The Distribution of Robust Distances

Johanna HARDIN and David M. ROCKE

Mahalanobis-type distances in which the shape matrix is derived from a consistent, high-breakdown robust multivariate location and scale estimator have an asymptotic chi-squared distribution as is the case with those derived from the ordinary covariance matrix. For example, Rousseeuw's minimum covariance determinant (MCD) is a robust estimator with a high breakdown. However, even in quite large samples, the chi-squared approximation to the distances of the sample data from the MCD center with respect to the MCD shape is poor. We provide an improved F approximation that gives accurate outlier rejection points for various sample sizes.

Key Words: Mahalanobis squared distance; Minimum covariance determinant; Outlier detection; Robust estimation.

1. INTRODUCTION

In one or two dimensions, outlying points that are sufficiently far from the main mass of data are easily identified from simple plots, but detection of outliers is more challenging in higher dimensions. In multivariate applications, with three or more dimensions, outliers can be difficult or impossible to identify from coordinate plots of observed data. Although the outliers may lie at a great distance from the main body of data in a certain projection, identification of this projection can be difficult.

Various methods for detecting multivariate outliers have been studied (Atkinson 1994; Barnett and Lewis 1994, Becker and Gather 1999, 2001; Davies and Gather 1993; Gather and Becker 1997; Gnanadesikan and Kettenring 1972; Hadi 1992, 1994; Hawkins 1980; Maronna and Yohai 1995; Penny 1995; Rocke and Woodruff 1996; Rousseeuw and van Zomeren 1990). One way to identify possible multivariate outliers is to calculate a distance from each point to a "center" of the data. An outlier would then be a point with a distance larger than some predetermined cutoff. A conventional measurement of quadratic distance

Johanna Hardin is Assistant Professor, Department of Mathematics and Computer Science, Pomona College, 610 N. College Avenue, Claremont, CA 91711 (E-mail: jo.hardin@pomona.edu). David M. Rocke is Professor, Division of Biostatistics and Department of Applied Science, University of California–Davis, One Shields Avenue, Davis, CA 95616 (E-mail: dmrocke@ucdavis.edu).

Journal of Computational and Graphical Statistics, Volume 14, Number 4, Pages 1–19
DOI: 10.1198/106186005X77685

from a point X to a location Y given a shape S, in the multivariate setting is

$$d_S^2(X,Y) = (X-Y)'S^{-1}(X-Y).$$

This quadratic form is often called the Mahalanobis squared distance (MSD). If there are only a few outliers, large values of $d_S^2(x_i, \bar{X})$, where $\bar{X}$ and S are the standard sample mean and covariance matrix, indicate that the point x_i is an outlier (Barnett and Lewis 1994). The distribution of the MSD with both the true location and shape parameters and the standard sample location and shape parameters is well known (Gnanadesikan and Kettenring 1972). However, the standard sample location and shape parameters are not robust to outliers, and the distributional fit to the distance breaks down when robust measures of location and shape are used in the MSD (Rousseeuw and van Zomeren 1991). Determining exact cutoff values for outlying distances continues to be a difficult problem.

In trying to detect a single outlier in a multivariate normal sample, $d_S^2(x_i, \bar{X})$ will identify a sufficiently outlying point. In data with clusters of outliers, however, the distance measure $d_S^2(x_i, \bar{X})$ breaks down (Rocke and Woodruff 1996). Datasets with multiple outliers or clusters of outliers are subject to problems of masking and swamping (Pearson and Chandra Sekar 1936). As an example, consider a dataset due to Hawkins, Bradu, and Kass (1984). These data consist of 75 points in dimension three. We can see only one outlying point, but 14 of the points were constructed to be outliers. By using the mean and variance of all the data, we have masked the remaining 13 outliers (see Figure 1).

Problems of masking and swamping can be resolved by using robust estimates of shape and location, which by definition are less affected by outliers. Outlying points are less likely to enter into the calculation of the robust statistics, so they will be less likely to influence the parameters used in the MSD. The inlying points, which all come from the underlying distribution, will completely determine the estimate of the location and shape of the data. We use Rousseeuw's minimum covariance determinant (MCD) (Rousseeuw 1985) to estimate the location and shape of the data. When using the MCD in the distance function, however, we no longer have well-known distributional information for the distances. Using the motivation that independent data distances have an F distribution, we apply an adjusted F distribution to the extreme sample points. The F distribution is more representative of the extreme points than the more commonly used χ^2 distribution.

For the remainder of the article we describe in detail how to use robust distances to determine outlying data points. Section 2 describes the minimum covariance determinant. Section 3 derives the method for determining critical values used in identifying outlying points. Section 4 provides the results of simulation studies, and we conclude the article with Section 5.

2. ROBUST ESTIMATORS FOR OUTLIER DETECTION

The estimation of multivariate location and shape is one of the most difficult problems in robust statistics (Campell 1980, 1982; Davies 1987; Devlin, Gnanadesikan, and Ketten-

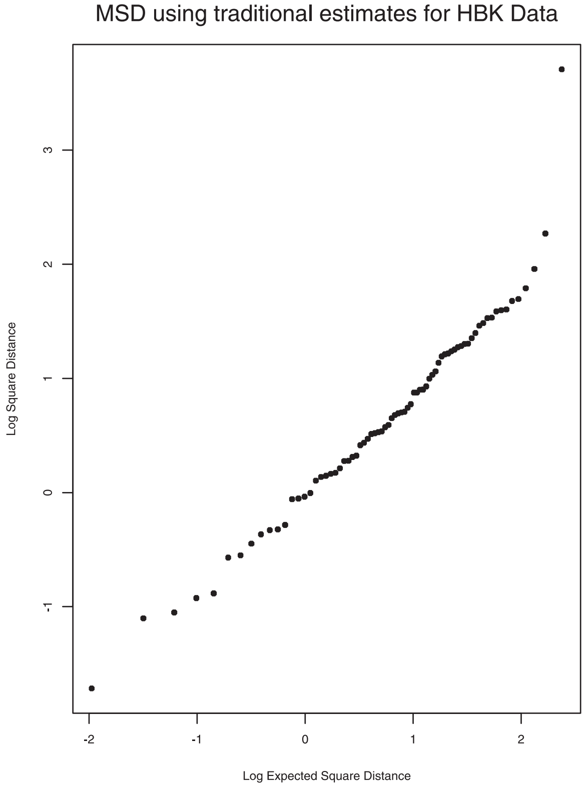

Figure 1. Mahalanobis squared distances for the HBK data plotted against the χ^2_3 expected order statistics using the ordinary mean and covariance matrix. There are by construction 14 introduced outliers; these are masked when the mean and covariance are used to determine distances.

ring 1981; Donoho 1982; Hampel, Ronchetti, Rousseeuw, and Stahel 1986; Huber 1981; Lopuhaä 1989; Maronna 1976; Rocke and Woodruff 1993; Rousseeuw 1985; Rousseeuw and Leroy 1987; Stahel 1981; Tyler 1983, 1991). The multivariate location and shape problem is difficult, because many known methods (including monotone M-estimators) will break down if the fraction of outliers is larger than $1/(p+1)$, where p is the dimension of the data (Donoho 1982; Maronna 1976; Stahel 1981) indicating that in high dimensions, a small amount of outliers can result in arbitrarily bad estimates.

2.1 Minimum Covariance Determinant

The MSD can take as its arguments any location and shape estimates. In this article we are interested in robust location and shape estimates, which are better suited for detecting outliers. In particular, we are interested in the MCD location and shape estimates. Given n data points, the MCD of those data is the mean and covariance matrix based on the sample of size h ($h \leq n$) that minimizes the determinant of the covariance matrix.

$$\text{MCD} = (\bar{X}_J^*, S_J^*)$$

where

$$\begin{aligned} J &= \{\text{set of } h \text{ points} : |S_J^*| \leq |S_K^*| \ \ \forall \text{ sets } K \ \text{ s.t. } |K| = h\} \\ \bar{X}_J^* &= \frac{1}{h}\sum_{i \in J} x_i \\ S_J^* &= \frac{1}{h}\sum_{i \in J}(x_i - \bar{X}_J^*)(x_i - \bar{X}_J^*)^\top. \end{aligned}$$

The value h can be thought of as the minimum number of points which must not be outlying. The MCD has its highest possible breakdown at $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ where $\lfloor \cdot \rfloor$ is the greatest integer function (Rousseeuw and Leroy 1987; Lopuhaä and Rousseeuw 1991). Because we are interested in outlier detection, we will use h at its highest possible breakdown; $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ in our calculations, and we refer to a sample of size h as a "half sample." The MCD is computed from the "closest" half sample, and therefore, the outlying points will have little effect on the MCD location or shape estimate. (Symmetric contamination will not affect the MCD estimates, though extreme asymmetric contamination may affect the MCD estimates, albeit less than classical estimates.) Calculating the MCD can be quite computationally intensive. Using our own R code, we implement the algorithm of Rousseeuw and Van Driessen, which is reasonably computationally efficient (Rousseeuw and Van Driessen 1999). (Calculations are done in R; note, however, if one were to use the built-in R function for calculating the MCD, the R function `cov.mcd()` reweights the MCD and scales the distances so that they do not show the elbow effect and will not be distributed according to the F distribution. One way to calculate the true MCD is to use the `cov.mcd()` function in S-Plus with the `$raw.mcd` output.)

2.2 AFFINE EQUIVARIANT ESTIMATORS

We are particularly interested in affine equivariant estimators of multivariate location and shape (Rousseeuw and Leroy 1987). Because MSDs are affine invariant, the properties and procedures that use the MSD can be calculated without loss of generality for standardized distributions. For the properties under normality, we can simulate N(0,I) data as a measure of random normally distributed data.

Large values of MSDs, using the MCD location ($\bar{X}^*$) and shape estimate (S^*), will be robust estimates of distance and will be more likely than $\overline{X}$ and S to correctly identify points as outlying. Recall the constructed data by Hawkins, Bradu, and Kass. Using the MCD estimates, the distances give a clear identification of the 14 outlying points (see Figure 2).

Not every dataset will give rise to an obvious separation between the outlying and nonoutlying points. Consider the data given by Daudin, Dauby and Trecourt and analyzed by Atkinson (Daudin, Duby, and Trecourt 1988; Atkinson 1994). The data are eight measurements on 85 bottles of milk. Using the robust MCD estimates, we are not subject to masking or swamping, but we are not sure which group of points should be considered

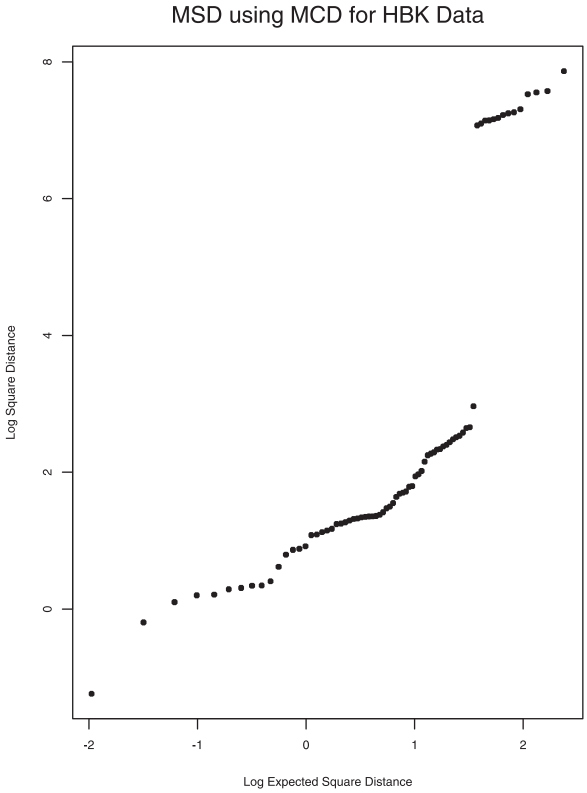

Figure 2. Mahalanobis squared distances for the HBK data plotted against the χ^2_3 expected order statistics using the MCD mean and covariance matrix. All 14 outlying points are clearly visible as outlying.

as outlying (see Figure 3). In Figure 2, points were identified as obvious outliers, but in many situations (including Figure 3) it will be important to construct a minimum outlying distance in order to determine outlyingness.

Finding a good approximation to the distribution of $d^2_{S^*}(X_i, \bar{X}^*)$ will lead to cutoff values that identify minimum outlying values, even for clusters of outliers. We argue that the $d^2_{S^*}(X_i, \bar{X}^*)$ will be approximately distributed as a multiple of an F statistic for the *outlying* points not included in the MCD calculation. This insight allows us to find cutoff values for outlying points using an estimation of the degrees of freedom associated with the F statistic. We will examine various cutoff values for MSD with MCD shape and location estimates for multivariate normal data given different values of n and p.

3. APPROXIMATE DISTANCE DISTRIBUTIONS

Mahalanobis squared distances give a one-dimensional measure of how far a point is from a location with respect to a shape. We describe an alternative approximation to the distribution of robust MSDs that we later show is superior to the standard χ^2 approximation. We first cite some known results on the exact distribution of squared distances under certain

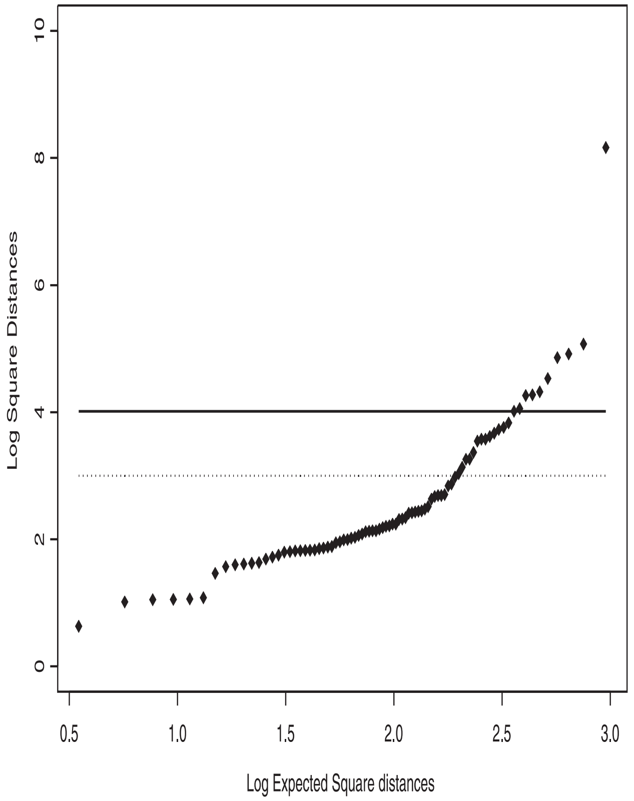

Figure 3. Mahalanobis squared distances for the Milk data plotted against the χ^2_8 expected order statistics using the MCD mean and covariance matrix. The dotted line shows the usual χ^2 cutoff (at .05), and the solid line shows the F-cutoff as developed above. One outlier is apparent, but how many outlying points are there? One? Five? Six?

conditions and then use these results to construct a new approximation for robust distances. Our method is suggested by heuristic arguments and strongly supported by computational results.

Consider n multivariate data points in $\mathcal{R}^p$, $X_i \sim N(\mu, \Sigma)$. Let S be an estimate of Σ such that, $mS \sim$ Wishart$_p(m, \Sigma)$. Below are three distributional results for distances based on multivariate normal data.

1. The first distance distribution is based on the true parameters μ and Σ. We know that if the data are normal, the distances have an exact χ^2_p distribution (Mardia, Kent, and Bibby 1979).

$$d^2_\Sigma(X_i, \mu) \quad \sim \quad \chi^2_p.$$

Which gives

$$\begin{aligned} \mathrm{E}[d^2_\Sigma(X_i, \mu)] &= p, \\ \mathrm{var}[d^2_\Sigma(X_i, \mu)] &= 2p. \end{aligned}$$

2. The second distance distribution is based on the standard mean and covariance estimates. These distances have an exact Beta distribution (Gnanadesikan and Kettenring 1972; Wilks 1962). When the distances are scaled to have the same mean as in case 1, they have smaller variance than the distances that take μ and Σ as arguments because fitting the mean and covariance allows the distances to be made smaller since the estimates accommodate random fluctuations in the data.
   Given

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ S &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\top}, \end{aligned}$$

   then

$$\frac{(n-1)^2}{n} d_S^2(X_i, \bar{X}) \sim \text{Beta}\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right).$$

   Which gives

$$\begin{aligned} \mathrm{E}\left[\frac{n d_S^2(X_i, \bar{X})}{(n-1)}\right] &= p \\ \mathrm{var}\left[\frac{n d_S^2(X_i, \bar{X})}{(n-1)}\right] &= 2p\frac{(n-p-1)}{(n+1)}. \end{aligned}$$

3. The third distance distribution is based on an estimate, S, of Σ that is independent of the X_i. S is an unbiased estimate of Σ based on a sample of size n. These distances have an exact F distribution when μ is the location argument (Mardia, Kent, and Bibby 1979), and an approximate F distribution when $\bar{X}$ is the location argument [using a Slutsky-type argument; see Serfling (1980)]. It is interesting to note here that this metric has a larger variance than the metric that takes μ and Σ as its parameters. This is because the independent variation in S adds to the variability of the distances which are in part functions of S.
   Given S and X_i independent

$$\frac{n-p}{(n-1)p} d_S^2(X_i, \mu) \sim F_{p,n-p}.$$

   Using a variant of Slutsky's Theorem

$$\frac{n-p}{(n-1)p} d_S^2(X_i, \bar{X}) \dot{\sim} F_{p,n-p}.$$

   Which gives

$$\begin{aligned} \mathrm{E}\left[\frac{(n-p-2)}{(n-1)} d_S^2(X_i, \bar{X})\right] &\doteq p \\ \mathrm{var}\left[\frac{(n-p-2)}{(n-1)} d_S^2(X_i, \bar{X})\right] &\doteq 2p\frac{(n-2)}{(n-p-4)}. \end{aligned}$$

The standard location and shape estimates ($\bar{X}$ and S) are fully within-sample estimates since all data points are used with equal weight in their calculations. The MCD location and shape estimates ($\bar{X}^*$ and S^*) behave partially like out-of-sample estimates because extreme observations will not be used to calculate the MCD (with high probability). Our interest is in the extreme points which enter into the within-sample (WS) calculations but not the partial out-of-sample (POS) calculations.

Since $\bar{X}$ and $\bar{X}^*$ are consistent estimators for μ, and since S and $c^{-1}S^*$ (for some constant c) are consistent estimators for Σ, we know that the WS and POS MSD are both asymptotically χ^2_p statistics (Mardia, Kent, and Bibby 1979; Serfling 1980). χ^2_p quantiles are often used for identifying MSD extrema even though use of χ^2_p quantiles will often lead to identifying too many points as outliers (Rousseeuw and van Zomeren 1991).

The main insight behind this article is that distances based on MCD estimates of location and shape will behave like Case 1 or 2 above for points that were used to calculate the MCD (equivalently, that have MSDs in the lower half of the empirical distribution of distances), and will behave more like Case 3 for extreme points. Approximating the extreme end of the distribution of robust distances using the tail of the F distribution of Case 3 depends on two approximations. First, note that the large Mahalanobis squared distances using the MCD location and shape behave very much like the squared distances from an independent sample. Second, we approximate the distribution of the shape estimate from an MCD by a Wishart by fitting the scale parameter and degrees of freedom.

The elbow pattern in robust MSDs described by Rousseeuw and van Zomeren (1991) can be seen in Figures 4 and 5, which show the mean ordered MSDs from the MCD in two different situations plotted against the χ^2_p quantiles. The distances that are in the smallest half of distances (coming from points that are included in the MCD subset) appear to follow a χ^2_p distribution since they lie on the line $y = x$, while the larger distances diverge in a systematic pattern.

To motivate the F distribution, let X be a sample of size n in $\mathcal{R}^p$ generated from $N(\mu, \Sigma)$ and let $(\overline{X}^*, S^*)$ be the location and shape estimates from the MCD with h points included. Let $h/n < \epsilon < \delta < 1$ be fixed. Define the following sets in $\mathcal{R}^p$:

$$\begin{aligned} R_1 &= \{X \in \mathcal{R}^p | (X-\mu)'\Sigma^{-1}(X-\mu) \le \chi^2_{p,\epsilon}\}, \\ R_2 &= \{X \in \mathcal{R}^p | \chi^2_{p,\epsilon} < (X-\mu)'\Sigma^{-1}(X-\mu) \le \chi^2_{p,\delta}\}, \end{aligned}$$

and

$$R_3 = \{X \in \mathcal{R}^p | (X-\mu)'\Sigma^{-1}(X-\mu) > \chi^2_{p,\delta}\}.$$

Note that these regions are based on the true ellipsoidal contours of the generating distribution and are not data dependent. Let X_{R3} be the set of points in the sample X lying in R_3. Let $(\overline{X}^{**}, S^{**})$ be the location and shape estimates from the MCD with h points included, where the set of possible points is restricted to those in $R_1 \cup R_2$. Finally, let Y be an independent sample with the same distribution as X and let Y_{R3} be the set of points in the sample Y lying in R_3. Then

1. The distribution of $y \in Y$ conditional on $y \in Y_{R3}$ is independent of $(\overline{X}^{**}, S^{**})$. This is obvious since they are derived from different samples.

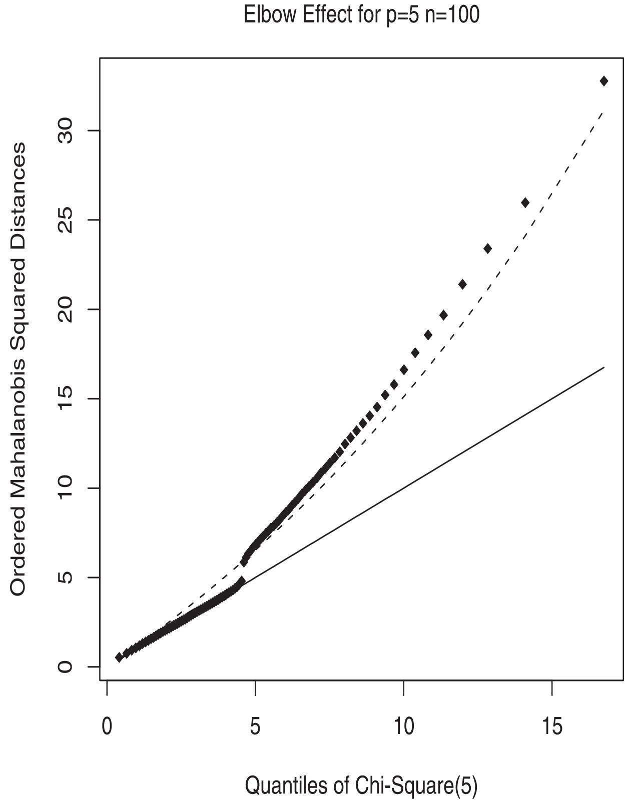

Figure 4. Mean Mahalanobis squared distances for simulated (n = 100, p = 5) data plotted against the χ^2_5 expected order statistics using the MCD mean and covariance matrix. The points that are in the MCD sample appear to have a χ^2_5 distribution, but the points not included are definitely not distributed χ^2_5. The dotted line represents the F distribution with simulated parameters. The distributions appear to fit the F distribution quite well.

2. The distribution of $x \in X$ conditional on $x \in X_{R3}$ is independent of $(\overline{X}^{**}, S^{**})$. This is true since $(\overline{X}^{**}, S^{**})$ is calculated only from points in $R_1 \cup R_2$.
3. $(\overline{X}^{**}, S^{**})$ and $(\overline{X}^{*}, S^{*})$ coincide whenever the set of points defining $(\overline{X}^{*}, S^{*})$ does not overlap with R_3. Conditional on this event and $x \in X_{R3}$, the distribution of $x \in X$ is independent of $(\overline{X}^{**}, S^{**})$.
4. Let p_n be the probability that $(\overline{X}^{**}, S^{**})$ and $(\overline{X}^{*}, S^{*})$ do not coincide. Then $p_n \to 0$ as $n \to \infty$. To see this, note that
   (a) $(\overline{X}^{*}, S^{*}) \to (\mu, c^{-1}\Sigma)$.
   (b) The h/n quantile of the distribution of MSDs from the MCD location and shape converges to $\chi^2_{p,h/n} < \chi^2_{p,\epsilon}$ with O$(n^{-1/2})$ standard deviation (where $\chi^2_{\nu,\alpha}$ is the α cutoff point for a χ^2_ν random variable).
   (c) Since $\delta > \epsilon$, the chance that the h/n quantile of the distribution of MSDs from the MCD location and shape exceeds $\chi^2_{p,\delta}$ is vanishingly small as $n \to \infty$.
   (d) The probability that the MSDs from the MCD for points in R_3 is larger than $\chi^2_{p,\delta}$ is asymptotically 1 (except for a margin of $O(n^{-1/2})$).

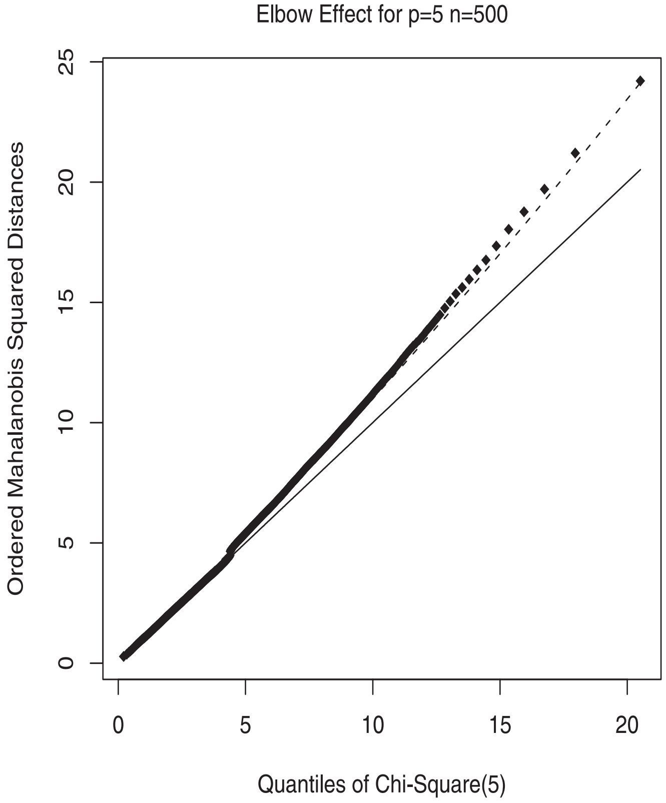

Figure 5. Mean Mahalanobis squared distances for simulated ($n = 500$, $p = 5$) data plotted against the χ^2_5 expected order statistics using the MCD mean and covariance matrix. Again, the points that are in the MCD sample appear to have a χ^2_5 distribution, but the points not included, and especially the furthest outlying points, are not distributed χ^2_5. Even in large samples, there is still an elbow effect. The dotted line represents the F distribution with simulated parameters. The distributions appear to fit the F distribution quite well.

These results motivate the use of the F distribution which relies on independence between the extreme points and the metric. The experimental independence of the extreme points and the MCD sample can also be seen in Figure 6. The picture shows average distances of two sets of independently simulated datasets whose distances were computed using the same MCD estimates. The first set contains the MCD sample, the second set was generated completely independently of the first sample and the MCD estimates. The extrema behave like the completely independently generated data.

The only remaining step in approximating the distribution of the extreme distances from the MCD is to approximate the distribution of the MCD shape by a Wishart, so that we can apply the F distribution result cited above. We will then be able to apply the following idea:

If X_i is multivariate normal data, and $\bar{X}^*$ and S^* are the MCD mean and covariance, then

1. $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$.

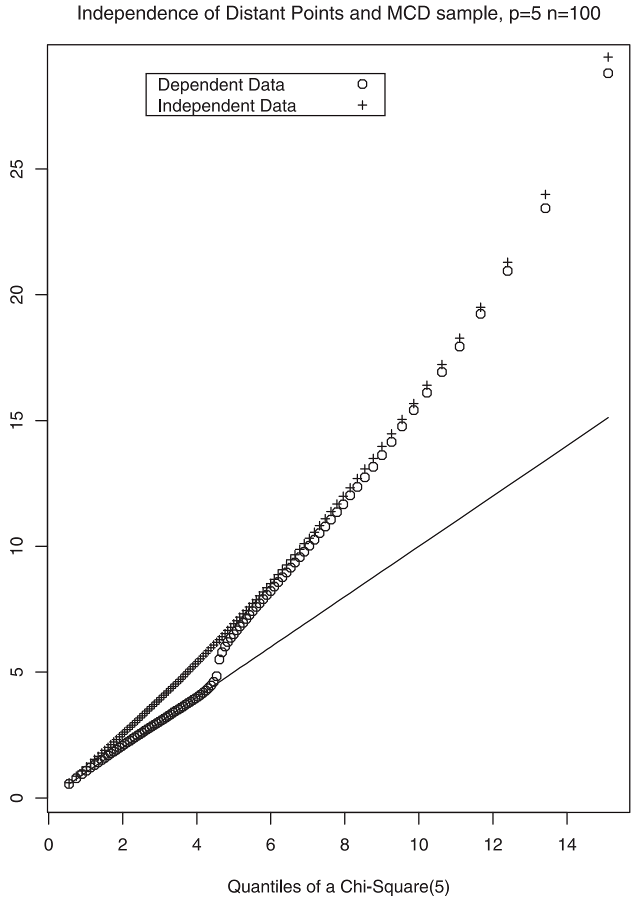

Figure 6. This figure illustrates the lack of dependence of extreme points on the MCD estimates. The distances for the dependent data set, the "o"'s, are calculated using the MCD estimates from the "o" data. Independent data are then simulated, the "+"'s, and the distances are calculated using the MCD estimates from the "o" data. For both sets of data, the points are averages of the ordered distances for 1,000 repetitions of dimension 5 size 100 data. It is apparent that the extreme distances are not affected by whether the MCD was calculated using the same sample or a different one.

2. The distribution of S^* can be approximated by

$$mc^{-1}S^* \mathrel{\dot\sim} \text{Wishart}_p(m, \Sigma), \tag{3.1}$$

where m is of unknown degrees of freedom, and c is a constant satisfying $\mathrm{E}[S^*] = c\Sigma$ [where the expectation holds for some c because S^* is an affine equivariant shape estimator of Σ in an elliptical family of distributions (Grübel and Rocke 1990)], and

3. The tail elements of X_i can be treated as if they were independent of S^*.

Then, using $\bar{X}^* \to \mu$,

$$\frac{c(m-p+1)}{pm} d^2_{S^*}(X_i, \bar{X}^*) \mathrel{\dot\sim} F_{p,m-p+1} \tag{3.2}$$

for the tail elements of X_i. That is, the tail elements will follow the tail of the F distribution.

Using the above F distribution to calculate cutoff values for distances based on the MCD sample is a robust way of identifying outliers. The only remaining problem, then, is to estimate c and m correctly.

3.1 FINDING THE DEGREES OF FREEDOM FOR THE F DISTRIBUTION

Using a method of moments identification by the coefficient of variation (CV), we can estimate the degrees of freedom associated with the approximate F distribution of $\frac{c(m-p+1)}{pm} d^2_{S^*}(X_i, \bar{X}^*)$. If for some m, S^* had a distribution that was a multiple of a Wishart, then it would be the case that

$$mc^{-1} s^*_{jj} \dot{\sim} \chi^2_m \sigma_{jj}, \tag{3.3}$$

where s^*_{jj} are the diagonal elements of S^*. Since the estimators are affine equivariant, we perform all calculations without loss of generality on $\mathrm{N}(0, I)$ data, in which case $\sigma_{jj} = 1$ and the diagonal elements are identically distributed (Grübel and Rocke 1990).

From (3.3),

$$\mathrm{E}[mc^{-1} s^*_{jj}] = m \Rightarrow \mathrm{E}[s^*_{jj}] = c,$$

and

$$\mathrm{var}[mc^{-1} s^*_{jj}] = 2m \Rightarrow \mathrm{var}[s^*_{jj}] = \frac{2c^2}{m}$$

which gives

$$\mathrm{CV} = \frac{\sqrt{\mathrm{var}[s^*_{jj}]}}{\mathrm{E}[s^*_{jj}]} = \frac{c\sqrt{2/m}}{c} = \sqrt{\frac{2}{m}}.$$

So we can estimate m by

$$\hat{m} = \frac{2}{\widehat{\mathrm{CV}}^2}$$

where CV ($\widehat{\mathrm{CV}}$) is the (estimated) coefficient of variation of the diagonal elements of the MCD shape estimator. The estimation can be done either from the asymptotics of the MCD shape matrix or by simulation. Note that the simulation will be used only to compute the mean and variance of the diagonal elements of the covariance matrix and not the distribution of the distance order statistics, which greatly simplifies the task. Since the diagonal elements are identically distributed and uncorrelated, we can simulate N independent copies of the $p \times p$ MCD shape matrix from the n data points in each independent sample, and then estimate c and m from the mean and coefficient of variation of the Np diagonal elements.

Alternatively, an asymptotic expression for c exists that works well even for small samples.

$$c = \frac{P(\chi^2_{p+2} < \chi^2_{(p,h/n)})}{h/n},$$

where χ^2_ν is a chi-square random variable with ν degrees of freedom, and $\chi^2_{\nu,\epsilon}$ is the ϵ cutoff point for a χ^2_ν random variable. This formula is easily derived and is apparently well known (also see Croux and Haesbroeck 1999).

For m there exists an asymptotic expression that is good in large samples and only moderately accurate in small samples due to Croux and Haesbroeck (1999) who used influence functions to determine an asymptotic expression for the variance elements of the MCD sample (see Appendix).

3.1.1 Interpolation to Find Parameters

As we will see from the next section, the results with simulated parameter values are quite good but take extensive computation time. The following interpolation formula can be used to modify the theoretical parameter value of the degrees of freedom.

$$m_{\text{pred}} = m_{\text{asy}} \cdot e^{(.725 - .00663p - .0780\ln(n))}, \tag{3.4}$$

where m_{pred} is the predicted degrees of freedom from adjusting the asymptotic degrees of freedom, and m_{asy} is given by Croux and Haesbroeck (1999) (details for the computation of m_{asy} are given in the appendix).

This formula was derived using linear regression on p, the dimension, and the logarithm of n, the sample size, with $\ln(m_{\text{sim}}/m_{\text{asy}})$ as the dependent variable, where m_{sim} is the simulated value of the degrees of freedom. The motivation behind the above linear model comes from knowing that the asymptotic value of m will be correct for large sample sizes (while depending weakly on p.) We could model our predicted value as

$$\frac{m_{\text{pred}}}{n} = \frac{m_{\text{asy}}(p)}{n} + b_1(p) \cdot n^{-\alpha}.$$

Where $m_{\text{asy}}(p)$ and $b_1(p)$ are functions of p, and $\alpha > 1$ is some power representing the relationship of p on the predicted value. Alternatively, we can model the relationship as

$$\begin{aligned} \frac{m_{\text{pred}}}{m_{\text{asy}}} &= \frac{c_1(p)}{n^{\alpha-1}} \qquad \text{or} \\ \ln\left(\frac{m_{\text{pred}}}{m_{\text{asy}}}\right) &= d_1(p) - (\alpha - 1)\ln n \qquad \text{(as above)}, \end{aligned}$$

where $c_1(p)$ and $d_1(p)$ are linear functions of p.

The regression used 36 data points with p=3, 5, 7, 10, 15, 20 and n=50, 100, 250, 500, 750, 1000. A subset of the results from the fitting are given below.

Dimension and size	m_{sim}	m_{asy}	m_{pred}
$p = 5$, $n = 50$	13.09	8.76	12.89
$p = 10$, $n = 100$	32.76	24.56	33.13
$p = 10$, $n = 500$	122.32	106.51	126.71
$p = 20$, $n = 1,000$	318.05	282.87	298.35

As seen from the table, the asymptotic expression can be adjusted to give values closer to the simulated degrees of freedom. The adjusted values have the benefit of being more accurate than the asymptotic values while requiring negligible computation to use.

4. RESULTS

A common and reasonable method for identifying clusters of outliers is to find robust distances and then compute distributional quantiles to determine cutoffs. By comparing simulated data to different percentile cutoffs we can determine the entire distribution of the tail elements of our robust distances. In order to assess the accuracy of the method, we compare the four distributional cutoff choices that have been described.

1. χ^2_p (which is known to reject too many points);
2. F (from (3.2)) with degrees of freedom m and scaling constant c calculated from the asymptotic formulas (m_{asy});
3. F (from (3.2)) with degrees of freedom m calculated from the adjusted asymptotic formulas and scaling constant c from the asymptotic formula (m_{pred}); and
4. F (from (3.2)) with degrees of freedom m and scaling constant c calculated from simulations (m_{sim}).

We examined the performance of these methods in the null case by a Monte Carlo study with $p = 5$, 10, 20 and $n = 50$, 100, 500, 1,000. First, simulations of the MCD shape estimators with 1000 trials were undertaken to obtain values for m and c for each pair of n and p. Then the cutoff values for 5%, 1%, and .1% rejection for each of the four distribution choices were calculated.

Next, 1,000 sets of independent data for each pair of dimension and size were simulated, and the number of points the cutoffs identified as outlying was counted. For the 5% nominal test, the percent identified as outliers is shown in Table 1. As expected, the chi-square cutoff points are far too liberal. The problem is worse in higher dimension but gets better with larger samples. The asymptotic cutoff points are an improvement on the chi-square cutoffs, but are too conservative, especially in small sample sizes. The performance at sample sizes of 500 and 1,000 is not bad. Both the adjusted and the simulated cutoff points are close to the nominal values and appear to accurately reject the correct percentage of points.

Results for 1% and .1% nominal tests are in Tables 2 and 3. Again, the chi-square cutoffs are too liberal, the asymptotic cutoffs are too conservative, and the adjusted and simulated cutoffs are quite good. Cutoffs from either the asymptotic or simulated methods can be used

Table 1. Each Entry Represents the Percent of Simulated Data That Were Above a Specific 5% Cutoff Value. (Ideally, an entry in a cell would be 5.) Directly underneath the value (in parentheses) is the standard error of the estimate, also given in units of percentages. The cutoff values were determined by dimension, size, and method of analysis. We can see that the chi-square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become more accurate as *n* increases. Both the adjusted and the simulation methods are very good for medium to large samples, and the simulation method has the best performance of the four for small samples.

Chi-square(p) cutoff values						*Asymptotic cutoff values*					
		n						*n*			
		50	*100*	*500*	*1,000*			*50*	*100*	*500*	*1,000*
p	5	26.29	16.64	7.30	6.08	*p*	5	.63	2.44	4.60	4.78
		(6.1)	(4.9)	(1.6)	(.9)			(1.6)	(2.2)	(1.3)	(.8)
	10	36.75	27.16	8.85	6.83		10	0.42	2.00	4.43	4.74
		(2.6)	(4.9)	(1.6)	(.9)			(1.2)	(1.9)	(1.1)	(.8)
	20	29.47	36.79	12.80	8.62		20	.14	1.07	3.87	4.53
		(1.0)	(1.7)	(1.7)	(1.1)			(.6)	(1.3)	(1.0)	(.7)

Adjusted asymptotic cutoff values						*Monte Carlo cutoff values*					
		n						*n*			
		50	*100*	*500*	*1,000*			*50*	*100*	*500*	*1,000*
p	5	4.93	5.18	5.05	4.95	*p*	5	5.11	5.09	4.99	4.92
		(4.4)	(3.2)	(1.3)	(.8)			(4.4)	(3.1)	(1.3)	(.8)
	10	7.12	5.71	5.02	4.96		10	5.71	5.56	4.91	4.95
		(5.1)	(3.3)	(1.2)	(.8)			(4.7)	(3.3)	(1.2)	(.8)
	20	8.41	5.27	4.49	4.70		20	6.58	4.98	4.73	4.90
		(4.4)	(2.9)	(1.1)	(.7)			(4.2)	(2.8)	(1.1)	(.8)

for rejecting outliers in multivariate data without fear that more than the nominal proportion of good data will be rejected (on the average). The method is a large improvement on the previously available methods.

From the tables, we can see that the asymptotic accuracy depends primarily on n and not on p. As expected, the asymptotic cutoff becomes more accurate as n increases. These results lead to the following recommendations:

1. For large values of n (at least 1,000 observations), asymptotic formulas may be used for cutoff values of outlying MCD distances.
2. For smaller values of n (fewer than 1,000 observations), the asymptotic formula for c can be used, but m should be adjusted using (3.4).
3. If ample computation time is available, simulation can be used to find the most accurate cutoffs. The simulation programs, in R, are available at http://pages.pomona.edu/~jsh04747/Research/Papers.htm.

An example of the application of this method is shown in Figure 3 (p. 6). The solid line shows the cutoff from method (2) at 1% significance, and the dotted line shows the comparable (and known to be overly liberal) χ^2 cutoff.

Table 2. Each Entry Represents the Percent of Simulated Data That Were Above a Specific 1% Cutoff Value. (Ideally, an entry in a cell would be 1.) Directly underneath the value (in parentheses) is the standard error of the estimate, also given in units of percentages. The cutoff values were determined by dimension, size, and method of analysis. Again, we see the same results, the chi-square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as *n* increases. Both the adjusted and the simulation methods are very good for medium to large samples, and the simulation method has the best performance of the four for small samples.

Chi-square(p) cutoff values						*Asymptotic cutoff values*					
		n						*n*			
		50	*100*	*500*	*1,000*			*50*	*100*	*500*	*1,000*
p	5	15.76	7.64	1.98	1.44	*p*	5	.03	.24	.86	.93
		(6.5)	(3.8)	(.8)	(.4)			(.3)	(.6)	(.5)	(.4)
	10	30.49	14.88	2.57	1.68		10	.01	.24	.83	.92
		(4.5)	(4.7)	(.8)	(.5)			(.2)	(.6)	(.5)	(.3)
	20	28.60	30.91	4.22	2.31		20	.00	.12	.07	.86
		(1.6)	(2.8)	(1.0)	(.5)			(.0)	(.4)	(.4)	(.3)

Adjusted asymptotic cutoff values						*Monte Carlo cutoff values*					
		n						*n*			
		50	*100*	*500*	*1,000*			*50*	*100*	*500*	*1,000*
p	5	.89	1.03	1.02	.99	*p*	5	.98	.99	1.00	.98
		(1.9)	(1.4)	(.6)	(.4)			(2.0)	(1.4)	(.6)	(.4)
	10	1.74	1.32	1.03	.99		10	1.22	1.27	.98	.99
		(2.5)	(1.5)	(.5)	(.3)			(2.1)	(1.5)	(.5)	(.3)
	20	2.99	1.25	.88	.91		20	2.00	1.15	.95	.97
		(3.3)	(1.4)	(.5)	(.3)			(2.7)	(1.4)	(.5)	(.3)

5. CONCLUSION

This article derived a new method for determining outlying points in a multivariate normal sample. The methods presented here are superior to the commonly used chi-square cutoff. Asymptotic values for the cutoffs work well in samples of size 1,000 or larger, while an adjustment formula gives good results down to relatively small sample sizes.

Because this work concerns clusters of outliers, there are implications for clustering as well as outlier identification. It is possible that robust distances may be able to identify outlying points in populations that are made up of two or more different clusters.

Also, the only robust method discussed in depth here is the MCD. The above methods also apply to other robust methods such as Rousseeuw's minimum volume ellipsoid, S-estimation, and M-estimation [for which similar Wishart parameters can be derived (Davies 1987; Lopuhaä 1989)].

Table 3. Each Entry Represents the Percent of Simulated Data That Were Above a Specific .1% Cutoff Value. (Ideally, an entry in a cell would be .1.) Directly underneath the value (in parentheses) is the standard error of the estimate, also given in units of percentages. The cutoff values were determined by dimension, size, and method of analysis. Again, we see the same results, the chi-square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as *n* increases. Both the adjusted and the simulation methods are very good for medium to large samples, and the simulation method has the best performance of the four for small samples.

Chi-square(p) cutoff values						Asymptotic cutoff values					
		n						*n*			
		50	*100*	*500*	*1,000*			*50*	*100*	*500*	*1,000*
p	5	8.25	2.85	.32	.20	*p*	5	.00	.01	.08	.09
		(5.5)	(2.4)	(.3)	(.2)			(.0)	(.1)	(.1)	(.1)
	10	21.76	6.69	.45	.23		10	.00	.01	.07	.09
		(6.1)	(3.5)	(.3)	(.2)			(.0)	(.1)	(.1)	(.1)
	20	26.90	21.38	.91	.36		20	.00	.01	.06	.08
		(2.2)	(3.9)	(.5)	(.2)			(.0)	(.1)	(.1)	(.1)

Adjusted asymptotic cutoff values						*Monte Carlo cutoff values*					
		n						*n*			
		50	*100*	*500*	*1,000*			*50*	*100*	*500*	*1,000*
p	5	.07	.09	.10	.10	*p*	5	.08	.09	.10	.10
		(.5)	(.4)	(.2)	(.1)			(.6)	(.4)	(.2)	(.1)
	10	.24	.15	.10	.10		10	.12	.15	.10	.10
		(.9)	(.5)	(.2)	(.1)			(.6)	(.5)	(.1)	(.1)
	20	0.62	0.16	.08	.09		20	.36	.14	.09	.10
		(1.4)	(.5)	(.1)	(.1)			(1.1)	(.5)	(.1)	(.1)

APPENDIX

In this appendix we provide for completeness the formulas due to Croux and Haesbroeck (1999) needed to estimate the asymptotic degrees of freedom parameter m of the Wishart approximation.

$$\alpha \quad = \quad \frac{n-h}{n} \tag{A.1}$$

where n is the sample size and $h = \left\lfloor \frac{(n+p+1)}{2} \right\rfloor$.

$$q_\alpha \quad \text{is} \quad \text{such that: } 1-\alpha = P(\chi^2_p \le q_\alpha) \tag{A.2}$$

$$c_\alpha \quad = \quad \frac{1-\alpha}{P(\chi^2_{p+2} \le q_\alpha)} \tag{A.3}$$

$$c_2 \quad = \quad \frac{-P(\chi^2_{p+2} \le q_\alpha)}{2} \tag{A.4}$$

$$c_3 = \frac{-P(\chi^2_{p+4} \leq q_\alpha)}{2} \tag{A.5}$$

$$c_4 = 3 \cdot c_3 \tag{A.6}$$

$$b_1 = \frac{c_\alpha(c_3 - c_4)}{1-\alpha} \tag{A.7}$$

$$b_2 = .5 + \frac{c_\alpha}{(1-\alpha)}\left(c_3 - \frac{q_\alpha}{p}\left(c_2 + \frac{(1-\alpha)}{2}\right)\right) \tag{A.8}$$

$$\begin{aligned} v_1 = {} & (1-\alpha)b_1^2(\alpha(\frac{c_\alpha q_\alpha}{p} - 1)^2 - 1) - 2c_3c_\alpha^2(3(b_1 - pb_2)^2 \\ & + (p+2)b_2(2b_1 - pb_2)) \end{aligned} \tag{A.9}$$

$$v_2 = n(b_1(b_1 - pb_2)(1-\alpha))^2c_\alpha^2 \tag{A.10}$$

$$v = \frac{v_1}{v_2} \tag{A.11}$$

$$\hat{m} = \frac{2}{c_\alpha^2 v}. \tag{A.12}$$

ACKNOWLEDGMENTS

The research reported in this article was supported by grants from the National Science Foundation (DMS 95-10511, DMS 96-26843, ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P42 ES04699).

[Received January 2003. Revised January 2005]

REFERENCES

Atkinson, A. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, 89, 1329–1339.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data*, New York: Wiley.

Becker, C., and Gather, U. (1999), "The Masking Breakdown Point of Multivariate Outlier Identification Rules," *Journal of the American Statisical Association*, 94, 947–955.

——— (2001), "The Largest Nonidentifiable Outlier: A Comparison of Multivariate Simultaneous Outlier Identification Rules," *Computational Statistics and Data Analysis*, 36, 119–127.

Campell, N. (1980), "Robust Procedures in Multivariate Analysis I: Robust Canonical Variate Analysis," *Applied Statistics*, 29, 1–8.

——— (1982), "Robust Procedures in Multivariate Analysis II: Robust Canonical Variate Analysis," *Applied Statistics*, 31, 1–8.

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190.

Daudin, J., Duby, C., and Trecourt, P. (1988), "Stability of Principal Component Analysis Studied by the Bootstrap Method," *Statistics*, 19, 241–258.

Davies, P. (1987), "Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, 15, 1269–1292.

Davies, P., and Gather, U. (1993), "The Identification of Multiple Outliers," *The Journal of the American Statistical Association*, 88, 782–792.

Devlin, S., Gnanadesikan, R., and Kettenring, J. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association*, 76, 354–362.

Donoho, D. (1982), "Breakdown Properties of Multivariate Location Estimators," unpublished PhD thesis, Harvard University, Department of Statistics.

Gather, U., and Becker, C. (1997), "Outlier Identification and Robust Methods," in *Handbook of Statistics, Vol 15, Robust Inferences*, eds. G. Maddala and C. Rao, Amsterdam: Elsevier, pp. 123–143.

Gnanadesikan, R., and Kettenring, J. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics*, 28, 81–124.

Grübel, R., and Rocke, D. (1990), "On the Cumulants of Affine Equivariant Estimators in Elliptical Families," *Journal of Multivariate Analysis*, 35, 203–222.

Hadi, A. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society*, Series B, 54, 761–771.

——— (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," *Journal of the Royal Statistical Society*, Series B, 56, 393–396.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Hawkins, D. (1980), *Identification of Outliers*, London: Chapman and Hall.

Hawkins, D., Bradu, D., and Kass, G. (1984), "Location of Several Outliers in Multiple-Regression Data Using Elemental Sets," *Technometrics*, 26, 197–208.

Huber, P. (1981), *Robust Statistics*, New York: Wiley.

Lopuhaä, H. (1989), "On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance," *The Annals of Statistics*, 17, 1662–1683.

Lopuhaä, H., and Rousseeuw, P. (1991), "Breakdown of Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics*, 19, 229–248.

Mardia, K., Kent, J., and Bibby, J. (1979), *Multivariate Analysis*, New York: Academic Press.

Maronna, R. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67.

Maronna, R., and Yohai, V. (1995), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association*, 90, 330–341.

Pearson, E., and Chandra Sekar, C. (1936), "The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations," *Biometrika*, 28, 308–320.

Penny, K. (1995), "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance," *Applied Statistics*, 45, 73–81.

Rocke, D., and Woodruff, D. (1993), "Computation of Robust Estimates of Multivariate Location and Shape," *Statistica Neerlandica*, 47, 27–42.

——— (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047–1061.

Rousseeuw, P. (1985), "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications, Volume B*, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Werz, The Netherlands: Dordrecht-Reidel, pp. 283–297.

Rousseeuw, P., and Leroy, A. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Rousseeuw, P., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Rousseeuw, P., and van Zomeren, B. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

——— (1991), "Robust Distances: Simulations and Cutoff Values," in *Directions in Robust Statistics and Diagnostics Part 2*, eds. W. Stahel and S. Weisberg, Berlin: Springer Verlag, pp. 195–203.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Stahel, W. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," unpublished PhD thesis, ETH Zurich.

Tyler, D. (1983), "Robust and Efficiency Properties of Scatter Matrices," *Biometrika*, 70, 411–420.

——— (1991), "Some Issues in the Robust Estimation of Multivariate Location and Scatter," in *Directions in Robust Statistics and Diagnositcs Part III*, Berlin: Springer-Verlag, pp. 327–336.

Wilks, S. (1962), *Mathematical Statistics*, New York: Wiley.