The Annals of Statistics
1992, Vol. 20, No. 4, 1828–1843

THE ASYMPTOTICS OF ROUSSEEUW'S MINIMUM VOLUME ELLIPSOID ESTIMATOR

BY LAURIE DAVIES

University of Essen

Rousseeuw's minimum volume estimator for multivariate location and dispersion parameters has the highest possible breakdown point for an affine equivariant estimator. In this paper we establish that it satisfies a local Hölder condition of order 1/2 and converges weakly at the rate of $n^{-1/3}$ to a non-Gaussian distribution.

1. Introduction and notation.

1.1. *Introduction.* It is well known that the mean and covariance matrix of a k-dimensional data set are highly susceptible to the influence of outliers. Indeed, one aberrant observation is sufficient to cause arbitrarily large changes in both estimators. This may be formalized by noticing that the finite-sample breakdown point in the sense of Donoho and Huber (1983) is $1/n$ for a sample of size n. It is of course a simple matter to obtain estimators with a breakdown point of 1 by, for example, taking the location estimator to be the zero vector and the dispersion estimator to be the identity matrix independently of the data. If, however, attention is restricted to affine equivariant estimators, then it can be shown that the breakdown point for both estimators cannot be greater than 1/2. In the one-dimensional situation it is possible to find M-estimators whose breakdown point is arbitrarily close to 1/2 but in higher dimensions this no longer holds. Maronna (1976) formally defined M-estimators for k-dimensional data and gave the upper bound of $1/(k+1)$ for the breakdown point. Other affine equivariant estimators such as those based on convex peeling or classical outlier rejection also have breakdown points bounded by $1/(k+1)$ [Donoho (1982)]. Chapter 7 of Rousseeuw and Leroy (1987) contains further examples and a discussion. The first affine equivariant estimators with the highest possible breakdown point were proposed independently by Stahel (1981) and Donoho (1982). A second such estimator was introduced by Rousseeuw [see Rousseeuw (1986)] and is defined as follows: Determine that ellipsoid $\hat{E}_n$ of minimal volume which contains at least $[n/2]+1$ data points. The center of the ellipsoid may be taken as a location estimator and the positive-definite symmetric matrix determining the shape of $\hat{E}_n$ gives an estimator for the dispersion of the data. For data points in general position, that is, with at most k points on any $(k-1)$-dimensional hyperplane, the finite-sample breakdown point of these estimators taken together is

Received March 1989; revised November 1991.
AMS 1980 *subject classifications*. Primary 62H12; secondary 62F12, 62F35.
Key words and phrases. Minimum volume ellipsoid, affine invariant metrics, Hölder conditions, cube root convergence.

$([n/2] - k + 1)/n$, which is close to the theoretical upper bound of $[(n - k + 1)/2]/n$ [see Davies (1987) and Lopuhaä and Rousseeuw (1991)]. It is this estimator which is the subject of the present paper.

The finite-sample breakdown point as a measure of the robustness of an estimator has become very popular. Its advantages are simplicity of definition, ease of explanation to nonstatisticians and its calculability for many interesting estimators. Nevertheless, it does have the following disadvantages. It is not a stochastic concept and has no analogue for theoretical distributions. It cannot therefore measure the change in the estimators if the underlying distribution of the data points is not that of the assumed model. Moreover, the finite-sample breakdown point does not measure the effect of "wobbling" the data as, when calculating the breakdown point, all but a certain number of the data points are held fixed. A good robust estimator should be continuous in a suitable topology on the space of distributions, small changes in the distribution giving rise to only small changes in the estimate. It is at least theoretically possible for an estimator to have a high finite-sample breakdown point but not to be continuous. Continuity of an estimator was part of Hampel's original definition of robustness and is a property not captured by the finite-sample breakdown point. Continuity with respect to the Prohorov metric has often been proposed as a desirable property of a robust estimator as it incorporates the idea of robustness against measurement inaccuracy. However, the Prohorov metric is a very strong metric and, in particular, empirical measures do not in general converge at the rate of $n^{-1/2}$ in the metric to the underlying distribution [Kersting (1978)]. Furthermore, it is difficult to calculate breakdown points in the Prohorov metric. This may be connected with the fact that the Prohorov metric is not affine invariant. We call a metric d on the space $\mathfrak{W}(\mathbb{R}^k)$ of distributions on the Borel sets $\mathfrak{B}(\mathbb{R}^k)$ of $\mathbb{R}^k$ affine invariant if $d(Q_1^A, Q_2^A) = d(Q_1, Q_2)$ for all affine transformations A: $\mathbb{R}^k \to \mathbb{R}^k$ and where Q^A is defined by $Q^A(B) = Q(A^{-1}(B))$. For affine invariant metrics and using the general definition of the breakdown point for metrics [see Huber (1981), pages 11–13], it is seen that the breakdown point is unaltered if the underlying measures are subject to an affine transformation. There remains the question as to which metric to choose. It has been suggested that one should first choose the metric and then try to prove continuity with respect to this metric. We take the other point of view, namely that the metric should depend on the problem at hand and, in particular, one should work with the weakest metric which will give the required results. We address such problems in Section 2 and introduce an affine invariant metric based on ellipsoids. As ellipsoids form a Vapnik–Cervonenkis class of subsets of $\mathbb{R}^k$, the metric is sufficiently weak to give an $n^{-1/2}$ rate of convergence of empirical measure to the underlying distribution [Pollard (1984), page 157, 21 Theorem, and page 150, 15 Equicontinuity Lemma].

Not only should a good robust estimator exist at the basic model, it should also exist in a neighborhood of the model and, in particular, it should exist for empirical measures. This problem is dealt with in Section 3 where it is shown that although an estimate may be constructed in the neighborhood of the model it is not in general unique. This indeed would seem to be a weakness of the minimum volume ellipsoid estimator as it means that off the model, although the *set* of possible choices will transform in an equivariant manner, there is no guarantee that a particular choice of the estimator will do so. Furthermore, the different choices of the estimator can give rise to different outliers. The practical consequences of such a lack of uniqueness may not be great, but it is helpful to be aware of the possibility.

In Section 3.2 we consider the form of the minimum volume ellipsoid for empirical measures and are able to characterize it. Such a characterization is of interest as it gives some insight into the problem of calculating the ellipsoid. It corresponds to characterizations of the Hampel–Rousseeuw least median of squares estimator [Hampel (1975) and Rousseeuw (1984)] given by Steele and Steiger (1986). It turns out that the description of the minimum volume ellipsoid shows that the calculation of the minimum volume ellipsoid is of polynomial complexity in n, the sample size, but is such that a direct calculation is probably only feasible for small data sets with $k = 2$ and $n \le 25$. For practical purposes it may not be necessary to calculate the minimum volume ellipsoid as any ellipsoid whose volume is considerably smaller than that of the ellipsoid based on the sample mean and covariance matrix will give useful information. We discuss this point in more detail below.

In Section 4 we consider the breakdown and continuity properties of the minimum volume ellipsoid estimator in terms of the metric defined in Section 2.

The estimator $(\hat{\mu}_n, \hat{\Sigma}_n)$ of location and dispersion based on $\hat{E}_n$ may be defined as the solution to the following problem. Determine $\mu \in \mathbb{R}^k$ and a symmetric positive-definite $k \times k$ matrix Σ so as to minimize $\det(\Sigma) = |\Sigma|$ subject to

$$(1) \qquad \int \left\{ (x-\mu)^T \Sigma^{-1} (x-\mu) \le 1 \right\} d\hat{P}_n(x) \ge \frac{[n/2]+1}{n},$$

where now and in the future { } will denote both a set and the indicator function of the set. If we replace the indicator function in (1) by a function ρ which is symmetric, nonincreasing on $\mathbb{R}_+$, continuous at 0 with $\rho(0) = 1$ and has bounded support, then we obtain a whole class of estimators, the so-called S-estimators which were first introduced by Rousseeuw and Yohai (1984) in the context of linear regression. The minimum volume ellipsoid estimator is thus an S-estimator. The properties of such estimators were investigated by Davies (1987) and in particular it was shown that for sufficiently smooth ρ-functions the resulting estimators are asymptotically normally distributed. The minimum volume ellipsoid estimator is not covered by this result and indeed there are reasons to suspect that it will have a different behaviour. If one specializes to the case $k = 1$, the minimum volume estimator reduces to the middle and the length of the shortest half-sample. Grübel (1988) showed that the length of the interval has an $n^{-1/2}$ rate of convergence and tends weakly to a Gaussian random variable. The middle of the shortest half-sample, however, has an $n^{-1/3}$ rate of convergence and the limiting distribution is nonnormal [Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972), Rousseeuw (1984), Shorack and Wellner (1986), pages 767–771, Kim and Pollard (1990) and Davies (1990)]. Kim and Pollard (1990) contains a general theory of cube root asymptotics and in Section 5 we apply this to obtain the asymptotic behavior of the minimum volume ellipsoid.

For practical applications of the minimum volume ellipsoid, we refer to Rousseeuw and van Zomeren (1990).

1.2. *Notation.* We shall employ the following notation. The set of symmetric strictly positive definite $k \times k$ matrices will be denoted by PDS(k) and the $k \times k$ orthogonal matrices by $\mathfrak{O}_k$. Elements of PDS(k) will be denoted by Σ and Δ and a diagonal matrix by Λ with diagonal elements $\Lambda_{ii} = \lambda_i$, $1 \le i \le k$, which in turn will be represented by λ in $\mathbb{R}^k$. I_k will denote the identity matrix. Determinants will be denoted by $|\ |$. The Borel sets of $\mathbb{R}^k$ will be denoted by $\mathfrak{B}(\mathbb{R}^k)$ and the set of all probability measures on $\mathfrak{B}(\mathbb{R}^k)$ by $\mathfrak{W}(\mathbb{R}^k)$. Lebesgue measure in $\mathbb{R}^k$ will be denoted by m_k and the Dirac measure at the point x by δ_x. If $Q \in \mathfrak{W}(\mathbb{R}^k)$, then $\hat{Q}_n$ will denote the empirical measure defined by n independently and identically distributed random variables with common distribution Q. The set of all nondegenerate ellipsoids in $\mathbb{R}^k$ will be denoted by $\mathfrak{E}$ and the ball with center x and radius r by $B_r(x)$. The Euclidean norm of any element x of $\mathbb{R}^k$ will be denoted by $\|x\|$.

2. Affine invariant metrics. As mentioned in Section 1.1 we consider metrics on $\mathfrak{W}(\mathbb{R}^k)$ which are invariant under all nonsingular affine transformations of $\mathbb{R}^k$ into $\mathbb{R}^k$. The total variation metric $d_{\mathfrak{T}\mathfrak{V}}$ is affine invariant but is much too strong, implying as it does a distance of 1 between any empirical distribution and a continuous theoretical distribution. We shall work with the weak affine invariant metrics $d_{\mathfrak{V}\mathfrak{E}}$ and $d_{\mathfrak{E}}$ defined as follows. We set

$$d_{\mathfrak{V}\mathfrak{E}}(Q_1, Q_2) = \sup_{E \in \mathfrak{E}} |Q_1(E) - Q_2(E)|.$$

This metric is a form of the total variation metric and does not reflect the idea of measurement error. It could be weakened by defining a corresponding Prohorov metric restricted to sets in $\mathfrak{E}$ but this would not be affine invariant as the definition of error is based on the Euclidean metric in $\mathbb{R}^k$. If, however, we replace the absolute error in the definition of the Prohorov metric by a proportional error, it is possible to retain both affine invariance and the formalization of measurement error. For $\eta > 0$ and the ellipsoid $E = \{x: (x-\mu)^T\Sigma^{-1}(x-\mu) \le c\}$, we define E^η by

$$E^\eta = \{x: (x-\mu)^T\Sigma^{-1}(x-\mu) \le c\exp(\eta)\}.$$

The metric $d_{\mathfrak{E}}$ is now defined by

$$d_{\mathfrak{E}}(Q_1, Q_2) = \inf\{\eta > 0: Q_1(E) \le Q_2(E^\eta) + \eta,$$
$$Q_2(E) \le Q_1(E^\eta) + \eta \text{ for all } E \in \mathfrak{E}\}.$$

We note that $d_{\mathfrak{E}}(Q_1, Q_2) \le d_{\mathfrak{V}\mathfrak{E}}(Q_1, Q_2)$ and we have the following result.

THEOREM 1. *The metrics $d_{\mathfrak{E}}$ and $d_{\mathfrak{B}\mathfrak{E}}$ defined above are affine invariant metrics on $\mathfrak{W}(\mathbb{R}^k)$.*

PROOF. This follows from the fact that the class of ellipsoids is invariant under affine transformations. □

One popular model in the theory of robust statistics is the gross error model. Given a measure Q in $\mathfrak{W}(\mathbb{R}^k)$, the mixture $Q_\varepsilon = (1-\varepsilon)Q + \varepsilon W$ with $0 < \varepsilon < 1$ and W in $\mathfrak{W}(\mathbb{R}^p)$ represents a $100\varepsilon\%$ contamination of Q. The next theorem establishes a connection between ε and the distance between Q_ε and Q.

THEOREM 2.

$$\sup_{W \in \mathfrak{W}(\mathbb{R}^k)} d_{\mathfrak{E}}(Q, Q_\varepsilon) = \varepsilon.$$

PROOF. The inequality $d_{\mathfrak{E}}(Q, Q_\varepsilon) \le \varepsilon$ is straightforward. In the other direction, it suffices to choose $W = \delta_x$, where x is some point with $Q(\{x\}) = 0$ and then to consider the ellipsoids $B_{1/n}(x)$ as $n \to \infty$. $Q(B_{1/n}(x))$ tends to 0 as $Q(\{x\}) = 0$ whilst $Q_\varepsilon(B_{1/n}(x))$ tends to ε. □

Choosing weak metrics also has another advantage which will not be directly exploited here. As $d_{\mathfrak{E}} \le d_{\mathfrak{B}\mathfrak{E}}$ and the class of ellipsoids is a Vapnik–Cervonenkis class, it follows that $d_{\mathfrak{E}}(\hat{Q}_n, Q) = O_p(n^{-1/2})$. Suppose now that a functional T is Fréchet differentiable at Q with respect to $d_{\mathfrak{E}}$. Then

$$T(\hat{Q}_n) - T(Q) = \int I(x: Q, T)\, d(\hat{Q}_n - Q) + o_p(n^{-1/2}),$$

which immediately implies a central limit theorem for $T(\hat{Q}_n)$ [Huber (1981), pages 34–40]. As the empirical distributions in general converge at a rate slower than $n^{-1/2}$ in the Prohorov metric [Kersting (1978)], Fréchet differentiability with respect to the Prohorov metric will not in general imply a central limit theorem.

3. Existence of the minimum volume ellipsoid estimator.

3.1. *General measures.* We consider a probability measure P on $\mathfrak{B}(\mathbb{R}^k)$ with the following properties. P has a density function f_P with respect to Lebesgue measure of the form $f_P(x) = f(\|x\|^2)$, where f: $\mathbb{R}_+ \to \mathbb{R}_+$ is nonincreasing and satisfies

$$1 = \int f(\|x\|^2)\, dx = kV_k \int_0^\infty f(r^2) r^{k-1}\, dr,$$

where $V_k = 2\pi^{k/2}/(k\Gamma(k/2))$ [see Stromberg (1981), pages 394 and 395].

Let ε, $0 < \varepsilon < 1$, be fixed and suppose that $\tau = \tau(\varepsilon)$ is such that $\int\{\|x\| \le \tau\} f(\|x\|^2)\,dx = 1 - \varepsilon$ and

$$f(r) > f(\tau^2) > f(r') \quad \text{for all } r \text{ and } r' \text{ with } r < \tau^2 < r'. \tag{2}$$

Let U be an orthogonal $k \times k$ matrix, μ a point in $\mathbb{R}^k$ and

$$\lambda^T = (\lambda_1, \dots, \lambda_k) \in (-1, \infty)^k.$$

For a fixed probability measure Q in $\mathfrak{W}(\mathbb{R}^k)$, we denote by $\mathfrak{P}(Q, \varepsilon)$ the set of solutions (μ, λ, U) of the problem of choosing μ, λ and U so as to minimize $\prod_1^k(1 + \lambda_i)$ subject to

$$\int\left\{\left\|(I_k + \Lambda)^{-1}U(x - \mu)\right\| \le \tau\right\} dQ(x) \ge 1 - \varepsilon.$$

If $\mathfrak{P}(Q, \varepsilon) \ne \varnothing$ we define

$$\mathfrak{L}(Q, \varepsilon) = \left\{\left(\mu, U^T(I_k + \Lambda)^{-2}U\right): (\mu, \lambda, U) \in \mathfrak{P}(Q, \varepsilon)\right\}$$

and set $\mathfrak{L}(Q, \varepsilon) = \varnothing$ otherwise.

If P is as above with f nonincreasing, (10) of Davies (1987) is satisfied. Because of (2) the function f has a point of strict decrease at τ^2. The function $\kappa(u) = \{|u| \le \tau^2\}$, however, does not have a point of strict decrease at τ^2 so that (11) of Davies (1987) is not satisfied. Use of this is made only in Lemma 4.4 of Davies (1987), where the following weaker definition of a point of decrease

$$\xi(u) \ge \xi(d) \ge \xi(v) \text{ and } \xi(u) > \xi(v) \text{ for all } u \text{ and } v \text{ satisfying } u < d < v$$

is sufficient. This weakening of (11) of Davies (1987) is also corrected in Davies (1990), Lemma 1, where a proof is given for the one-dimensional version. This weaker condition is satisfied because of (2) and given this it follows from Theorem 1 of Davies (1987) that $\mathfrak{L}(P, \varepsilon) = \{(0, I_k)\}$. We now show that for measures Q not too far from P in the metric $d_{\mathfrak{E}}$ the problem $\mathfrak{P}(Q, \varepsilon)$ has solutions which lie in a compact set of $\mathbb{R}^k \times \mathrm{PDS}(k)$. On denoting the open ball $\{Q: d_{\mathfrak{E}}(P, Q) < \eta\}$ by $b_{\mathfrak{E}}(P, \eta)$, we have the following result.

THEOREM 3. *If $\eta_0 < \min(\varepsilon, 1 - \varepsilon)$, then $\mathfrak{L}(Q, \varepsilon) \ne \varnothing$ for all Q in $b_{\mathfrak{E}}(P, \eta_0)$. Furthermore, each $\mathfrak{L}(Q, \varepsilon)$ is compact and there exists a compact subset $K(\eta_0)$ of $\mathbb{R}^k \times PDS(k)$ such that $\mathfrak{L}(Q, \varepsilon) \subset K$ for all Q in $b_{\mathfrak{E}}(P, \eta_0)$.*

PROOF. We defer the proof to the Appendix. □

It is easy to show that in any neighbourhood of P there exist distributions Q such that $\mathfrak{L}(Q, \varepsilon)$ contains more than one element and for which no affine equivariant choice can be made. This is one of the main weaknesses of the minimum volume ellipsoid estimator but its importance should not be exaggerated. For Q close to P all choices will be close together and for empirical measures the minimum volume ellipsoid will, in general, be unique. A unique choice based on some ordering, for example, the lexographical ordering can be made and for such a choice we define the minimum volume ellipsoid estimator T_{MVE} at Q to be this element of $\mathfrak{L}(Q, \varepsilon)$. With this convention T_{MVE}: $b_{\mathfrak{E}}(P, \eta_0) \to \mathbb{R}^k \times \mathrm{PDS}(k)$ is well defined.

3.2. *Empirical measures.* Let $(X_j)_1^\infty$ be i.i.d. random variables defined on a probability space $(\Omega, \mathscr{F}, W)$ and with a common distribution P as in Section 3.1. We now consider the minimum volume ellipsoid estimator evaluated at $\hat{P}_n$ and show that if $n \geq k + 1$, then $\mathfrak{L}(\hat{P}_n, \varepsilon)$ contains exactly one point with probability 1. We require the following lemma.

LEMMA 1. *Let $(x_j)_1^n$ with $n \geq k + 1$ be points in $\mathbb{R}^k$ with the property that no more than k such points lie on any hyperplane of dimension less than k. Then the following hold:*

(a) *There exists a unique ellipsoid $E(\mu_M, \Sigma_M) = \{x: (x - \mu_M)^T \Sigma_M^{-1}(x - \mu_M) \leq 1\}$ which contains the $(x_j)_1^n$ and such that $|\Sigma_M| < |\Sigma|$ for any other ellipsoid $E(\mu, \Sigma)$ which contains the $(x_j)_1^n$.*
(b) *At least $k + 1$ of the $(x_j)_1^n$ lie on the surface of $E(\mu_M, \Sigma_M)$ and $E(\mu_M, \Sigma_M)$ is the minimum volume ellipsoid for the points on its surface.*

PROOF. The statement of the lemma together with indications of its proof may be found in Silverman and Titterington (1980), Titterington (1975) and Sibson (1972). □

THEOREM 4. *Suppose $n \geq k + 1$ and $(X_j)_1^n$ are i.i.d. random variables with common distribution P. Then the following hold:*

(a) *$\mathfrak{L}(P_n, \varepsilon)$ contains exactly one element $(\hat{\mu}_n, \hat{\Sigma}_n)$ with probability* 1.
(b) *$\{x: (x - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1}(x - \hat{\mu}_n) \leq \tau^2\}$ is the minimal covering ellipsoid for the data points $(X_j^*)_1^N$ which lie on its surface.*
(c) *$k + 1 \leq N \leq k(k + 3)/2$ with probability* 1.

PROOF. An ellipsoid in $\mathbb{R}^k$ is uniquely determined by $k(k + 3)/2$ points on its surface. Using this and the fact that P has a Lebesgue density, it follows that with probability 1 no more than $k(k + 3)/2$ data points lie on the surface of any ellipsoid. Parts (b) and (c) of the theorem follow now from Lemma 1 on noting that the conditions of that lemma are satisfied with probability 1. Uniqueness will follow if we can show that minimum covering ellipsoids with different surface points have different volumes. As the X_j are assumed to have a density, most readers will accept this without any qualms and we therefore defer the proof of the uniqueness part of the theorem to the Appendix.

4. Breakdown and continuity.

4.1. *Breakdown.* We define the breakdown point $\varepsilon_{\mathrm{MVE}}^*(P, d_{\mathfrak{E}})$ of the minimum volume estimator T_{MVE} at P as the infimum over all $\eta > 0$ with the following property. Either there exists a $Q \in b_{\mathfrak{G}}(P, \eta)$ with $\mathfrak{P}(Q, \varepsilon) = \varnothing$ or

$$\sup_{Q \in b_{\mathfrak{G}}(P,\eta)} \left(\sup\left\{ \|\mu\| + \sum_1^k \left(\lambda_j + \lambda_j^{-1} \right) : (\mu, \Lambda, U) \in \mathfrak{P}(Q, \eta) \right\} \right) = \infty.$$

The estimator T_{MVE} is therefore considered to have broken down if there are solutions with an arbitrarily large mean vector or if the eigenvalues of the dispersion matrix become arbitrarily small or arbitrarily large.

THEOREM 5.

$$\varepsilon^*_{\mathrm{MVE}}(P, d_{\mathfrak{G}}) = \min(\varepsilon, 1 - \varepsilon).$$

PROOF. Theorem 3 gives $\varepsilon^*_{\mathrm{MVE}}(P, d_{\mathfrak{G}}) \geq \min(\varepsilon, 1-\varepsilon)$. In the opposite direction we put $Q_\eta = (1-\eta)P + \eta\delta_x$. If $\eta > \varepsilon$, then $x \in E$ for any ellipsoid E with $Q_\eta(E) \geq 1 - \varepsilon$ and breakdown is achieved by letting x tend to ∞. This implies $\varepsilon^*_{\mathrm{MVE}}(P, d_{\mathfrak{G}}) \leq \varepsilon$. If now $\eta > (1-\varepsilon)$, then $\mathfrak{P}(Q_\eta, \varepsilon)$ has the degenerate solution $E = \{x\}$ and again T_{MVE} has broken down implying $\varepsilon^*_{\mathrm{MVE}}(P, d_{\mathfrak{G}}) \leq 1 - \varepsilon$. □

4.2. *Continuity.* As mentioned in Section 1.1 a good robust estimator should not only have a high breakdown point but should also be continuous. We prove below, Theorems 7 and 8, that the $T_{\mathrm{MVE}}(\hat{\mathbb{P}}_n)$ converges at a rate of $n^{-1/3}$ to $T_{\mathrm{MVE}}(\mathbb{P})$. If T_{MVE} were locally Lipschitz at $\mathbb{P}$, we would have

$$\left| T_{\mathrm{MVE}}(\hat{\mathbb{P}}_n) - T_{\mathrm{MVE}}(\mathbb{P}) \right| \leq c d_{\mathfrak{G}}(\hat{\mathbb{P}}_n, \mathbb{P})$$

for some constant c depending on $\mathbb{P}$. As $d_{\mathfrak{G}}(\hat{\mathbb{P}}_n, \mathbb{P}) = O_p(n^{-1/2})$ it would follow that $T_{\mathrm{MVE}}(\hat{\mathbb{P}}_n)$ converges at a rate of $n^{-1/2}$ to $T_{\mathrm{MVE}}(\mathbb{P})$. This argument shows that T_{MVE} cannot be Lipschitz continuous at $\mathbb{P}$. A simple extension of the argument, using the results of Theorems 7 and 8, shows that T_{MVE} can at best satisfy a local Hölder condition of order $2/3$ at $\mathbb{P}$. We now show that in fact it satisfies a Hölder condition of order $1/2$ and that $1/2$ is the correct order.

LEMMA 2.

$$\begin{aligned}
&\int \left\{ \left\| (I_k + \Lambda)^{-1} U(x - \mu) \right\| \leq \tau \right\} dP \\
&\quad = \int \{\|x\| \leq \tau\}\, dP + c_0 \left(\sum_1^k \lambda_j - \frac{1}{2}\|\lambda\|^2 \right) \\
&\qquad + \left(c_1 \|\lambda\|^2 + c_2 \|\mu\|^2 \right)(1 + o(1)) + O\left(\left(\sum_1^k \lambda_j \right)^2 \right),
\end{aligned}$$

where the constants c_0 to c_2 are as follows:

$$\begin{aligned}
c_0 &= V_k \tau^k f(\tau^2) > 0, \\
c_1 &= V_k \tau^{k+2} f^{(1)}(\tau^2)/(k+2) < 0, \\
c_2 &= V_k \tau^k f^{(1)}(\tau^2) < 0, \\
V_k &= 2\pi^{k/2}/(k\Gamma(k/2)).
\end{aligned}$$

PROOF. The claims of the lemma follow from a standard Taylor expansion. □

THEOREM 6. *Let* T_{MVE} *be as above and suppose that* f *is continuously differentiable at* τ^2 *with* $f^{(1)}(\tau^2) < 0$. *Then* T_{MVE} *satisfies an exact Hölder condition of order* $1/2$ *locally at* P.

PROOF. We write $T_{\mathrm{MVE}}(Q) = (\mu, U^T(I_k + \Lambda)^{-2}U)$. If $\eta > d_{\mathfrak{C}}(P,Q)$ we have

$$1 - \varepsilon + \eta = \int \{\|x\| \le \tau\gamma(\eta)\}\, dP \le \int \{\|x\| \le \tau\gamma(\eta)e^{\eta}\}\, dQ + \eta$$

with $\gamma(\eta) = 1 + O(\eta)$ as P has a bounded density function. This implies

$$\prod_1^k (1 + \lambda_j) \le 1 + O(d_{\mathfrak{C}}(P,Q)). \tag{3}$$

We have

$$\begin{aligned} 1 - \varepsilon &\le \int \Big\{\big\|(I_k + \Lambda)^{-1}U(x - \mu)\big\| \le \tau\Big\}\, dQ \\ &\le \int \Big\{\big\|(I_k + \Lambda)^{-1}U(x - \mu)\big\| \le \tau e^{\eta}\Big\}\, dP + \eta. \end{aligned} \tag{4}$$

Let $(Q_n)_1^\infty$ be a sequence such that $\lim_{n\to\infty} Q_n = P$ and choose a subsequence, which we continue to denote by $(Q_n)_1^\infty$, such that $\lim_{n\to\infty}(\mu_n, \Lambda_n, U_n) = (\mu^*, \Lambda^*, U^*)$. Then

$$1 - \varepsilon \le \int \Big\{\big\|(I_k + \Lambda^*)^{-1}U^*(x - \mu)\big\| \le \tau\Big\}\, dP,$$

giving $\prod_1^k(1 + \lambda_i^*) \ge 1$. This together with (3) implies $\prod_1^k(1 + \lambda_i^*) = 1$ and the uniqueness of $T_{\mathrm{MVE}}(P)$ implies $(\mu^*, \Lambda^*) = (0, 0)$. As this holds for every subsequence $\lim_{n\to\infty} T_{\mathrm{MVE}}(Q_n) = T_{\mathrm{MVE}}(P)$, showing that T_{MVE} is continuous at P.

From (3) and $\lambda_j = o(1)$, $1 \le j \le k$, as $d_{\mathfrak{C}}(P,Q) \to 0$, we may deduce

$$\sum_1^k \lambda_j - \frac{1}{2}\|\lambda\|^2 \le O(d_{\mathfrak{C}}(P,Q)) + o(\|\lambda\|^2). \tag{5}$$

In the other direction we note that (4) implies

$$1 - \varepsilon \le \int \Big\{\big\|(I_k + \Lambda)^{-1}U(x - \mu)\big\| \le \tau\Big\}\, dP + O(\eta).$$

From Lemma 2 we deduce

$$\begin{aligned} &\int \Big\{\big\|(I_k + \Lambda)^{-1}U(x - \mu)\big\| \le \tau\Big\}\, dP \\ &\quad = \int \{\|x\| \le \tau\}\, dP + c_0\Big(\sum_1^k \lambda_j - \frac{1}{2}\|\lambda\|^2\Big)(1 + o(1)) \\ &\qquad - \big(c_1\|\lambda\|^2 + c_2\|\mu\|^2\big)(1 + o(1)). \end{aligned}$$

From this we conclude

$$-\left(c_1\|\lambda\|^2 + c_2\|\mu\|^2\right)(1 + o(1)) \le O\left(d_{\mathfrak{E}}(P,Q)\right) + c_0\left(\sum_1^k \lambda_j - \frac{1}{2}\|\lambda\|^2\right)(1 + o(1)).$$

As c_1 and c_2 are negative and c_0 positive, this together with (5) yields $\|\lambda\|^2 + \|\mu\|^2 \le 0(d_{\mathfrak{E}}(P,Q))$ and hence

$$\|T_{\mathrm{MVE}}(Q) - T_{\mathrm{MVE}}(P)\| = O\left(d_{\mathfrak{E}}(P,Q)^{1/2}\right).$$

To finish the proof of the theorem, we must show that for all η sufficiently small there exists a measure Q_η with $d_{\mathfrak{E}}(P, Q_\eta) \le \eta$ and $\|T_{\mathrm{MVE}}(Q_\eta) - T_{\mathrm{MVE}}(P)\| \ge A\eta^{1/2}$ for some fixed constant $A > 0$. To ease the notation, we consider the one-dimensional case $k = 1$ only. The general case follows from this. We set $Q_\eta = (1 - \eta)P + \eta\delta_{\tau+\alpha}$ for some α. The proof of Theorem 5 shows that for $\alpha \ne 0$, $d_{\mathfrak{E}}(P, Q_\eta) = \eta$. We write $F(u) = \int_{-\infty}^{u} f(v^2)\,dv$ and define $\alpha = \alpha(\eta)$ by $F(\tau + \alpha(\eta)) - F(-\tau + \alpha(\eta)) = (1 - \varepsilon - \eta)/(1 - \eta)$ for $\eta < 1 - \varepsilon$. A Taylor expansion shows that $\alpha(\eta)^2 = A\eta(1 + o(1))$ with $A = -1/(2\tau f^{(1)}(\tau^2)) > 0$. As f is nonincreasing and strictly decreasing at τ^2, it follows that the shortest interval $[a, b]$ with $Q_\eta([a, b]) \ge 1 - \varepsilon$ is given by $[a, b] = [-\tau + \alpha(\eta), \tau + \alpha(\eta)]$. The midpoint of the interval is $\alpha(\eta)$ and hence $\|T_{\mathrm{MVE}}(Q_\eta) - T_{\mathrm{MVE}}(P)\| = A\eta^{1/2}(1 + o(1))$, which completes the proof of the theorem. □

5. Weak convergence. We now consider the empirical process $\hat{P}_n(E)$ indexed by the ellipsoids $E \in \mathfrak{E}$. As set of ellipsoids forms a Vapnik–Cervonenkis class, we have $d_{\mathfrak{E}}(P, \hat{P}_n) \le d_{\mathfrak{B}\mathfrak{E}}(P, \hat{P}_n) = O_p(n^{-1/2})$, which together with Theorem 6 gives $\|T_{\mathrm{MVE}}(\hat{P}_n) - T_{\mathrm{MVE}}(P)\| = O_p(n^{-1/4})$. We now show that this result can be improved to $\|T_{\mathrm{MVE}}(\hat{P}_n) - T_{\mathrm{MVE}}(P)\| = O_p(n^{-1/3})$.

We parametrize the ellipsoids as follows. We write

$$E(\Lambda, \mu, U, v) = \left\{x\colon \left\|(I_k + \Lambda)^{-1}U(x - \mu)\right\| \le \tau(1 + v)\right\}$$

with $-1 < v < \infty$, $\prod_1^k(1 + \lambda_j) = 1$ and the convention that λ_k is defined in terms of $\lambda_1, \dots, \lambda_{k-1}$. This latter equality implies that for small λ, $\sum_1^k \lambda_j = \|\lambda\|^2/2 + O(\|\lambda\|^3)$. Given an ellipsoid E, the values of Λ and μ are uniquely determined but the orthogonal matrix U is not. The weak convergence results we prove below are for processes indexed by ellipsoids and are to be interpreted in this sense.

We set $\theta = (\Lambda, \mu, U, v)$ and write $\hat{E}_n$ for the ellipsoid which is a solution of the empirical problem, $\hat{E}_n = E(\hat{\theta}_n) = E(\hat{\Lambda}_n, \hat{\mu}_n, \hat{U}_n, \hat{v}_n)$. We note that $(\hat{\Lambda}_n, \hat{\mu}_n, \hat{v}_n) \to 0$ almost surely because of Theorem 6 and, as $\mathfrak{E}$ is a Vapnik–Cervonenkis class,

$$V_n(E) = \sqrt{n}\left(\hat{P}_n(E) - P(E)\right) \Rightarrow V(E),\ E \in \mathfrak{E},$$

where V is a continuous bounded Gaussian process on $\mathfrak{E}$. From Lemma 2 we have

$$\begin{aligned}(6)\quad &\int\left\{\left\|(I_k+\Lambda)^{-1}U(x-\mu)\right\|\le\tau\right\}dP\\ &\quad=P(\{\|x\|\le\tau\})+c_0v+c_2\|\mu\|^2+c_3\|\lambda\|^2+o(\|\theta\|^2)+O(v^2)\end{aligned}$$

with $c_3=c_2-c_0/2<0$.

THEOREM 7. *Let* $\hat{v}_n$ *and* V *be as above. Then* $\sqrt{n}\,\hat{v}_n\Rightarrow -V(B_\tau(0))/c_0$.

PROOF. Arguing as in Kim and Pollard (1990), we obtain

$$\begin{aligned}1-\varepsilon\le\hat{P}_n(\hat{E}_n)&=P(\hat{E}_n)+\frac{1}{\sqrt{n}}V_n(\hat{E}_n)\\ &=P(\hat{E}_n)+\frac{1}{\sqrt{n}}V_n(B_\tau(0))+o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

From Lemmas 4 and 8 of Davies (1987), it follows that

$$P(\hat{E}_n)\le P\left(E\left(\hat{\Lambda}_n,0,\hat{U}_n,\hat{v}_n\right)\right)\le P\left(E\left(0,0,\hat{U}_n,\hat{v}_n\right)\right)=1-\varepsilon+c_0\hat{v}_n\left(1+o_p(1)\right)$$

and hence $-c_0\sqrt{n}\,\hat{v}_n(1+o_p(1))\le V_n(B_\tau(0))+o_p(1)$.

In the other direction we note that as $\hat{E}_n$ minimizes the volume of the ellipsoid,

$$\begin{aligned}1-\varepsilon\ge\hat{P}_n(E(0,0,I_k,\hat{v}_n))&=P(E(0,0,I_k,\hat{v}_n))+\frac{1}{\sqrt{n}}V_n(B_\tau(0))+o_p\left(\frac{1}{\sqrt{n}}\right)\\ &=1-\varepsilon+c_0\hat{v}_n\left(1+o_p(1)\right)+\frac{1}{\sqrt{n}}V_n(B_\tau(0))+o_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

and hence $-c_0\sqrt{n}\,\hat{v}_n(1+o_p(1))\ge V_n(B_\tau(0))+o_p(1)$. The two inequalities for $\hat{v}_n$ now imply the theorem. □

For $k=1$ this is contained in a result of Grübel (1988) and, in the context of linear regression, Davies (1990).

The asymptotic behavior of the ellipsoid $E(\hat{\Lambda}_n,\hat{\mu}_n,\hat{U}_n,\hat{v}_n)$ is covered by the next theorem. In order to state it we write $E(\hat{\Lambda}_n,\hat{\mu}_n,\hat{U}_n,\hat{v}_n)=E(\hat{\Xi}_n,\hat{\mu}_n,\hat{v}_n)$, where

$$\hat{\Xi}_n=\hat{U}_n^T\hat{\Lambda}_n\hat{U}_n$$

and

$$E\left(\hat{\Xi}_n,\hat{\mu}_n,\hat{v}_n\right)=\left\{x:\left\|\left(I_k+\hat{\Xi}_n\right)^{-1}(x-\mu)\right\|\le\tau(1+v)\right\}.$$

We write

$$\mathcal{X} = \left\{ \Xi : \Xi \text{ is a symmetric } k \times k \text{ matrix} \right.$$

$$\left. \text{with eigenvalues } \Lambda = (\lambda_j)_1^k \text{ satisfying } \sum_1^k \lambda_j = 0 \right\}.$$

Any symmetric matrix Ξ with eigenvalues Λ may be written in the form

$$\Xi = U^T \Lambda U, \tag{7}$$

where U is an orthogonal matrix.

THEOREM 8. *Let $(X_i)_1^\infty$ be independently and identically distributed random variables with common distribution P as given above. Then $(n^{1/3}\hat{\Xi}_n, n^{1/3}\hat{\mu}_n) \Rightarrow (\Xi^*, \mu^*)$, where (Ξ^*, μ^*) maximizes $Z(\Xi, \mu) + c_2\|\mu\|^2 + c_3\|\lambda\|^2$ and Z is a nondegenerate zero-mean continuous Gaussian process defined on $\{(\Xi, \mu) \colon \Xi \in \mathcal{X},\ \mu \in \mathbb{R}^k\}$ with covariance structure given by*

$$E\big(Z(\Xi, \mu) Z(\hat{\Xi}, \hat{\mu})\big)$$
$$= f(\tau^2) \lim_{\alpha \downarrow 0} \frac{1}{\alpha} m_p\Big(E(\alpha\Xi, \alpha\mu, 0) E\big(\alpha\hat{\Xi}, \alpha\hat{\mu}, 0\big) - E(\alpha\Xi, \alpha\mu, 0) B_\tau(0)$$
$$- E\big(\alpha\hat{\Xi}, \alpha\hat{\mu}, 0\big) B_\tau(0) + B_\tau(0)\Big).$$

Furthermore, the process $Z(\Xi, \mu)$ is independent of $V(B_\tau(0))$ of Theorem 7.

PROOF. The proof is based on the results and arguments of Kim and Pollard (1990). We write $\theta = (\Xi, \mu, v)$ and define a norm by $\|\theta\| = \|\Xi\| + \|\mu\| + |v|$. We set

$$F_R = \sup\{|E(\Xi, \mu, v) - E(0, 0, v)| \colon \|\Xi\| + \|\mu\| + |v| \le R\}$$

and consider the process $\hat{P}_n(E(\Xi, \mu, v) - E(0, 0, v))$. The conditions of Lemma 4.1 of Kim and Pollard (1990) are satisfied and we may deduce that for any $\eta > 0$,

$$\hat{P}_n(E(\Xi, \mu, v)) - E(0, 0, v)) \le P(E(\Xi, \mu, v) - E(0, 0, v))$$
$$+ \eta\big(\|\lambda\|^2 + \|\mu\|^2 + v^2\big) + O_p(n^{-2/3}).$$

As $\hat{P}_n(E(\hat{\Xi}_n, \hat{\mu}_n, \hat{v}_n)) \ge \hat{P}_n(E(0, 0, \hat{v}_n))$ and $\hat{v}_n = O_p(n^{-1/2})$, it follows that $\|\hat{\lambda}_n\| + \|\hat{\mu}_n\| = O_p(n^{-1/3})$ where we have used (6). We set $D(\Xi, \mu, v) = E(\Xi, \mu, v) - E(0, 0, v)$ and consider the rescaled process

$$Z_n(\Xi, \mu, v) = n^{2/3}\big(P_n\big(D(n^{-1/3}\Xi, n^{-1/3}\mu, n^{-1/2}v)\big)$$
$$- P\big(D(n^{-1/3}\Xi, n^{-1/3}\mu, n^{-1/2}v)\big)\big).$$

As the ellipsoids form a Vapnik–Cervonenkis class, one may check along the lines of Lemma 4.6 of Kim and Pollard (1990) that the process $Z_n(\Xi, \mu, v)$

satisfies the stochastic equicontinuity condition for weak convergence. It remains to check that the finite-dimensional distributions converge. A short calculation shows that it is sufficient to prove the existence of

$$\lim_{\alpha \downarrow 0} \frac{1}{\alpha} P\Big(\Big| E(\alpha\Xi, \alpha\mu, 0) - E\big(\alpha\tilde{\Xi}, \alpha\tilde{\mu}, 0\big)\Big|\Big). \tag{8}$$

Direct calculations give

$$\begin{aligned}
&P\Big(E(\alpha\Xi, \alpha\mu, 0) \setminus E\big(\alpha\tilde{\Xi}, \alpha\tilde{\mu}, 0\big)\Big) \\
&\quad = \int f(\|x\|^2)\Big\{\big\|(I_k + \alpha\Xi)^{-1}(x - \alpha\mu)\big\| \le \tau\Big\} \\
&\qquad \times \Big\{\Big\|\big(I_k + \alpha\tilde{\Xi}\big)^{-1}(x - \alpha\tilde{\mu})\Big\| > \tau\Big\}\, dx \\
&\quad = \big(f(\tau^2) + o(1)\big) m_k\big(\{\|x - \alpha\gamma\| \le \tau + O(\alpha^2)\} \\
&\qquad\qquad \times \{\|(I_k + \alpha\Delta)x\| > \tau + O(\alpha^2)\}\big),
\end{aligned}$$

where $\gamma = U(\mu - \tilde{\mu})$ and $\Delta = \tilde{U}U^{\mathrm{T}}\Lambda - \tilde{\Lambda}\tilde{U}U^{\mathrm{T}}$, m_k is Lebesgue measure in $\mathbb{R}^k$ and we have used the representation (7). The set $\{\|x - \alpha\gamma\| \le \tau + O(\alpha^2)\}$ $\{\|(I_k + \alpha\Delta)x\| > \tau + O(\alpha^2)\}$ may be written as the sum of sets of the form

$$\begin{aligned}
&\Big\{x\colon \big(c^2 - \|x^{k-1}\|^2\big)^{1/2} + \alpha H_1\big(x^{k-1}, \gamma, \Delta\big) + O(\alpha^2) \le x_k \\
&\qquad \le \big(c^2 - \|x^{k-1}\|^2\big)^{1/2} + \alpha H_2\big(x^{k-1}, \gamma, \Delta\big) + O(\alpha^2)\Big\}
\end{aligned}$$

for some functions H_1 and H_2, where $x^{k-1} = (x_1, \dots, x_{k-1})$. The Lebesgue measure of such a set is given by $\alpha\int_{\|x^{k-1}\| \le \tau}(H_2 - H_1)\, dm_{k-1} + O(\alpha^2)$ and this implies the existence of the limit in (8). Furthermore, a direct calculation shows that $E((Z(\Xi, \mu) - Z(\tilde{\Xi}, \tilde{\mu}))^2) = 0$ if and only if $\Xi = \tilde{\Xi}$ and $\mu = \tilde{\mu}$. Combining all this, we obtain

$$\begin{aligned}
\hat{P}_n\big(E(n^{-1/3}, n^{-1/3}\mu, n^{-1/2}v)\big) &= n^{-2/3}\big(c_2\|\mu\|^2 + c_3\|\lambda\|^2 + Z_n(\Xi, \mu)\big) \\
&\quad + \big(\hat{P}_n(E(0,0,0)) - P(E(0,0,0))\big) \\
&\quad + 1 - \varepsilon + c_0 v/\sqrt{n} + o_p(n^{-2/3}),
\end{aligned}$$

where $Z_n \Rightarrow Z$ and $\sqrt{n}\,(\hat{P}_n(E(0,0,0)) - P(E(0,0,0))) \Rightarrow V(B_\tau(0))$. Again following the arguments in Kim and Pollard (1990) on the uniqueness of argmax and the weak convergence to argmax, we obtain the first part of the theorem.

To prove the last claim, we note that

$$\begin{aligned}
&E\big(V(B_\tau(0))Z(\Xi, \mu)\big) \\
&\quad = \lim_{n \to \infty} n^{1/6}\big(P\big(B_\tau(0) \cap E(n^{-1/3}\Xi, n^{-1/3}\mu, 0)\big)\big) \\
&\qquad - P(B_\tau(0)) - P(B_\tau(0))P\big(E(n^{-1/3}\Xi, n^{-1/3}\mu, 0) + P(B_\tau(0))^2\big).
\end{aligned}$$

From Lemma 2 we obtain $P(E(n^{-1/3}\Xi, n^{-1/3}\mu, 0)) = P(B_\tau(0)) + O(n^{-1/3})$ and hence it is sufficient to prove that

$$\lim_{n\to\infty} n^{1/6}\left|P\big(B_\tau(0) \cap E(n^{-1/3}\Xi, n^{-1/3}\mu, 0)\big) - P\big(B_\tau(0)\big)\right| = 0.$$

This, however, follows from $E(n^{-1/3}\Xi, n^{-1/3}\mu, 0) \supset \{x\colon \|x\| \le \tau - O(n^{-1/3})\}$ and hence

$$P\big(B_\tau(0) \cap E(n^{-1/3}\Xi, n^{-1/3}\mu, 0)\big) \ge P\big(B_\tau(0)\big) - O\big(n^{-1/3}\big),$$

which proves the theorem. □

APPENDIX

Proof of Theorem 3. As $1 - \varepsilon + \eta_0 < 1$ it follows from the properties of P that for some τ', $0 < \tau' < \infty$,

$$1 - \varepsilon + \eta_0 \le \int\{\|x\| \le \tau'\}\, dP \le \int\{\|x\| \le \tau' e^{\eta_0}\}\, dQ + \eta_0$$

and hence $\int\{\|x\| \le \tau' e^{\eta_0}\}\, dQ \ge 1 - \varepsilon$. If we set

$$\mathfrak{C}(Q,\varepsilon) = \left\{(\mu,\lambda,U)\colon \int\left\{\left\|(I_k + \Lambda)^{-1}U(x-\mu)\right\| \le \tau\right\} dQ(x) \ge 1 - \varepsilon\right\},$$

we have $\mathfrak{C}(Q,\varepsilon) \ne \varnothing$ and that we may therefore restrict the search for elements of $\mathfrak{P}(Q,\varepsilon)$ to the set $\mathfrak{C}^*(Q,\varepsilon) = \{(\mu,\Lambda,U)\colon (\mu,\Lambda,U) \in \mathfrak{C}(Q,\varepsilon)\prod_1^k(1+\lambda_j) \le \gamma\}$ for some $\gamma > 0$. If, for some μ, Λ and U, $1 - \varepsilon \le \int\{\|(I_k + \Lambda)^{-1}U(x-\mu)\| \le \tau\}\, dQ(x)$, the definition of $d_{\mathfrak{C}}$ gives

$$\begin{aligned} 1 - \varepsilon &\le \int\left\{\left\|(I_k + \Lambda)^{-1}U(x-\mu)\right\| \le \tau e^{\eta_0}\right\} dP + \eta_0 \\ &= \int\left\{\left\|(I_k + \Lambda)^{-1}(x-\tilde{\mu})\right\| \le \tau e^{\eta_0}\right\} dP + \eta_0 \end{aligned} \tag{7}$$

with $\tilde{\mu} = U\mu$. If now $\lambda_i \to -1$ for some i or $\|\mu\| \to \infty$ the integral on the right-hand side of (7) converges to 0, giving $1 - \varepsilon \le \eta_0$. Thus if $\eta_0 < 1 - \varepsilon$, $\|\mu\|$ must be bounded above and the λ_i must be bounded away from -1.

The fact that $\mathfrak{L}(Q,\varepsilon)$ is closed follows from a simple argument using the upper semicontinuity of the function $r \to \{|r| \le \tau\}$. □

Proof of Theorem 4(a). To prove uniqueness, it remains to show that for any two different sets $(X_j)_{j\in S_r}$ and $(X_j)_{j\in S'_{r'}}$ of sizes r and r' with $k+1 \le r$, $r' \le k(k+3)/2$ the minimal covering ellipsoids which pass through these points have different volumes with probability 1. Without loss of generality we may suppose that $S_r = \{1,\dots,r\}$, $1 \notin S'_{r'}$, $S_r \cup S'_{r'} = \{1,\dots,p\}$ with $p > r$.

For any points $(x_j)_1^p$ with the property that no $k+1$ lie on a hyperplane of dimension less than k, we denote by $(\mu((x_j)_{j\in S_r}), \Sigma((x_j)_{j\in S_r}))$ the parametrization of the minimal covering ellipsoid through the points $(x_j)_{j\in S_r}$.

Using the separability of the space of $k \times k$ matrices, it is not difficult to show that $\mu((x_j)_{j \in S_r})$ and $\Sigma((x_j)_{j \in S_r})$ are Borel measurable functions of $(x_j)_{j \in S_r}$.

As the random variables $(X_j)_1^n$ are independently and identically distributed, a simple conditioning argument will show that $|\Sigma((X_j)_{j \in S_r})| \neq |\Sigma((X_j)_{j \in S'_{r'}})|$ almost surely if we can prove $P(\{x_1: |\Sigma((x_j)_{j \in S_r})| = |\Sigma((x_j)_{j \in S'_{r'}})|\}) = 0$. As $1 \notin S'_{r'}$ it follows that it is sufficient to prove $P(\{x_1: |\Sigma((x_j)_1^r)| = v\}) = 0$ for all $v \in \mathbb{R}_+$. Suppose then that $P(\{x_1: |\Sigma((x_j)_1^r)| = v\}) > 0$ and let $\bar{x} = (1/r - 1)\Sigma_2^r x_j$. Then $\bar{x}$ lies in the interior of the minimal covering ellipsoid and hence so does the point $x_\lambda = (1 - \alpha)x_1 + \alpha\bar{x}$ for $0 < \alpha < 1$. On writing $V(x_2, \dots, x_r) = \{x_1: |\Sigma((x_j)_1^r)| = v\}$ it follows from the fact that P has a continuous density f that there exists a δ, $0 < \delta < 1$, such that

$$P\big(\{x_1: (1 - \alpha)x_1 + \alpha\bar{x} \in V(x_2, \dots, x_r)\}\big) > 0 \tag{8}$$

for all α, $0 < \alpha < \delta$. As the minimal covering ellipsoid is unique and $(1 - \alpha)x_1 + \alpha\bar{x}$ lies inside the minimal covering ellipsoid for $(1 - \alpha')x_1 + \alpha'\bar{x}, x_2, \dots, x_r$, for $0 < \alpha' < \alpha < \delta$, it follows that the sets $\{x_1: (1 - \alpha)x_1 + \alpha\bar{x} \in V(x_2, \dots, x_r)\}$ are disjoint for different α's, $0 < \alpha < \delta$. This implies that not all such sets can have positive probability which contradicts (8). We may therefore conclude that with probability 1 the minimal covering ellipsoids for different subsets are different, proving the theorem. □

Acknowledgments. I would like to acknowledge the help of two referees and an Associate Editor who found several mistakes in the first version of the paper and whose comments led to a considerable improvement.

Note added in proof. After the final version of the article had been accepted for publication the article Nolan (1991) appeared. Nolan also considers the asymptotics of the minimum volume ellipsoid as described by Theorem 8 of the present paper. Her work is motivated by problems of modality in high dimensions.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton Univ. Press.

DAVIES, L. (1990). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.* **18** 1651–1675.

DAVIES, L. (1992). An efficient Fréchet differentiable high breakdown multivariate location and dispersion estimator. *J. Multivariate Anal.* **40** 311–327.

DAVIES, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292.

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Qualifying paper, Harvard Univ.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann in Honor of His Sixty-fifth Birthday* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, Calif.

GRÜBEL, R. (1988). The length of the shorth. *Ann. Statist.* **16** 619–628.

HAMPEL, F. R. (1975). Beyond location parameters: Robust concepts and methods. *Bull. Inst. Internat. Statist.* **46** 375–382.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

KERSTING, G. D. (1978). Die Geschwindigkeit der Glivenko–Cantelli-Konvergenz gemessen in der Prohorov-Metrik. *Math. Z.* **163** 65–102.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219.

LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248.

MARONNA, R. A. (1976). Robust *M*-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

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

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

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

ROUSSEEUW, P. J. (1986). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications* (W. Grossman, G. Pflug, I. Vincze and W. Wertz, eds.) 283–297. Reidel, Dordrecht.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

ROUSSEEUW, P. J. and YOHAI, V. (1984). Robust regression by means of *S*-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.* **26** 256–272. Springer, New York.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points (with comment). *J. Amer. Statist. Assoc.* **85** 633–651.

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

SIBSON, R. (1972). Discussion of "Results in the theory and construction of D-optimal experimental designs" by H. P. Wynn. *J. Roy. Statist. Soc. Ser. B* **34** 174–175.

SILVERMAN, B. W. and TITTERINGTON, D. M. (1980). Minimum covering ellipses. *SIAM J. Sci. Statist. Comput.* **4** 401–409.

STAHEL, W. A. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. thesis, ETH Zürich, Switzerland.

STEELE, J. M. and STEIGER, W. L. (1986). Algorithms and complexity for least median of squares regression. *Discrete Appl. Math.* **14** 93–100.

STROMBERG, K. R. (1981). *An Introduction to Classical and Real Analysis*. Wadsworth, Belmont, Calif.

TITTERINGTON, D. M. (1975). Optimal design: Some geometrical aspects of *D*-optimality. *Biometrika* **62** 313–320.

FACHBEREICH MATHEMATIK
UNIVERSITÄT GESAMTHOCHSCHULE ESSEN
D-4300 ESSEN 1
GERMANY