Multivariate Generalizations of the Median and Trimmed Mean, I

By

David L. Donoho
Miriam Gasko

Department of Statistics
University of California, Berkeley

Technical Report No. 133
December 1987

This work was supported by an NSF graduate fellowship, by NSF grant DMS-84-51753, and by NIH grant 5T32HL07365-10.

Department of Statistics
University of California
Berkeley, California

Multivariate Generalizations of the Median and Trimmed Mean, I

David L. Donoho
Miriam Gasko

Department of Statistics
University of California, Berkeley

ABSTRACT

We describe multivariate generalizations of the Median, Trimmed Mean, and W-estimates. The estimates are based on a geometric construction involving "projection pursuit". They are distinguished from other generalizations by the fact that they are both affine equivariant (coordinate-free) and have high breakdown point. Such a combination is not easy to find. For example, we show that various estimators based on rejecting apparent outliers and taking the mean of the remaining observations, have breakdown points not larger $1/(d+1)$ in dimension d, whereas the methods we propose obtain larger breakdown points in high dimensions. As an example, our version 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.

A sequel discusses asymptotic properties of these estimators, such as consistency and limiting distribution.

AMS Subject Classification. Primary 62F35. Secondary 62G05, 62H12.

Key Words and Phrases. Multivariate order statistics. Location estimates. Robustness. Breakdown point. Projection Pursuit. Halfspace Distance. Glivenko-Cantelli Property. Outlier Rejection.

Acknowledgements. 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 Rousseuw, and John Tukey for various discussions. Adele Cutler developed the computer program to plot the depth contours shown here. This work was supported by an NSF graduate fellowship, by NSF grant DMS-84-51753, and by NIH grant 5T32HL07365-10.

1. Introduction

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

$$depth_1(x;X) = \min(\#\{i : X_i \le x\},\ \#\{i : X_i \ge x\})$$

(see also Tukey (1977)). The d-dimensional depth of a point $x \in R^d$ in a d-dimensional dataset is the least depth of x in any 1-dimensional projection or "view" of the dataset. In detail, if we let u denote a vector in R^d of unit norm, then the dataset $\{u^T X_i\}$ is a one-dimensional projection of the dataset X, and we define

$$\begin{aligned} depth_d(x;X) &= \min_{|u|=1} depth_1(u^T x;\ \{u^T X_i\}) \\ &= \min_{|u|=1} \#\{i : u^T X_i \ge u^T x\} \end{aligned} \tag{1.1}$$

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 order statistic. Of course, in dimension one, 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* $\sim n/2$.

While Tukey's proposal does not seem to be widely known, it raises a number of interesting possibilities. First, it gives a way of defining the median in 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* $\ge n/10$, say.

The resulting notions of median, covariance estimate, and trimmed mean have a pair of properties pointed out by Donoho (1982). The first is that they are affine equivariant — i.e. they commute with translations and linear transformations of the data. The second is that these estimators 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.

This combination of properties (equivariance and robustness) is interesting because many other ways of defining location estimators lack one or both of these properties. Maronna (1976) and Huber (1976) found that affine-equivariant M-estimates of location / scatter have breakdown points bounded 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.

Thus 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 1-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

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

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

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

This is a measure of how outlying x is in the worst 1-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

$$T_w(X) = \Sigma w_i X_i / \Sigma w_i \tag{1.4}$$

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).

This paper is a shortened and somewhat improved version of Donoho (1982). It covers the results just mentioned. A companion paper discusses the asymptotic properties of these estimates -- con-

sistency, limiting distribution -- as well as the robustness against infinitesimal perturbations. It will emerge for example, that our generalization of the median has very similar properties to the 1-dimensional median in terms of equivariance, breakdown point, conditions for consistency and root-n consistency, and influence function. Whereas this paper is elementary in outlook and methods -- appropriate to the results we discuss -- the sequel uses a more demanding vocabulary and technique.

Contents of Paper

Section 2 covers properties of the *depth* and of outlyingness r. Section 3 covers the breakdown properties of the various estimators. Section 4 shows how other methods of constructing robust estimators do not provide the same breakdown properties. Section 5 discusses related results on breakdown properties; the need for a computational breakthrough in order to make our estimates practical; and some applications of the technology proposed here to other problems in multivariate analysis.

All proofs are contained in section 6. The so-called halfspace distance, from the theory of empirical processes, plays a key role in some of our proofs.

2. Depth and Outlyingness.

In this section, we mention some of the basic properties of depth and outlyingness.

Lemma 2.1.

depth is affine invariant:

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

for every b and every nonsingular linear transformation A.

In other words, depth is independent of the coordinate system chosen.

Let D_k be the set of all $x \in R^d$ with $depth(x;X) \geq k$. 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 equation (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

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

and

$$k^+(X) = \max_i depth(X_i;X)$$

these are the maximum depth at any $x \in R^d$ 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 3 in any plane, etc. Let $\lceil a \rceil$ denote the nearest integer $\geq a$, and let $\lfloor a \rfloor$ denote the nearest integer $\leq a$.

Proposition 2.3.

If X is in general position, the maximum depth $k^*(X)$ lies between $\left\lceil \frac{n}{d+1} - \frac{d}{2} \right\rceil$ *and* $\left\lceil \frac{n}{2} \right\rceil$.

The proof of the lower bound is the most difficult proof in this paper. We don't know of any case where it is attained. We conjecture that the sharp bound is $\left\lceil \frac{n}{d+1} \right\rceil$. This larger bound is attained if the dataset is a strategically nested set of d-simplices. See the discussion and figure in section 4.1 below. About $k^+(X)$ one can in general say only that $1 \le k^+(X) \le k^*(X)$, 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 $X^{(n)} = \{X_1, \cdots, X_n\}$ 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 $n \cdot (1/2 - \varepsilon)$. 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. However, we defer discussion of such facts to the sequel.

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 $X^{(n)}=\{X_1, \cdots, X_n\}$ 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 distrubtion Φ_d on $\mathbf{R}^d$, 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 1-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.

Depth contours can be informative even when ellipsoidal symmetry fails to hold. For example, suppose that the probability distribution P is centrosymmetric about x_0. Then, given a sequence of random samples from P, the $\lfloor n\alpha \rfloor$-depth contours converge to certain convex sets which are symmetric about x_0. Detailed information about the limiting behavior of depth contours is provided by the sequel to this paper.

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 1-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 2 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,... up to depth 8. There are no values of depth 10 (= $n/2$) because of the slight asymmetry in the sample.

Figure 2.1 about here

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 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.

Figure 2.2 about here

Outlyingness

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

(1) *Outlyingess 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"* $O_r = \{x : r_d(x;X) \le r\}$ are convex and nested: $O_{r+h} \subset O_r$, $h > 0$.

(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 illustrates the outlyingness contours for the dataset used in the earlier figures. They are similar in shape to the covariance ellipse of the "good" data; the two outliers both have large outlyingnesses.

Figure 2.3 about here

3. Breakdown properties of $T*$, T_α, and T_w

Using the notions of depth and outlyingness, it is possible to define d-dimensional analogs of 1-dimensional location estimates. For the analog of the median, we have the *deepest point*, $T*$, defined by

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

(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 results given below; we propose "averaging"

$$T_*(X) = Ave\,\{x:\ depth(x;X) = \max_x depth(x;X)\}.)$$

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

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

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 generally have $n^{-1/2}$ rates of convergence to their limiting values. These facts are discussed in the sequel to this paper.

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.

Donoho (1982) found these three estimates to have good breakdown properties. The breakdown point is, intuitively, the smallest amount of contamination necessary to upset an estimator entirely.

Our formal definition of the breakdown point is as follows (Donoho, 1982). 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:

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

For example, the breakdown point of the mean $Ave(X)$ is $1/n$, while that of the 1-dimensional median $Med(X)$ 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).

Adapting results in Donoho (1982) yields a number of facts about the breakdown point of our estimates. It turns out to be most natural to begin by studying the estimator $T_{(k)}(X) = Ave\{X_i : depth(X_i; X) \geq k\}$. For this estimator, the depth-trimming does not change with 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 $X^{(n)} = \{X_1, \cdots, X_n\}$ be a sample of size n from an absolutely continuous, centrosymmetric distribution on $\mathbf{R}^d$, 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. For this breakdown point we have*

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

as $n \rightarrow \infty$.

The limitation $\alpha < 1/3$ is real. In fact, no amount of depth trimming can give a breakdown point bigger than 1/3.

Proposition 3.3. *Let $X^{(n)} = \{X_1, \cdots, X_n\}$ be a sample of size n from a centrosymmetric distribution*

on $\mathbf{R}^d$, 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 $k^*/n \to_{a.s.} \beta \le 1/2$. Then the argument for Proposition 3.2 will show that for large n, T_α is well-defined and has a well-defined breakdown point for $\alpha < \beta/(1+\beta)$; and that the limiting breakdown point is α. As for T_* the following lower bound is always available -- i.e. without using probability or asymptotics.

Proposition 3.4. *Let* X *be in general position.*

$$\varepsilon^*(T_*, X) \ge \frac{1}{d+1}\left[\frac{n-d(d+1)/2}{n-d/2}\right].$$

This is near sharp, so that without additional hypotheses on X, we can only conclude that the breakdown point of T_* is roughly $1/(d+1)$ or better.

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* $X^{(n)} = \{X_1, \cdots, X_n\}$ *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*

$$\frac{n-2d+1}{2n-2d+1}.$$

This may well be the best possible result. It is relatively easy to show that *no* affine equivariant estimator can exceed a breakdown point of

$$\frac{n-d+1}{2n-d+1}.$$

4. Methods which do not attain high breakdown point

The significance of these results comes from the fact, indicated earlier, that it is not easy to find estimators with a high breakdown point in high dimensions. The Maronna/Huber results establish this fact for M-estimators. Donoho (1982) gives several other examples of affine-equivariant estimators that seem, at first glance, "robust" but which are do not have high breakdown points.

[A] Iterative ellipsoidal trimming (Gnanadesikan and Kettenring, 1972), followed by mean.

[B] Sequential deletion of apparent outliers (Dempster and Gasko, 1981), followed by mean.

[C] Convex Hull peeling (Bebbington, 1978) followed by mean.

[D] 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 tsection we discuss why this happens in cases [B] and [C].

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 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, as Donoho (1982) explains in detail, 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*

$$\varepsilon^* \leq \frac{1}{d+1}\left\lceil \frac{n+d+1}{n+2} \right\rceil$$

The proof is very simple; 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.

Figure 4.1 about here

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 -- i.e. 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, Donoho (1982) 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.

We remark that a result similar to Proposition 4.1 also holds for ellipsoidal peeling, for similar reasons.

One connection between depth trimming and peeling seems worth pointing out. Let $peel(x;X)$ 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, $peel(x;X)=1$; if x is a bounding point of the convex hull of what remains after one peeling step, the $peel(x;X)=2$; etc. In analogy with the deepest point, we may define the "maximally-peeled mean"

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

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

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

so that the breakdown point of depth trimming is always larger than that of peeling. This can be strengthened to

$$\varepsilon^*(T_p,X) - O(\frac{d}{n}) \;\le\; \frac{1}{d+1} \;\le\; \varepsilon^*(T_*,X) + O(\frac{d^2}{n}). \tag{4.3}$$

Thus, 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

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

Figure 4.2 about here

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

$$D^2(X_i;X) = (X_i - Ave(X))^T Cov^{-1}(X)(X_i - Ave(X)), \tag{4.4}$$

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 be 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 $D^2 \simeq d$. 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 $Ave(X)$ and $Cov(X)$ used in (4.4). 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 $Cov(X)$ and $Ave(X)$ used in (4.4), we employ "leave-one-out" estimates. That is, let $Ave(X_{(-i)})$ and $Cov(X_{(-i)})$ denote estimates of mean and covariance formed without suing the i-th data point. Then, if these are used in place of $Ave(X)$ and $Cov(X)$ in (4.4), a breakdown bound similar to that of Proposition 4.2 still applies.

Remark

Ideas like peeling and cleaning, while they do not give high breakdown, may not give particularly good estimators in any of the traditional senses, either. One simply doesn't know -- it is difficult to analyze such sequential-deletion methods. In contrast, the high-breakdown methods introduced here can be expressed as functionals of the empirical distribution, and so it is possible to analyze their asymptotic properties in detail, as the sequel to this paper shows.

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 does not take very much sophistication for an estimator to have 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 (1985) has shown that if one constrains the allowed contamination so that no two contaminating points can be at close angular distance, then M-estimates can not 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.

For a general discussion of the relation between affine equivariance and robustness, consult Donoho, Rousseuw, and Stahel (Forthcoming).

5.2. Other methods of attaining high breakdown

Donoho (1982) showed it was possible to attain high breakdown via suitably-chosen minimum-distance estimates based on the so-called halfspace distance. Donoho and Liu (1986) have shown that this is a general phenomenon: in situations of invariance, minimum distance estimators have the best attainable breakdown point.

5.3. On the need for projection pursuit

The examples in this paper, and the minimum-distance estimates just referenced, all depend on projection pursuit in some way. Is it necessary to use projection pursuit to get high breakdown? Donoho et al. (1985) point out that it is not, but that in some sense the information needed to avoid breakdown is contained in the projections, and so projection pursuit very naturally is related to the breakdown problem.

5.4. Regression

Phenomena similar to those occuring in the location problem happen also in regression. According to Maronna and Yohai (1978), the breakdown point of regression M-estimators is never better than $1/\sqrt{d}$. Rousseuw (1984) exhbited an estimator which could do much better -- which in fact had a breakdown point approaching 1/2. He called this the least median-of-squares estimator.

Several of our remarks about location apply also to regression. For example, the least median of squares estimator has a projection pursuit character. The arguments of Donoho and Liu (1986) can be used to show that in regression as well as in location, a suitable minimum distance estimator will have the best possible breakdown point. This minimum distance estimator also has a projection-pursuit character.

5.5. Covariance Estimation

Donoho (1982) showed that the obvious analogs of T_α and T_w for covariance estimation:

$$C_\alpha(X) = Ave\,\{X_i X_i^T : depth\,(X_i;X) \geq n\,\alpha\}$$

$$C_w(X) = \frac{\Sigma_i X_i X_i^T w\,(X_i;X)}{\Sigma_i w\,(X_i;X)}$$

are affine equivariant and could have high breakdown points if X and w satisfied certain conditions. Compare also Stahel (1981).

5.6. Discriminant Analysis

There is a curious but perhaps illuminating link between the depth contours and linear discriminant analysis. Suppose we have two datasets X and Y (of sizes n and m) and want to form a linear rule which best discriminates between them. Suppose this rule should misclassify no more than $n\,\alpha$ of the X's as Y's; we want, subject to this constraint, to misclassify as few Y's as possible. Let β denote the best achievable misclassification rate for Y's subject to the constraint on misclassifying X's.

A solution to this problem is a linear rule which defines a hyperplane having no more than $n\,\alpha$ X's on one side, and $m\,\beta$ Y's on the other. This hyperplane can be obtained from the depth contours of X and Y. It is a hyperplane tangent to the $n\,\alpha$ depth contour of the X-dataset and also to the $m\,\beta$ depth contour of the Y-dataset.

Figure 5.1 about here

It may help to see an illustration with real data. Figure 5.2 presents depth contours for data on Chemical and Overt diabetics obtained from Jerry Halpern and the late Rupert Miller of Stanford University (Reaven and Miller, 1979). Contours are drawn for three different groups: overt diabetics ("o"), chemical diabetics ("c") and normal patients ("n"). Contours are drawn at depth 4,6,8,10, etc. Because in no case do the contours of depth 4 for one group overlap with the contours of depth 4 for another group, we can see that each cloud can be separated from either of the others by a linear rule making at most 4 misclassifications.

Figure 5.2 about here

The fact that depth contours are involved in optimal linear discrimination between two groups makes it at least plausible that they are good at separating outliers from good data. This motivates their usefulness in robust estimation.

5.7. 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.

Souvane and Steele (1986) have developed a number of promising techniques for speeding up this sort of computation, so there is perhaps prospect of doing better in the future.

6. Proofs

6.1. Notation and Background

Halfspaces, Empirical Distributions, and Depths. Below, $H_{u,x}$ is the *halfspace* $\{y : u^T y \le u^T x\}$, with interior $int H_{u,x} = \{y : u^T y < u^T x\}$ and boundary $bdry H_{u,x} = \{y : u^T y = u^T x\}$. Given data X_i, $i = 1, \ldots, n$, P_n is the *empirical distribution*, defined by $P_n(S) = n^{-1} \#\{i : X_i \in S\}$ 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

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

see (Steele, 1977) or a book on empirical processes, such as Pollard (1984). See also the sequel, where the halfspace metric is employed throughout.

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 following connection to depth:

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

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

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

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

Three lemmas about the behaviour of Π are useful below; 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. *If P is absolutely continuous* $\max_x \Pi(x) \ge 1/(d+1)$.

Properly applied, these lemmas imply that the maximum depth is about $n/2$ under centrosymmetry, and is always about $n/(d+1)$ or larger.

Sets and Datasets. A *dataset* X is, for us, an object which does not quite fit in with traditional mathematical concepts. We like to think of it as a bag containing slips of paper, each slip with one data point X_i written on it. We can count the number of elements: $\#X$ is just the number of slips of paper in the bag. We can merge two datasets X and Y: $X \cup Y$ is gotten by pouring all the slips from both bags together into one big bag. We index the slips of paper in the bag in order to keep track of them, so that X_1 is the first slip of paper, X_2 the second, and so on; but the indexing is arbitrary.

A dataset is something like a set. Thus, there is little risk of confusion if we say things like "outside the convex hull of X" or "if $y \notin X$". For the first statement we mean: "outside the convex hull of the set in R^d consisting of every point named on some slip of paper in the bag". For the second statement we mean: "if the value y is not named on any slip of paper in the bag".

On the other hand, a dataset is not a set. Sets have elements without multiplicity. In a dataset, a given value may be written down on several slips of paper, and so occur with multiplicity. Thus we can say "let the dataset Y consist of m repetitions of Y_1".

Despite this distinction, we use the traditional set notation: for datasets we write $X = \{X_i\}$, and for mergers we write $X \cup Y$. We have not been able to find an acceptable substitute, and although we are abusing notation, we believe that in what follows there is little risk of confusion. Note that 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:

$$depth(x;X) \le depth(x;X \cup Y). \tag{6.8}$$

It is used several times below.

6.2. Proofs for Section 2.

Proof of Lemma 2.1. Membership in a halfspace is coordinate-free: $X_i \in H_{u,x}$ iff $AX_i + b \in AH_{u,x} + b$ for every b and every nonsingular A. Consequently,

$$\#\{i : X_i \in H_{u,x}\} = \#\{i : AX_i + b \in A\,H_{u,x} + b\}$$

for every u and x, and so

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

By the second line of (1.1), this gives

$$depth(x; X) = depth(Ax + b; \{AX_i + b\}).$$

Proof of Lemma 2.2. A depth contour is the intersection of half spaces 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 Lemma 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,

$$\max_t depth_1(t; \{v^T X_i\}) = \lceil n/2 \rceil.$$

But

$$\begin{aligned} depth_d(x; X) &= \min_{|u|=1} depth_1(u^T x; \{u^T x_i\}) \\ &\le depth_1(v^T x; \{v^T X_i\}) \\ &\le \max_t depth_1(t; \{v^T X_i\}) = \lceil n/2 \rceil. \end{aligned}$$

So $k^*(X) \le \lceil n/2 \rceil$.

It remains to establish the lower bound.

Let $P_{n,h}$ be the empirical distribution convolved with an isotropic Gaussian measure of variance h^2, h small. Then $P_{n,h}$ is absolutely continuous. Let $\Pi_{n,h}$ denote the projected probability defined from $P_{n,h}$ as in (6.3). It follows from Lemma 6.3 (proved below) that $\max_x \Pi_{n,h}(x)$ is at least $1/(d+1)$. Now the same reasoning that leads to (6.6) gives

$$\sup_x |\Pi_{n,h}(x) - \Pi_n(x)| \le \mu_H(P_{n,h}, P_n)$$

and from this we have that

$$|\max_x \Pi_{n,h}(x) - \max_x \Pi_n(x)| \le \mu_H(P_{n,h}, P_n). \tag{6.9}$$

At this point we note that if $X^{(n)}$ is in general position, then

$$\lim_{h \to 0} \mu_H(P_{n,h}, P_n) = d/2n. \tag{6.10}$$

Indeed, the limit is just half the largest fraction of mass in any jump of any 1-dimensional distribution $F_{n,u}(t)=P_n(H_{u,\alpha_u})$. By general position, no such jump has more than d points.

Combining now (6.9) and (6.10) we have

$$\max_x \Pi_n(x) \geq 1/(d+1)-d/2n.$$

So from (6.5),

$$k^*(X^{(n)}) \geq n/(d+1)-d/2.$$

Proof of Lemma 6.1. 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 $H_{u,x_0} \cup H_{-u,x_0} = \boldsymbol{R}^d$, we have $2P(H_{u,x_0}) = P(H_{u,x_0})+P(H_{-u,x_0}) \geq 1$, so that $P(H_{u,x_0}) \geq 1/2$.

Let P be absolutely continuous; then $P(bdry\, H_{u,x_0}) = 0$ 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 6.2. Now

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

where $f_u(x)=P(H_{u,x})$. By an indicator function argument such as that in the next paragraph, f_u is a continuous function of x. Π is thus the infimum of a collection of continuous functions; it is upper semicontinuous.

We now show that Π is lower semicontinuous. The two semicontinuities then imply the desired result. Let $x_n \to x_0$ and let u_n be a sequence of directions satisfying $P(H_{u_n,x_n}) \leq \Pi(x_n) + \frac{1}{n}$. As the u_n all lie on the unit sphere in $\boldsymbol{R}^d$, they contain a cluster point. Extracting a subsequence if necessary, we may assume that u_n converges, to u, say. Now

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

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

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

Thus Π is lower semicontinuous.

Proof of Lemma 6.3. Fix $R \gg 1$. Let P_h denote the convolution of P with a Gaussian of width h. Then it is easy to see that

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

Also, $P_h(H_{u,x})$ 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 $|x| \leq R$,

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

It turns out to be sufficient to establish the result for P_h. Indeed, by an upper semicontinuity argument

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

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

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

for each h, the result for P follows. In the remainder of the proof we drop the h subscript, although we depend on properties (6.11)-(6.12) for the proof we present.

We remark that by an argument like that for depth, 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, i.e. that

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

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

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

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 $\sup_u P(H_{u,0})$. Suppose this is not true; 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 $C \subset S$, of opening less than 90 degrees, with north pole v. Continuity in u gives

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

and also

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

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

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

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)

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

and so in particular

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

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

On the other hand

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

By monotonicity and an earlier display. 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

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

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

Here the index i runs through a possibly infinite, possibly uncountable set.

Our plan is now to find a subcollection of $\mathbf{H}_0$ having only $d+1$ halfspaces, but with properties (P1) and (P2). We begin by extracting a a subcollection $\mathbf{H}_I$ of $\mathbf{H}_0$ still satisfying (P1)-(P2) but containing only finitely many halfspaces. We argue as follows. Let $S_i = \{v: u_i^T v \geq 0, \, |v|=1\}$ be the hemisphere of directions with pole u_i. Then (P2) says that $\{S_i\}$ covers the sphere $\{|v|=1\}$ in R^d. The sphere is compact, so the Heine-Borel theorem says that $\{S_i\}$ contains a finite subcover $\{S_i: i \in I\}$, where I is a finite set of indices. Let $\mathbf{H}_I = \{H_{u_i,0}: i \in I\}$ be the corresponding collection of halfspaces. As $\{S_i: i \in I\}$ has the covering property, $\mathbf{H}_I$ has property (P2). As $\mathbf{H}_I$ inherits (P1) from $\mathbf{H}_0$, we have now a finite collection of halfspaces with properties (P1)-(P2).

We now claim there is a subcollection $\mathbf{H}_J$ of $\mathbf{H}_I$ with no more than $d+1$ halfspaces that satisfies (P1)-(P2). By Lemma 6.4 below, placing the condition (P2) on $\mathbf{H}_I$ is equivalent to saying that 0 is contained in the finite polyhedron

$$K_I = Hull(\{u_i, i \in I\}),$$

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 convex combination of $d+1$ or fewer of the extreme points $\{u_i: i \in I\}$. Let J be the set of indices of the u_i's used in this combination. Then we have

$$0 = \sum_{j \in J} \theta_j u_j. \tag{6.16}$$

with $\theta_j > 0$, $\sum_J \theta_j = 1$. Put now $K_J = Hull(\{u_j: j \in J\})$. Then by (6.16) $0 \in K_J$ and 0 is not extreme in

K_J. It follows by another application of Lemma 6.4 below that

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

Define now

$$\mathbf{H}_J = \{H_{u_j,0}: j \in J\}.$$

This is 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 non-extremality of 0 in K_J. We now note that property (P2) is equivalent to

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

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

Because $H_{-u_j,0}$ is the complement of $int\, H_{u_j,0}$, $P(H_{-u_j,0}) = 1 - P(int\, H_{u_j,0})$. Invoking absolute continuity, $P(bdry\, H_{u_j,0}) = 0$, and applying (P1), we conclude that $P(H_{-u_j,0}) = \pi^*$. By this, (6.17) and (6.18) we have

$$\begin{aligned} 1 = P(R^d) &= P(\cup_{j \in J} H_{-u_j,0}) \\ &\leq \sum_{j \in J} P(H_{-u_j,0}) \\ &= \#J \cdot \pi^* \leq (d+1)\pi^*. \end{aligned}$$

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

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

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

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

Proof. (B) implies (A). Suppose that 0 is in the hull of the u_i. Then by definition of the convex hull of a finite set,

$$0 = \sum_i \theta_i u_i$$

where $\theta_i \geq 0$ and $\Sigma \theta_i = 1$. Then

$$\max_i u_i^T v \geq \Sigma \theta_i u_i^T v$$
$$= 0^T v_0 = 0.$$

(A) implies (B). Suppose (A) held, but 0 were not in the hull of the u_i. Then by a separating hyperplane argument, there would be some v with

$$u_i^T v > 0 \text{ for all } i.$$

But then putting $v' = -v$ we get

$$\max_i u_i^T v' < 0,$$

contradicting (A). So 0 must lie in the hull of the u_i. Also, 0 cannot be extreme; as the set $\{u_i\}$ is finite, it is contains all extreme points of its hull. But by hypothesis, no u_i is zero.

Proof of Proposition 2.4.

As P is centrosymmetric and absolutely continuous, Lemma 6.1 implies $\Pi(x_0) = 1/2$.

Recalling (6.5)-(6.7) we have

$$n^{-1} \mathit{depth}\,(x; X^{(n)}) = \Pi_n(x_0) \rightarrow_{a.s.} \Pi(x_0) = 1/2.$$

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

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

Combining the last 2 displays we have

$$n^{-1} k^*(X^{(n)}) \rightarrow_{a.s.} \frac{1}{2}.$$

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.

Now by (6.3) and (6.4)

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

Because P is absolutely continuous, we may apply Lemma 6.2 to conclude that $\Pi(X_{i_n}) \rightarrow_{a.s.} \Pi(x_0)$. Then by the Glivenko-Cantelli property (6.2) we have

$$\liminf_{n \rightarrow \infty} n^{-1} k^+(X^{(n)}) \geq_{a.s.} \liminf_{n \rightarrow \infty} \Pi(X_{i_n}) = \Pi(X_0) = \frac{1}{2}.$$

As $k^+ \le k^*$ we conclude $n^{-1}k^+ \to_{a.s.} \frac{1}{2}$.

Proof of Lemma 2.5.

The proof for the Gaussian case is presented in the sequel to this paper following Lemma 2.2 there. The proof for any other elliptically symmetric distribution is similar.

6.3. 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, i.e. if $k^+(X \cup Y) \ge k^+(X)$. This inequality follows from (6.8).

Now we show $\varepsilon^* \ge \frac{k}{n+k}$. For T_k to breakdown 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 $(Y_1; X \cup Y) \ge k$ 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 $\#Y \ge k$, and the contamination fraction must be at least $k/(n+k)$.

Finally, we show $\varepsilon^* \le \frac{k}{n+k}$ i.e., 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 depth$(Y_i : X \cup Y) \ge k$, $i = 1 \cdots k$. 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 $\beta \in (3\alpha/2, 1/2)$. By Lemma 2.3 $k^+(X^{(n)})/n \to_{a.s.} 1/2$.

This implies that there is a positive random variable $n_0(\beta)$ which is almost surely finite with

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

Let Y consist of m contaminating points, $m \le n/2$. Then for $n > n_0(\beta)$ $k^+(X \cup Y) \ge k^+(X) > \beta n$. Now $T_\alpha(X \cup Y)$ is well-defined iff $k^+(X \cup Y) \ge \lfloor \alpha(n+m) \rfloor$. But $\beta n > \lfloor \alpha(n+m) \rfloor$ for $m \le n/2$; 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 $m \ge k$, i.e.

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

For $\alpha < 1/3$, $m = n/2$ is always a solution of this inequality; hence the restriction that Y have cardinality $\le n/2$, imposed earlier, does not prevent solving (6.19). The smallest value of m solving this inequality is either

$$m = \lfloor \frac{\alpha}{1-\alpha} n \rfloor \text{ or } m = \lceil \frac{\alpha}{1-\alpha} n \rceil.$$

It follows that for $n > n_0(\beta)$, the breakdown point is well defined and

$$\varepsilon^*(T_\alpha, X^{(n)}) = \frac{m}{n+m} = \alpha + O(\frac{1}{n}).$$

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

$$\begin{aligned} m \ge \textit{depth}\,(T_*(X \cup Y), X \cup Y) &= k^*(X \cup Y) \\ &\ge k^*(X), \end{aligned}$$

the last inequality following from (6.8). As $m \ge k^*(X)$, $\varepsilon^* = m/(n+m) \ge k^*(X)/(n+k^*(X))$. Now $k^*(X) = \frac{n}{2}(1 + o_{a.s.}(1))$ so $k^*(X)/(n+k^*(X)) \to_{a.s.} 1/3$. Hence $\liminf_n \varepsilon^* \ge_{a.s.} 1/3$.

Next, we show that the limiting breakdown point is *at most* 1/3. Let x_0 be the point of centrosymetry of P; put $k^0 = \max_u N(H_{u,x_0})$ and $m = k^0 + 2d + 1$. We will prove in a moment that

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

Before doing this, we observe that $m = \frac{n}{2}(1 + o_{a.s.}(1))$. Indeed, $N(H_{u,x_0}) = n P_n(H_{u,x_0})$, and, by absolute continuity and centrosymetry of P, $P(H_{u,x_0}) = 1/2$ for all u. Thus

$$\begin{aligned} k^0/n &\leq 1/2 + \sup_u |P_n(H_{u,x_0}) - P(H_{u,x_0})| \\ &\leq 1/2 + \mu_H(P_n, P). \end{aligned}$$

Hence

$$\begin{aligned} m &\leq n/2(1 + \mu_H(P_n,P)/2) + 2d + 1 \\ \textit{and}\quad m &\geq n/2(1 - \mu_H(P_n,P)/2) + 2d + 1 \\ \textit{so}\quad m &= n/2(1 + o_{a.s.}(1)) \end{aligned}$$

by the Glivenko-Cantelli property (6.2) of μ_H. Consequently, if (6.20) holds

$$\varepsilon^* \leq \frac{\frac{n}{2}(1 + o_{a.s.}(1))}{n + \frac{n}{2}(1 + o_{a.s.}(1))}$$

and so $\limsup_{n\to\infty} \varepsilon^*(T_*, X^{(n)}) \leq_{a.s.} 1/3$.

It remains to prove (6.20). Let y be an arbitrary point in $\boldsymbol{R}^d$, and let $Y^{(m)}$ be a dataset consisting of m exact repetitions of y. Now

$$\textit{depth}(y, X \cup Y) \geq m.$$

We claim that y is the deepest point for $X \cup Y$:

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

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

To establish (6.21) we proceed as follows. Let N be the counting measure $N(\mathrm{S}) = \#\{i : X_i \in \mathrm{S}\}$ and M be the equivalent for Y:

$$M(\mathrm{S}) = \begin{cases} m \text{ if } y \in \mathrm{S} \\ 0 \textit{ else} \end{cases}.$$

Now

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

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

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 $N(\mathit{int} H_{u,x}) \leq k^0$, and $y \notin \mathit{int} H_{u,x}$. 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

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

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

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

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

Proof. Unless $y \in \mathit{bdry} H_{u,x}$ 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 non-membership in $H_{w,x}$ as it does in $H_{u,x}$. Thus,

$$N(H_{w,x} \Delta H_{u,x}) \leq N(\mathit{bdry}\, H_{u,x}) + N(\mathit{bdry}\, H_{w,x})$$

where Δ denotes symmetric difference. As X is in general position, $N(\mathit{bdry}\, H_{u,x}) \leq d$, 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

$$\delta = \min \{|u^T (X_i - x)| : u^T (X_i - x) \neq 0\}.$$

Let $M = \max |X_i - x|$. As X is a finite set, $\delta > 0$, $M < \infty$. Pick $\alpha \in (0, \delta/M)$. Put

$$w_0 = u + \alpha(y - x)$$

and

$$w = \frac{w_0}{|w_0|}.$$

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,

$$\begin{aligned} |w^T (X_i - x) - u^T (X_i - x)| &= |(w^T - u^T)(X_i - x)| \\ &\leq |w - u|\,|X_i - x|. \\ &\leq |w_0 - u| \max_i |X_i - x| \leq \alpha M \leq \delta \end{aligned}$$

But $|u^T (X_i - x)| \geq \delta$ if $X_i \notin bdry\, H_{u,x}$. Thus

$$\text{if } |u^T (X_i - x)| \neq 0, \quad \operatorname{sgn} w^T (X_i - x) = \operatorname{sgn} u^T (X_i - x)$$

It follows that 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_*, $m \geq k^*(X \cup Y)$. By proposition 2.3, $k^*(X \cup Y) \geq \frac{n+m}{d} - \frac{d}{2}$. Combining these two inequalities with some simple algebra gives the result.

Proof of Proposition 3.5.

The proof given by Donoho (1982) has been published in (Huber, 1985).

6.4. Proofs for section 4

Proof of Proposition 4.2. We show that with the $1/(d + 1)$ fraction of contamination, breakdown occurs in the worst possible way, namely, every contaminated point is judged less discrepant than any of the original data.

First we prove two lemmas.

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

$$Max_i\, D^2(V_i : V) \geq Dim\ (span\ (V)) \tag{6.22}$$

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

Proof.

D^2 is affine invariant, so without loss of generality assume that $Ave\ (V) = 0$ and $Cov\ (V)$ acts as an identity on $span\ (V)$ and 0 elsewhere. Then $D^2(V_i\,; V) = V_i' V$ and

$$\begin{aligned} Ave_i\, D^2(V_i\,; V) = Ave\ (V_i' V) &= Ave\ (trace\ (V_i\, V_i')) \\ &= trace\ (Ave\ (V_i\, V_i')) \\ &= trace\ (Cov\ (V)). \end{aligned}$$

By assumption $trace\ (Cov\ (V)) = Dim\ (Span\ (V))$, and so (6.21) follows from $Max \geq Ave$.

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.

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

The inequality is strict if Range($Cov\ (W)$) = $Span\ (W \cup Y)$.

Proof. The basic updating formulas for Ave($W \cup Y$) and Cov($W \cup Y$) are

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

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

Write $D^2(Y_1 : W \cup Y)$ as

$$\sup_{u \in Span\,(W \cup Y)} \frac{(u'(Y_1 - Ave\ (W \cup Y)))^2}{u'\, Cov\ (W \cup Y)\, u}.$$

Put $e = Y_1 - Ave\ (W)$ and use the updating formulas to change this to

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

or

$$\sup_u \frac{n}{n+m} / \{\frac{m}{n+m} + \frac{u'\, Cov\ (W) u}{(u'e)^2}\}$$

which is less than n/m, strictly so if $u'\, Cov\ (W) u > a > 0$ for all u of norm one, i.e. if

Range($Cov\ (W)$) = $Span\ (W \cup Y)$.

Proof of Prop 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

$$\text{The first n deleted points come from } X; \tag{6.23}$$

$$\text{The remaining points come from } Y; \tag{6.24}$$

so all the X_i's are judged more discrepant than any contaminating point Y_j. 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) require that 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,

$$D^2(Y_1; W \cup Y) \leq d\,(\frac{n-k}{n})$$

with strict inequality if W is in general position, i.e. if $n - k \geq d + 1$. If $n - k \leq d$, then Dim($Span\ (W \cup Y)$) = $n - k$. In either case,

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

Applying Lemma 6.7 with $V = W \cup Y$,

$$D^2(Y_1; W \cup Y) < \max_i D^2(V_i; V).$$

Evidently, the maximum on the RHS is not attained at any point in Y; it must be attained in W. Hence, (6.25), and the proof is complete.

7. References

Bebbington, A.C. (1978). A method of bivariate trimming for robust estimation of the correlation coefficient. *Appl. Statist.* **27**, 221-226.

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, Department of Statistics, Harvard University.

Donoho, D.L. and Huber, P.J. (1982). The notion of breakdown point. in *Festschrift for Erich L. Lehmann*, Bickel et al. Eds.: Wadsworth, Belmont CA. 157-184.

Donoho, D., Johnstone, I., Rousseuw, P., and Stahel, W. (1985). Comments on Huber's Paper. *Ann. Statist.* **13**, 495-500.

Donoho, D.L., and Liu, R.C. (1985). The automatic robustness of minimum distance estimators. To appear, *Ann. Statist.*

Donoho, D., Rousseuw, P., and Stahel, W. (Forthcoming) Group equivariance, exact fit, and high breakdown. Manuscript.

Gnanadesikan, R. and Kettering, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81-124.

Huber, P.J. (1985). Projection pursuit. *Ann. Statist.* **13**, 435-474.

Mosteller, C. F., and Tukey, J.W. (1977). *Data Analysis and Regression.* Addison-Wesley.

Reaven, G.M. and Miller, R. (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia.* **16**, 17-24.

Rousseuw, P.J. (1984) Least median of squares regression. *Journal of American Statistical Association.* **79**, 871-880.

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

Souvaine, D. and Steele, J.M. (1986). Time and space efficient algorithms for least-median-of-squares regression. Manuscript.

Steele, J.M. (1977). Empirical discrepancies and subadditive processes. *Ann. Prob.* **6**, 118-127.

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

Tukey, J.W. (1974a). *T6: Order Statistics.* in mimeographed notes for Statistics 411, Princeton University.

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

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading MA.

Tyler, D. (1985). Breakdown properties of M-estimators. Manuscript.

List of Figures.

Figure 2.1

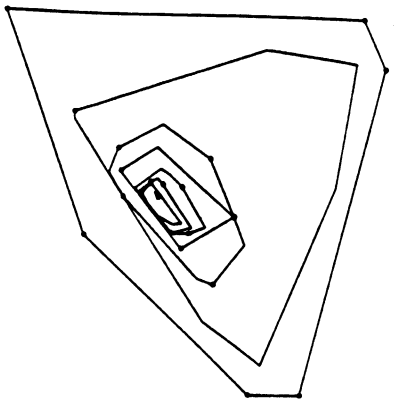
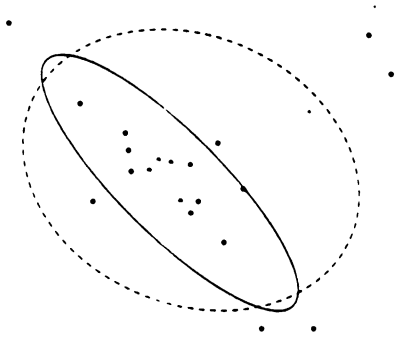


Figure 2.2



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

Figure 2.3

Figure 4.2

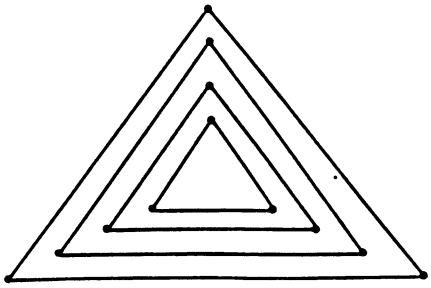


Figure 4.1

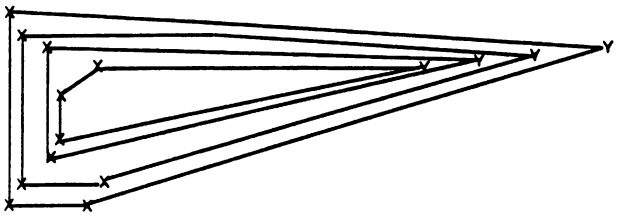


Figure 5.1

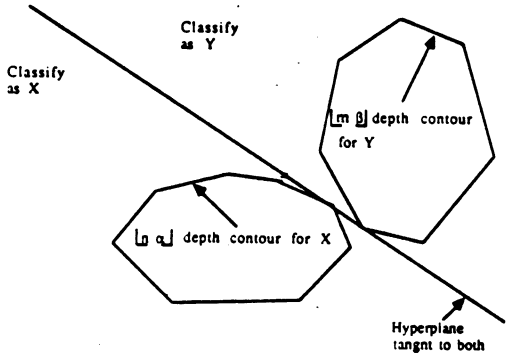


Figure 5.2