The Annals of Statistics
1992, Vol. 20, No. 4, 1803–1827

BREAKDOWN PROPERTIES OF LOCATION ESTIMATES BASED ON HALFSPACE DEPTH AND PROJECTED OUTLYINGNESS[1]

BY DAVID L. DONOHO AND MIRIAM GASKO

University of California, Berkeley, and San Jose State University

We describe multivariate generalizations of the median, trimmed mean and W estimates. The estimates are based on a geometric construction related to "projection pursuit." They are both affine equivariant (coordinate-free) and have high breakdown point. The generalization of the median has a breakdown point of at least $1/(d+1)$ in dimension d and the breakdown point can be as high as $1/3$ under symmetry. In contrast, various estimators based on rejecting apparent outliers and taking the mean of the remaining observations have breakdown points not larger than $1/(d+1)$ in dimension d.

1. Introduction. In 1974, Tukey (1974a, b) introduced the notion of the depth of a point in a multivariate dataset as follows. The depth of a value x in a one-dimensional dataset $X = \{X_1, \ldots, X_n\}$ is the minimum of the number of data points on the left and on the right of x:

$$\text{depth}_1(x; X) = \min\bigl(\#\{i\colon X_i \le x\}, \#\{i\colon X_i \ge x\}\bigr)$$

[see also Tukey (1977)]. The depth of a point $x \in \mathbb{R}^d$ in a d-dimensional dataset is the least depth of x in any one-dimensional projection or view of the dataset. In detail, if we let u denote a vector in $\mathbb{R}^d$ of unit norm, then the dataset $\{u^T X_i\}$ is a one-dimensional projection of the dataset X and we define

$$\begin{aligned}\text{depth}_d(x; X) &= \min_{|u|=1} \text{depth}_1\bigl(u^T x; \{u^T X_i\}\bigr) \\ &= \min_{|u|=1} \#\{i\colon u^T X_i \ge u^T x\}.\end{aligned} \tag{1.1}$$

Tukey considered the use of contours of depth for indicating the shape of two-dimensional datasets and suggested that depth might allow one to define a reasonable multivariate analog of rank statistic. Of course, in dimension 1, the sample minimum and maximum are the data points of depth 1, the upper and lower quartiles of depth $\sim n/4$ and the median of depth $\sim n/2$.

Received December 1987; revised February 1992.

[1]Research supported by an NSF graduate fellowship, NSF Grant DMS-84-51753 and Public Health Service Traineeship 5T32HL07365-10.

AMS 1980 *subject classifications*. Primary 62F35; secondary 62G05, 62H12.

Key words and phrases. Multivariate depth and outlyingness, location estimates, robustness, breakdown point, projection pursuit, halfspace distance, Glivenko–Cantelli property, outlier rejection.

Tukey's proposal raises a number of interesting possibilities. First, it gives a possible definition of the median of a multivariate dataset. Since in $d = 1$, the median is a "deepest" x value, a deepest x value in higher dimensions can be thought of as a multidimensional median. Second, the contour of depth $= n/4$ (say) is a convex region whose shape indicates the scale and correlation of the data in a manner analogous to the way a standard probability content ellipse for a Normal distribution indicates its scale and correlation. Third, one can define trimmed means, averaging those points of depth $\geq n/10$, say.

The resulting notions of median, trimmed mean and covariance estimate have two important properties. First, they are affine equivariant, that is, they commute with translations and linear transformations of the data. Second, they are robust in high dimensions. Indeed, the depth-trimmed mean and the deepest point can have high breakdown points—as high as $1/3$—in high dimensions (Donoho, 1982).

This combination of properties (equivariance and robustness) is interesting because many classical ways of defining location estimators lack one or both of these properties. Maronna (1976) and Huber (1977) found that affine-equivariant M estimates of location/scatter have breakdown points bounded above by $1/d$ in dimension d. This means that in high dimensions, such "robust" estimators can be upset by a relatively small fraction of outliers strategically placed.

The notion of depth leads to estimators which are affine equivariant and have high breakdown point. By considering why depth is successful in this regard, it becomes apparent that the idea of looking at all one-dimensional views of a dataset—projection pursuit—can be used in other ways as well. In dimension 1, a measure of the outlyingness of a value x with respect to a dataset X is given by the robust measure

$$r_1(x; X) = |x - \mathrm{Med}(X)|/\mathrm{MAD}(X), \tag{1.2}$$

where Med denotes median and MAD denotes median absolute deviation from the median. As an analog in dimension d, one could use

$$r_d(x; X) = \max_{|u|=1} r_1\left(u^T x; \{u^T X_i\}\right). \tag{1.3}$$

This is a measure of how outlying x is in the worst one-dimensional projection or view of the dataset. The measure r_d can be used to develop a robust estimator generalizing what Mosteller and Tukey (1977) call a W estimator. Their definition is for dimension 1, and such an estimate takes the form

$$T_w(X) = \sum w_i X_i \Big/ \sum w_i, \tag{1.4}$$

where the weights $w_i = w(r_1(X_i; X))$ are generated by a weight function $w(r)$ which downweights outlying observations. The obvious generalization to $d > 1$, simply replacing r_1 by r_d, works and defines an affine equivariant estimator of multivariate location. Under very mild conditions on the dataset X, T_w has a breakdown point close to $1/2$, even in high dimensions. This is the best one can hope for in an equivariant estimator, and it means that quite heavy

contamination is necessary in order to upset T_w completely. This result is due to Stahel (1981) and, independently, to Donoho (1982).

The particular combination of properties emphasized above—equivariance and high breakdown point—has been of interest to a number of researchers recently; see, for example, Rousseeuw (1984), Rousseeuw and Yohai (1985), Lopuhaä and Rousseeuw (1991), Davies (1987) and Lopuhaä (1988). Results about estimators defined by depth and outlyingness which motivated this recent work have appeared only in thesis/qualifying paper form. Because of the very simple and geometric form of depth and outlyingness, we present here a published discussion of their breakdown properties.

The paper is arranged as follows. Section 2 covers general properties of the depth and of outlyingness r. Section 3 covers the breakdown properties of location estimators built using them. Section 4 shows how some natural methods of constructing robust estimators do not provide the same breakdown properties. Section 5 discusses recent research on breakdown properties and the need for a computational breakthrough in order to make high-breakdown estimates practical.

All proofs are contained in Section 6. The so-called halfspace distance, from the theory of empirical processes, plays a key role in our proofs of asymptotic results. The paper of Donoho (1982) will be referred to as [D] and the technical report of Donoho and Gasko (1987) will be referred to as [DG].

2. Depth and outlyingness. To begin with, depth is independent of the coordinate system chosen.

LEMMA 2.1. depth *is affine invariant*:

$$\text{depth}(Ax + b; \{AX_i + b\}) = \text{depth}(x; X) \tag{2.1}$$

for every b *and every nonsingular linear transformation* A.

Let D_k be the set of all $x \in \mathbb{R}^d$ with $\text{depth}(x; X) \geq k$. We call D_k the contour of depth k, although a stricter usage might reserve this phrase for the boundary of D_k. By the second line of (1.1), we have, equivalently, that D_k is the intersection of all the d-dimensional halfspaces containing $n + 1 - k$ points of the dataset X.

LEMMA 2.2. *The* depth *contours form a sequence of nested convex sets*: *Each* D_k *is convex and* $D_{k+1} \subset D_k$.

How many contours are there? That is, what is the maximum depth for a given dataset X? In $d = 1$, of course, the median is about $n/2$ deep. In $d > 1$, the maximum depth can be smaller than $n/2$; this depends on the shape of the dataset. We introduce some notation. Let

$$k^*(X) = \max_x \text{depth}(x; X)$$

and

$$k^+(X) = \max_i \text{depth}(X_i; X),$$

which are the maximum depth at any $x \in \mathbb{R}^d$ and at any $X_i \in X$, respectively. We say that a dataset is in general position if no more than d points lie in any $(d-1)$-dimensional affine subspace. In particular, a dataset in general position has no ties, no more than two points on any line, no more than three in any plane and so forth. Let $\lceil a \rceil$ denote the nearest integer greater than or equal to a and let $\lfloor a \rfloor$ denote the nearest integer less than or equal to a.

PROPOSITION 2.3. *If X is in general position, the maximum* depth $k^*(X)$ *lies between* $\lceil n/(d+1) \rceil$ *and* $\lceil n/2 \rceil$.

This lower bound is attained if the dataset is a strategically nested set of d simplices. See the discussion and figure in Section 4.1.

About $k^+(X)$ one can say in general only that $1 \leq k^+(X) \leq k^*(X)$, both possibilities occurring. If the dataset is nearly symmetric, the maximum depth will be much larger than $n/(d+1)$, in fact approximately $n/2$. We say that a probability distribution P is centrosymmetric about x_0 if $P(x_0 + S) = P(x_0 - S)$ for all measurable sets S.

PROPOSITION 2.4. *Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be a sample from an absolutely continuous, centrosymmetric probability distribution. Then $n^{-1}k^*(X^{(n)})$ converges in probability and almost surely to $1/2$ with increasing n. If, in addition, P has a positive density at x_0, then $n^{-1}k^+(X^{(n)})$ converges in probability and almost surely to $1/2$.*

In short, *if X is nearly symmetric, then the maximum depth is nearly* $1/2$. Actually, this principle is general and does not depend on probabilistic or asymptotic machinery. For example, using the language of Section 6.1, we can say that if the data have an empirical distribution lying within ε distance of some centrosymmetric distribution according to the "halfspace" metric, then the maximum depth $k^*(X)$ is at least $n(1/2 - \varepsilon)$. This shows that the distance from symmetry explicitly controls $k^*(X)$. It shows more. Using known facts about asymptotic properties of halfspace distance, one can easily show that

$$k^*(X^{(n)}) = n/2 - O_P(n^{1/2})$$

when $X^{(n)}$ is a random sample from an absolutely continuous, centrosymmetric distribution.

So there can be as many as $n/2$ depth contours if the dataset is nearly symmetric, but far fewer for highly asymmetric datasets.

What shape do depth contours have? This depends on the data. For example, if the data arise as a random sample from an ellipsoidal distribution, the contours are good estimates of the ellipsoid's shape.

LEMMA 2.5. *Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be a random sample from an elliptically symmetric distribution. The $\lfloor n\alpha \rfloor$-depth contour of $X^{(n)}$ converges, as*

$n \to \infty$, almost surely and in probability, to an ellipsoid of the same shape as that of the parent distribution, and a scale which depends on α.

("Convergence of contours" here refers to convergence of sets in Hausdorff distance.) For example, if the sample comes from the standard Gaussian distribution Φ_d on $\mathbb{R}^d$, the limiting shape of the $\lfloor n\alpha \rfloor$ contour will be a sphere of radius $R_\alpha = \Phi^{-1}(1 - \alpha)$, where Φ^{-1} denotes the inverse of the one-dimensional Gaussian distribution function. Thus, the contours of depth can play much the same role as the covariance ellipsoid in indicating the shape and orientation of data arising from ellipsoidal distributions.

In short, the contours of depth are convex and nested, they are coordinate-free, they track the shape of the dataset in a quite acceptable fashion for datasets with ellipsoidal symmetry and the maximum depth behaves as in the one-dimensional case for datasets with centrosymmetry.

For a picture of depths, see Figure 2.1. This shows the pattern of depths for a dataset consisting of 18 observations from a Normal distribution with one covariance and two outlying observations from a Normal distribution with another covariance. The figure shows a sequence of nested convex sets, giving the contours of depth $1, 2, \ldots$ up to depth 8. There are no values of depth 10 $(= n/2)$ because of the slight asymmetry in the sample.

The interest of depths from the point of view of robustness is clear from Figure 2.2. That figure presents, for the same dataset, the standard covariance estimate computed from the full dataset and the estimate computed from the 18 "good" observations. Comparing Figures 2.1 and 2.2, it is clear that the inner contours of depth reflect the covariance of the "good" data much better than does the covariance of the full dataset. A fact underlying some results of Section 3 is that, by adding k "bad" data points to a dataset, one can corrupt at most the k-outermost depth contours; the ones inside must still reflect the

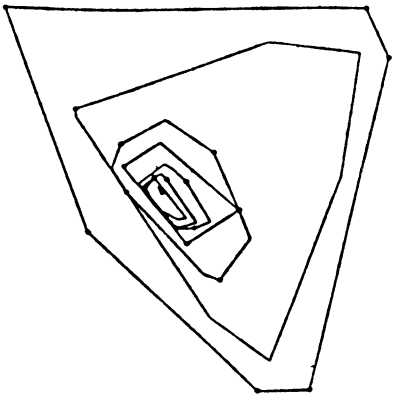
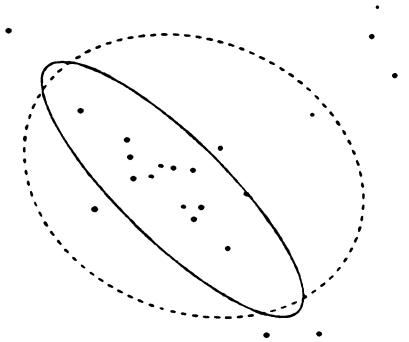


FIG. 2.1.



solid line = 18 points, dotted line = all 20 points.

FIG. 2.2.

shape of the "good" data. Thus statistics based only on data of depth $> k$ turn out to be robust against contamination by k or fewer outliers.

Outlyingness. The results just stated for depths have analogs for the outlyingness r_d:

1. Outlyingness is affine invariant:
$$r_d(Ax + b; \{AX_i + b\}) = r_d(x; X)$$
for every b and every nonsingular A.
2. The outlyingness "contours" $O_r = \{x: r_d(x; X) \leq r\}$ are convex and nested: $O_r \subset O_{r+h}$, $h > 0$.
3. Under random sampling from a centrosymmetric distribution, the minimum outlyingness is close to zero, with high probability, for large n.
4. Under random sampling from an ellipsoidal distribution P, the outlyingness contours converge to ellipsoids with the same shape as the ellipsoid of P.

Figure 2.3 displays the outlyingness for the data used in the earlier figures. The level sets are similar in shape to the covariance ellipse of the good data; the two outliers both have large outlyingness.

3. Breakdown properties of T_*, T_α and T_w. Using the notions of depth and outlyingness, we may define d-dimensional analogs of one-dimensional location estimates. The analog of the median is the deepest point T_* defined by

$$T_*(X) = \arg\max_x \text{depth}(x; X). \tag{3.1}$$

(When the depth does not have a unique maximum, any sensible rule for selecting among the maximum-depth values may be used without affecting the

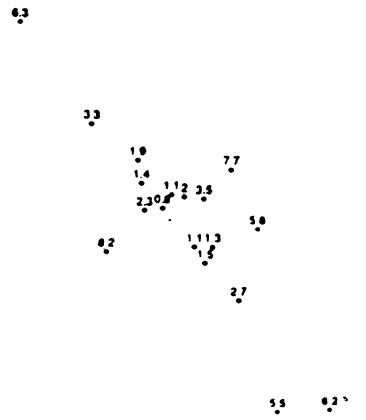


FIG. 2.3.

results given below; we propose "averaging"

$$T_*(X) = \text{Ave}\Big\{x\colon \text{depth}(x; X) = \max_x \text{depth}(x; X)\Big\}.$$

The analog of the α-trimmed mean is the α-depth-trimmed mean T_α, the average of all points which are at least $n\alpha$ deep in the sample

$$T_\alpha(X) = \text{Ave}\{X_i \in X\colon \text{depth}(X_i; X) \geq n\alpha\}. \tag{3.2}$$

The generalized W estimate was defined by (1.4).

These estimators have decent asymptotic properties. For example, they are consistent estimators of the center of symmetry of any centrosymmetric distribution, and they have $n^{-1/2}$ rates of convergence to their limiting values under weak regularity conditions.

It is easy to see that these estimators satisfy the affine equivariance condition

$$T(\{AX_i + b\}) = AT(X) + b$$

for every b and every nonsingular linear transformation A. Put otherwise, this means that they select the same point of space independent of the coordinate system put on the space. These three estimates have good breakdown properties. The breakdown point is, intuitively, the smallest amount of contamination necessary to upset an estimator entirely.

Our formal definition of finite-sample breakdown point is as in [D]. Let $X^{(n)}$ denote a given dataset of size n, at which the breakdown point is to be evaluated. Let T be the estimator of interest. Consider adjoining to $X^{(n)}$ another dataset $Y^{(m)}$ of size m. If, by strategic choice of $Y^{(m)}$, we can make $T(X^{(n)} \cup Y^{(m)}) - T(X^{(n)})$ arbitrarily large, we say that the estimator breaks down under contamination fraction $m/(n + m)$. The breakdown point $\varepsilon^*(T, X)$ is the smallest contamination fraction under which the estimator

breaks down:

$$\varepsilon^* = \min\left\{ \frac{m}{n+m} : \sup_{Y^{(m)}} |T(X \cup Y^{(m)}) - T(X)| = \infty \right\}.$$

For example, the breakdown point of the mean $\mathrm{Ave}(X)$ is $1/(n+1)$, while that of the one-dimensional median $\mathrm{Med}(X)$ is $1/2$. In colloquial terms, it takes only one (sufficiently) bad observation to corrupt an average, whereas it takes about 50% bad observations to corrupt the median. We note (Donoho, 1982) that for translation equivariant estimators, $\varepsilon^* \leq 1/2$, so the median has the best achievable breakdown point among location estimates. A fuller discussion of the breakdown concept is available in Donoho and Huber (1982), Rousseeuw (1985) and Lopuhaä and Rousseeuw (1991). In particular, these references discuss the replacement breakdown, which is different from the convention in this paper, and which has certain advantages; compare Rousseeuw and Leroy [(1987), page 117].

In discussing breakdown points, we begin by studying the estimator

$$T_{(k)}(X) = \mathrm{Ave}\{X_i : \mathrm{depth}(X_i; X) \geq k\}.$$

For this estimator, the criterion for depth-trimming is fixed at k, independently of sample size.

LEMMA 3.1. *If $k^+(X) \geq k$, then $T_{(k)}$ is well defined, its breakdown point is well defined and*

$$\varepsilon^*(T_{(k)}, X) = \frac{k}{n+k}.$$

The lemma shows that k^+ controls what robustness is possible using $T_{(k)}$. Now, as $T_\alpha(X^{(n)}) = T_{(\lfloor \alpha n \rfloor)}(X^{(n)})$, we can use this to get a result for T_α. The key idea is that, by Proposition 2.4, $k^+ \approx n/2$ under centrosymmetry.

PROPOSITION 3.2. *Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be a sample of size n from an absolutely continuous, centrosymmetric distribution on $\mathbb{R}^d$, with $d > 2$. Let $\alpha < 1/3$. With probability 1, for all n large enough, T_α is well defined and the breakdown point of $T_\alpha(X^{(n)})$ is well defined and*

$$\varepsilon^*(T_\alpha, X^{(n)}) \to \alpha, \quad a.s.$$

The limitation $\alpha < 1/3$ is real. In fact, even the maximal degree of depth trimming offered by T_* cannot give a breakdown point bigger than $1/3$.

PROPOSITION 3.3. *Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be a sample of size n from an absolutely continuous centrosymmetric distribution on $\mathbb{R}^d$, where $d > 2$. The breakdown point of $T_*(X^{(n)})$ converges almost surely to $1/3$ as $n \to \infty$.*

What happens if P is not centrosymmetric? Suppose that $k^*/n \to \beta \leq 1/2$, a.s. Then the argument for Proposition 3.2 will show that for each $\alpha < \beta/(1 + \beta)$, T_α is well defined and has a well-defined breakdown point for sufficiently large n and that the breakdown point has a.s. limit α. As for T_*, the following lower bound is always available, that is, without using probability or asymptotics.

PROPOSITION 3.4. *Let X be in a general position*:

$$\varepsilon^*(T_*, X) \geq \frac{1}{d+1}.$$

The outlyingness-weighted mean T_w has better breakdown properties, which do not depend on any near symmetry of X or on any probabilistic arguments.

PROPOSITION 3.5. *Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be a collection of points in general position. Suppose that $r \cdot w(r)$ is bounded and positive. Then the breakdown point of $T_w(X^{(n)})$ is $(n - 2d + 1)/(2n - 2d + 1)$.*

This is nearly the best possible result. It is relatively easy to show that *no* affine equivariant estimator can exceed a breakdown point of $(n - d + 1)/(2n - d + 1)$. For discussion of the maximal finite-sample breakdown points of affine equivariant estimators, see Lopuhaä and Rousseeuw (1991).

4. Methods which do not attain high breakdown point. It is not completely straightforward to construct estimators with a high breakdown point in high dimensions. The Maronna–Huber results establish this fact for M estimators. [D] gives several other examples of affine-equivariant estimators that seem, at first glance, "robust" but which do not have high breakdown points.

1. Iterative ellipsoidal trimming [Gnanadesikan and Kettenring (1972)] followed by mean.
2. Sequential deletion of apparent outliers [Dempster and Gasko (1981)] followed by mean.
3. Convex hull peeling [Bebbington (1978)] followed by mean.
4. Ellipsoidal peeling [Titterington (1978)] followed by mean.

It turns out in each case (but for different reasons) that these procedures *never* have a breakdown point exceeding $1/(d + 1)$. In this section we discuss why this happens in cases 2 and 3.

4.1. *Convex hull peeling.* Convex peeling is an intuitive and pretty idea. One takes the points lying on the boundary of a sample's convex hull, discards them, takes the boundary points of the remaining sample, peels those away and so on, until one decides that any outliers must have been removed, at

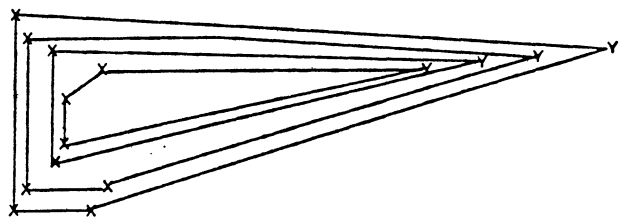


FIG. 4.1.

which point the mean of the remaining observations is taken as one's estimate of location.

Since the set of boundary points of X is affine invariant (affine transformations preserve membership in the boundary of the convex hull), so is the peeling procedure itself. If the rule for terminating the peeling iteration is affine invariant, the resulting peeled mean is affine equivariant. This procedure has close links to depth trimming, and many people who hear depth trimming described mistake it for convex peeling. Actually, the procedure has very different breakdown properties.

PROPOSITION 4.1. *If X is in general position, the breakdown point of any peeled mean is no better than*

$$\varepsilon^* \leq \frac{1}{d+1}\left(\frac{n+d+1}{n+2}\right).$$

The proof in [D] is very simple and we sketch it here. Note that each stage of peeling removes at least $d + 1$ points from the dataset, because a set of data points in general position has at least $d + 1$ extreme points. On the other hand, it is possible to arrange the contamination Y in such a fashion that the points removed at each stage of peeling contain only one point from Y; see Figure 4.1. In such a case, the peeling procedure removes at least d good data for every bad data point it succeeds in removing. Therefore if the fraction of bad points slightly exceeds $1/(d + 1)$, the set of observations remaining after peeling must contain bad points. On the other hand, as the picture shows, these bad points can be arbitrarily far from the X data without affecting the property that d good points are removed for every bad point. This means that the average of the points remaining after peeling can be arbitrarily far from the average of the X's, that is, breakdown.

Actually, this bound may be somewhat more favorable than what actually occurs in practice. If X represents a sample of size n from the Gaussian, [D] reports that the breakdown point appears to tend to zero as n increases. Intuitively, this is because peeling removes many more than the minimum $d + 1$ observations at each stage. Again, with strategically chosen contamination, only one of these need be a contaminating point, and so peeling has to

remove many more than d good points for each bad point successfully removed.

A result similar to Proposition 4.1 also holds for ellipsoidal peeling, for similar reasons [D]. The minimum volume ellipsoid containing a set of points has at least $d + 1$ points of the set on its surface. Only one of these need be bad.

One connection between depth trimming and peeling seems worth pointing out. Let $\text{peel}(x; X)$ denote the last stage in the peeling of X at which x is in the convex hull of the peeled sample. Thus if x is a boundary point of the convex hull, $\text{peel}(x; X) = 1$; if x is a bounding point of the convex hull of what remains after one peeling step, the $\text{peel}(x; X) = 2$; and so forth. In analogy with the deepest point, we may define the "maximally peeled mean"

$$T_p(X) = \text{Ave}\left\{X_i : \text{peel}(X_i; X) = \max_i \text{peel}(X_i; X)\right\}. \tag{4.1}$$

Referring to Donoho (1982), one can see that we must have

$$\varepsilon^*(T_p, X) \leq \varepsilon^*(T_*, X), \tag{4.2}$$

so that the breakdown point of depth trimming is always larger than that of peeling. At root, this derives from the inequality

$$\max_x \text{peel}(x; X) \leq \max_x \text{depth}(x; X), \tag{4.3}$$

which itself derives from the (easy) inequality $\text{peel}(x; X) \leq \text{depth}(x; X)$.

We also have

$$\varepsilon^*(T_p, X) - O\left(\frac{d}{n}\right) \leq \frac{1}{d+1} \leq \varepsilon^*(T_*, X). \tag{4.4}$$

Except for remainder terms, the best breakdown point of peeling is no better than the worst breakdown point of depth trimming.

For an example of an X giving approximate equality in (4.3), see Figure 4.2. This figure portrays a dataset of points at the vertices of a collection of nested simplices. In this case,

$$\max_x \text{peel}(x; X) = 4 = \max_x \text{depth}(x; X).$$

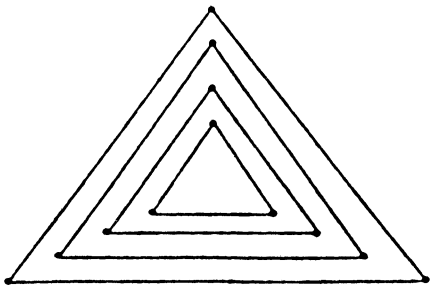


FIG. 4.2.

4.2. *Data cleaning.* Another method for robustifying the mean in high dimensions is based on sequential deletion of outliers. Using an affine-invariant discrepancy such as the Mahalanobis distance

$$(4.5) \qquad D^2(X_i; X) = (X_i - \text{Ave}(X))^T \text{Cov}^{-1}(X)(X_i - \text{Ave}(X)),$$

one identifies the observation which is most discrepant relative to the dataset X and removes it. Then one identifies the next most discrepant observation, using an average and covariance estimated from the data remaining after the first point was deleted and so on. At each stage, one identifies the most discrepant data point relative to the remaining data. At some point, one decides that all the outliers have been cleaned out of the data and takes the average of the remaining points.

Note that since D^2 is affine invariant, the resulting "cleaned mean" is affine equivariant, provided the rule for terminating the cleaning is affine invariant. However, the procedure again has low breakdown point.

PROPOSITION 4.2. *If X is in general position, the breakdown point of any cleaned mean is not larger than $1/(d + 1)$.*

The proof actually shows that with this amount of contamination, one can arrange the contaminating Y_j's so that every good point is cleaned out of the sample before any bad point is, even though the Y_j are arbitrarily outlying in some absolute (coordinate dependent) sense. Thus breakdown occurs in the worst possible way.

Our informal explanation for this goes as follows. In dimension d, "most" good data points will have $D^2 \approx d$. If a tight cluster of at least n/d outliers is placed far away from the good data, the D^2 for points in the cluster, one can check, is less than d because of the influence of this cluster on the $\text{Ave}(X)$ and $\text{Cov}(X)$ used in (4.5). Thus, the good points appear more discrepant than the bad ones.

We remark that the situation does *not* markedly improve if, instead of the estimates $\text{Cov}(X)$ and $\text{Ave}(X)$ used in (4.5), we employ "leave-one-out" estimates. That is, let $\text{Ave}(X_{(-i)})$ and $\text{Cov}(X_{(-i)})$ denote estimates of mean and covariance formed without using the ith data point Then, if these are used in place of $\text{Ave}(X)$ and $\text{Cov}(X)$ in (4.5), a breakdown bound similar to that of Proposition 4.2 still applies.

5. Discussion.

5.1. *Combining high breakdown with affine equivariance.* If one is willing to relax the affine equivariance condition (3.3) to, say, rigid-motion equivariance or simply location equivariance, it is easy to find estimators with a high breakdown point. For example, the simple coordinatewise median is location equivariant and has breakdown point $1/2$ in any dimension. The difficulty comes in being both coordinate-free and robust. When one is willing to adopt a

specific coordinate system, it is much easier to identify outliers than if one does not commit to such a specific choice.

In another direction, Tyler (1986) has shown that if one constrains the allowed contamination so that no two contaminating points can be at close angular distance, then M estimates cannot be broken down easily. But this is again a form of coordinate dependence, since the constraint on the contamination makes reference to a specific choice of coordinates.

5.2. *Other methods of attaining high breakdown.* Rousseeuw (1985) showed that the center of the minimum volume ellipsoid containing at least half the data provided a method with breakdown point of nearly $1/2$ in high dimensions. See the discussion in Lopuhaä and Rousseeuw (1991).

Oja (1983) introduced a notion of multivariate median based on simplicial volumes which is affine equivariant. It turns out to have interesting breakdown properties; see Lopuhaä and Rousseeuw (1991) and Niinimaa, Oja and Tableman (1989). [Liu (1990) introduced a different notion of multivariate depth, called simplicial depth. Unfortunately, this interesting method appears to have a rather low breakdown point.] See also Small (1987, 1990).

Simple geometric methods of obtaining high breakdown, such as the depth and outlyingness-based methods discussed here, the Rousseeuw's minimum volume ellipsoid approach or the Oja median may have inefficient statistical performance, in the sense of large sample asymptotics. Accordingly, attention has turned toward more sophisticated and less geometric estimators.

[D] showed it was possible to combine affine equivariance and high breakdown via suitably chosen minimum-distance estimates based on the so-called halfspace distance. Donoho and Liu (1988) have shown that this is a general phenomenon: In situations of invariance, certain minimum distance estimators have the best attainable breakdown point. Minimum distance estimates "automatically" possess root-n consistency, but they are not generally fully efficient, when the model is true.

Rousseeuw and Yohai (1984), Yohai (1987) and Davies (1987) have developed S-estimation techniques for combining high breakdown point with affine equivariance and asymptotic efficiency; see also Lopuhaä (1988). These techniques have the advantage of extending naturally from location estimation to other settings like regression fitting.

5.3. *Computational difficulty.* Some sort of computational breakthrough is necessary to make the estimators, as defined here, really practical. Adele Cutler has prepared, for $d = 2$, a program which computes the contour of depth $\lfloor n\alpha \rfloor$ in $O(n^2 \log n)$ time. The algorithm is based on the observation that for calculating depths, it is sufficient to restrict the search over projections in (1.1) and (1.3) to a finite number of projections, namely, to those projections which map d points of the dataset into the same value. In general, unfortunately, the algorithm runs in $O(n^{d+1} \log n)$ time in dimension d, so this approach is impractical for dimensions greater than 4 or 5. Souvaine and Steele (1987) have developed a number of promising techniques for speeding

up this sort of computation, based on properties of arrangements of hyperplanes in computational geometry.

Most high-breakdown methods are currently based on a computational approximation initiated by Rousseeuw (1984). One gives up the hope of exactly computing the suprema indicated at various places in the definition of one's estimator and in practice replaces the suprema by maxima over a discrete set of cases obtained by random search.

6. Proofs.

6.1. *Notation and background. Sets and datasets.* A dataset X is, for us, a collection of elements in which multiplicity counts. Despite this distinction from the traditional notion of set, we use the traditional set notation: For datasets we write $X = \{X_i\}$ and for mergers we write $X \cup Y$. Although we are abusing notation, we believe there is little risk of confusion. The only letters we use for datasets are X, Y and W. As an illustration, the reader may wish to prove the following fact about mergers:

$$\text{depth}(x; X) \leq \text{depth}(x; X \cup Y). \tag{6.0}$$

It is used several times below.

Halfspaces, empirical distributions and depths. Below, $H_{u,x}$ is the halfspace $\{y\colon u^T y \leq u^T x\}$, with interior $\text{int}\, H_{u,x} = \{y\colon u^T y < u^T x\}$ and boundary $\text{bdry}\, H_{u,x} = \{y\colon u^T y = u^T x\}$. Given data X_i, $i = 1, \ldots, n$, P_n is the empirical distribution, defined by $P_n(S) = n^{-1}\#\{i\colon X_i \in S\}$ for every measurable set S. The halfspace metric μ_H is used to compare empirical and theoretical distributions:

$$\mu_H(P_n, P) = \sup_{u,x} |P_n(H_{u,x}) - P(H_{u,x})|. \tag{6.1}$$

This is the largest discrepancy between P_n and P on any halfspace. We remark that μ_H has the Glivenko–Cantelli property: If $\{X_i\}$ are iid P, then

$$\mu_H(P_n, P) \to 0 \quad \text{a.s. as } n \to \infty. \tag{6.2}$$

See Steele (1978) or a book on empirical processes, such as Pollard (1984).

In order to discuss the limiting behavior of depth in large samples, we introduce the projected probability

$$\Pi(x) = \inf_u P(H_{u,x}). \tag{6.3}$$

This is the minimal probability attached to any halfspace containing x. We note that for the empirical version of Π,

$$\Pi_n(x) = \inf_u P_n(H_{u,x}), \tag{6.4}$$

we have the connection to depth

$$n^{-1}\,\text{depth}(x; X^{(n)}) = \Pi_n(x). \tag{6.5}$$

It is therefore of interest that we have the inequality

$$\sup_x |\Pi_n(x) - \Pi(x)| \le \mu_H(P_n, P) \tag{6.6}$$

which implies, for example, that

$$n^{-1}\,\text{depth}(x, X^{(n)}) \to \Pi(x) \quad \text{a.s.}, \tag{6.7}$$

thus Π represents the large-sample limit of n^{-1} depth.

Three lemmas about the behaviour of Π are useful. We state them here and prove them in Section 6.2.

LEMMA 6.1. *Π is an upper semicontinuous function of x. If P is absolutely continuous, Π is a continuous function of x.*

LEMMA 6.2. *If P is centrosymmetric about x_0, $\Pi(x_0) \ge 1/2$. If, in addition, P is absolutely continuous, $\Pi(x_0) = 1/2$.*

LEMMA 6.3. $\max_x \Pi(x) \ge 1/(d+1)$.

These lemmas, together with (6.5), imply that the maximum depth is about $n/2$ under centrosymmetry and is always $n/(d+1)$ or larger.

6.2. *Proofs of Lemmas* 6.1–6.3.

PROOF OF LEMMA 6.1. Note initially that for each closed halfspace H, the linear functional $P \to P(H)$ is upper semicontinuous for weak convergence. Also, as $\nu \to 0$, the measure $P(\cdot - \nu)$ converges weakly to P. As $H_{u,x} = H_{u,w} + (x - w)$, we get that $f_u(x) = P(H_{u,x})$ is upper semicontinuous (u.s.c.) in x. Now

$$\Pi(x) = \inf_u P(H_{u,x}) = \inf_u f_u(x). \tag{6.8}$$

Π is thus the infimum of a collection of u.s.c. functions; it is upper semicontinuous.

We now show that if P is absolutely continuous, then Π is also lower semicontinuous, hence continuous. Let $x_n \to x_0$ and let u_n be a sequence of directions satisfying $P(H_{u_n,x_n}) \le \Pi(x_n) + 1/n$. As the u_n all lie on the unit sphere in $\mathbb{R}^d$, they contain a cluster point. Extracting a subsequence if necessary, we may assume that u_n converges, to u, say. Now

$$P(H_{u,x_0}) - P(H_{u_n,x_n}) = \int I_{H_{u,x_0}} - I_{H_{u_n,x_n}}\, dP, \tag{6.9}$$

where I_S is the indicator function of the set S. The difference in indicator functions is dominated in absolute value by the constant 1, and, as $u_n \to u$, $x_n \to x_0$, the difference tends to zero almost everywhere $[P]$. By the dominated convergence theorem, it follows that $P(H_{u,x_0}) - P(H_{u_n,x_n}) \to 0$ as $u_n \to u$,

$x_n \to x_0$. We conclude that

$$\liminf_{n\to\infty} \Pi(x_n) = \liminf_{n\to\infty} P(H_{u_n,x_n}) = P(H_{u,x_0}) \geq \Pi(x_0). \tag{6.10}$$

Thus Π is lower semicontinuous. □

PROOF OF LEMMA 6.2. As $H_{u,x_0} = x_0 + H_{u,0}$, centrosymmetry of P about x_0 gives $P(H_{u,x_0}) = P(x_0 + H_{u,0}) = P(x_0 - H_{u,0}) = P(H_{-u,x_0})$. As $H_{u,x_0} \cup H_{-u,x_0} = \mathbb{R}^d$, we have $2P(H_{u,x_0}) = P(H_{u,x_0}) + P(H_{-u,x_0}) \geq 1$, so that $P(H_{u,x_0}) \geq 1/2$. If P is absolutely continuous, then $P(\text{bdry}\, H_{u,x_0}) = 0$ for every halfspace. Hence $P(H_{u,x_0}) + P(H_{-u,x_0}) = 1$, and so $P(H_{u,x_0}) = 1/2$. □

PROOF OF LEMMA 6.3. Let P_h denote the convolution of P with a Gaussian of width h. It is sufficient to establish the result for P_h. Indeed, P_h converges weakly to P as $h \to 0$; and, as in Lemma 6.1, for each closed halfspace H, the mapping $f_u(x, P) = P(H_{u,x})$ is upper semicontinuous for the product topology (Euclidean convergence, weak convergence). As Π is the infimum of f_u's, it is also upper semicontinuous and so

$$\sup_x \Pi(x) \geq \limsup_{h\to 0} \sup_x \Pi_h(x),$$

where $\Pi_h(x) = \inf_u P_h(H_{u,x})$. So if we show that

$$\sup_x \Pi_h(x) \geq \frac{1}{d+1}$$

for each h, the result for P follows.

P_h has two special properties. First, every projection has a strictly positive continuous density and so for each $R > 0$,

$$\beta \equiv \inf_u \inf_{|t|\leq R} \frac{d}{dt} P_h(H_{u,ut}) > 0. \tag{6.11}$$

Second, $P_h(H_{u,x})$ is uniformly continuous in u and x. Thus, for example, given $\varepsilon > 0$, we have $\delta > 0$ so that $|u - u_0| \leq \delta$ and $|x| \leq R$,

$$\left|P_h(H_{u,x}) - P_h(H_{u_0,x})\right| \leq \varepsilon. \tag{6.12}$$

In the remainder of the proof we drop the h subscript, although we depend on properties (6.11) and (6.12) for the proof we present.

The "contour" $\{x: \Pi(x) \geq \pi\}$ is the intersection of all halfspaces containing at least $1 - \pi$ of the probability of P. Hence Π has convex contours. As Π is continuous (Lemma 6.1) and has convex contours (which are easily seen to be bounded), there is a maximizer of Π. Let us suppose that 0 is a maximizer, that is, that

$$\Pi(0) = \pi^* = \sup_x \Pi(x). \qquad \square$$

CLAIM. For every direction v, $|v| = 1$, which we can consider moving away from 0, there exists a halfspace $H_{u,0}$ so that

$$P(H_{u,0}) = 1 - \pi^* \tag{P1}$$

and

(P2) $$P(H_{u,\alpha v}) \le P(H_{u,0}) \quad \text{for all } \alpha > 0.$$

PROOF. (P1) and (P2) are a consequence of the fact that 0 maximizes Π. They assert that for every v there is a u in the closed hemisphere with north pole v which attains $\sup_u P(H_{u,0})$. If this is not true, then we will derive a contradiction.

If the claim is not true, there exists a hemisphere S,with north pole v, say, that contains no maximizers of $P(H_{u,0})$. Consequently, by the reflection symmetry $P(H_{u,0}) = 1 - P(H_{-u,0})$, all minimizers are contained in S. Moreover, by the reflection symmetry of the boundary of a hemisphere, no maximizers and no minimizers of $P(H_{u,0})$ are contained in the boundary of S. Finally, by continuity of $P(H_{u,0})$ all minimizers are actually contained strictly in the interior of the hemisphere, in a polar cap $\mathbf{C} \subset S$, of opening less than 90°, with north pole v. Continuity in u gives

$$\inf_{\{u:\, u^T v \le 0\}} P(H_{u,0}) > \inf_{u \in \mathbf{C}} P(H_{u,0}) = \pi^*$$

and also

(6A) $$\inf_{u \in S \setminus \mathbf{C}} P(H_{u,0}) > \inf_{u \in \mathbf{C}} P(H_{u,0}) = \pi^*.$$

Continuity of $P(H_{u,x})$ in u and x then gives that for small enough $\alpha > 0$,

$$\inf_{\{u:\, u^T v \le 0\}} P(H_{u,\alpha v}) > \inf_{\{u:\, u^T v \ge 0\}} P(H_{u,\alpha v}),$$

so that

(6.13) $$\Pi(\alpha v) = \inf_{\{u:\, u^T v \ge 0\}} P(H_{u,\alpha v}).$$

On the other hand, it is easy to see that for u in the hemisphere with north pole v, $P(H_{u,\alpha v})$ is a monotone increasing function of α. In fact, we have by (6.11),

(6B) $$P(H_{u,\alpha v}) - P(H_{u,0}) \ge \beta \alpha u^T v$$

and so in particular

(6.14) $$\inf_{u \in \mathbf{C}} P(H_{u,\alpha v}) \ge \pi^* + \beta \alpha \cos(\gamma),$$

where γ is the opening of $\mathbf{C}$ ($\gamma < \pi/2$).

On the other hand

(6.15) $$\inf_{u \in S \setminus \mathbf{C}} P(H_{u,\alpha v}) > \inf_{u \in S \setminus \mathbf{C}} P(H_{u,0}) > \pi^*,$$

by monotonicity, (6A) and (6B). Combining (6.13)–(6.15), we conclude that

$$\Pi(\alpha v) > \pi^* = \Pi(0)$$

which contradicts the assumption that 0 is a deepest point. This contradiction establishes the claim.

We may recast the claim as follows. There exists a collection $\mathbf{H}_0 = \{H_{u_i,0}\}$ so that

$$P(H_{u_i,0}) = 1 - \pi^* \quad \text{for all } i \tag{P1}$$

$$\sup_i u_i^T v \geq 0 \quad \text{for all } v. \tag{P2}$$

Here the index i runs through a possibly infinite, possibly uncountable set. By judicious application of the Heine–Borel theorem, [DG] shows that there is a subcollection $\mathbf{H}_I$ of $\mathbf{H}_0$ still satisfying (P1) and (P2) but containing only finitely many halfspaces.

We now claim there is a subcollection $\mathbf{H}_J$ of $\mathbf{H}_I$ with no more than $d + 1$ halfspaces that satisfies (P1) and (P2). By Lemma 6.4, placing the condition (P2) on $\mathbf{H}_I$ is equivalent to saying that 0 is contained in the finite polyhedron $K_I = \text{Hull}(\{u_i,\ i \in I\})$ and that 0 is not an extreme point of that polyhedron.

By Caratheodory's theorem [Rockefellar (1970), page 155], if 0 has this property, then 0 can be expressed as a strict convex combination of $d + 1$ or fewer of the extreme points $\{u_i\colon i \in I\}$. Let J be the set of indices of the u_i's used in this combination. Then we have $0 = \sum_{j \in J} \theta_j u_j$ with $\theta_j > 0$, $\sum_J \theta_j = 1$. Now put $K_J = \text{Hull}(\{u_j\colon j \in J\})$. Then $0 \in K_J$ and 0 is not extreme in K_J. It follows by another application of Lemma 6.4 that

$$\max_{j \in J} u_j^T v \geq 0 \quad \text{for all } v. \tag{6.16}$$

Hence if we define $\mathbf{H}_J = \{H_{u_j,0}\colon j \in J\}$, we get a collection of halfspaces with properties (P1) and (P2), having cardinality

$$2 \leq \#J \leq d + 1, \tag{6.17}$$

the upper bound being furnished by Caratheodory's theorem, the lower bound by nonextremality of 0 in K_J. We now note that property (P2) is equivalent to

$$\bigcup_{j \in J} H_{-u_j,0} = \mathbb{R}^d. \tag{6.18}$$

Indeed, $-x \in H_{-u_j,0}$ if and only if $u_j^T x \geq 0$. Thus $-x$ is in some $H_{-u_j,0}$ if and only if $\max_j u_j^T x \geq 0$.

Because $H_{-u_j,0}$ is the complement of $\text{int}\, H_{u_j,0}$,

$$P(H_{-u_j,0}) = 1 - P(\text{int}\, H_{u_j,0}).$$

Invoking absolute continuity, $P(\text{bdry}\, H_{u_j,0}) = 0$, and applying (P1), we conclude that $P(H_{-u_j,0}) = \pi^*$. By this, (6.17) and (6.18) we have

$$1 = P(\mathbb{R}^d) = P\Big(\bigcup_{j \in J} H_{-u_j,0}\Big) \leq \sum_{j \in J} P(H_{-u_j,0}) = \#J \cdot \pi^* \leq (d + 1)\pi^*.$$

Thus $\pi^* \geq 1/(d + 1)$ as claimed. □

LEMMA 6.4. *Let* $\{u_i\}$ *be a finite collection of points in* $\mathbb{R}^d$ *none of which are zero. The following two properties are equivalent*:

(A) $$\max_i u_i^T v \geq 0 \quad \textit{for all } v.$$

(B) $\text{Hull}(\{u_i\})$ *contains* 0, *but* 0 *is not an extreme point*.

PROOF. See [DG]. □

6.3. *Proofs for Section* 2.

PROOF OF LEMMA 2.1. Membership of a point in a halfspace is coordinate-free: $X_i \in H_{u,x}$ if and only if $AX_i + b \in AH_{u,x} + b$ for every b and every nonsingular A. Consequently,

$$\min_{|u|=1} \#\{i\colon X_i \in H_{u,x}\} = \min_{|u|=1} \#\{i\colon AX_i + b \in AH_{u,x} + b\}.$$

By the second line of (1.1), this gives (2.1). □

PROOF OF LEMMA 2.2. A depth contour is the intersection of halfspaces and so is convex. Recall the definition of the depth contour D_k as the intersection of all halfspaces containing at least $n + 1 - k$ points. Now D_{k+1} is the intersection of all halfspaces containing at least $n - k$ points. Every halfspace containing $n + 1 - k$ contains $n - k$, so D_k is the intersection of a subfamily of the family defining D_{k+1}. As points in D_k satisfy a subset of the conditions which points in D_{k+1} must satisfy, $D_{k+1} \subset D_k$. □

PROOF OF PROPOSITION 2.3. For X in general position, there exists a projection direction v for which there are no ties in the projected dataset $\{v^T X_i\}$. In this projection, the maximum depth is $\lceil n/2 \rceil$. But

$$\text{depth}_d(x; X) = \min_{|u|=1} \text{depth}_1\big(u^T x; \{u^T x_i\}\big) \leq \text{depth}_1\big(v^T x; \{v^T X_i\}\big) \leq \lceil n/2 \rceil.$$

So $k^*(X) \leq \lceil n/2 \rceil$. The lower bound follows from Lemma 6.3 and the identity (6.5). □

PROOF OF PROPOSITION 2.4. As P is centrosymmetric and absolutely continuous, Lemma 6.1 implies $\Pi(x_0) = 1/2$. (6.5)–(6.7) give

$$n^{-1} \text{depth}\big(x_0; X^{(n)}\big) = \Pi_n(x_0) \to \Pi(x_0) = 1/2 \quad \text{a.s.}$$

Now P is absolutely continuous, so with probability 1, $X^{(n)}$ is in general position. Thus, by Lemma 2.3,

$$\left\lceil \frac{n}{2} \right\rceil \geq k^*(X^{(n)}) \geq \text{depth}\big(x_0; X^{(n)}\big).$$

Combining the last two displays we have $n^{-1}k^*(X^{(n)}) \to 1/2$ a.s.

Consider now $k^+(X^{(n)})$. Let X_{i_n} be the closest among $X_1, \ldots, X_n$ to x_0. By the positive density of P at x_0 and the Borel–Cantelli lemma, $\{X_{i_n}\}_{n=1}^{\infty}$

converges to x_0 almost surely. By (6.3) and (6.4),

$$n^{-1}k^+(X^{(n)}) \geq n^{-1}\,\mathrm{depth}\big(X_{i_n}; X^{(n)}\big) = \Pi_n\big(X_{i_n}\big) \geq \Pi\big(X_{i_n}\big) - \mu_H(P, P_n).$$

Because P is absolutely continuous, we may apply Lemma 6.2 to conclude that $\Pi(X_{i_n}) \to \Pi(x_0)$ a.s. The Glivenko–Cantelli property (6.2) and $k^+ \leq k^*$ yield $n^{-1}k^+ \to 1/2$ a.s. □

PROOF OF LEMMA 2.5. The proof for the Gaussian case amounts to the fact that $\Pi_n \to \Pi$ uniformly in x (6.6) and the formula $\Pi(x) = \Phi(-\sqrt{x^T\Sigma x})$, where Σ is the covariance of P and Φ is the the standard normal distribution function. The proof for other elliptically symmetric distributions is similar. □

6.4. *Proofs for Section* 3.

PROOF OF LEMMA 3.1. As X contains points of depth k, $T_{(k)}$ is well defined. Now the breakdown point of $T_{(k)}$ is well defined just in case $T_{(k)}(X \cup Y)$ is well defined for all Y. This will be the case if $X \cup Y$ contains points of depth k, for every choice Y, that is, if $k^+(X \cup Y) \geq k^+(X)$. This inequality follows from (6.8).

Now we show $\varepsilon^* \geq k/(n+k)$. For $T_{(k)}$ to break down at X, the contamination $Y = \{Y_i\}$ must be such that $T_{(k)}(X \cup Y)$ lies outside any fixed bounded set —for example, outside the convex hull of X. In order to place $T_{(k)}$ outside the convex hull of X, it must be possible to arrange the contamination Y so that there will be a contaminating point, say Y_1, with depth $(Y_1; X \cup Y) \geq k$ outside the convex hull of X. By the separating hyperplane theorem there will then be a direction u separating all the X_i's from Y_1: $\max_i u^T X_i < u^T Y_1$. But Y_1 is of depth k in $X \cup Y$, so that there must be at least k members in the combined dataset $X \cup Y$ whose projection on u lies to the right of Y_1. As none of these can be in X (by the last display), they must be in Y. Hence $\#Y \geq k$, and the contamination fraction must be at least $k/(n+k)$.

Finally, we show $\varepsilon^* \leq k/(n+k)$, that is, that k is a sufficient amount of contamination. Place $Y_1 \cdots Y_k$ on the same site. For every u, $u^T Y_1 = u^T Y_2 = \cdots = u^T Y_k$; therefore, $\mathrm{depth}(Y_i\colon\ X \cup Y) \geq k$, $i = 1, \ldots, k$. Thus $T_{(k)}(X \cup Y)$ is an average over a set containing all of Y. However, as we could choose Y_1 to have an arbitrarily large norm, $T_{(k)}(X \cup Y)$ can be made arbitrarily large. □

PROOF OF PROPOSITION 3.2. For $\alpha < 1/3$, pick $\beta \in (3/2\alpha, 1/2)$. By Lemma 2.3 $k^+(X^{(n)})/n \to 1/2$ a.s. This implies that there is a positive random variable $n_0(\beta)$ which is almost surely finite with

$$k^+(X^{(n)})/n > \beta \quad \text{for } n > n_0(\beta).$$

Let Y consist of m contaminating points, $m \leq n/2$. Then for $n > n_0(\beta)$, $k^+(X \cup Y) \geq k^+(X) > \beta n$. Now $T_\alpha(X \cup Y)$ is well defined if and only if $k^+(X \cup Y) \geq \lfloor \alpha(n+m) \rfloor$. But $\beta n > \lfloor \alpha(n+m) \rfloor$ for $m \leq n/2$, so $T_\alpha(X \cup Y)$ is well defined for $n > n_0(\beta)$.

For n, m fixed, $T_\alpha(X \cup Y) = T_{(k)}(X \cup Y)$, where $k = \lfloor \alpha(n+m) \rfloor$. By Proposition 3.1, Y can be chosen so that $T_{(k)}$ breaks down if and only if the contamination amount $m \geq k$, that is,

$$m \geq \lfloor \alpha(n+m) \rfloor. \tag{6.19}$$

For $\alpha < 1/3$, $m = n/2$ is always a solution of this inequality. Hence the restriction that Y have cardinality less than or equal to $n/2$, imposed earlier, does not prevent solving (6.19). The smallest value of m solving this inequality is either $m = \lfloor (\alpha/(1-\alpha))n \rfloor$ or $m = \lceil (\alpha/(1-\alpha))n \rceil$. It follows that for $n > n_0(\beta)$, the breakdown point is well defined and

$$\varepsilon^*\left(T_\alpha, X^{(n)}\right) = \frac{m}{n+m} = \alpha + O\left(\frac{1}{n}\right) \quad \text{a.s.} \qquad \square$$

PROOF OF PROPOSITION 3.3. First, we show that the limiting breakdown point is at least $1/3$. Now, m contaminating points are sufficient to cause breakdown only if they are sufficient to place $T_*(X \cup Y)$ outside the convex hull of the points in X. But, by the separating hyperplane argument of Lemma 3.1, if $T_*(X \cup Y)$ is outside the hull of X, the number of contaminating points must be at least the depth of $T_*(X \cup Y)$. Hence, for m to cause breakdown we must have

$$m \geq \operatorname{depth}(T_*(X \cup Y), X \cup Y) = k^*(X \cup Y) \geq k^*(X),$$

the last inequality following from (6.0). Thus $\varepsilon^* = m/(n+m) \geq k^*(X)/(n + k^*(X))$. Now $k^*(X) = n/2(1 + o_{\text{a.s.}}(1))$ so $k^*(X)/(n + k^*(X)) \to 1/3$ a.s. Hence $\liminf_n \varepsilon^* \geq 1/3$ a.s.

Next, we show that the limiting breakdown point is at most $1/3$. Let N be the counting measure $N(S) = \#\{i\colon X_i \in S\}$. Let x_0 be the point of centrosymmetry of P and put $k^0 = \max_u N(H_{u,x_0})$. Observe that $k^0 = n/2(1 + o_{\text{a.s.}}(1))$. Indeed, $N(H_{u,x_0}) = nP_n(H_{u,x_0})$ and, by absolute continuity and centrosymmetry of P, $P(H_{u,x_0}) = 1/2$ for all u. Thus

$$|k^0/n - 1/2| \leq \sup_u \left| P_n(H_{u,x_0}) - P(H_{u,x_0}) \right| \leq \mu_H(P_n, P) \to 0$$

by the Glivenko–Cantelli property (6.2) of μ_H.

Set $m = k^0 + 2d + 1$. We will prove in a moment that

$$\varepsilon^* \leq \frac{m}{n+m}. \tag{6.20}$$

As $m = n/2(1 + o_{\text{a.s.}}(1))$, (6.20) implies $\limsup_{n\to\infty} \varepsilon^*(T_*, X^{(n)}) \leq 1/3$ a.s.

It remains to prove (6.20). Let y be an arbitrary point in $\mathbb{R}^d$ and let $Y^{(m)}$ be a dataset consisting of m exact repetitions of y. Now $\operatorname{depth}(y, X \cup Y) \geq m$. We claim that y is the deepest point for $X \cup Y$:

$$\operatorname{depth}(x, X \cup Y) < m \qquad x \neq y. \tag{6.21}$$

As y is arbitrary, this will prove that $T_*(X \cup Y) = y$ has a solution for any $y \in \mathbb{R}^d$, and so T_* breaks down under contamination of size m.

Let M be the counting measure for Y: $M(S) = m$ if $y \in S$, $= 0$ else. Now

$$\operatorname{depth}(x; X \cup Y) = \inf_u (N(H_{u,x}) + M(H_{u,x})),$$

$$\leq \inf\{N(H_{u,x}) : M(H_{u,x}) = 0\}.$$

Now $N(H_{u,x_0}) \leq k^0$ for all u by definition of k^0. Invoking Lemma 6.5, there exists a particular u with $N(\operatorname{int} H_{u,x}) \leq k^0$ and $y \notin \operatorname{int} H_{u,x}$. Then, by Lemma 6.6, there exists a w with $N(H_{w,x}) \leq k^0 + 2d$ and $y \notin H_{w,x}$. As $y \notin H_{w,x}$, $M(H_{w,x}) = 0$, and so

$$\inf\{N(H_{u,x}) : M(H_{u,x}) = 0\} \leq N(H_{w,x}) \leq k^0 + 2d.$$

Combining the last two displays, together with $m > k^0 + 2d$, gives (6.21) and completes the proof of Proposition 3.3. □

LEMMA 6.5. *Let x be arbitrary and let x_0 be a point with $N(H_{u,x_0}) \leq k^0$ for every u. There is a u so that $N(\operatorname{int} H_{u,x}) \leq k^0$ and $\operatorname{int} H_{u,x}$ does not contain y.*

PROOF. Pick v so that $v^T x = v^T x_0$. Then $H_{v,x} = H_{v,x_0}$, and

$$N(H_{v,x}) = N(H_{v,x_0}) \leq \sup_w N(H_{w,x_0}) \leq k^0.$$

By the same argument $N(H_{-v,x}) < k^0$. As $\operatorname{int}(H_{v,x})$ and $\operatorname{int}(H_{-v,x})$ are disjoint, one of the two sets does not intersect y. Let u be one of v or $-v$, the choice being made so that $\operatorname{int} H_{u,x}$ does not intersect y. □

LEMMA 6.6. *Let X be in general position, $N(\operatorname{int} H_{u,x}) \leq k^0$, $y \notin \operatorname{int} H_{u,x}$. Then there exists w so*

$$N(H_{w,x}) \leq k^0 + 2d, \qquad y \notin H_{w,x}.$$

PROOF. Unless $y \in \operatorname{bdry} H_{u,x}$, there is nothing to prove. Hence we assume $u^T(y - x) = 0$. We will show that there is a w close to u so that $H_{w,x}$ has essentially the same properties as $H_{u,x}$ and does not contain y.

We say that w agrees with u if $(u^T X_i)(w^T X_i) \geq 0$ for all i. If w agrees with u, every point in X which is not on the boundary of $H_{u,x}$ or on the boundary of $H_{w,x}$ has the same membership or nonmembership in $H_{w,x}$ as it does in $H_{u,x}$. Thus, $N(H_{w,x} \Delta H_{u,x}) \leq N(\operatorname{bdry} H_{u,x}) + N(\operatorname{bdry} H_{w,x})$, where Δ denotes symmetric difference. As X is in general position, $N(\operatorname{bdry} H_{u,x}) \leq d$, so if w agrees with u,

$$N(H_{w,x}) \leq N(H_{u,x}) + N(H_{u,x} \Delta H_{w,x}) \leq N(H_{u,x}) + 2d \leq k^0 + 2d.$$

The lemma is therefore proved if we can show there is a w agreeing with u for which $y \notin H_{w,x}$. Let $\delta = \min\{|u^T(X_i - x)| : u^T(X_i - x) \neq 0\}$ and $M = \max|X_i - x|$. As X is a finite set, $\delta > 0$, $M < \infty$. Pick $\alpha \in (0, \delta/M)$. Set $w_0 = u + \alpha(y - x)$ and $w = w_0/|w_0|$. Now using $u^T(y - x) = 0$, we have by construction $w^T(y - x) > 0$. Thus $y \notin H_{w,x}$. On the other hand, one calcu-

lates that

$$\left|w^T(X_i - x) - u^T(X_i - x)\right| \le |w_0 - u|\max_i|X_i - x| \le \alpha M < \delta.$$

As $|u^T(X_i - x)| \ge \delta$ if $X_i \notin \text{bdry } H_{u,x}$, we have that

$$\text{if } \left|u^T(X_i - x)\right| \ne 0, \qquad \operatorname{sgn} w^T(X_i - x) = \operatorname{sgn} u^T(X_i - x).$$

In short w agrees with u and $y \notin H_{w,x}$. □

PROOF OF PROPOSITION 3.4. As in the last proposition, if m points are enough to break down T_*, $m \ge k^*(X \cup Y)$. By Proposition 2.3, $k^*(X \cup Y) \ge (n + m)/(d + 1)$. Hence $m/(n + m) \ge 1/(d + 1)$. □

PROOF OF PROPOSITION 3.5. The proof given in [D] has been published in Huber (1985). □

6.5. *Proofs for Section* 4.

PROOF OF PROPOSITION 4.1. We use two lemmas.

LEMMA 6.7. *Let $V = \{V_i\}$ be a nonempty dataset*

$$\text{(6.22)} \qquad \operatorname*{Max}_i D^2(V_i; V) \ge \operatorname*{Ave}_i D^2(V_i; V) = \text{Dim}(\text{Span}(V)),$$

where Dim(Span(V)) *is the dimension of the smallest affine subspace containing all the points of V.*

PROOF. See the technical report [DG]. □

LEMMA 6.8. *Let W be a nonempty dataset and let Y consist of a number of points all at the same site, Y_1, say.*

$$D^2(Y_1; W \cup Y) \le \frac{\#W}{\#Y}.$$

The inequality is strict if $\text{Range}(\text{Cov}(W)) = \mathbb{R}^d$.

PROOF. The basic updating formulas for $\text{Ave}(W \cup Y)$ and $\text{Cov}(W \cup Y)$ are

$$\text{Ave}(W \cup Y) = \frac{n}{n + m}\text{Ave}(W) + \frac{m}{n + m}Y_1$$

$$\text{Cov}(W \cup Y) = \frac{n}{n + m}\text{Cov}(W) + \frac{mn}{(n + m)^2}(Y_1 - \text{Ave}(W))(Y_1 - \text{Ave}(W)).$$

Put $e = Y_1 - \text{Ave}(W)$ and use the updating formulas to write $D^2(Y_1; W \cup Y)$ as

$$\sup_u \frac{n}{n + m} \Bigg/ \left\{ \frac{m}{n + m} + \frac{u' \text{Cov}(W) u}{(u'e)^2} \right\}.$$

This is less than n/m, strictly so if $u' \operatorname{Cov}(W)u > a > 0$ for all u of norm 1, that is, if $\operatorname{Range}(\operatorname{Cov}(W)) = \mathbb{R}^d$. □

PROOF OF PROPOSITION 4.2. Place the contamination $Y_1, \ldots, Y_m$ all at the same site, Y_1, say. It will be shown that if $m \geq n/d$, Y_1 may be chosen to be any point not in X and yet iterative deletion applied to $X \cup Y$ will produce

(6.23) The first n deleted points come from X.

(6.24) The remaining points come from Y.

Then, whatever rule we use for terminating the iterative deletion, the resulting estimate will be an average of terms including Y_j's. As Y_1 may be chosen to have an arbitrarily large norm, the estimator breaks down. Proposition 4.2 then follows from $m \geq n/d$.

$X^{(k)}$ will denote the part of X remaining in $X \cup Y$ after k deletions have been made. (6.23) and (6.24) will follow if for $1 \leq k \leq n$, $D^2(Y_1; X^{(k)} \cup Y) < \max_i D^2(X_i; X^{(k)} \cup Y)$. In fact, an even stronger result is true for any nonempty subset W of X:

$$D^2(Y_1; W \cup Y) < \max_i D^2(W_i; W \cup Y). \tag{6.25}$$

If $\#W = n - k$, then from $m \geq n/d$, using Lemma 6.8, $D^2(Y_1; W \cup Y) \leq d((n-k)/n)$. If $n - k \geq d + 1$, then $\operatorname{Span}(W \cup Y) = \mathbb{R}^d$ and the inequality is strict. If $n - k \leq d$, then $\operatorname{Dim}(\operatorname{Span}(W \cup Y)) = n - k$. In either case,

$$D^2(Y_1; W \cup Y) < \operatorname{Dim}(\operatorname{Span}(W \cup Y)).$$

Applying Lemma 6.7 with $V = W \cup Y$ gives (6.25). □

Acknowledgments. Part of this work was reported in the first author's Ph.D. qualifying paper at Harvard University. The authors are grateful to Persi Diaconis, Peter Huber, Peter Rousseeuw and John Tukey for various discussions. An Associate Editor improved substantially the proofs of Propositions 2.3 and 6.3. Adele Cutler developed the computer program to plot the depth contours shown here.

REFERENCES

BEBBINGTON, A. C. (1978). A method of bivariate trimming for robust estimation of the correlation coefficient. *J. Roy. Statist. Soc. Ser. C* **27** 221–226.

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

DEMPSTER, A. P. and GASKO, M. G. (1981). New tools for residual analysis. *Ann. Statist.* **9** 945–959.

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. (1982). 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.

DONOHO, D., JOHNSTONE, I., ROUSSEUW, P. and STAHEL, W. (1985). Comment on "Projection pursuit" by P. J. Huber. *Ann. Statist.* **13** 496–500.

DONOHO, D. L. and LIU, R. C. (1988). The "automatic" robustness of minimum distance estimators. *Ann. Statist.* **16** 552–586.

GASKO, M. and DONOHO, D. L. (1987). Multivariate generalization of the median and trimmed sum. I. Technical Report 133, Dept. Statistics, Univ. California, Berkeley.

GNANADESIKAN, R. and KETTENRING, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28** 81–124.

HUBER, P. J. (1977). Robust covariances. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds.) 165–191. Academic, New York.

HUBER, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414

LOPUHAÄ, H. P. (1988) Highly efficient estimates of multivariate location with high breakdown point. Technical report 88-184, Delft Univ. of Technology.

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-estimates of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

MOSTELLER, C. F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, Mass.

NIINIMAA, A., OJA, H. and TABLEMAN, M. (1990). On the finite-sample breakdown point of the bivariate median. *Statist. Probab. Lett.* **10** 325–328.

OJA, H. (1983). Descriptive statistics for multivariate distributions, *Statist. Probab. Lett.* **1** 327–332.

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

ROCKEFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.

ROUSSEEUW, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, Vol. B* (W. Grossman, G. Ch. Pflug, I. Vincze and W. Wertz, eds.) 283–297. Reidel, Dordrecht.

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

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outliner 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.

SMALL, C. G. (1987). Measures of centrality for multivariate and directional distributions. *Canad. J. Statist.* **15** 31–39.

SMALL, C. G. (1989). A survey of multidimensional medians. *Internat. Statist. Rev.* **58** 263–277.

STAHEL, W. A. (1981). Robust estimation: infinitesimal optimality and covariance matrix estimators. Ph.D. thesis, ETH, Zurich (in German).

SOUVAINE, D. and STEELE, J. M. (1987). Time and space efficient algorithms for least-median-of-squares regression. *J. Amer. Statist. Assoc.* **82** 794–801.

STEELE, J. M. (1978). Empirical discrepancies and subadditive processes. *Ann. Probab.* **6** 118–127.

TITTERINGTON, D. M. (1978). Estimation of correlation coefficients by ellipsoidal trimming. *J. Roy. Statist. Soc. Ser. C* **27** 227–234.

TUKEY, J. W. (1974a). T6: Order Statistics. In mimeographed notes for Statistics 411, Princeton Univ.

TUKEY, J. W. (1974b). Address to International Congress of Mathematicians, Vancouver.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.

TYLER, D. (1986). Breakdown properties of *M*-estimators of multivariate scatter. Research report, Rutgers Univ.

YOHAI, V. (1987). High breakdown point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF MARKETING AND
QUANTITATIVE STUDIES
SAN JOSE STATE UNIVERSITY
1 WASHINGTON SQUARE
SAN JOSE, CALIFORNIA 95192-0069