The Annals of Statistics
1996, Vol. 24, No. 3, 1327–1345

ROBUSTNESS PROPERTIES OF *S*-ESTIMATORS OF MULTIVARIATE LOCATION AND SHAPE IN HIGH DIMENSION[1]

BY DAVID M. ROCKE

University of California, Davis

For the problem of robust estimation of multivariate location and shape, defining S-estimators using scale transformations of a fixed ρ function regardless of the dimension, as is usually done, leads to a perverse outcome: estimators in high dimension can have a breakdown point approaching 50%, but still fail to reject as outliers points that are large distances from the main mass of points. This leads to a form of nonrobustness that has important practical consequences. In this paper, estimators are defined that improve on known S-estimators in having all of the following properties: (1) maximal breakdown for the given sample size and dimension; (2) ability completely to reject as outliers points that are far from the main mass of points; (3) convergence to good solutions with a modest amount of computation from a nonrobust starting point for large (though not near 50%) contamination. However, to attain maximal breakdown, these estimates, like other known maximal breakdown estimators, require large amounts of computational effort. This greater ability of the new estimators to reject outliers comes at a modest cost in efficiency and gross error sensitivity and at a greater, but finite, cost in local shift sensitivity.

1. Introduction. Maronna (1976), in introducing M-estimators of multivariate location and shape, showed that such estimates with monotone ψ function have a breakdown point of at most $1/(p+1)$, where p is the dimension of the data. Subsequently, most other robust estimators of multivariate location and shape were shown independently by Donoho (1982) and Stahel (1981) to have the same bound. They each proposed a projection-based estimator that has breakdown approaching 1/2 [also see Donoho and Huber (1983); Hampel, Ronchetti, Rousseeuw and Stahel (1986); Huber (1985)]. Since that time, several high-breakdown estimators have been defined, with the greatest attention being paid to Rousseeuw's minimum volume ellipsoid (MVE) estimator, Rousseeuw and Yohai's S-estimators and Huber's M-estimators [Campbell (1980, 1982); Davies (1987); Hampel, Ronchetti, Rousseeuw and Stahel (1986); Huber (1981); Kent and Tyler (1991); Lopuhaä (1989); Lopuhaä and Rousseeuw (1991); Maronna (1976); Rousseeuw (1985); Rousseeuw and Leroy (1987); Rousseeuw and Yohai (1984); Rousseeuw and van Zomeren (1990a, b, 1991); Tyler (1983, 1988, 1991)]. At first, Donoho's and Stahel's results were

Received March 1992; revised March 1993.

[1] Work supported in part by the National Institute of Environmental Health Sciences, NIH Grant P42-ES04699 and NSF Grants DMS-93-01344, DMS-94-06193, and DMS-95-10511.

AMS 1991 *subject classifications*. 62H12, 62F35.

Key words and phrases. Breakdown point, efficiency, gross error sensitivity, local shift sensitivity, outlier rejection.

taken to mean that even redescending M-estimators had breakdown bounded by $1/(p+1)$; however, it has become clear that, with a high-breakdown point start, these too can have high breakdown [Lopuhaä (1989); Tyler (1991)].

The main point of this paper concerns the definition of S-estimators in terms of the function ρ. Although this is often not explicitly stated, it is clear that the function ρ is intended not to vary with the dimension, by analogy with the multivariate normal distribution, in which $\rho(x) = 0.5x^2$ for any dimension. The only accommodation to the dimension is to scale the Mahalanobis distance d, so that one considers $\rho(d/c)$, and c varies with the dimension to maintain the desired breakdown. We show that this implies that the estimator in high dimension loses the ability to reject as outliers points that are very far from the main mass of points, even if they are few. We also show that this phenomenon is not necessary, and provide two new examples of classes of S-estimators with high breakdown and better outlier rejection properties. We give an example in which the application of this idea allows construction of an M-estimator that has more robust behavior than the standard biweight S-estimator.

2. Breakdown and outlier rejection. An S-estimate of multivariate location and shape is defined as that vector t and positive definite symmetric (PDS) matrix C which minimize $|C|$ subject to

$$n^{-1}\sum \rho([(x_i - t)^\top C^{-1}(x_i - t)]^{1/2}) = b_0, \tag{2.1}$$

where ρ is a nondecreasing function on $[0, \infty)$. We write this as

$$n^{-1}\sum \rho(d_i) = b_0. \tag{2.2}$$

The function ρ is usually differentiable (the major exception being the MVE, in which ρ is 0 or 1). For this to have nonzero breakdown, ρ must be bounded; the breakdown point is given by the ratio of b_0 to the maximum of ρ [Lopuhaä and Rousseeuw (1991)].

The function ρ is usually chosen to be a scaled version of a base function ρ_0 such as the biweight, which reaches its maximum at c_0. We can then write the constraint (2.2) as

$$n^{-1}\sum \rho(d_i/c) = b_0, \tag{2.3}$$

where b_0 and c are chosen so that $\mathrm{E}(\rho(d/c)) = b_0$ and $b_0 = r\rho(c_0)$ for breakdown r. Often a value of r near 0.5 is used to obtain very high breakdown.

It is known [Lopuhaä (1989); Rocke (1993)] that such an S-estimate is also a root (t, C) of an M-estimation iteration in which weighted estimation

$$t_i^{(j)} = \frac{\sum w(\tilde{d}_i^{(j)})x_i}{\sum w(\tilde{d}_i^{(j)})}, \tag{2.4}$$

$$C^{(j)} = \frac{p\sum w(\tilde{d}_i^{(j)})(x_i - t^{(j)})(x_i - t^{(j)})^\top}{\sum v(\tilde{d}_i^{(j)})} \tag{2.5}$$

is alternated with determination of c to satisfy (2.3). Here

$$w(\tilde{d}) = \psi(\tilde{d})/\tilde{d}, \tag{2.6}$$

$$\psi(\tilde{d}) = \rho'(\tilde{d}), \tag{2.7}$$

$$v(\tilde{d}) = \tilde{d}\psi(\tilde{d}) \tag{2.8}$$

and the $\tilde{d}$ are the Mahalanobis distances of the points after rescaling to satisfy (2.3).

Perhaps the main point of conducting an analysis using a very high breakdown estimator is to avoid letting the outlier have much influence on the estimates. Since the weight given to points whose distance lies beyond c is zero, one might expect that points that are a great distance from the main body of points will receive a weight of zero. That is certainly true in estimation on the real line, where the 50% breakdown biweight S-estimator gives zero weight to any point that lies more than 1.55 estimated standard deviations from the middle.

As it happens, this behavior changes dramatically as the dimension rises. In 20 dimensions, the square distances under normality have a χ^2_{20} distribution with mean 20 and standard deviation 6.32. In order for a point to receive zero weight from a 50% breakdown biweight S-estimator, it must lie a square distance of 94.5 from the middle. Under normality, such distances occur with very low probability ($\sim 10^{-11}$) so that even points much closer to the center are clear outliers. Yet these clear outliers are given positive weight in the analysis.

This suggests a new concept in robustness theory which is called here the *asymptotic rejection probability*. This begins with the rejection point which is defined in Hampel, Ronchetti, Rousseeuw and Stahel [(1986), page 88] as the smallest distance at which all points at larger distances have zero influence. We then determine the chance, in large samples, that when the estimator is presented with all "good" data, a randomly chosen point lies beyond the rejection point. Although this should be small for the sake of efficiency, it is useful to be able to reject as outliers (by giving zero weight) points that are very improbable under the null model. We now define formally the property of the estimator that concerns this ability to reject outliers completely.

DEFINITION 1. Consider a redescending M- or S-estimator, in which $c_0 = \inf\{d_0 \mid w(d) = 0,\ \forall d > d_0\}$, where w is given by (2.6). The asymptotic rejection probability (ARP) α of this estimator is then defined as the probability in large samples under a reference distribution (usually multivariate normal) that a Mahalanobis distance exceeds c_0. If the estimator is normed to the normal distribution, this is $1 - F_{\chi^2(p)}(c_0^2)$.

We now show that the ARP of the usual class of S-estimators, in which the ρ function varies by dimension only by scaling, tends to zero as p rises.

THEOREM 1. *Let ρ: $\Re^+ \mapsto \Re^+$ be a continuously differentiable, function which is increasing on $[0, c_0]$ and constant at $\rho(d) = \rho(c_0)$ for all $d \geq c_0$. Given a data set of n points in $\Re^p$, let (t, C) be defined by minimizing $|C|$ subject to*

$$n^{-1} \sum_{i=1}^{n} \rho([(x_i - t)^\top C^{-1}(x_i - t)]^{1/2}/c_p) = n^{-1} \sum_{i=1}^{n} \rho(d_i/c_p) = b_p, \tag{2.9}$$

where $b_p = E(\rho(d/c_p))$ and where c_p is chosen so that the breakdown point of the estimator is r for every p [so that $b_p = r\rho(c_p)$]. The expectation is taken over a fixed elliptical distribution with finite fourth moments. Let α_p be the ARP of this estimator when the dimension is p and the data have the given multivariate elliptical distribution. Then $\lim_{p\to\infty} \alpha_p = 0$. If the reference distribution is multivariate normal, then for large dimension,

$$\ln(\alpha_p) \approx 0.5p(\ln(k) - k + 1), \tag{2.10}$$

where $k = 1/M^2$ and $\rho(M) = r\rho(c_0)$.

PROOF. Because the Mahalanobis distance is affine equivariant, we may analyze, without loss of generality, the case in which the elliptical distribution is centered at zero and has identity shape. Also, without loss of generality, let $c_0 = 1$. The components $x_{i1}, x_{i2}, \ldots, x_{ip}$ of a data point x_i are identically distributed and uncorrelated (but not independent unless the distribution is normal). Suppose $\mathrm{E}(x_{ij}^2) = A$ and $\mathrm{Var}(x_{ij}^2) = B$ (these would be $A = 1$ and $B = 2$, respectively, under normality). Then $d^2 = \sum_{j=1}^{p} x_{ij}^2$ is asymptotically (in p) normal with mean pA and variance pB.

In large dimension

$$\begin{aligned} \mathrm{E}(\rho(d/c_p)) &= \mathrm{E}\bigl(\rho(\sqrt{d^2}/c_p)\bigr) \\ &\approx \rho\bigl(\sqrt{pA}/c_p\bigr). \end{aligned} \tag{2.11}$$

For this to have breakdown r we need $\mathrm{E}(\rho(d/c_p)) = r\rho(1)$ [Lopuhaä and Rousseeuw (1991)]. Let M be the unique value satisfying $\rho(M) = r\rho(1)$. Then, for large p, $\sqrt{pA}/c_p \approx M$, so $c_p \approx \sqrt{pA}/M$ and

$$\begin{aligned} \Pr(d/c_p \geq 1) &= \Pr(d^2 \geq c_p^2) \\ &= \Pr\biggl(\frac{d^2 - pA}{\sqrt{pB}} \geq \frac{c_p^2 - pA}{\sqrt{pB}}\biggr) \\ &= \Pr\biggl(\frac{d^2 - pA}{\sqrt{pB}} \geq \frac{pA/M^2 - pA}{\sqrt{pB}}\biggr) \\ &= \Pr\biggl(\frac{d^2 - pA}{\sqrt{pB}} \geq \sqrt{p}A\frac{1/M^2 - 1}{\sqrt{B}}\biggr) \\ &= \to 0. \end{aligned}$$

If the reference distribution is multivariate normal, then $A = 1$ and in large dimension and large samples, using the first term of an asymptotic expansion for the χ^2-distribution [Abramowitz and Stegun (1972), Section 26.4.12], we have

$$\begin{aligned}\alpha_p &\approx \Pr(d^2 \geq kp)\\ &\approx \frac{(kp)^{p/2-1}e^{-kp/2}}{2^{p/2-1}\Gamma(p/2)},\end{aligned} \tag{2.12}$$

where $k = 1/M^2$, so that, using Stirling's formula,

$$\ln(\alpha_p) \approx 0.5p(\ln(k) - k + 1) - 0.5\ln(p) - \ln(k) - 0.5\ln(\pi), \tag{2.13}$$

$$\approx 0.5p(\ln(k) - k + 1) \tag{2.14}$$

as required. □

REMARK 1. The conclusions of this theorem also hold for M-estimators if the central part of the weight function is matched to the center of the distribution of d in any of a number of ways. For example, if $c/2$ is chosen to match the median of the distances, then this is the case.

REMARK 2. The stated conditions for this theorem are somewhat stronger than required. The main condition is that there exist M satisfying $\rho(M) = r\rho(c_0)$, and that ρ be continuous at M. The MVE does not satisfy this condition, which allows this estimator to have an ARP of 0.5.

2.1. *Example: the biweight.* Suppose

$$\psi_b(d; c) = \begin{cases} d(1 - (d/c)^2)^2, & 0 \leq d \leq c,\\ 0, & d > c,\end{cases} \tag{2.15}$$

so that

$$\rho_b(d; c) = \begin{cases} d^2/2 - d^4/(2c^2) + d^6/(6c^4), & 0 \leq d \leq c,\\ c^2/6, & d > c,\end{cases} \tag{2.16}$$

and

$$w_b(d; c) = \begin{cases} (1 - (d/c)^2)^2, & 0 \leq d \leq c,\\ 0, & d > c.\end{cases} \tag{2.17}$$

Figure 1 (solid line) shows the asymptotic rejection point of the S-estimator with breakdown 0.5 using the biweight ρ. For example, when $p = 20$, then $c = 9.72$ and the ARP is $1 - F_{\chi^2(20)}(94.5) = 10^{-11}$. Otherwise put, the "average" multivariate normal data point has $d^2 = 20$; the point at which a data point is unequivocally declared an outlier is $d^2 = 94.5$, which is 12 standard deviations

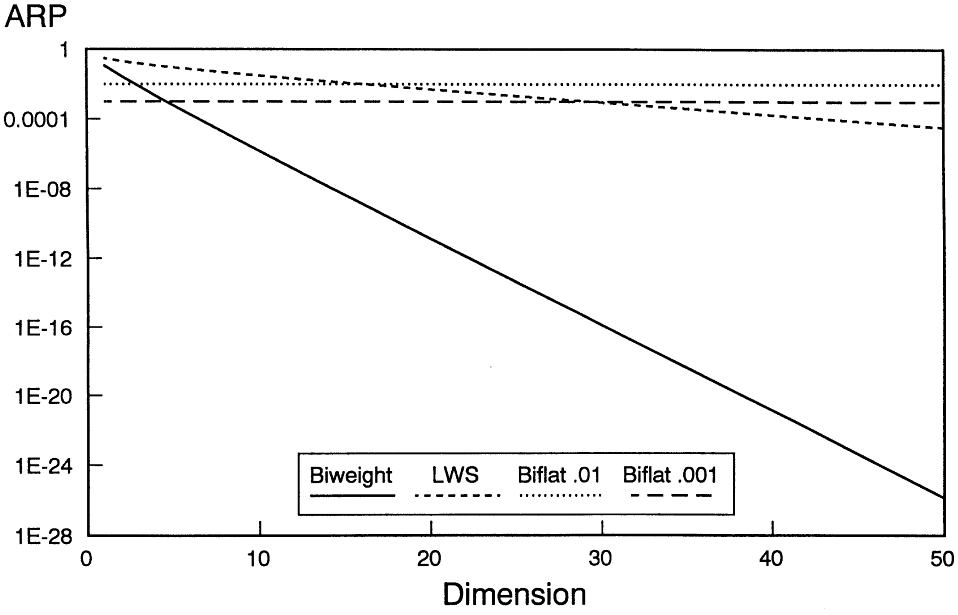

FIG. 1. *Asymptotic rejection probabilities of four S-estimators.*

away. Thus, although points sufficiently far away fail to influence the estimate (achieving high breakdown), outliers may drag the estimates a long distance away, even though that distance is bounded.

2.2. *Two improved S-estimation procedures.* The perverse behavior shown above for standard S-estimators is not a necessary consequence of the definition of S-estimator. First, an S-estimator is exhibited whose weight function is the same as the biweight's, except it has been translated so that it begins rising from zero at a point M, rather than 0. It is defined by a two-parameter class of ρ functions. This will be called the *t-biweight* (for translated biweight).

Let

$$(2.18)\quad w_t(d;c,M) = \begin{cases} 1, & 0 \le d < M, \\ (1-((d-M)/c)^2)^2, & M \le d \le M+c, \\ 0, & d > M+c, \end{cases}$$

$$(2.19)\quad \psi_t(d;c,M) = \begin{cases} d, & 0 \le d < M, \\ d(1-((d-M)/c)^2)^2, & M \le d \le M+c, \\ 0, & d > M+c, \end{cases}$$

$$(2.20)\quad \rho_t(d;c,M)=\begin{cases} d^2/2, & 0\le d<M,\\ M^2/2 & \\ \quad -M^2(M^4-5M^2c^2+15c^4)/(30c^4) & \\ \quad +d^2(1/2+M^4/(2c^4)-M^2/c^2) & \\ \quad +d^3(4M/(3c^2)-4M^3/(3c^4)) & \\ \quad +d^4(3M^2/(2c^4)-1/(2c^2)) & \\ \quad -4Md^5/(5c^4)+d^6/(6c^4), & M\le d\le M+c,\\ M^2/2+c(5c+16M)/30, & d>M+c. \end{cases}$$

The limit of this as $c\to 0$ is the least Winsorized squares (LWS) estimate in which

$$(2.21)\qquad w_{\mathrm{LWS}}(d;c)=\begin{cases}1, & 0\le d\le c,\\ 0, & d>c,\end{cases}$$

$$(2.22)\qquad \psi_{\mathrm{LWS}}(d;c)=\begin{cases}d, & 0\le d\le c,\\ 0, & d>c,\end{cases}$$

$$(2.23)\qquad \rho_{\mathrm{LWS}}(d;c)=\begin{cases}d^2/2, & 0\le d\le c,\\ c^2/2, & d>c.\end{cases}$$

(This is least *metrically* Winsorized squares, as distinguished from that based on order statistics, as is used in least trimmed squares [Rousseeuw and Leroy (1987)].)

Now the two parameters c and M of the t-biweight can be chosen to give the desired breakdown point and ARP, subject to the following constraint.

THEOREM 2. *An S-estimator with breakdown point r in dimension p using the t-biweight ρ function with parameters c and M has* ARP *no larger than that of* LWS *with the same breakdown. The* ARP *of this latter estimator also goes to* 0 *as $p\to\infty$, but much more slowly than the biweight. In particular, for multivariate normal data and for large p, the ARP α_p of the biweight satisfies*

$$(2.24)\qquad \ln(\alpha_p)\approx -1.13p$$

and the ARP *of* LWS *satisfies*

$$(2.25)\qquad \ln(\alpha_p)\approx -0.15p.$$

PROOF. We show in the Appendix that

$$(2.26)\qquad \frac{\rho_t(d;c,M)}{\rho_t(c+M;c,M)}\ge\frac{\rho_{\mathrm{LWS}}(d;c+M)}{\rho_{\mathrm{LWS}}(c+M;c+M)}.$$

Since the two functions are normed to have the same maximum, the fact that the t-biweight curve lies above the LWS curve means that each fractile of the

t-biweight lies to the left of the equivalent fractile of LWS. This means that the ARP is smaller for the t-biweight than for LWS. The relative behavior of the ARP for the biweight and LWS estimators, each with breakdown 0.5, are shown in Figure 1. Using Theorem 1, together with the fact that $M = 0.4542c_0$ for the biweight ρ and $M = 0.7071c_0$ for the LWS ρ, we arrive at the stated expressions. The expression (2.24) is fairly accurate even for dimensions as low as 30–50. The expression (2.25) is less exact, since the linear term in p is smaller, so (2.13) should be used if an approximation is needed. □

An estimator that can be set to have a given breakdown point and ARP for any dimension is given by the following. We call it the biflat, because the ρ function is flat for small values and for large values of d. The weight function is similar to the biweight in between. Note that Davies (1987) suggests, apparently for technical reasons, a ρ function that is flat in a neighborhood of zero.

Let

$$(2.27)\qquad \psi_f(d; c, M) = \begin{cases} 0, & 0 \le d < M - c, \\ (1 - ((d - M)/c)^2)^2, & M - c \le d \le M + c, \\ 0, & d > M + c, \end{cases}$$

$$(2.28)\qquad w_f(d; c, M) = \begin{cases} 0, & 0 \le d < M - c, \\ (1 - ((d - M)/c)^2)^2/d, & M - c \le d \le M + c, \\ 0, & d > M + c, \end{cases}$$

$$(2.29)\qquad \rho(d) = \begin{cases} 0, & 0 \le d < M - c, \\ (8c/15 + 2M^3/(3c^2) & \\ \quad - M^5/(5c^4) - M) & \\ \quad + d(1 + M^4/c^4 - 2M^2/c^2) & \\ \quad + d^2(2M/c^2 - 2M^3/c^4) & \\ \quad + d^3(2M^2/c^4 - 2/(3c^2)) & \\ \quad - Md^4/c^4 + d^5/5c^4, & M - c \le d \le M + c, \\ 16c/15, & d > M + c. \end{cases}$$

In this case, the constraints on c and M can be simply satisfied by setting $M + c = (F^{-1}_{\chi^2(p)}(\alpha))^{1/2}$, which forces the ARP to α, and then solving for the breakdown point, which can be done by monotonicity so long as $\mathrm{E}(\rho) > 8c/15$ with $M = c = 0.5F^{-1}_{\chi^2(p)}(\alpha)$. For $\alpha = 0.01$, this requires $p \ge 4$ and for $\alpha = 0.001$ this requires $p \ge 7$. The problem is that, for $p = 3$, for example, no member of the class with breakdown 0.5 has an ARP as bad as 0.01. In these cases, one can use the 50% breakdown estimator with $M = c$ and whatever ARP is produced, or one can reasonably use the biweight.

As an example, Figure 2 shows the weight functions for the biweight, t-biweight and biflat estimators, each with breakdown 0.5 for dimension 10. The latter two have been set to have ARP of 0.01, whereas the ARP of the biweight is 10^{-6}. Also shown is the density of the square root of a χ^2_{10} variate. The t-biweight has weight near 1 for most of the data (when normally distributed), so the efficiency will be high. The biflat has the ability to downweight inliers as well as outliers, and can be set at an arbitrary ARP even in high dimensions. Both these should be useful in different circumstances. The ordinary biweight seems to have little advantage over the other two, and has the severe problem of failing to reject outliers. This problem and the differences among the estimators are even more pronounced when the dimension is higher than 10.

3. Efficiency, gross error sensitivity and local shift sensitivity. The biweight, LWS and biflat estimators have ARP properties studied in the last section. In this section, three other properties of these estimators are studied: the normal theory statistical efficiency, the gross error sensitivity and the local shift sensitivity. Limiting the ARP does have consequences for the efficiency and gross error sensitivity, but they seem to be modest compared to the serious consequences for outlier rejection of using the biweight in high dimension. Consequences for local shift sensitivity are greater, but the LWS exhibits a

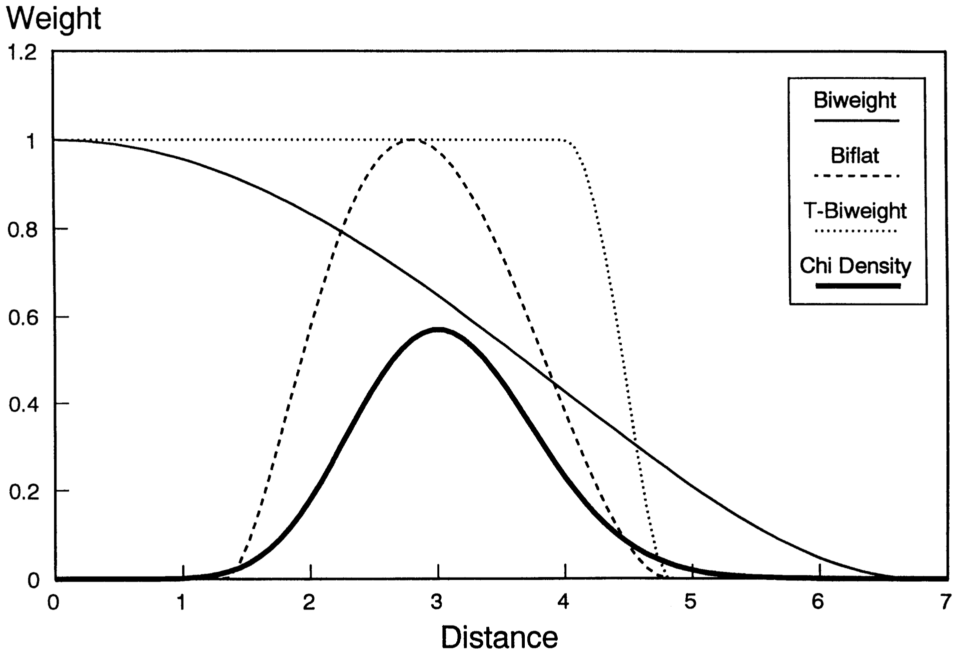

FIG. 2. *Three weight functions.*

large improvement in local behavior over the MVE, and the biflat exhibits an equally large improvement in local behavior over the LWS.

Figure 3 shows the asymptotic (in n) efficiency of the location estimators for multivariate normal data as a function of p. (Behavior of the shape estimation values is similar.) Both the biweight and the LWS estimator have efficiency that approaches 1 as $p \to \infty$, while the biflat in high dimension has an efficiency that depends on the ARP setting. Thus the biflat estimator can be chosen in high dimension to fit any relative valuation of efficiency versus ARP; the smaller the ARP that can be tolerated, the larger the efficiency at the normal model. The theorem below gives the large dimension values for these efficiencies.

The gross error sensitivity is the limit as $\varepsilon \to 0$ of ε^{-1} times the largest displacement of the estimate obtainable by contamination of size ε [Hampel, Ronchetti, Rousseeuw and Stahel (1986)]. In order to compare these across dimension, this is then divided by $\sqrt{p}$, which is the approximate mean distance of points from the center. Also, the average random displacement of the sample mean from the true value is of order $\sqrt{p}$. Figure 4 shows the gross error sensitivity values for the biweight, LWS and two biflat estimators. While the values for the other estimators are larger than that for the biweight, the difference seems modest, especially with the biflat with $\alpha = 0.001$.

The local shift sensitivity is defined in Hampel, Ronchetti, Rousseeuw and Stahel (1986) to be the limit as $\epsilon \to 0$ and $\Delta \to 0$ of $\epsilon^{-1}\Delta^{-1}$ times the largest distance the estimate can be moved by shifting a fraction ϵ of the data a

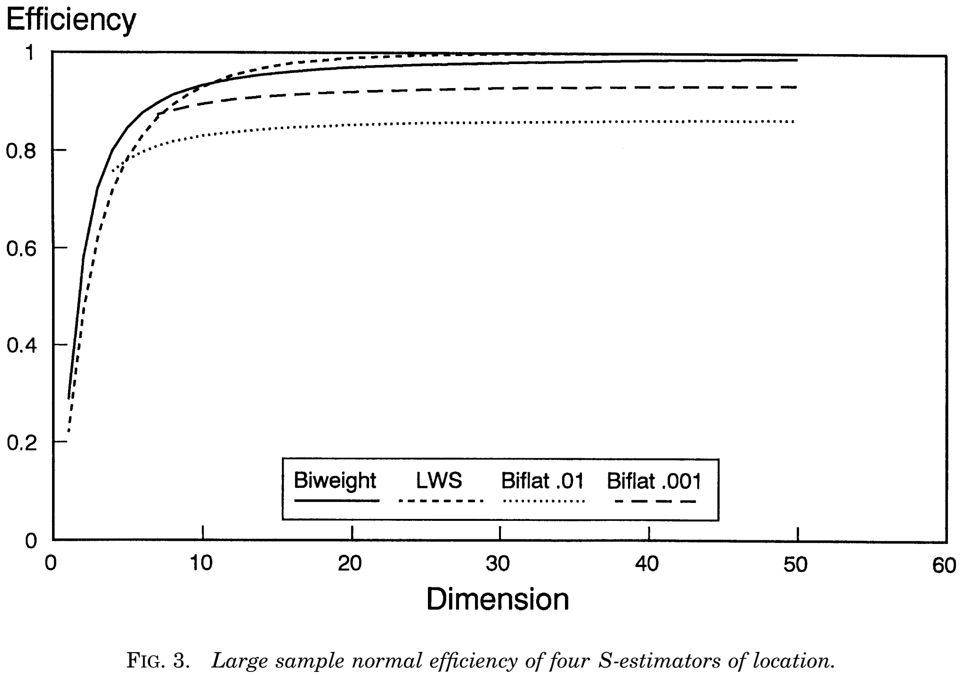

FIG. 3. *Large sample normal efficiency of four S-estimators of location.*

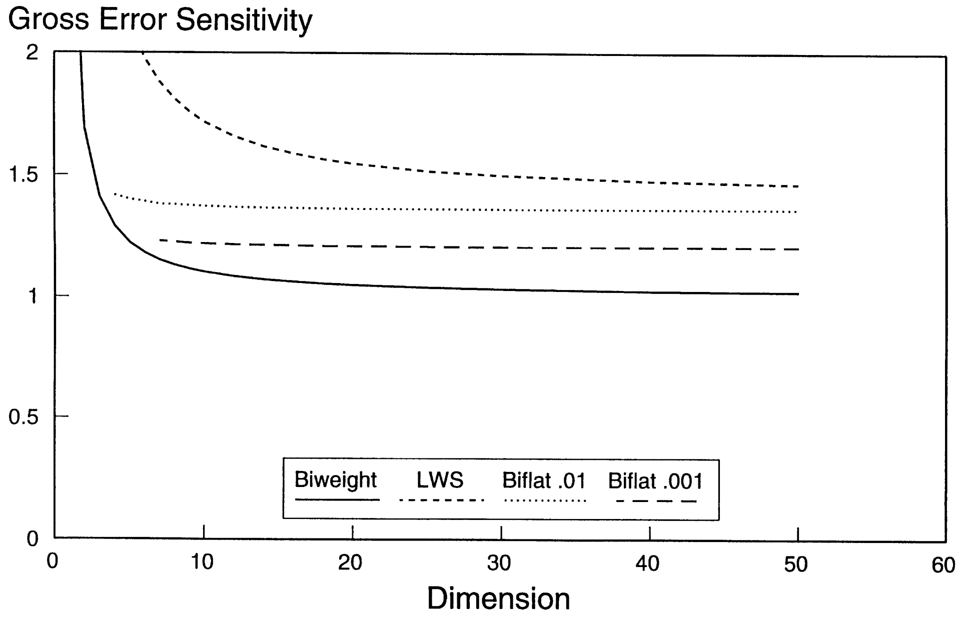

FIG. 4. *Large sample scaled gross error sensitivity of four S-estimators of location.*

distance Δ. For reasons that will be clear shortly, this will be described in terms of nonzero quantities ϵ and Δ, and in terms of movement of a single point in finite samples, so that $\epsilon = n^{-1}$. First consider the location problem on the real line. It is easily seen that the sample mean has local shift sensitivity of 1, since moving one of n points a distance of Δ moves the sample mean by Δ/n. For the sample median, moving the central point (n odd) moves the median by Δ, so that the local shift sensitivity is $O(n)$ and is thus infinite. The MVE in one dimension is the shorth [Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972)] and this can be moved essentially from the midrange of the left half of the data to the midrange of the right half of the data by a shift of Δ in one point, so that the local shift sensitivity is $O(n/\Delta)$. This is doubly infinite, both in n and in Δ, indicating that the local shift sensitivity of the shorth is worse than that of the median, even though both have infinite local shift sensitivity. The LWS estimate has a maximum shift obtained by moving a single point from just where the weight is one to just where it is zero, a simple computation shows that this is $O(1/\Delta)$. The LWS estimate thus has local shift sensitivity that is infinite (like the median), but is better behaved than the shorth (MVE). The biflat has finite local shift sensitivity, but one that increases with the dimension and is higher than that of the biweight, as shown in Figure 5.

The following theorem gives the large p values for these robustness measures.

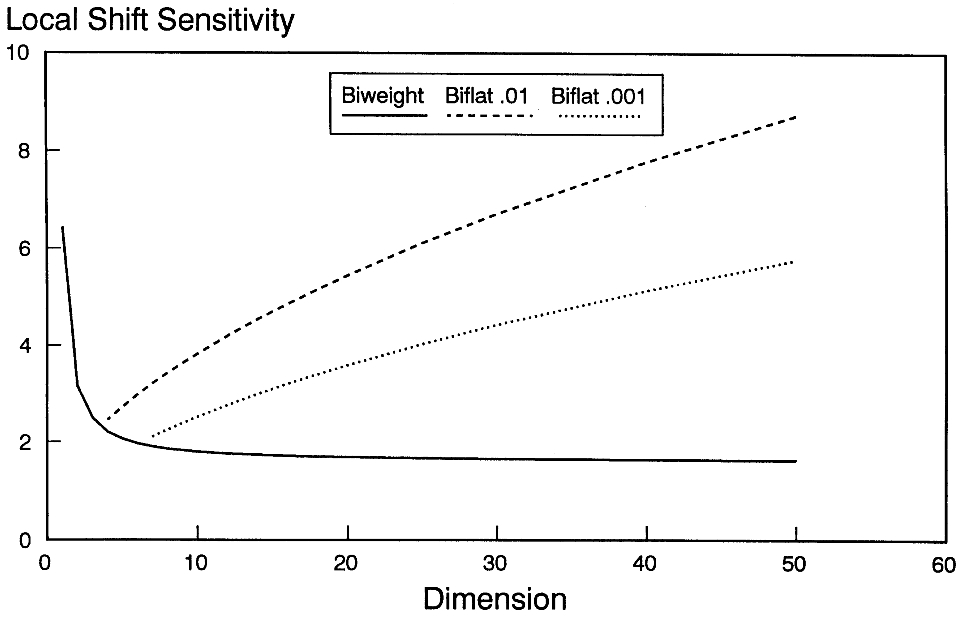

FIG. 5. *Large sample local shift sensitivity of three S-estimators of location.*

THEOREM 3. *For large p, the normal efficiency, gross error sensitivity and local shift sensitivity of the biweight, MVE, LWS and the biflat for $\alpha = 0.01$ and $\alpha = 0.001$, each with* 50% *breakdown, are given by Table* 1.

PROOF. For the biweight, the part of the weight function with nonnegligible probability becomes a smaller and smaller fraction of the support, so that, for large p, the biweight is essentially an unweighted mean, which thus has asymptotic efficiency 1. A similar statement holds even more strongly for the LWS estimator, since the weight function is constant at 1 from 0 to c. The efficiency values for the biflat are less straightforward, since the rejection point does not continue to move further out in the tail, so the values were derived by a direct numerical calculation using a normal approximation for the dis-

TABLE 1

Estimator	Efficiency	Gross error sensitivity	Local shift sensitivity
Biweight	1	1.00	1.59
MVE	0	∞	$O(n/\Delta)$
LWS	1	1.41	$O(1/\Delta)$
Biflat, $\alpha = 0.01$	0.87	1.36	$1.26\sqrt{p}$
Biflat, $\alpha = 0.001$	0.94	1.21	$0.85\sqrt{p}$

tance. Note that, for large p, the constants are approximately $M = \sqrt{p}$ and $c = z_\alpha/\sqrt{2}$.

To determine the gross error sensitivity for the biweight, note that the maximum value of ψ occurs at $d = c/\sqrt{5}$ and the achieved maximum is $16c\sqrt{5}/125$. Using the large p approximation $c = \sqrt{p}/M$, we obtain $16\sqrt{5p}/(125M)$. Now, for large p, the influence function of the location estimate is $\psi(d)/\omega$, where $\omega = \mathrm{E}(w(d))$ [Lopuhaä (1989); Maronna (1976)], and $\omega = (1 - 2M^2 + M^4) + 2M^4/p$ for large p, so the large p gross error sensitivity is approximately

$$\frac{16\sqrt{5p}}{125M(1 - 2M^2 + M^4)}.$$

When scaled by $\sqrt{p}$ and using $M = 0.454202$, we arrive at the stated number.

For LWS, the expectation of $w(d)$ for large p is 1, and the maximum of ψ is c at $d = c$. Using the approximation $c = \sqrt{p}/M$, we obtain a scaled gross error sensitivity of $1/M$. Since $M \approx 1/\sqrt{2}$, the large p gross error sensitivity is $\sqrt{2}$. The values for the biflat were computed numerically using the normal asymptotic distribution of d.

Local shift sensitivity of the biweight is determined by noting that the maximum slope of ψ is 1, which occurs at $d = 0$. The local shift sensitivity for large p is thus $1/(1 - 2M^2 + M^4) = 1.5874$. The LWS local shift sensitivities are infinite and the MVE local shift sensitivity is doubly infinite, as in the dimension 1 case. The local shift sensitivities for the biflat were computed numerically using the normal asymptotic distribution of d. □

A comparison of smooth ρ function S-estimators to the MVE estimator seems called for at this point. In large samples, the MVE has breakdown 50% and ARP of 50%. These extremely robust properties are balanced by zero asymptotic efficiency, infinite gross error sensitivity, doubly infinite local shift sensitivity, difficult computation and poor small sample properties. While it has been suggested for use as a starting value for S-estimation, the minimum covariance determinant (MCD) estimator [Rousseeuw (1985)] seems superior both in efficiency and in computability [Woodruff and Rocke (1993, 1994)]. In any case, the properties of the MVE and MCD as estimators in their own right may be of less interest than their performance in achieving a high-breakdown starting value for iterative estimators [Woodruff and Rocke (1994); Rocke and Woodruff (1996)] which offer the potential of good efficiency as well as good robustness properties.

The biflat, on the other hand, has ARP bounded away from zero, with small disadvantages compared to the biweight in efficiency and gross error sensitivity. The local shift sensitivity is measurably worse than the biweight, but still finite, unlike the MVE, or even the LWS and sample median. Even the LWS estimator, which has infinite local shift sensitivity, is still "infinitely better" in this property than the MVE.

The biflat does suffer from a computational problem in small samples when the contamination is heavy. Suppose that the majority of the data were normally distributed with mean 0 and covariance I and that the remainder were displaced a distance η. Using the covariance of the good data, the distribution of the Mahalanobis distances would be bimodal with a region of small probability in between. It could easily happen in that case that the "bulge" of the biflat weight function could fall in the empty region, leading to instability.

This instability happens in part because the S-estimator standardization constraint is based on the mean of $\rho(d)$. If this is replaced by a constraint using the median of $\rho(d)$, this problem disappears, since it insures that there are data where the weight function is high. In fact, one can standardize by fixing the $\lfloor (n+p+1)/2 \rfloor$ order statistic of the set of Mahalanobis square distances to the $\lfloor (n+p+1)/2 \rfloor/(n+1)$ quantile of a χ^2_p. As in the case of the MVE [Lopuhaä and Rousseeuw (1991)], this insures a high breakdown.

The resulting estimator is no longer an S-estimator, since the location and shape iterations use a ρ function that is not used for standardization. This M-type estimator appears to be statistically and computationally robust. The same scaling can be used with any M-estimator; we call it "median"scaling, although the order statistic used is not quite the median. The efficiency, gross error sensitivity and local shift sensitivity properties should remain unchanged. In Rocke and Woodruff (1996), an M-type estimator based on the t-biweight is successfully applied to the detection of multivariate outliers.

4. An example.

We illustrate in this section the difference that it can make to take account of points made in this paper. We take a data set of 50 points in 10 dimensions distributed as independent standard normal. Consider a succession of data sets in which i of the 50 points have 5 added to each value (this is a distance of $\sqrt{250}$ away from the main body of data). We iterate the biweight S-estimator and the median scaled biflat M-estimator to convergence from two starting points. The first is the mean and covariance of the data; this is the worst case for such estimators. The second is the mean and covariance of the unperturbed data; this approximates the best case in which a good, robust starting point is found.

The quality of the resulting estimate will be measure by the largest eigenvalue of the estimated shape matrix. Other measures of the goodness of either the location or shape estimate yield comparable results, since the main point is whether the estimate breaks down and follows the outliers or resists and stays with the good data.

Figure 6 shows the results of a 40% breakdown biweight S-estimator and a 40% breakdown, 1%-ARP biflat M-estimator with median scaling applied to each of 22 data sets having from 0 to 21 outliers in 50. Here, the starting point is the mean and covariance of the good data. As promised by the nominal breakdown, both estimators give good results until the contamination nears 40%. In this case, the biweight breaks down at 36% contamination and the biflat breaks down at 42%, but this could be mere finite-sample fluctuations.

Figure 7 shows the real advantage of the biflat. From a bad starting point (mean and covariance of all the data), the biweight breaks down at 10% outliers (perhaps not coincidentally, this is the reciprocal of the dimension). The

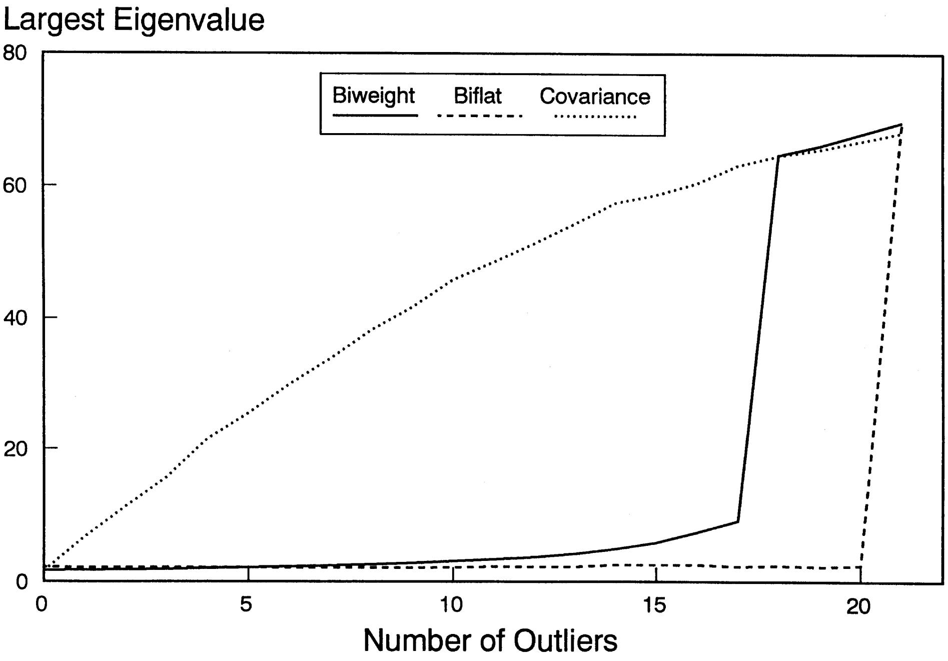

FIG. 6. *Estimation and breakdown: good start.*

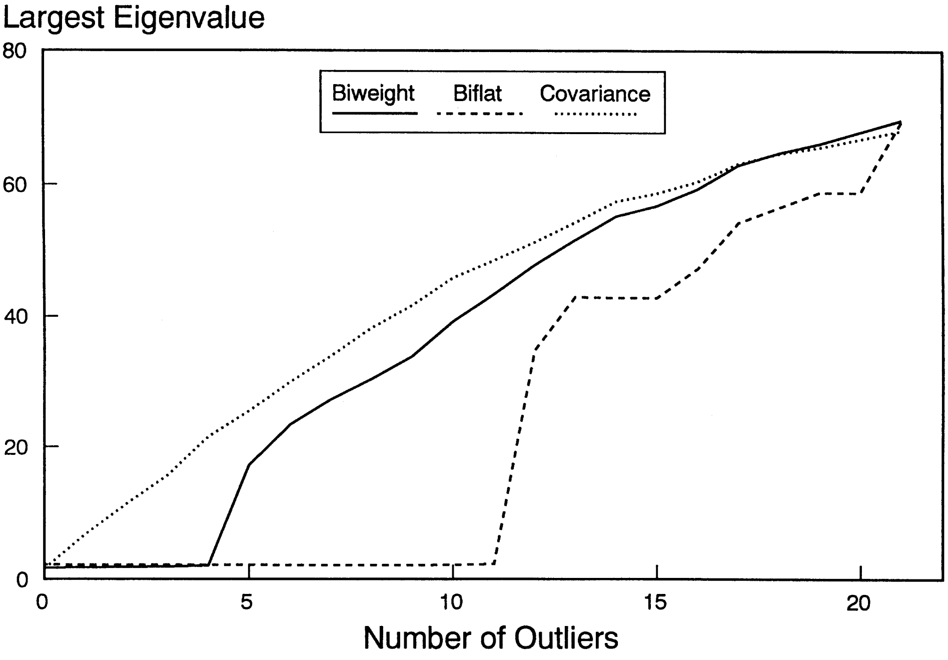

FIG. 7. *Estimation and breakdown: bad start.*

biflat, on the other hand, converges to a good solution until the contamination reaches 24%. This remarkable behavior gives hope that this class of estimators will allow robust estimation in the presence of substantial contamination without excessive computation. Even here, however, large amounts of contamination will require a robust starting point such as the MCD, of which the successful computation in even moderate dimension requires considerable sophistication in algorithms and large amounts of computer time [Woodruff and Rocke (1993, 1994); Rocke and Woodruff (1993, 1996)].

Similar good results were obtained using the t-biweight M-estimator with median scaling. The t-biweight S-estimator performed about the same as the biweight S-estimator. In larger dimension, however, some experimentation suggests that the t-biweight will do much better at resisting outliers than the biweight.

5. Conclusion. In this paper, it has been shown that 50% breakdown S- (and M-) estimates as usually defined do not apply zero weight to obvious outliers in high dimension due to the use of an inflexible family of ρ functions. This problem is repaired by use of a two-parameter class of ρ functions, so that both the breakdown and the asymptotic rejection probability can be chosen. This greater ability to reject outliers comes at a modest cost in efficiency and gross error sensitivity and at a greater, but finite, cost in local shift sensitivity.

APPENDIX

PROOF OF THEOREM 2. We show that, for all $c \geq 0,\ M \geq 0,\ d \geq 0$,

$$\text{(A.1)} \qquad \frac{\rho_t(d; c, M)}{\rho_t(c+M; c, M)} \geq \frac{\rho_{\text{LWS}}(d; c+M)}{\rho_{\text{LWS}}(c+M; c+M)}.$$

Without loss of generality, we norm the problem so that $c + M = 1$, so we must show that

$$\text{(A.2)} \qquad \frac{\rho(d; \beta)}{\rho(1, \beta)} \geq \frac{\rho_{\text{LWS}}(d; 1)}{\rho_{\text{LWS}}(1; 1)},$$

where

$$\text{(A.3)} \quad \psi(d; \beta) = \begin{cases} d, & 0 \leq d < \beta, \\ d(1 - ((d-\beta)/(1-\beta))^2)^2, & \beta \leq d \leq 1, \\ 0, & d > 1, \end{cases}$$

and

$$\text{(A.4)} \quad \rho(d; \beta) = \begin{cases} d^2/2, & 0 \leq d < \beta, \\ \beta^2/2 + \int_\beta^d x(1 - ((x-\beta)/(1-\beta))^2)^2\, dx, & \beta \leq d \leq 1, \\ (4\beta^2 + 6\beta + 5)/30, & d > 1. \end{cases}$$

This is obviously true for $d \geq 1$ and is immediate for $0 \leq d \leq \beta$, so we need to show that

$$\frac{\rho(d;\beta)}{\rho(1,\beta)} \geq \frac{d^2/2}{1/2} \tag{A.5}$$

for all $0 \leq \beta \leq 1$ and $\beta \leq d \leq 1$; that is,

$$\beta^2/2 + \int_\beta^d x(1-((x-\beta)/(1-\beta))^2)^2\,dx \geq d^2\rho(1,\beta), \tag{A.6}$$

$$\int_\beta^d x(1-((x-\beta)/(1-\beta))^2)^2\,dx \geq d^2(4\beta^2+6\beta+5)/30 - \beta^2/2. \tag{A.7}$$

Now, letting $y = (x-\beta)/(1-\beta)$ and $z = (d-\beta)/(1-\beta)$ the integral becomes

$$\begin{aligned}&\int_\beta^d x(1-((x-\beta)/(1-\beta))^2)^2\,dx\\ &\qquad = \int_0^z (\beta+(1-\beta)y)(1-y^2)^2(1-\beta)\,dy\end{aligned} \tag{A.8}$$

$$= \beta(1-\beta)\int_0^z (1-y^2)^2\,dy + (1-\beta)^2\int_0^z y(1-y^2)^2\,dy \tag{A.9}$$

$$= \beta(1-\beta)(z-2z^3/3+z^5/5) + (1-\beta)^2[1-(1-z^2)^3]/6 \tag{A.10}$$

This makes (A.7)

$$\begin{aligned}&\beta(1-\beta)(z-2z^3/3+z^5/5) + (1-\beta)^2[1-(1-z^2)^3]/6\\ &\qquad \geq (\beta+(1-\beta)z)^2(4\beta^2+6\beta+5)/30 - \beta^2/2,\end{aligned} \tag{A.11}$$

which can also be expressed as

$$\begin{aligned}&(1-\beta)(1-z)[(10\beta^2+4\beta^3) + z(20\beta-2\beta^2-4\beta^3) + z^2(10+4\beta)\\ &\qquad + z^3(10-16\beta) - z^4(5+\beta) - z^5(5-5\beta)]/30 \geq 0,\end{aligned} \tag{A.12}$$

$$\begin{aligned}&(10\beta^2+4\beta^3) + z(20\beta-2\beta^2-4\beta^3) + z^2(10+4\beta)\\ &\qquad + z^3(10-16\beta) - z^4(5+\beta) - z^5(5-5\beta) \geq 0.\end{aligned} \tag{A.13}$$

We now show that the left-hand side of (A.13) is increasing in β for every $0 \leq z \leq 1$. The derivative of (A.13) with respect to β is

$$(20\beta+12\beta^2) + z(20-4\beta-12\beta^2) + 4z^2 - 16z^3 - z^4 + 5z^5. \tag{A.14}$$

This derivative is nonnegative since (1) it is nonnegative at $\beta = 0$, where it is $20z+4z^2-16z^3-z^4+5z^5$, since $20+4z-16z^2-z^3+5z^4 > 20+4(0)-16(1)-1+5(0) = 4$ and (2) it is increasing in β because its derivative $20+24\beta-z(4+24\beta)$ is clearly at least 16.

Thus, we may show the inequality (A.13) by showing it is true for $\beta = 0$, in which case it becomes

$$10z^2+10z^3-5z^4-5z^5 = z^2(10+10z-5z^2-5z^3) \geq 0, \tag{A.15}$$

where the last inequality follows since $10 + 10z - 5z^2 - 5z^3 > 10 + 10(0) - 5(1) - 5(1) = 0$. □

Acknowledgments. The helpful comments of two referees, an Associate Editor and the Editor are gratefully acknowledged.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of Mathematical Functions.* Dover, New York.

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.

CAMPBELL, N. A. (1980). Robust procedures in multivariate analysis I: robust covariance estimation. *J. Roy. Statist. Soc. Ser. C* **29** 231–237.

CAMPBELL, N. A. (1982). Robust procedures in multivariate analysis I: robust canonical variate analysis. *J. Roy. Statist. Soc. Ser. C* **31** 1–8.

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

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickell, K. A. Doksum and J. L. Hodges, eds.) 157–184. Wadsworth, Belmont, CA.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

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

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

KENT, J. T. and TYLER, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19** 2102–2119.

LOPUHAÄ, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683.

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.

ROCKE, D. M. (1993). On M- and S-estimators of multivariate location and shape. Unpublished manuscript.

ROCKE, D. M. and WOODRUFF, D. L. (1993). Computation of robust estimates of multivariate location and shape. *Statist. Neerlandica* **47** 27–42.

ROCKE, D. M. and WOODRUFF, D. L. (1996). Identification of outliers in multvariate data. *J. Amer. Statist. Assoc.* To appear.

ROUSSEEUW, P. J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications* **B** (W. Grossmann, G. Pflug, I. Vincze and W. Werz, 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, Berlin.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990a). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* **85** 633–639.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990b). Rejoinder. *J. Amer. Statist. Assoc.* **85** 648–651.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1991). Robust distances: simulations and cutoff values. In *Directions in Robust Statistics and Diagnostics* **2** (W. Stahel and S. Weisberg, eds.) 195–203. Springer, New York.

STAHEL, W. A. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. dissertation, ETH, Zurich.

TYLER, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika* **70** 411–420.

TYLER, D. E. (1988). Some results on the existence, uniqueness, and computation of the *M*-estimates of multivariate location and scatter. *SIAM J. Sci. Statist. Comput.* **9** 354–362.

TYLER, D. E. (1991). Some issues in the robust estimation of multivariate location and scatter. In *Directions in Robust Statistics and Diagnostics* **2** (W. Stahel and S. Weisberg, eds.) 327–336. Springer, New York.

WOODRUFF, D. L. and ROCKE, D. M. (1993). Heuristic search algorithms for the minimum volume ellipsoid. *J. Comput. Graphical Statist.* **2** 69–95.

WOODRUFF, D. L. and ROCKE, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *J. Amer. Statist. Assoc.* **89** 888–896.

GRADUATE SCHOOL OF MANAGEMENT
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616
E-MAIL: dmrocke@ucdavis.edu