitzki (1989) and Hartigan (1987) examine set statistics that are offshoots of the $\alpha$-level excess-mass ellipsoid. What is interesting in both cases is that these set statistics offer rates of convergence other than the $n^{1/3}$ seen already. A rigorous treatment of the asymptotic properties of these statistics follows from an extension of Kim and Pollard's (1990) main result. In this section we extend their result to cover general rates of convergence and apply it to the set statistics of Muller and Sawitzki (1989) and Hartigan (1987).

**5.1. Example.** Muller and Sawitzki (1989) consider the following difference, in one dimension,

$$D_n(\alpha) = \sup_{\{I,J: I \cap J = \emptyset, I,J \in \mathcal{I}\}} P_n(I \cup J) - \alpha V(I \cup J) - \sup_{I \in \mathcal{I}} P_n(I) - \alpha V(I).$$

They use $D_n(\alpha)$ to indicate multimodality of the distribution $P$. They bound it by

$$\max\left(\sup_{I \subset I_\alpha^c} P_n(I) - \alpha V(I), \sup_{I \subset I_\alpha} -P_n(I) + \alpha V(I)\right), \tag{8}$$

which they then bound by a Kolmogorov-Smirnov statistic with a $\sqrt{n}$ rate of convergence. Marron and Nolan will investigate the exact asymptotic distribution of $D_n(\alpha)$ in $\mathbb{R}^d$ in a future paper.

If $P$ is strongly unimodal the first term in (8) should be close to

$$\max(\sup_{t>0} P_n[b_\alpha, b_\alpha + t] - \alpha t, \quad \sup_{t>0} P_n[a_\alpha - t, a_\alpha] - \alpha t)$$

where $I_\alpha = [a_\alpha, b_\alpha]$. Here we find the asymptotic distribution of the related multidimensional set statistic

$$\arg\sup_{t\geq 0} P_n(B_t) - \alpha V(B_t) \tag{9}$$

where $B_t = S((1+t,0),t)$ is the ball with center $(1+t,0)$ and radius $t$. We take $P$ to be the standard bivariate normal and $\alpha = e^{-1/2}/2\pi$. We show that $t_n$, the value of $t$ that maximizes (9), converges at a $n^{1/4}$ rate to a nondegenerate limit. □

**5.2. Example.** Hartigan illustrates the slow rate of convergence that occurs in (1) with a heuristic argument that finds the limit distribution of the Hausdorff distance:

$$\rho(S_\alpha \cup T(s_n), S_\alpha) = s_n,$$

where $T(s_n)$ is the triangular cap formed by taking the convex hull of $S_\alpha$ and the point $(1+s_n, 0)$, $s_n \geq 0$. He considers the case where $P$ is the standard bivariate normal and $S_\alpha = S(0,1)$. We make his example rigorous by finding the limit distribution of $s_n$, where $s_n$ is chosen to maximize

$$\sup_{s\geq 0} P_n(S_\alpha \cup T(s)) - \alpha V(S_\alpha \cup T(s)). \quad \square$$

The following theorem extends Theorem 1.1 of Kim and Pollard (1990) to cover these examples. We borrow their notation and format for the statement of the result.

**5.1 Theorem.** *Let $\{f(\cdot,\theta) : \theta \in \Theta\}$ be a class of functions indexed by a subset $\Theta$ of $\mathbb{R}$. Let $\{\theta_n\}$ be a sequence of estimators of $\theta_0 \in \Theta$ based on a random sample from a distribution $P$ such that*

*(i) $P_n f(\cdot,\theta_n) \geq \sup_\theta P_n f(\cdot,\theta)$.*

*Suppose that*

*(ii) $\theta_n$ is consistent for $\theta_0$, the unique maximizer of $Pf(\cdot,\theta)$*

*(iii) $\theta_0$ is an interior point of $\Theta$.*

*Let the functions be standardized such that $f(\cdot,\theta_0)\equiv 0$. If the classes $\mathcal{F}_R = \{f(\cdot,\theta)\colon |\theta-\theta_0| < R\}$, for $R$ near 0, are uniformly manageable for the natural envelope $F_R = \sup_{\mathcal{F}_R}|f(\cdot,\theta)|$ and satisfy*

*(iv) $Pf(\cdot,\theta) = \kappa|\theta-\theta_0|^\alpha + o(|\theta-\theta_0|^\alpha)$ for $\theta$ near $\theta_0$, $\alpha > 1$, $\kappa < 0$.*

*(v) $C(s,t) = \lim_{\delta\to 0}\delta^{-\beta}Pf(\cdot,\theta_0+s\delta)f(\cdot,\theta_0+t\delta)$ exists for each $s,t\in\mathbb{R}^d$ and $\lim_{\delta\to 0}\delta^{-\beta}Pf(\cdot,\theta_0+s\delta)^2\,\{|f(\cdot,\theta_0+s\delta)| > \epsilon\delta^{-\beta}\} = 0$ for each $\epsilon > 0$ and $s\in\mathbb{R}$, for some $0\le\beta<2\alpha$.*

*(vi) $PF_R^2 = O(R^\beta)$ as $R\to 0$ and for each $\epsilon > 0$ there is a constant $M$ such that $PF_R^2\{F_R > M\} < \epsilon R^\beta$ for $R$ near 0.*

*(vii) $P|f(\cdot,\theta_1) - f(\cdot,\theta_2)| = O(|\theta_1^\beta - \theta_2^\beta|)$ near 0*

*then the sequence $n^{\frac{\alpha}{2\alpha-\beta}}\,P_n f(\cdot,\theta_0+sn^{-\frac{1}{2\alpha-\beta}})$ converges in distribution to a Gaussian process $Y(s)$ with continuous sample paths, expected value $\kappa|s|^\alpha$ and covariance kernel $C$. If $Y$ has nondegenerate increments then $n^{\frac{1}{2\alpha-\beta}}(\theta_n-\theta_0)$ converges weakly to the (almost surely unique) random vector that maximizes $Y$.* □

Before proving the above result, we apply it to the two examples. Rather than rigorously checking all the conditions of the theorem in the examples, we simply determine $\alpha$, $\beta$, and the rate of convergence. See the proofs in Section 6 for a discussion of uniform manageability and other conditions.

**5.1. Example** (continued). Here we let $\theta = t$ and

$$f(\cdot,t) = S_t - \alpha\nu_2 t^2.$$

Then $Pf(\cdot,t)$ is maximized at $t=0$. A change of variables followed by a three term Taylor series expansion of $Pf(\cdot,t)$ about 0 shows

$$\begin{aligned} Pf(\cdot,t) &= t^2\int_{\{\boldsymbol{x}'\boldsymbol{x}\le 1\}} g(t^2\boldsymbol{x}'\boldsymbol{x} + 2x_1(t+t^2) + (1+t)^2)d\boldsymbol{x} - \alpha\nu_2 t^2 \\ &= 2\pi|t|^3 g'(1) + o(|t|^3). \end{aligned}$$

Also,

$$Pf(\cdot,t)f(\cdot,s) = PS_tS_s + o(s^2t^2).$$

Apply Theorem 5.1 with $\alpha = 3$ and $\beta = 2$ to the collection of functions $\{S_t - \alpha\nu_2 t^2 : t\ge 0\}$. Therefore

$$n^{3/4}P_n f(\cdot,tn^{-1/4})$$

converges in distribution to a Gaussian process with expectation $2\pi g'(1)|t|^3$ and covariance kernel $C(s,t) = \alpha\nu_2 \min(t^2, s^2)$. As $g'(1) < 0$, $n^{1/4}t_n$ converges to the unique maximizer of the limit process. □

**5.2. Example** (continued). In this example

$$\begin{aligned} f(\cdot, s) &= S_\alpha \cup T(s) - \alpha V(S_\alpha \cup T(s)) - S_\alpha + \alpha V(S_\alpha) \\ &= T(s) - \alpha V(T(s)). \end{aligned}$$

As above, $Pf(\cdot, s)$ is maximized for $s = 0$. However,

$$\begin{aligned} Pf(\cdot, s) &= \int_{T(s)} \frac{1}{2\pi}(e^{-r^2/2} - e^{-1/2}) r dr d\theta \\ &= \frac{-\alpha}{20}|s|^{5/2} + o(s^{5/2}). \end{aligned}$$

Take $\alpha$ to be $5/2$. For some constant $c$,

$$PT(\cdot, s)T(\cdot, t) = \alpha c \min(s,t)^{3/2}. \tag{10}$$

So $\beta = 3/2$. By Theorem 5.1, $n^{5/7}P_n f(\cdot, sn^{-2/7})$ converges to a Gaussian process with expectation $-\frac{\alpha}{20}|s|^{5/2}$ and covariance kernel (10), and $n^{2/7}s_n$ converges to the maximizer of the limit Gaussian process. □

We close the section with a proof of the Theorem. The argument follows that of Theorem 1.1 of Kim and Pollard (1990). There, the limit process has a quadratic drift, and the normalization is $n^{2/3}$ for the process and $n^{1/3}$ for the parameter. References to their lemmas are made as needed.

**Proof:** We first show that $\theta_n - \theta_0 = O_p(n^{-1/\gamma})$, where $\gamma = 2\alpha - \beta$. Then it is shown that the stochastic process $Z_n$ defined by

$$Z_n(s) = n^{\alpha/\gamma}[Pf(\cdot, sn^{-1/\gamma}) + (P_n - P)f(\cdot, sn^{-1/\gamma})] \ , \tag{11}$$

converges to a Gaussian process $Z(s)$ with continuous sample paths, expectation $\kappa|s|^\alpha$ and an almost surely unique maximum. This in turn implies $n^{1/\gamma}(\theta_n - \theta_0)$ converges in distribution to $\arg\max_s Z(s)$.

Consider the stochastic process $(P_n - P)$ indexed by $\mathcal{F}_R = \{f(\cdot, \theta) : 0 \leq |\theta - \theta_0| \leq R\}$. By the assumption of uniform manageability it is stochastically equicontinuous. Stochastic equicontinuity says $(P_n - P)f(\cdot, \theta_n) = o_p(n^{-1/2})$

(Lemma VII.15 Pollard, 1984). This fact and the following inequality from (i) and (iv):

$$\kappa|\theta_n - \theta_0|^\alpha + o_p(|\theta_n - \theta_0|^\alpha) + (P_n - P)(f(\theta_n)) \geq 0, \tag{12}$$

imply that $\theta_n - \theta_0 = O_p(n^{-1/2\alpha})$. For, otherwise the lower bound of 0 is violated.

To further refine the stochastic order of $\theta_n - \theta_0$ we use the following maximal inequality, based on a inequality of Marcus and Pisier (1981),

(13)

$$\begin{aligned} \mathbf{P}\sup_{\mathcal{F}_R}|\sqrt{n}(P_n - P)f(\cdot,\theta)| &\leq \mathbf{P}\sqrt{P_n f^2(\cdot,R)}J(\sup_{\mathcal{F}_R} P_n f(\cdot,\theta)^2/P_n f(\cdot,R)^2) \\ &\leq cR^{\beta/2} \end{aligned}$$

Here $J$ is a continuous, increasing function with $J(0) = 0$ and $J(1) < \infty$. The last upper bound is due to the boundedness of $J(1)$ and the upper bound from (vi). Inequalities (12) and (13) imply that $\theta_n - \theta_0 = O_p(n^{-1/\gamma})$. This is seen from the two statements: if $\theta_n - \theta_0 = O_p(\delta_n)$ then by (13), $(P_n - P)f(\cdot,\theta_n) = O_p(n^{-1/2}\delta_n^{\beta/2})]$; and if $(P_n - P)f(\cdot,\theta_n) = O_p(n^{-1/2}\delta_n^{\beta/2})$ then by (12), $\theta_n - \theta_0 = O_p(n^{-1/\gamma}\delta_n^{\beta/2\alpha})$. The coefficient $\frac{1}{2\alpha-\beta} = \frac{1}{2} \times \frac{1}{\alpha}\sum_{k=0}^{\infty}(\frac{1}{\alpha} \times \frac{\beta}{2})^k$.

Now that the rate is established, the parameter $\theta - \theta_0$ can be rescaled by $n^{1/\gamma}$. Let $\theta - \theta_0 = sn^{-1/\gamma}$ and $\mathcal{F}_n = \{f(\cdot, sn^{-1/\gamma})\colon 0 \leq s \leq M\}$. Also rescale the process by $n^{\alpha/\gamma}$ to get $Z_n(s)$ as in (11). Convergence of $Z_n$ to $Z$ follows from the convergence of the finite dimensional distributions and a stochastic equicontinuity argument. For fixed $s$, note that by (v),

$$\begin{aligned} &\operatorname{cov}(n^{\alpha/\gamma}(P_n - P)f(\cdot, sn^{-1/\gamma}), n^{\alpha/\gamma}(P_n - P)f(\cdot, \tilde{s}n^{-1/\gamma})) \\ &\quad = n^{\alpha/\gamma}P[f(\cdot, sn^{-1/\gamma})f(\cdot, \tilde{s}n^{-1/\gamma})] - n^{\alpha/\gamma}P[f(\cdot, sn^{-1/\gamma})]P[f(\cdot, \tilde{s}n^{-2/7})] \\ &\quad \to C(s,t) \end{aligned}$$

The Lindeberg-Feller CLT and (v) show that $Z_n(s)$ converges in distribution to $Z(s)$ a normal random variable with mean $\kappa|s|^\alpha$ and variance $C(s,s)$.

Stochastic equicontinuity follows from Lemma (4.6) of Kim and Pollard (1990) adjusted to reflect the facts: the expectation of $Z(s)$ is $|s|^\alpha$ rather than $s^2$; the normalization in (11) is $n^{\alpha/\gamma}$ rather than $n^{2/3}$; and $Pf^2(\cdot,\theta) = O(|\theta - \theta_0|^\beta)$ rather than $O(|\theta - \theta_0|)$. Change the conditions (ii), (iii) and (iv) of Kim and Pollard's Lemma (4.6) to reflect these differences:

(ii)′ $Pf(\cdot, R)^2 = O(R^\beta)$ as $R \to 0$;
(iii)′ $P|f(\cdot, s_1) - f(\cdot, s_2)| = O(|s_1^\beta - s_2^\beta|)$ near 0;
(iv)′ for $\epsilon > 0$ there is a $K$ such that

$$Pf(\cdot, R)^2\{f(\cdot, R) > K\} < \epsilon R^\beta \text{ for } R \text{ near } 0 .$$

Then, according to this new version of the lemma,

$$n^{\alpha/\gamma}\mathbf{P}\sup_{[\delta_n]}|(P_n - P)(f(\cdot, s_1 n^{-1/\gamma}) - f(\cdot, s_2 n^{-1/\gamma}))| = o(1),$$

where $[\delta_n] = \{(s_1, s_2) : |s_1 - s_2| \leq \delta_n \text{ and } 0 \leq s_1, s_2 \leq M\}$.

Finally, in order for $n^{1/\gamma}(\theta_n - \theta_0)$ to converge in distribution to arg max $Z(s)$, the process $Z$ must have an almost surely unique maximum and $Z(s) \to -\infty$ as $s \to \infty$, almost surely (Theorem 2.7, Kim and Pollard 1990). The uniqueness of the max of $Z$ is ensured provided $\text{var}(Z(s) - Z(t)) \neq 0$ for $s \neq t$ (Lemma 2.6, Kim and Pollard, 1990). The second property follows from:

$$\mathbf{P}\{\limsup_{s\to\infty} |s|^{-\alpha}W(s) > \epsilon\} = 0,$$

where $Z(s) = \kappa|s|^\alpha + W(s)$. Lemma 2.5 of Kim and Pollard with $|t|^2$ replaced by $|s|^\alpha$ and $k^{1/2}$ replaced by $k^{\beta/2}$ yields this result. Therefore $\theta_n - \theta_0$ attains the claimed distribution. □

## 6 Proofs of Lemmas 4.1, 4.2 and 4.3

The proofs of the weak convergence results of section 4 are found here. To prove them we use the following result of Kim and Pollard (1990; Theorem 1.1).

**6.1 Theorem.** *Let $\{f(\cdot, \theta); \theta \in \Theta\}$ be a class of functions indexed by a subset $\Theta$ of $\mathbb{R}^d$. Let $\{\theta_n\}$ be a sequence of estimators of $\theta_0 \in \Theta$ based on a random sample from a distribution $P$ such that*

*(i) $P_n f(\cdot, \theta_n) \geq \sup_\theta P_n f(\cdot; \theta) - o_p(n^{-2/3})$.*

*Suppose that*

*(iia) $\theta_n$ is consistent for $\theta_0$*

*(iib) $\theta_0$ is the unique maximizer of $Pf(\cdot, \theta)$*

*(iii) $\theta_0$ is an interior point of $\Theta$.*

*Let the functions be standardized such that $f(\cdot, \theta_0) \equiv 0$. If the classes $\mathcal{F}_R = \{f(\cdot, \theta)\colon \|\theta\| < R\}$, for $R$ near 0, are uniformly manageable for the natural envelope $F_R = \sup_{\mathcal{F}_R} |f(\cdot, \theta)|$ and satisfy* *(iv) $Pf(\cdot,\theta)$ is twice differentiable with second derivative $V$ at $\theta_0$;*

*(v) $C(s,t) = \lim_{\delta\to 0} \delta^{-1} Pf(\cdot,\theta_0+s\delta)f(\cdot,\theta_0+t\delta)$ exists for each $s,t \in \mathbb{R}^d$, and $\lim_{\delta\to 0}\delta^{-1}Pf(\cdot,\theta_0+s\delta)^2\ \{|f(\cdot,\theta_0+s\delta)| > \epsilon\delta^{-1}\} = 0$ for each $\epsilon > 0$ and each $s \in \mathbb{R}^d$;*

*(vi) $PF_R^2 = O(R)$ as $R \to 0$ and for each $\epsilon > 0$ there is a constant $K$ such that $PF_R^2\{F_R > K\} < \epsilon R$ for $R$ near $0$.*

*(vii) $P|f(\cdot,\theta_1) - f(\cdot,\theta_2)| = O(\|\theta_1 - \theta_2\|)$ near $\theta_0$,*

*then the sequence of processes $n^{2/3}P_n f(\cdot,\theta_0 + sn^{-1/3})$ converges in distribution to a Gaussian process $Y(s)$ with continuous sample paths, expected value $\frac{1}{2}s'Vs$ and covariance kernel $C$. If $V$ is negative definite and if $Y$ has nondegenerate increments then $n^{1/3}(\theta_n - \theta_0)$ converges weakly to the (almost surely unique) random vector that maximizes $Y$.* □

**Proof of 4.1:** To prove Lemma 4.1, apply Theorem 6.1 to the collection of functions

$$\mathcal{F} = \{f : f(\boldsymbol{x},\boldsymbol{\mu},r) = \{\boldsymbol{x} \in S(\boldsymbol{\mu},r)\} - \{\boldsymbol{x} \in S(0,r_0)\} - \alpha v_d r^d + \alpha v_d r_0^d,$$
$$\boldsymbol{x},\boldsymbol{\mu} \in \mathbb{R}^d, r \in \mathbb{R}^+\}$$

Here $\{\boldsymbol{x} \in S(\boldsymbol{\mu},r)\}$ should be interpreted as the indicator function for the self same set. The $\boldsymbol{\theta}$ of Theorem 6.1 is $(\boldsymbol{\mu},r)$ in our application. The subclass $\mathcal{F}_R$ of $\mathcal{F}$ has envelope $F_R$:

$$\{\boldsymbol{x} \in S(0,r_0+R)\} - \{\boldsymbol{x} \in S(0,r_0-R)\} + \alpha v_d(R+r_0)^d - \alpha v_d(r_0-R)^d.$$

The requirement that the class $\mathcal{F}_R$ be manageable for envelope $F_R$ is a metric entropy condition on $\mathcal{F}_R$. Pollard (1989) provides sufficient conditions for manageability. Of particular interest here are the following conditions for manageability:

(i) A collection of indicator functions for a Vapnik-Cervonenkis class of sets is manageable for the envelope constructed from the supremum over the indicator functions.

(ii) The collection of constant functions $\{g : g(x) \equiv c,\ 0 \le c \le C\}$ is manageable for the envelope $C$, because the class of sets $\{A_c = \{(x,y)\colon x \in \mathbb{R},\ 0 \le y \le c\}\}$ is a Vapnik-Cervonenkis class of sets.

(iii) If $\mathcal{F}$ and $\mathcal{G}$ are manageable for envelopes $F, G$ respectively, then $\{f+g : f \in \mathcal{F} \text{ and } g \in \mathcal{G}\}$ is manageable for the evelope $F+G$.

Uniform manageability implies manageability for a family of classes $\{\mathcal{F}_R\}$, where the bounds in the metric entropy for each $\mathcal{F}_R$ depend on $R$ only through the envelope $F_R$. In our case, the collection of spheres $\{S(\boldsymbol{\mu}, r) : |\boldsymbol{\mu}| + |r - r_0| < R\}$ is a Vapnik-Cervonenkis class of sets, and so is manageable for the envelope $S(0, r_0 + R)$. The constant functions $\{\alpha v_d(r^d - r_0^d)\text{: } |r - r_0| < R\}$ are also manageable, and so by (iii) $\mathcal{F}_R$ is manageable for $F_R$.

Conditions (i) through (iii) of Theorem 6.1 are immediate consequences of the assumptions of the Lemma 4.1. Condition (vi) is easily met because $PF_R^2 \leq 2p_0((r_0 + R)^d - r_0^d)$, where $p_0 = \sup_r g(r)$. The same is true for (vii). It is (iv) and (v), the expectation and covariance structure of the limit process, that need to be established.

Recall $P(S(\boldsymbol{\mu}, r)) - \alpha V(S(\boldsymbol{\mu}, r))$ is:

$$J(\boldsymbol{\mu}, r) = \int_{\{\|\boldsymbol{x} - \boldsymbol{\mu}\| \leq r\}} (g(\boldsymbol{x}'\boldsymbol{x}) - \alpha) d\boldsymbol{x}.$$

A change of variables gives

$$J(\boldsymbol{\mu}, r) = \frac{r^d}{r_0^d} \int_{\{\|\boldsymbol{y}\| \leq r_0\}} (g(\|\frac{r}{r_0}\boldsymbol{y} + \boldsymbol{\mu}\|^2) - \alpha) d\boldsymbol{y}.$$

Then

$$\frac{\partial J(\boldsymbol{\mu}, r)}{\partial \mu_i} = \frac{r^d}{r_0^d} \int_{\{\|\boldsymbol{y}\| \leq r_0\}} g^{(1)}(\|\frac{r}{r_0}\boldsymbol{y} + \boldsymbol{\mu}\|^2) 2[\frac{r}{r_0} y_i + \mu_i] d\boldsymbol{y}$$

and

$$\frac{\partial J(\boldsymbol{\mu}, r)}{\partial r} = dr^{-1} J(\boldsymbol{\mu}, r) + \frac{r^d}{r_0^{d+1}} \int_{\{\|\boldsymbol{y}\| \leq r_0\}} g^{(1)}(\|\frac{r}{r_0}\boldsymbol{y} + \boldsymbol{\mu}\|^2) 2[\frac{r}{r_0}\|\boldsymbol{y}\|^2 + \boldsymbol{y}'\boldsymbol{\mu}] d\boldsymbol{y}.$$

Evaluate these integrals at $\boldsymbol{\mu} = 0$, $r = r_0$ to see that they are both 0. In particular, integration by parts gives

$$\begin{aligned} \frac{1}{r_0} \int_{\{\|\boldsymbol{y}\| \leq r_0\}} g^{(1)}(\|\boldsymbol{y}\|^2) \|\boldsymbol{y}\|^2 d\boldsymbol{y} &= \frac{1}{r_0} \sum_{i=1}^{d} \int_{\{\|\boldsymbol{y}\| \leq r_0\}} g^{(1)}(\|\boldsymbol{y}\|^2) 2y_i^2 d\boldsymbol{y} \\ &= \frac{d}{r_0} [g(r_0^2) r_0^d v_d - \int_{\{\|\boldsymbol{y}\| \leq r_0\}} g(\|\boldsymbol{y}\|^2) d\boldsymbol{y}] \\ &= -\frac{d}{r_0} J(0, r_0). \end{aligned}$$

Symmetry of $g$ implies the second order partial derivatives $\partial^2 J(\boldsymbol{\mu}, r)/\partial\mu_i\partial\mu_j$ are 0 at $\boldsymbol{\mu} = 0$, $r = r_0$ for $i \neq j$. For the same reason $\partial^2 J(\boldsymbol{\mu}, r)/\partial r\partial\mu_i|_{r=r_0,\mu=0}$ is also 0. The only nonzero terms are:

$$
\begin{aligned}
\text{(14)} \qquad \frac{\partial^2 J(\boldsymbol{\mu},r)}{\partial\mu_i^2}\Big|_{\substack{\boldsymbol{\mu}=0\\ r=r_0}} &= \textstyle\int_{\{\boldsymbol{y}'\boldsymbol{y}\le r_0^2\}} g^{(2)}(\boldsymbol{y}'\boldsymbol{y})4y_i^2 d\boldsymbol{y} + \textstyle\int_{\{\boldsymbol{y}'\boldsymbol{y}\le r_0^2\}} g^{(1)}(\boldsymbol{y}'\boldsymbol{y})2d\boldsymbol{y} \\
&= 2r_0^d g^{(1)}(r_0^2)v_d
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial^2 J(\boldsymbol{\mu},r)}{\partial r^2}\Big|_{\substack{\boldsymbol{\mu}=0\\ r=r_0}} &= d(d-1)r_0^{-2}J(0,r_0) + r_0^2(2d+1)\int_{\{\boldsymbol{y}'\boldsymbol{y}\le r_0^2\}} g^{(1)}(\boldsymbol{y}'\boldsymbol{y})2\boldsymbol{y}'\boldsymbol{y}d\boldsymbol{y} + \\
&\quad r_0^6\int_{\{\boldsymbol{y}'\boldsymbol{y}\le r_0^2\}} g^{(2)}(\boldsymbol{y}'\boldsymbol{y})(\boldsymbol{y}'\boldsymbol{y})^2 d\boldsymbol{y} \\
&= 2r_0^d g^{(1)}(r_0^2)w_d(d^2-2d).
\end{aligned}
$$

This establishes condition (iv) and gives the expectation of the limit process.

Finally, we establish (v), and find the covariance kernel of the process. Reparametrize $(\boldsymbol{\mu}, r)$ as $(\boldsymbol{t}\delta, r_0 + s\delta)$ for some positive scalar $\delta$. Then

$$
\begin{aligned}
&\lim_{\delta\to 0}\delta^{-1}Pf(\boldsymbol{t}\delta, r_0+s\delta)f(\tilde{\boldsymbol{t}}\delta, r_0+\tilde{s}\delta) \\
&= \lim_{\delta\to 0}\delta^{-1}P[S(\boldsymbol{t}\delta, r_0+s\delta) - S(0,r_0)]\,[S(\tilde{\boldsymbol{t}}\delta, r_0+\tilde{s}\delta) - S(0,r_0)] \\
&= \alpha\lim_{\delta\to 0}\delta^{-1}[V(A(\boldsymbol{t}\delta, s\delta, 0, 0)) + V(A(0,0,\tilde{\boldsymbol{t}}\delta, \tilde{s}\delta)) - V(A(\boldsymbol{t}\delta, s\delta, \tilde{\boldsymbol{t}}\delta, \tilde{s}\delta))]\ ,
\end{aligned}
$$

where $A(\boldsymbol{x}, \beta, \boldsymbol{y}, \gamma)$ is the set $S(\boldsymbol{x}, r_0+\beta)\cap S(\boldsymbol{y}, r_0+\gamma)^c$. For the first equality, the quadratic and linear terms in $\alpha v_d[(r_0+s\delta)^d - r_0^d]$ are negligible, because $[(r_0+s\delta)^d - r_0^d] = O(\delta)$ and $P|S(\boldsymbol{t}\delta, r_0+s\delta) - S(0,r_0)| = o(1)$. The second equality above is due to the fact that $p(\boldsymbol{x}) = \alpha$ for $\boldsymbol{x} \in \partial S(0, r_0)$. The covariance reduces to limiting volumes of symmetric differences of spheres. We find the volume of $A(0,0,\boldsymbol{t}\delta, s\delta)$. The other two can be found by analogous argument.

$$\delta^{-1}V(A(0,0,\boldsymbol{t}\delta, s\delta)) = \delta^{-1}\int_{\{\|\boldsymbol{x}\|\le r_0\} - \{\|\boldsymbol{x}\|\le r_0\}\cap\{\|\boldsymbol{x}-\boldsymbol{t}\delta\|\le r_0+s\delta\}} d\boldsymbol{x}.$$

Symmetry allows the replacement of $\boldsymbol{t}\delta$ by $\|\boldsymbol{t}\delta\|e_1$ where $e_1 = (1, 0, \ldots, 0)$. From this representation it is evident that the integral above is 0 if $s \ge \|\boldsymbol{t}\|$.

It is also evident that the integral is $v_d|s|dr_0^{d-1}$ if $s \leq -||t||$. For the remainder of the argument we assume $-||t|| < s < ||t||$.

$$\delta^{-1}V(A(0,0,t\delta,s\delta)) = \delta^{-1}\int_{\{-r_0 \leq x_1 \leq -r_0 - s\delta + ||t||\delta,\ ||x|| \leq r_0\}} dx$$

$$+\delta^{-1}\int_{\{-r_0 - s\delta + ||t||\delta \leq x_1 \leq -r_0 s/||t||\}} [\left(\sqrt{r_0^2 - x_1^2}\right)^{d-1} - \left(\sqrt{(r_0+s\delta)^2 - (x_1 - ||t||\delta)^2}\right)^{d-1}]dx_1$$

The first integral converges to 0 as $\delta \to 0$. The second integral has the nondegenerate limit:

$$-(d-1)v_{d-1}r_0^{d-1}s\int_{-1}^{-s/||t||}(1-y^2)^{\frac{d-3}{2}}dy + (d-1)v_{d-1}r_0^{d-1}||t||(1-s^2/||t||^2)^{\frac{d-1}{2}}$$

The covariance kernel for the limit process has been established. This concludes the proof of Lemma 4.1. □

**Proof of Lemma 4.2:** To prove this lemma return to the proof of Lemma 4.1 and substitute $g(x'\sum^{-1} x)$ for $g(x'x)$ in $J(\mu, r)$. The first derivative is still 0, by the maximization property of $J(0, r_0)$. The second derivative is nearly the same as well (see 14)). Unfortunately it does not simplify as in the proof of Lemma 4.1. □

**Proof of Lemma 4.3:** The collection of functions to which we apply Theorem 6.1 is only slightly different from that of Lemma 4.1. Simply replace the spheres by ellipses to get

$$\mathcal{F} = \{f : f(x, \mu, \tilde{\lambda}) = \{x \in E(\mu, \tilde{\lambda})\} - \{x \in E(0, \tilde{\lambda}_0)\} - \alpha v_d(\det\Lambda - \det\Lambda_0),$$

$$\mu \in \mathbb{R}^d,\ \ \tilde{\lambda} \in \mathbb{R}^{d + \binom{d}{2}}$$

The work that remains is to find the quadratic drift by computing the second derivative of $Pf(\cdot, \mu, \tilde{\lambda})$.

First write $PE(\mu, \tilde{\lambda}) - \alpha V(E(\mu, \tilde{\lambda}))$ as $J(\mu, \tilde{\lambda})$. Then by a change of variables,

$$J(\mu, \tilde{\lambda}) = \det(\Lambda)\int_{\{||y|| \leq 1\}} g((\mu + \Lambda y)'\Sigma_0^{-1}(\mu + \Lambda y)) - \alpha dy\ .$$

The first derivatives are 0, by the maximization property of $(0, \tilde{\boldsymbol{\lambda}}_0, 0)$.

$$\frac{\partial J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})}{\partial \mu_i} = \det(\Lambda) \int_{\{\boldsymbol{y}'\boldsymbol{y} \le 1\}} g^{(1)}((\boldsymbol{\mu}+\Lambda \boldsymbol{y})'\Sigma_0^{-1}(\boldsymbol{\mu}+\Lambda \boldsymbol{y}))[\frac{2\mu_i}{\sigma_i} + \frac{2}{\sigma_i}\sum_j \lambda_{ij} y_i] d\boldsymbol{y}$$

$$\frac{\partial J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})}{\partial \lambda_{ii}}|_{\mu=0} = \det(\Lambda_{ii}) \int_{\{\boldsymbol{y}'\boldsymbol{y} \le 1\}} g(\boldsymbol{y}'\Lambda\Sigma_0^{-1}\Lambda \boldsymbol{y}) - \alpha d\boldsymbol{y} +$$
$$\det(\Lambda) \int g^{(1)}(\boldsymbol{y}'\Lambda\Sigma_0^{-1}\Lambda\boldsymbol{y})[\frac{2\lambda_{ii}}{\sigma_i}\sum_j y_i y_j] d\boldsymbol{y}$$

where $\Lambda_{ij}$ is the matrix found by dropping the ith row and jth column from $\Lambda$.

$$\frac{\partial J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})}{\partial \lambda_{ij}}|_{\mu=0} = 2(-1)^{i+j}\det(\Lambda_{ij}) \int g(\boldsymbol{y}'\Lambda\Sigma_0^{-1}\Lambda\boldsymbol{y}) - 2\boldsymbol{y} +$$
$$2\det(\Lambda) \int g^{(1)}(\boldsymbol{y}'\Lambda\Sigma_0^{-1}\Lambda\boldsymbol{y})[\frac{\lambda_{ij} y_i^2}{\sigma_i} + \frac{\lambda_{ij} y_j^2}{\sigma_j} +$$
$$\sum_{k \ne j} \frac{\lambda_{ik} y_k y_j}{\sigma_i} + \sum_{\ell \ne i} \frac{\lambda_{\ell j} y_\ell y_i}{\sigma_j}].$$

The partial $\partial^2 J/\partial\mu_i\partial\lambda_{jk}$ and $\partial^2 J/\partial\mu_i\partial\mu_j$ are 0 at $\boldsymbol{\mu} = 0$, $\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}_0$, because of the symmetry of $g(\boldsymbol{y}'\boldsymbol{y})$ in $y_i$. The second partial derivatives $\partial^2 J/\partial\mu_i^2$ are similar to the spherical case of Lemma 4.1.

$$\begin{aligned} \frac{\partial^2 J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})}{\partial \mu_i^2}\Big|_{\substack{\mu=0 \\ \Lambda=\Sigma_0^{1/2}}} &= \det(\Sigma_0^{1/2}) \int_{\{\boldsymbol{y}'\boldsymbol{y} \le 1\}} \frac{4y_i^2}{\sigma_i} g^{(2)}(\boldsymbol{y}'\boldsymbol{y}) + \frac{2}{\sigma_i} g^{(1)}(\boldsymbol{y}'\boldsymbol{y}) d\boldsymbol{y} \\ &= \frac{2}{\sigma_i}\det(\Sigma_0^{1/2}) g^{(1)}(1) v_d. \end{aligned}$$

Finally we find the derivatives for the square root matrix.

$$\begin{aligned} \frac{\partial^2 J(\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}})}{\partial \lambda_{ii}^2}\Big|_{\substack{\mu=0 \\ \Lambda=\Sigma_0^{1/2}}} &= \det(\Sigma_0^{1/2}) \frac{2}{\sigma_i} \int_{\{\boldsymbol{y}'\boldsymbol{y} \le 1\}} 3y_i^2 g^{(1)}(\boldsymbol{y}'\boldsymbol{y}) + 2y_i^4 g^{(2)}(\boldsymbol{y}'\boldsymbol{y}) d\boldsymbol{y} \\ &= \det(\Sigma_0^{1/2}) \frac{6}{\sigma_i} g^{(1)}(1) w_d \end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial^2 J(\boldsymbol{\mu},\tilde{\boldsymbol{\lambda}})}{\partial \lambda_{ij}^2}\Big|_{\substack{\mu=0\\ \Lambda=\Sigma_0^{1/2}}} &= -2\det(\Sigma_0^{1/2})\Big[\int_{\{\boldsymbol{y}'\boldsymbol{y}\le 1\}} \frac{g(\boldsymbol{y}'\boldsymbol{y})-\alpha}{\sqrt{\sigma_i\sigma_j}} + g^{(1)}(\boldsymbol{y}'\boldsymbol{y})\Big(\frac{y_i^2}{\sigma_i}+\frac{y_j^2}{\sigma_j} \\
&+ 2y_i^2y_j^2\Big(\frac{1}{\sqrt{\sigma_i}}+\frac{1}{\sqrt{\sigma_j}}\Big)^2\Big) \\
&= 2\det(\Sigma_0^{1/2})\Big(\frac{1}{\sqrt{\sigma_i}}+\frac{1}{\sqrt{\sigma_j}}\Big)^2 g^{(1)}(1)w_d.
\end{aligned}
$$

This concludes the proof of Lemma 4.3. □

## References


Alexander, K.S. (1987). Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Th. Rel. Fields* **75**: 379-423.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). *Robust Estimates of Location.* Princeton NJ: Princeton University Press.

Chernoff, H. (1964). Estimation of the mode. *Ann. Math. Statist.* **16**: 31-41.

Devroye, L. (1983). Moment inequalities for random variables in computational geometry. *Computing* **30**: 111-119.

Eddy, W.F., and Hartigan, J.A. (1977). Uniform convergence of the empirical distribution function over convex sets. *Ann. Statist.* **5**: 370-374.

Grübel, R. (1988). The length of the shorth. *Ann. Statist.* **16**: 619-628.

Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *JASA* **82**: 267-270.

Kim, J., and Pollard, D. (1990). Cube root asymptotics. *Ann. statist.* **18**: 191-219.

Müller, D.W., and Sawitzki, G. (1989). Excess mass estimates and tests for multimodality. Preprint.

Pollard, D. (1984). *Convergence of Stochastic Processes.* New York, NY: Springer-Verlag.

Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Sci.* **4**: 341-366.

Preparata, F.P., and Shamos, M.I. (1985). *Computational Geometry: An Introduction.* New York, NY: Springer-Verlag.

Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Sager, T.W. (1979). An iterative method for estimating a multivariate mode and isopleth. *JASA* **74**: 329-339.

Silverman, B.W., and Titterington, D.M. (1980). Minimum covering ellipses. *SIAM J. Sci. Stat. Comput.* **1**: 401-409.

Titterington, D.M. (1978). Estimation of correlation coefficients by ellipsoidal trimming. *Appl. Statist.* **27**: 227-234.

Tong, Y.L. (1980). *Probability Inequalities in Multivariate Distributions.* New York, NY: Academic Press.

Venter, J.H. (1967). On estimation of the mode. *Ann. Math. Statist.* **38**: 1446-1455.

Wong, M.A., and Schaack, C. (1985). Diagnostics for assessing multimodality. Computer Sciences and Statistics: The Interface, L. Billard (ed). North-Holland Press. 287-289.

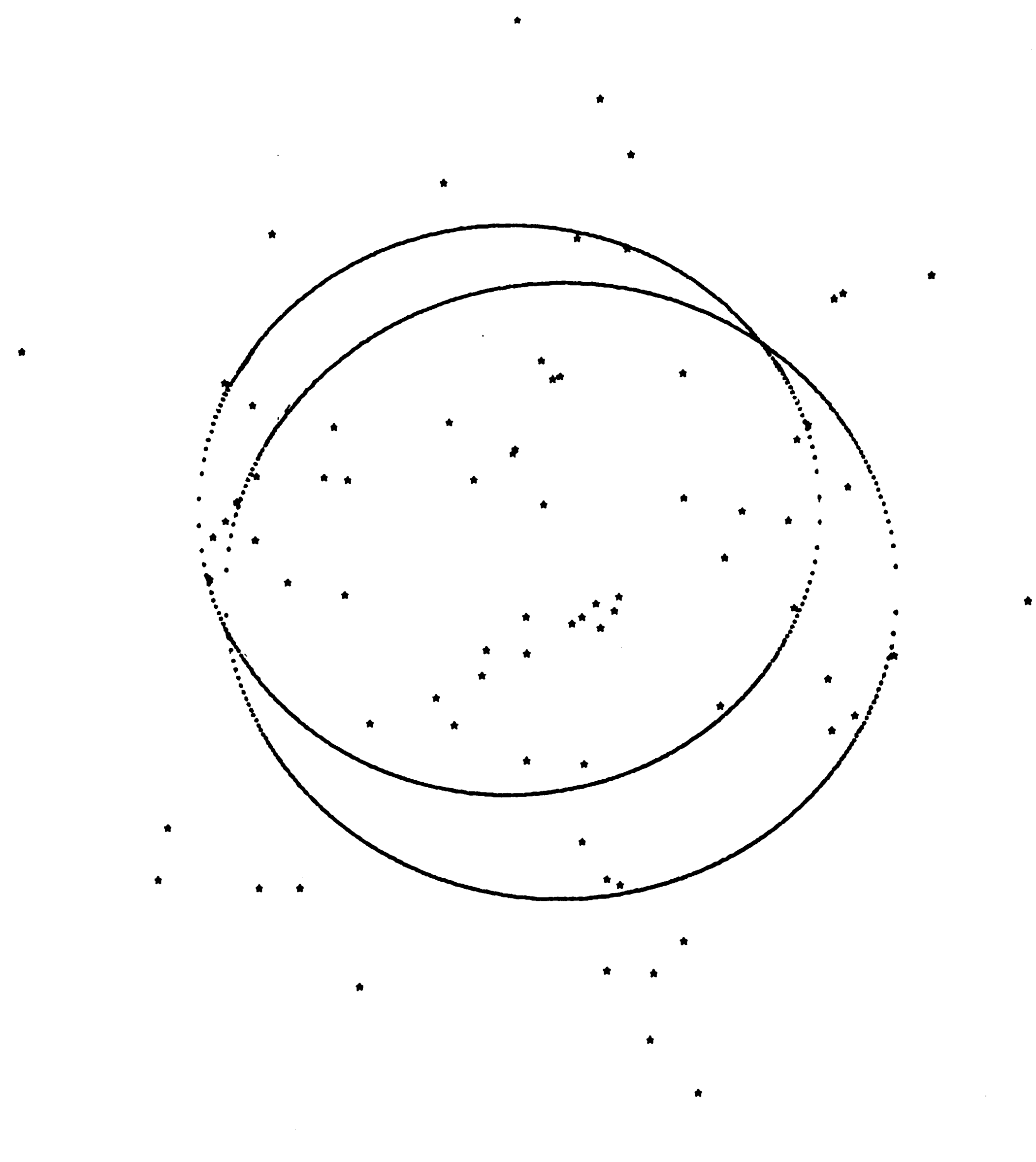

Figure 1

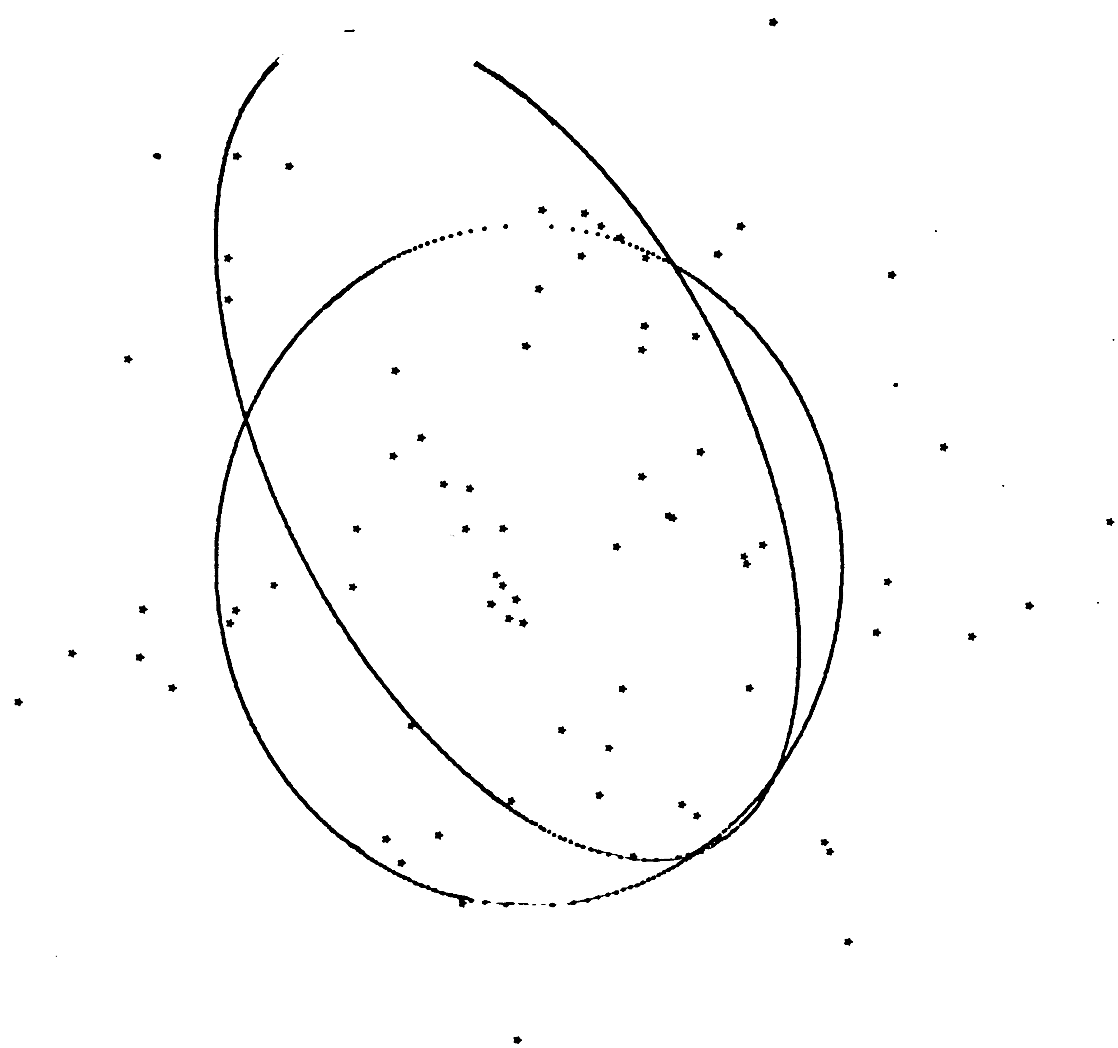

Figure 2

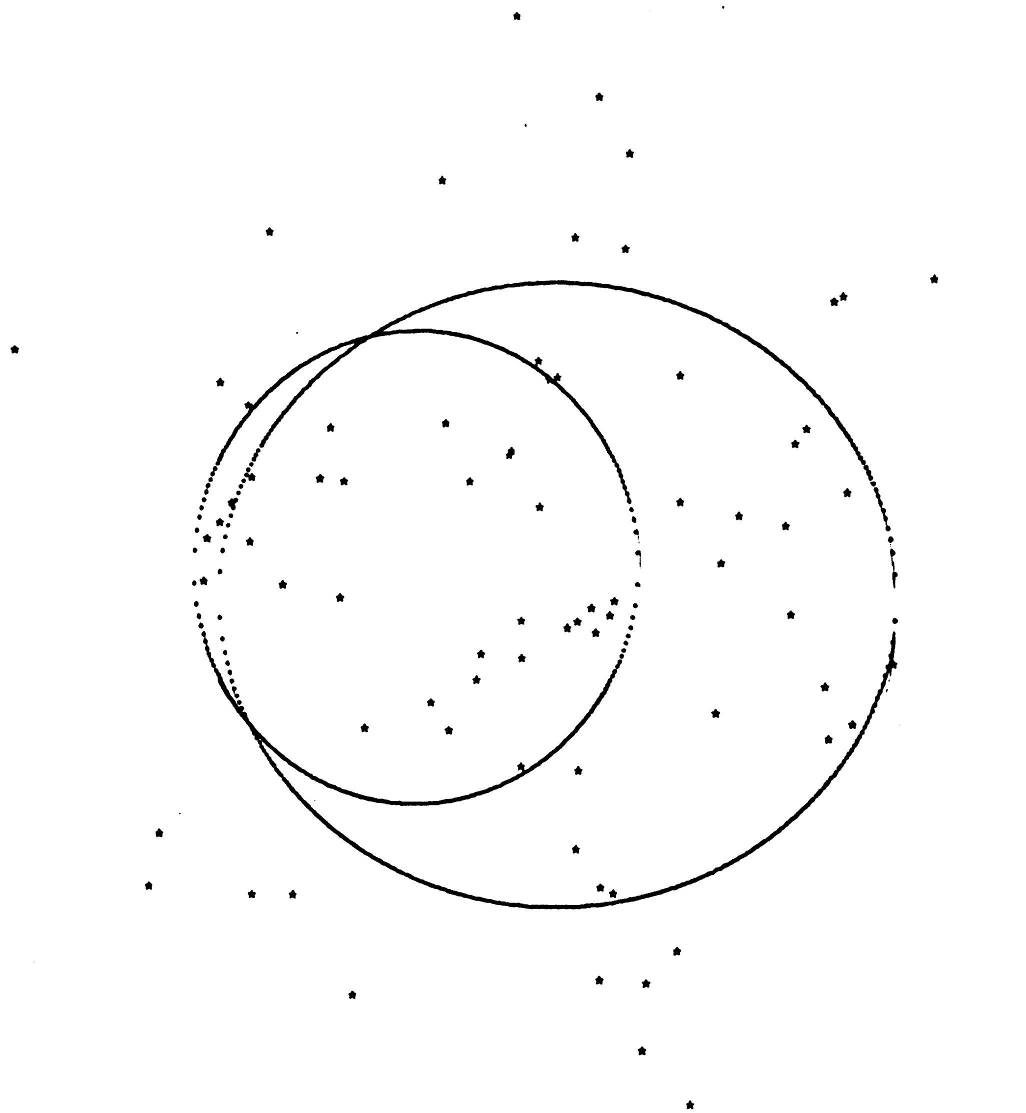

Figure 3

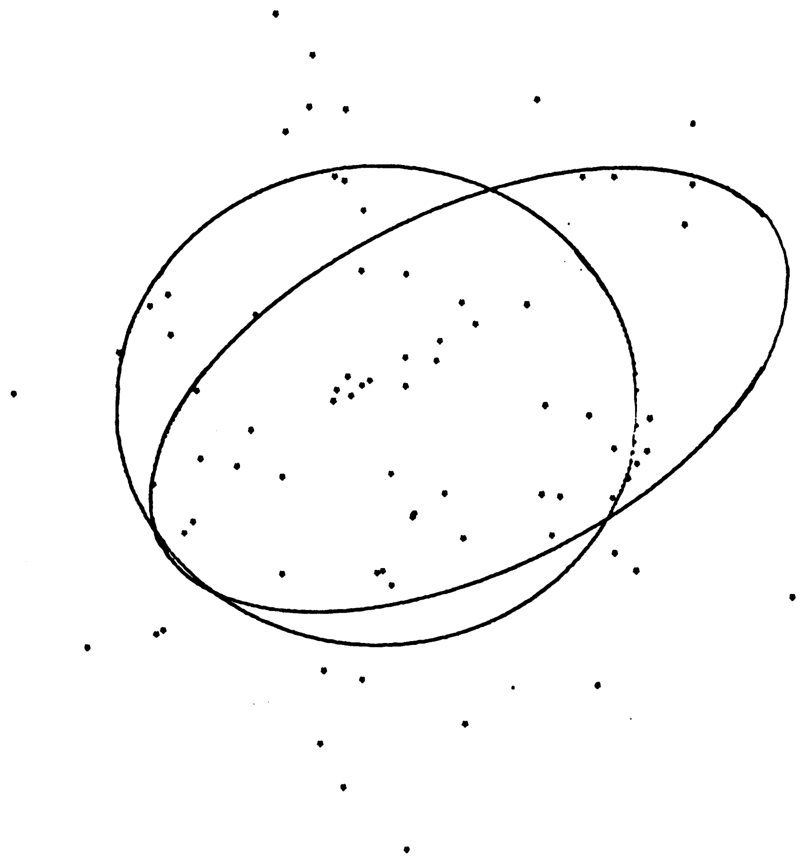

Figure 2