Title
Identification of Outliers in Multivariate Data

Journal
Journal of the American Statistical Association, 91

Authors
Rocke, David
Woodruff, David

Publication Date
1996

Peer reviewed

Identification of Outliers in Multivariate Data

David M. ROCKE and David L. WOODRUFF

New insights are given into why the problem of detecting multivariate outliers can be difficult and why the difficulty increases with the dimension of the data. Significant improvements in methods for detecting outliers are described, and extensive simulation experiments demonstrate that a hybrid method extends the practical boundaries of outlier detection capabilities. Based on simulation results and examples from the literature, the question of what levels of contamination can be detected by this algorithm as a function of dimension, computation time, sample size, contamination fraction, and distance of the contamination from the main body of data is investigated. Software to implement the methods is available from the authors and STATLIB.

KEY WORDS: Heuristic search; M estimation; Minimum covariance determinant; S estimation.

1. INTRODUCTION

Although methods of detecting sporadic outliers in multivariate data have existed for many years (see Hawkins 1980), the problem of detecting clusters of outliers can be extremely difficult. This essentially requires robust estimation of multivariate location and shape, and most estimators are known to fail when the fraction of contamination is greater than $(1/(p+1))$, where p is the dimension of the data. Thus detecting outliers or a disparate population that compose more than a small fraction of the data has been impractical in high dimension.

In this article we give new insights into why the problem of detecting multivariate outliers is so difficult and why the difficulty increases with the dimension of the data. We then describe significant improvements in methods for detecting outliers and demonstrate, using extensive experiments, that a hybrid method extends the practical boundaries of outlier detection capabilities. Determination of the exact envelope is complicated by the fact the probability of detecting outliers depends on many factors, such as the computer time expended, dimension, number of data points, fraction of data contaminated, type of contamination, and algorithm parameters. Nonetheless, we are able to specify approximately what levels of contamination can be detected by this algorithm under a variety of conditions.

The estimation of multivariate location and shape is one of the most difficult problems in robust statistics (Campbell 1980, 1982; Davies 1987; Devlin, Gnanadesikan, and Kettenring 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). For some statistical procedures, it is relatively straightforward to obtain estimates that are resistant to a reasonable fraction of outliers—for example, one-dimensional location (Andrews et al. 1972) and regression with error-free predictors (Huber 1981). The multivariate location and shape problem is more difficult, because most known methods 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). This means that in high dimension, a very small fraction of outliers can result in very bad estimates.

We are particularly interested in obtaining estimates that are *affine equivariant*. A location estimator $\mathbf{t}_n \in \mathbb{R}^p$ is affine equivariant if and only if for any vector $\mathbf{b} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{A}$,

$$\mathbf{t}_n(\mathbf{Ax} + \mathbf{b}) = \mathbf{A}\mathbf{t}_n(\mathbf{X}) + \mathbf{b}.$$

A shape estimator $\mathbf{C}_n \in \mathrm{PDS}(p)$, the set of $p \times p$ positive definite symmetric (PDS) matrices, is affine equivariant if and only if for any vector $\mathbf{b} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{A}$,

$$\mathbf{C}_n(\mathbf{AX} + \mathbf{b}) = \mathbf{A}\mathbf{C}_n(\mathbf{X})\mathbf{A}^T.$$

This implies, for example, that stretching or rotating measurement scales will change the estimates appropriately. Dropping the requirement of affine equivariance does increase the number of estimators that are available, and certainly there may be cases where a non–affine-equivariant estimator provides superior performance, but it is also important to have robust, computable, affine-equivariant estimators available for use.

Methods have been reported in the literature for a number of approaches for finding robust estimates of multivariate location and shape (and thus for identifying outliers). Combinatorial estimators, such as the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators of Rousseeuw (Hampel et al. 1986; Rousseeuw 1985; Rousseeuw and Leroy 1987), have been addressed with random search (Rousseeuw and Leroy 1987: MINVOL), steepest descent with random restarts (Hawkins 1993, 1994: FSA), and heuristic search optimization efforts (Woodruff and Rocke 1993, 1994). Smooth estimators such as maximum likelihood and M estimators (Campbell 1980, 1982; Huber 1981; Kent and Tyler 1991; Lopuhaä 1992;

David M. Rocke is Professor and David L. Woodruff is Associate Professor, Center for Statistics in Science and Technology, Graduate School of Management, University of California, Davis, CA 95616. This research was supported by grants from the National Science Foundation (DMS-93.01344, DMS-94.06193, and DMS-95.10511), E. I. du Pont de Nemours and Company, and the National Institute of Environmental Health Sciences, National Institutes of Health (P42 ES04699). The authors are grateful for helpful comments by the referees that markedly improved both the article and the underlying algorithm.

Maronna 1976; Rocke 1996; Tyler 1983, 1988, 1991) and S estimators (Davies 1987: Hampel et al. 1986; Lopuhaä 1989; Rousseeuw and Leroy 1987) can be computed with a straightforward iteration from a good starting point (Rocke and Woodruff 1993) or using an ad hoc search for the global minimum (Ruppert 1992: SURREAL). Sequential point addition estimators (FORWARD) have been defined algorithmically by Atkinson (1992) and Hadi (1992) working separately. Hadi (1992) suggested using a non–affine-equivariant starting point, but the point addition portion of the algorithm is affine equivariant and is nearly the same as the point addition portion of Atkinson's completely affine-equivariant algorithm. Maronna and Yohai (1995) reported some computational results for the Stahel–Donoho projection estimator (Stahel 1981; Donoho 1982); however, the method appears suitable only for small data sets in low dimension (their largest case is $n = 30, p = 6$). We have omitted any further analysis of this estimator, due to the current lack of a computational method suitable for higher dimension.

In the remainder of the article, we discuss that nature of multivariate outliers, with a special view to what sorts of outliers are worth studying. We show that outliers with the same shape as the main data are in some sense the hardest to find, and that the more compact the outliers, the harder they are to find. We adopt shift outliers as a reasonable target, being of the hardest shape but of a feasible size to locate. We also study more briefly outliers that are more compact as well as shifted, and also pure radial outliers.

We then analyze the comparative performance of the new hybrid algorithm and previous methods. Our algorithm, which uses search techniques from both FSA (Hawkins 1993a) and FORWARD (Atkinson 1993, 1995; Atkinson and Mulira 1993), as well as from our own previous work (Rocke 1996; Rocke and Woodruff 1993; Woodruff and Rocke 1993, 1994), proves as a package to be superior to other methods suggested for multivariate outlier identification. Finally, we investigate the question of what problems can be practically tackled with our methods.

2. THE NATURE OF MULTIVARIATE OUTLIERS

In this section we develop theory that leads to a characterization of classes of data with outliers that are, in a well-defined sense, the hardest to find. Armed with this, we are in a position to conduct experiments that support claims about the worst-case performance of algorithms. To create this characterization, we investigate the difficulties of locating multivariate outliers.

First, to frame the problem as this article deals with it, we assume that there is a fraction greater than one-half of the data from a well-behaved multivariate population; for example, multivariate normal. Of course, in practical cases, data transformations may be required before this plausibly holds. In addition to the well-behaved data, other data do not fit the pattern of this well-behaved majority; these may arise from a distinct population or may be measurement errors. We sometimes call the majority of the data that come from that well-behaved population the *good data*, and the remainder the *bad data*. There is supposed to be no implication that the bad data are necessarily errors—they may just arise from a distinct subpopulation—but the locution is convenient.

A second aspect of our viewpoint on this problem is that we aspire to methods that are affine equivariant, so that measurement scale changes or other linear transformations do not alter the behavior of analysis methods. An implication of this viewpoint is that Mahalanobis distances become very important, because these are among the few potentially affine-invariant outlier identification criteria.

Definition 1. Let $\mathbf{\Omega}$ be a positive definite symmetric $p \times p$ matrix. The *Mahalanobis distance* between points $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^p$ with respect to $\mathbf{\Omega}$ is defined by

$$d^2_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Omega}^{-1} (\mathbf{x} - \mathbf{y}). \tag{1}$$

We refer to the distance and the matrix that defines it interchangeably as a *metric*.

For data like those we consider here, the *true metric* is the covariance matrix of the population from which the good data arise, and a *good metric* is one close to the true metric. In particular, when the covariance of the whole sample differs greatly from the covariance of the good data, a good metric is one that resembles the latter rather than the former. The term *all-data metric* refers to the metric induced by the covariance matrix of the entire sample; this may be "good" or "bad," depending on the amount and type of contamination.

We find it convenient to distinguish the size and shape of a metric as follows.

Definition 2. Let $\mathbf{\Omega}$ be a matrix defining a metric. The *size* of the metric is the determinant $|\mathbf{\Omega}|$. The *shape* of the metric is the equivalence class of metrics Ξ such that $\mathbf{\Omega}/|\mathbf{\Omega}|^{1/p} = \Xi/|\Xi|^{1/p}$. Equivalently, we may identify the shape as the member of the equivalence class with determinant 1; that is, $\mathbf{\Omega}/|\mathbf{\Omega}|^{1/p}$.

This leads to similar definition of shape and size for samples.

Definition 3. Let $\mathbf{X}$ be an $n \times p$ matrix representing a sample of n points in $\mathbb{R}^p$. Let $\mathbf{S} = n^{-1}(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})$ be the sample covariance matrix. The *size* or *scale* of $\mathbf{X}$ is the determinant $|\mathbf{S}|$ of its covariance matrix, and the *shape* of $\mathbf{X}$ is $\mathbf{S}/|\mathbf{S}|^{1/p}$. By extension, we refer to the size and shape of other covariance-like estimators, such as the robust ones that are the subject of this article.

We now consider the question of what classes of outliers are hard to find. We begin by examining the case in which a good metric is available. This is the goal of most affine-equivariant outlier identification methods—find a good metric so that the outliers will reveal themselves. The following lemma is a routine application of multivariate computations.

Lemma 1. Consider a sample of n points in $\mathbb{R}^p$. Let the "good" data have mean μ_0 and covariance Σ_0. Let the "bad" data have mean $\mu_0 + \mu$ and covariance matrix Ω, and let this comprise a fraction ε of the overall data. Then the ex-

pected sample mean and covariance matrix are

$$E(\bar{\mathbf{x}}) = \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu} \tag{2}$$

and

$$E(\mathbf{S}) = (1-\varepsilon)\boldsymbol{\Sigma}_0 + \varepsilon\boldsymbol{\Omega} + \varepsilon(1-\varepsilon)\boldsymbol{\mu}\boldsymbol{\mu}^T \tag{3}$$

Theorem 1. Consider a sample of n points in $\mathbb{R}^p$. Let the "good" data have mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$. Let the "bad" data have mean $\boldsymbol{\mu}_0 + \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$. Consider the Mahalanobis square distance $d^2_{\boldsymbol{\Sigma}_0}(\mathbf{x}, \boldsymbol{\mu}_0)$ of a point from the true mean using the true metric. Then for a fixed location displacement $\boldsymbol{\mu}$ and size $|\boldsymbol{\Omega}|$ of the outliers, the expectation of the Mahalanobis square distance of a bad point from the true mean is least when the shape of $\boldsymbol{\Omega}$ is the same as the shape of $\boldsymbol{\Sigma}_0$. This is thus the worst case from a detection viewpoint.

Theorem 1 suggests that the hardest kind of outliers to find, when a good metric is available, is the kind that has a covariance matrix with the same shape as the good data. For this situation, this reduces the infinitely variable kinds of outliers to a single kind. If this kind of outlier can then be detected, then other kinds should be as well. Thus we intend to focus on a situation in which there are good data drawn from a multivariate normal distribution and bad data drawn from the same distribution and then displaced. These are often called *shift outliers* (Hawkins 1980; Rocke and Woodruff 1993).

Shift outliers may be contrasted with classes of outliers that may be easy to detect, in the sense of appearing disparate even with the metric obtained by using all the data. For easily detected outliers, less elaborate techniques are sufficient—examining the Mahalanobis distances from the mean of the data using the covariance matrix of the data will suffice. Although we have seen that the shape for bad data that maximizes their masking is the shape of the good data, we have not yet addressed the issue of size. The following theorem shows how easy detection is a consequence of the number and size of the contamination.

Theorem 2. Consider a sample of n points in $\mathbb{R}^p$. Let the "good" data be multivariate normal with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$. Let the "bad" data be multivariate normal with mean $\boldsymbol{\mu}_0 + \boldsymbol{\mu}$, where $|\boldsymbol{\mu}| = \eta$, and covariance matrix $\boldsymbol{\Omega} = \lambda\boldsymbol{\Sigma}_0$, and let this comprise a fraction ε of the overall data. Let $\boldsymbol{\Sigma}$ be the expected covariance matrix of the mixed sample as in Lemma 1 and consider $d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu})$ the Mahalanobis square distance in the all-data metric between a data point $\mathbf{x}$ and the overall population mean. Then,

1. If $\mathbf{x}$ is a good point, then

$$E(d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu})) = \frac{1+\varepsilon^2\eta^2}{1-\varepsilon+\lambda\varepsilon+\varepsilon(1-\varepsilon)\eta^2} + \frac{p-1}{-\varepsilon+\lambda\varepsilon} \xrightarrow[n\to\infty]{} \frac{\varepsilon}{\cdot} + \frac{p-1}{1-\varepsilon+\lambda\varepsilon}$$

2. If $\mathbf{x}$ is a bad point, then

$$E(d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu})) = \frac{\lambda^2 + (1-\varepsilon)^2\eta^2}{1-\varepsilon+\lambda\varepsilon+\varepsilon(1-\varepsilon)\eta^2} + \frac{\lambda(p-1)}{1-\varepsilon+\lambda\varepsilon} \xrightarrow[n\to\infty]{} \frac{1-\varepsilon}{\varepsilon} + \frac{\lambda(p-1)}{1-\varepsilon+\lambda\varepsilon}$$

3. The difference in the value of $E(d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu}))$ for a bad point and the value for a good point for large η is an increasing function of λ, so that $\lambda = 0$ is the worst case.

4. If $\lambda = 0$, so that the outliers form a point mass, and if η is large, then the value of $E(d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu}))$ for a bad point is less than the value for a good point whenever $\varepsilon > 1/(p+1)$.

5. If $\lambda = 1$ (pure shift outliers), and if η is large, then the value of $E(d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu}))$ for a bad point is always larger than the value for a good point. However, for large p, the standardized distribution of the distance of a good point and the standardized distribution of the distance of a bad point converge.

6. For large η, the value of λ at which $E(d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_0 + \varepsilon\boldsymbol{\mu}))$ has the same value for good points and bad points is

$$\lambda_0 \equiv \frac{(1-\varepsilon)(\varepsilon p - (1-\varepsilon))}{\varepsilon((1-\varepsilon)p - \varepsilon)} \tag{4}$$

whenever this is positive.

Remark 1. If a good starting estimate for the shape of the good data can be found, then the hardest kind of contamination to discover is that which has the same shape as the good data. Because substantial contamination can be found only by constructing a relatively good shape estimate, this is the most difficult case for such search methods.

Remark 2. Although point-mass contamination is the most difficult to detect by the Mahalanobis distance from the sample mean, it is easy to detect in other ways, such as pair-wise distances. Our hybrid algorithm has a preestimation phase that involves eliminating any exact duplicates. This will avoid problems with accidentally replicated data, for example, and will prevent exact point-mass outliers from being troublesome.

Remark 3. Although pure shift outliers might seem to be detectable, given that their mean Mahalanobis distance from the sample mean is larger than that of the good points, no method is known that can find the outliers with complete assurance. This is because the overlap in the distributions of distances can be very substantial if the amount of contamination is large. Although shift outliers are realistic, and are sufficiently challenging to separate the performance of different methods, we also examined some intermediate cases in which $0 < \lambda < 1$, particularly what we call *crossover* outliers, in which λ is chosen in accordance with Equation (4). These proved even more difficult (as predicted) for all methods examined.

As we see later, pure shift outliers are sufficient to baffle some previously proposed methods like the random search algorithm in the program MINVOL (Rousseeuw 1985). Others like those proposed by Atkinson (1993) turn out to be better than random search. However, the hybrid method proposed here dominates all other methods examined. Of course, it should be kept in mind that the algorithm of Hawkins (1992) has been incorporated into our hybrid algorithm and that of Atkinson (1993) is also used at one stage in the computations. We also examined some other types of outliers, including multiple clusters and outliers that have been reduced in scale as well as shifted.

Because we are interested mainly in high dimension, we rely primarily on extensive computational experiments to compare methods, rather than the standard, low-dimensional examples often used in the literature. However, we did examine the performance of the method on some of these standard examples, the results of which are reported in Section 5. For the reasons outlined in this section, the experiments involve mainly shift outliers, although we examined other cases to check for any sensitivity to this specification. Dimensions as large as 40 were examined so that high-dimensional cases would be represented, even though the computation times can rise rapidly with the dimension. Previously, the literature has concentrated almost exclusively on dimensions less than 10, and usually no larger than 5. Methods that appear satisfactory for a problem with 3 dimensions and 20 data points can be completely impractical for even somewhat larger problems (Woodruff and Rocke 1993). We examine a range of contamination fractions from $1/(p+1)$, which is the smallest nontrivial amount of contamination, to 40% or 45%, which can be almost impossible to find. There is a theoretical limit on the number of contaminated points that can be found, even in principle; the number of good points must be at least $h = \lfloor (n+p+1)/2 \rfloor$ (Lopuhaä and Rousseeuw 1991).

3. AFFINE-EQUIVARIANT METHODS FOR OUTLIER DETECTION

All known methods for this problem consist of the following two phases:

- Phase I: Estimate a location and shape.
- Phase II: Scale the shape estimate so that it can be used to suggest which points are far enough from the location estimate to be considered possible outliers.

We now discuss these two phases and the steps within them in reverse order.

3.1 Phase II

The output from Phase I of a multivariate outlier identification procedure is a location and shape, and thus a set of distances of points from the location using Definition 1. From this it is clear which points are the most distant, but not whether any of the distances is too large to be consistent with the absence of outliers.

A first step in answering the latter question is to apply some sort of consistency adjustment; for example, multiplying all the distances by the ratio of the median distance and the square root of the 50% point of a χ^2_p distribution ($\sqrt{\chi^2_{p;.5}}$). Because any affine-equivariant location and shape estimation method gives an unbiased location estimator and a shape estimator that has expectation a multiple of the true shape for elliptically symmetric distributions (Grübel and Rocke 1990), the square distances are asymptotically some multiple of χ^2_p for normal data. Thus standardization is sufficient to ensure that the distances are asymptotically χ^2_p. Alternatively, one can scale the shape matrix so that it is consistent for the covariance matrix of a multivariate normal distribution, which has the same effect. These equivalent forms of standardization were used by, for example, Maronna and Yohai (1995) and Rousseeuw and van Zomeren (1990). We prefer to standardize the distances by scaling the hth order statistic of the distances to the h/n quantile of $\sqrt{\chi^2_p}$, where $h = \lfloor (n+p+1)/2 \rfloor$.

However, even in rather large samples, the asymptotic χ^2 approximation is rather poor when robust estimators of location and shape are used, so more accurate small-sample methods need to be used. Usually, this is from some sort of simulation. For example, Rousseeuw and van Zomeren (1991) determined the 97.5% point of the distances by simulation, and Atkinson (1994) standardized by the total of the square distances, which is well known to be $p(n-1)$ for the mean and covariance estimator. We follow a similar procedure; by simulation, we determine the empirical $(1-\alpha_1)$ point of the distances using the given estimation method on multivariate normal samples. This gives a cutoff point L_{α_1} such that only a fraction α_1 of points on the average will have distances above the cutoff point if the given procedure is used on multivariate normal samples.

Although a standardization such as the foregoing can ensure that the type 1 error is maintained in multivariate normal samples, it can lead to insensitivity when large numbers of outliers exist. Suppose, for example, that a sample is given in dimension 20 in which 600 points are multivariate normal and 400 are outliers. Suppose that the estimation method is capable of finding a shape matrix that is essentially that of the good data. If the median square distance is scaled to the median of a χ^2_{20}, then the 500th distance will be set to $\sqrt{\chi^2_{20;.5}} = 4.397$. However, the 500th distance is actually at the 5/6 point of the good data, so the distance should be $\sqrt{\chi^2_{20;5/6}} = 5.096$, and thus all the distances of the good points are too small by a factor of .86. Now a cutoff for distances of $\sqrt{\chi^2_{20;.999}} = 6.732$ would result in an average of only 1 false outlier per 1,000 with normal samples, but in the present case, the probability of a good point exceeding it is roughly the chance that a χ^2_{20} exceeds $[(6.732)(5.096)/4.397]^2 = 60.87$, which is about 5×10^{-6}. Thus many points that are unequivocally outliers would not be declared discordant because of this bias induced by the presence of large numbers of outliers. The appropriate cutoff for this sample is $(6.732)(4.397)/5.096 = 5.801$ (not 6.732), and any point with a distance in the interval

[5.801, 6.732] probably should be declared an outlier, but will not be so declared under the rule.

To alleviate this problem, we add a second step to Phase II. We take the points not declared outliers in the first step of Phase II and calculate the mean and covariance matrix of those points. If the data were really free of outliers, then this should comprise a fraction $1 - \alpha_1$ of the sample; but if outliers are present, then the fraction may be much smaller. The covariance of the nearest $1 - \alpha_1$ fraction of a multivariate normal sample has as expectation a multiple of the true covariance matrix, where the multiple is

$$k(p, \alpha_1) = F_{\chi^2_{p+2}}(\chi^2_{p;1-\alpha_1})/(1 - \alpha_1) \tag{5}$$

so if we inflate the covariance matrix thus obtained, then we asymptotically unbias the calculation, while retaining sensitivity to outliers. Finally, we reject points beyond the $\chi^2_{p;1-\alpha_2}$ point using this new shape estimator. The entire Phase II process can be summarized as follows:

Phase II: Scaling and outlier determination.

Step 0. On entry, we have data consisting of n points in dimension p and Phase I estimates $\hat{\mu}$ and $\hat{\Sigma}$, with the shape matrix standardized so that the hth ordered distance is equal to $\sqrt{\chi^2_{p;h/n}}$, where $h = \lfloor (n + p + 1)/2 \rfloor$.

Step 1. Determine by simulation a cutoff point L_{α_1} so that when multivariate normal samples of size n in dimension p are submitted to the Phase I process, a fraction α_1 of the points on the average lie beyond L_{α_1}.

Step 2. Form a new shape matrix as $k^{-1}(p, \alpha_1)\mathbf{S}$, where $\mathbf{S}$ is the covariance matrix of all points whose distance in the first step is less than L_{α_1}. The new location estimator is the mean of those points.

Step 3. Reject as outliers any point whose distance using the revised location and shape is larger than $\chi^2_{p;1-\alpha_2}$.

We have generally used $\alpha_1 = \alpha_2 = .01$, but smaller numbers may be used at the cost of more simulation time if one wishes to reject fewer good points.

3.2 Phase I

The established methods for this problem fall into two classes: combinatorial and smooth. Combinatorial estimators construct estimates of location and shape from a subset of the data that itself is hoped to be at least mostly outlier-free. Smooth estimators attempt to satisfy a continuous equation by iteration from a starting point. Unless iteration from the whole sample mean and covariance suffices—an easy case—this requires either a direct search or use of a prior combinatorial estimator as a starting point.

Our proposed method is outlined as Phase I. Step 2 is included so that the resulting algorithm is sure to be permutation invariant *in expectation*. As a practical matter, this step is not important. The rest of this section is devoted to describing the other steps in more detail in reverse order. We also compare the method to methods in the literature, some of which are embedded in the algorithm. We refer to the complete method as the *hybrid algorithm* because it uses both combinatorial and smooth features, as well as incorporating several other useful heuristics.

Phase I: Hybrid robust estimation with roughly T seconds allowed (large T)

- Step 0. On entry, we have data consisting of n points in dimension p, a total available CPU time T in seconds, and a function $\gamma(p)$ that determines the size of partition to be used for any dimension p.
- Step 1. Remove any exact duplicate points.
- Step 2. Randomize the order of the data points.
- Step 3. Partition the data into $\lfloor n/\gamma(p) \rfloor$ cells indexed by j.
- Step 4. For each cell:
  a. Spend $T/\lfloor n/\gamma(p) \rfloor$ seconds on a search for the MCD (Hawkins 1993b; Woodruff and Rocke 1994).
  b. Use the MCD as a starting point for a sequential point addition algorithm (Atkinson 1992; Hadi 1992), using the entire sample of size n starting from the $p + 1$ points that have the smallest distance from the MCD location using the MCD metric.
  c. Use this result as the starting point for translated biweight M estimation (Rocke 1996), using the entire sample of size n. This yields estimates $\hat{\mu}_j$ and $\hat{\Sigma}_j$ of location and shape.
- Step 5. Select the index j for which $|\hat{\Sigma}_j|$ is least, and set $\hat{\mu} = \hat{\mu}_j$ and $\hat{\Sigma} = \hat{\Sigma}_j$.

3.3 M and S Estimation

An S estimate of multivariate location and shape is defined as that vector $\mathbf{t}$ and PDS matrix $\mathbf{C}$ that minimizes $|\mathbf{C}|$ subject to

$$n^{-1} \sum \rho([(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^- (\mathbf{x}_i - \mathbf{t})]^{1/2}) = b_0 \tag{6}$$

which we write as

$$n^{-1} \sum \rho(d_i) = b_0. \tag{7}$$

It has been shown by Lopuhaä (1989) that S estimators are in the class of M estimators with standardizing constraints with weight functions $v_1(d) = w(d), v_2(d) = pw(d)$, and $v_3(d) = v(d)$, where $\psi(d) = \rho'(d), w(d) = \psi(d)/d$, and $v(d) = \psi(d)d$, with constraint (7) (Rocke and Woodruff 1993).

Rocke (1996) showed that S estimators in high dimension can be sensitive to outliers even if the breakdown point is set to be near 50%. We use the translated biweight (or t-biweight) M estimation method defined by Rocke (1996), with a standardization step consisting of equating the median of $\rho(d_i)$ with the median under normality. This is then not an S estimate, but is instead a constrained M estimate.

The convergence criterion for the algorithm is subject to choice. We use the maximum change in the weights to decide on termination. The specifics of the iteration for

Table 1. Percentage of Successful Estimation Runs With Biweight S Estimation and t-Biweight M Estimation

	ε	Time (sec)	Biweight %
50	.30	3	
50	.30	30	
50	.35	3	
50	35	30	
200	.30	9	
200	.30	36	
200	.35	9	
200	.35	36	

NOTE: The last two columns are the fraction of runs out of 20 that the indicated method found the "good" root of the estimating equations. The experiments were in dimension 10 with outliers at a distance of $d = 2$.

both M and S estimators was given by Rocke and Woodruff (1993).

In accord with the theory of Rocke (1996), we have found that using the t-biweight M estimator greatly improves the performance of the hybrid algorithm compared to using biweight S estimation, at least when the outliers lie relatively close in ($d = 2$, as defined in Sec. 4). When $d = 4$, the smooth estimation method used made no important difference. Some detailed evidence is given in Table 1. The situation here is that twenty replicates of shift outliers at $d = 2$ in dimension 10 and with indicated sample size, fraction of outliers, and computation time allowed. The response is the percentage of replicates for which the indicated estimator achieved the good root. Note that the t-biweight performance exceeds that of the biweight S estimate by large amounts in every case. A large number of additional experiments confirm this important difference in performance. We use the t-biweight M estimator in Step 4(c) of Phase I.

3.4 Search and Partitioning

The simple iteration scheme for M estimation fails without a good starting point. An M estimator that begins iteration using an estimate based on all of the data breaks down with $1/(p + 1)$ of the data contaminated (Maronna 1976). Two methods of addressing this problem seem possible. One is to look directly for the global minimizer of the S criterion. The other is to find a good starting point for the iteration by using a preliminary combinatorial estimator.

Ruppert (1992) proposed an algorithm called Surreal for direct search for the global minimizer of an S estimator used in multiple regression. He reported computational experiments that demonstrated the effectiveness of the Surreal for this purpose. In the same paper, he also proposed an extension of the method to robust estimation of multivariate location and shape. It appears that Surreal is not as effective for this problem as for regression. In dimension 10, Surreal rarely found the good root when the fraction of contamination was greater than about 12%. Because this was not competitive with other algorithms examined, detailed results are not presented.

We also have examined direct search as a method of finding the good root for S or M estimation and have found that it seems better to use a preliminary combinatorial estimator such as the MCD (Rousseeuw 1985). As pointed out by Woodruff and Rocke (1994), using the MCD to find a good starting point presents severe computational difficulties. Regardless of which algorithms are used to compute them, combinatorial estimators such as the MCD search a space that increases exponentially with the sample size and the dimension. In fact, when using the MCD as a first stage in a two-stage estimator, one can have the perverse situation of being made worse off by having more data. To cope with this problem, the data must be partitioned so that the search space for the MCD is kept in a reasonable range. After some modest experimentation, we settled on a cell size of $\gamma = 5p$. This possibly may be too small for high dimension, but determining the optimal value was beyond the scope of this article.

As shown by Woodruff and Rocke (1994), using data partitioning in this fashion allows for acquisition of the good root with high probability with a computational time increasing only linearly with n (instead of exponentially). We use data partitioning in Step 3 of Phase I.

3.5 Sequential Point Addition

Working separately, Atkinson (1992) and Hadi (1992) have proposed algorithms that begin with an estimate of shape and location based on $(p + 1)$ points and then select successively larger sets. The set with $k + 1$ points consists of those points whose Mahalanobis distances from the mean of the k set using the covariance of the k set as a metric are smallest.

Hadi suggested using coordinatewise medians as a preliminary location estimator and the covariance of the whole data with that as center for a preliminary shape estimator. The initial set of $p + 1$ points consists of those whose Mahalanobis distance from the preliminary location estimator is least, using the preliminary shape estimator as a metric. The algorithm proposed by Hadi breaks down if the contamination is extremely far away from the good data in the correct metric. Also, the coordinatewise median is not affine equivariant and consequently can work extremely well on a suite of data sets, but then perform horribly on the same data after an affine transformation. For example, with outliers clustered on a diagonal and not very far from the good data, our experiments suggested near-perfect outlier detection. If the same data are transformed to have a covariance that is the identity matrix, then the performance is degraded significantly.

Atkinson's method is affine equivariant. He suggested restarting the procedure many times with randomly selected sets of $p + 1$ points. For each trial, sequential addition is performed and for each stage in the sequential addition, the covariance matrix is calculated, and the resulting shape matrix is expanded (or contracted), so that half (or $(n + p + 1)/2$) of the points are included in the ellipsoid defined by the current location and shape. The estimate over all trials and over all stages of each trial in which the scaled shape matrix has minimum determinant may be taken as the robust estimate of the shape and location of the data. As we see later, Atkinson's algorithm is a large improvement over MinVol. In our tables and graphs, we refer to this procedure, following Atkinson, as the forward algorithm, or Forward for short.

We found that including a sequential addition step between the search for the MCD and the smooth estimator improved the results in many cases. We ran more than 200 simulated data sets in dimension 20 with n values of 200, 400, and 800 and various fractions of "bad" data from .2–.4. In these experiments, inclusion of Step 4b in the Phase I algorithm resulted in an improved estimate in over 70% of the data sets. In many cases the improvement was very modest and did not affect Phase II results. Nonetheless, inclusion of the step seems well worth the small amount of computer time required to execute it (small relative to the time required for the MCD search). We use sequential point addition in Step 4b of Phase I.

3.6 Minimum Covariance Determinant

Faced with a subsample of contaminated data, our experiments indicate that the best way to find a good starting point for sequential point addition (or for M iteration) is to search for the MCD. It was originally thought that the MVE would be preferable for computational reasons (see Rousseeuw and Van Zomeren 1990), even though the MCD has greater asymptotic efficiency. This was based on the notion that MVE algorithms would make use of elemental subsets. Woodruff and Rocke (1993) demonstrated that heuristic search algorithms that use larger subsample sizes perform better. Given this fact, there is no longer any reason to prefer the MVE to the MCD. Simulations done by Woodruff and Rocke (1994) strongly support the contention that the MCD is in fact the better estimator to use.

The MCD for any set of data is defined by the half sample whose covariance matrix has minimum determinant. It is convenient to search for MCD half-samples moving from half sample to half sample by removing one point in the current half sample and adding one point not currently in the half sample. Neighborhoods defined in this way can form the basis of a steepest descent to a local minimum. Hawkins (1993b) suggested using steepest descent with random restarts, which he called FSA. Woodruff and Rocke (1994) advocated using a steepest descent–based meta-heuristic called tabu search (TS) (Glover 1989, 1990).

Our experiments indicate that FSA can outperform the simple TS algorithm given by Woodruff and Rocke, especially when not much time is allocated for the search. A much more complicated ghost image processing algorithm (Woodruff 1995) performs better than FSA given large amounts of time and data, but it does not perform better with small amounts of time and, furthermore, in our tests, the improved performance does not seem to be sufficient to make a major difference in Phase II performance for the search durations of interest to us. Given the lack of a qualitative difference in the performance envelope and the relative elegance of pure steepest descent with random restarts, we used the FSA algorithm in Step 3a of Phase I in the tests described here.

4. COMPUTATIONAL EXPERIMENTS

The results given in Section 2 allow the construction of a less arbitrary set of simulations than might otherwise have been thought possible. Our primary model is shift outliers, in which the good data are defined to be multivariate standard normal and the bad data to be multivariate unit normal with a shifted mean. We also take a more abbreviated look at outliers in which the covariance matrix is multiplied by λ_0 of Equation (4) so that the expected distance is equalized between a good point and a bad point in the metric of all the data (crossover outliers).

We measure the amount of shift in terms of the unit of measurement $Q_p = \sqrt{\chi^2_{p;.999}}$, which is more or less the radius of a sphere around the mean that contains almost all of the good points. If the outliers are centered at a distance of $2Q_p$, then these spheres should not overlap. We implement outliers at a distance of dQ_p by adding dQ_p^* to each component, where $Q_p^* = \sqrt{\chi^2_{p;.999}/p}$. This places the outliers at the correct distance out on a diagonal. In the experiments here with $\lambda = 1$ and $\lambda = \lambda_0$, we use $d = 2$, which we call close outliers, and $d = 4$, which we call far outliers. To obtain *radial* outliers, we generate each outlier separately as a multivariate normal with mean m, where m is in a different random direction for each outlier and $|m| = dQ_p$.

This generation mechanism is sufficient for affine-equivariant methods; but for non–affine-equivariant methods, the data should then be standardized so that the entire sample has mean $\mathbf{O}$ and covariance $\mathbf{I}$. This can be done using the singular value decomposition as follows. Let $\mathbf{S}$ be the covariance matrix of the whole sample of good and bad data. This can be written as $\mathbf{S} = \mathbf{Q}^T\mathbf{D}\mathbf{Q}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{D}$ is the diagonal matrix of eigenvalues. If $\mathbf{X}$ is the centered sample, then the sample $\mathbf{X}\mathbf{Q}^T\mathbf{D}^{-1/2}\mathbf{Q}$ has the desired properties.

One convenient aspect of using shift outliers (or crossover outliers) in this problem is that in our experience, smooth methods such as M and S estimation usually have at most two roots: one that can be found by iterating from the good data (the good root) and one that occurs when iterating from all the data (the bad root). For small amounts of contamination or very large amounts of contamination, these roots may not be distinct, but only when they differ is the problem interesting. This leads naturally to a fairly strict criterion of success for a Phase I algorithm. If the method yields a location $\hat{\mu}$ and a metric $\hat{\Sigma}$, then the method is successful if the largest value of $d_{\hat{\Sigma}}(\mathbf{x}, \hat{\mu})$ for a good point is smaller than the smallest value for a bad point. For the overall performance of both phases, we use the less-strict type I and type II error measurement.

4.1 Null Behavior

Table 2 gives some simulation results to support the good behavior of the proposed two-phase method when the data are multivariate normal. Each line of the table is based on 20 instances in which the entire algorithm, Phase I and Phase II, was applied to multivariate normal data. The third column is the fraction out of the total of $20np$ points that were rejected as outliers. It is easily seen that these numbers are all quite near the nominal rejection fraction of .01.

Table 2. Actual Type 1 Errors When the Nominal Type 1 Error is .01

p	n	Fraction rejected
	50	
	100	
	200	
	200	
	400	
	800	
	800	
	1,600	
	3,200	

NOTE: Each line of the table represents 20 replicates applying the algorithm to multivariate standard normal data. The third column is the fraction of the $20np$ points that were labeled as outliers by the algorithm.

4.2 A Comparison of Algorithms

In this section we compare results using three different strategies for Phase I: the hybrid algorithm, random search over elemental subsets (Rousseeuw 1985: MinVol), and the forward algorithm (Atkinson 1992: Forward). In all cases Phase II is as given in Section 3.1. The steepest descent algorithm (FSA) (Hawkins 1993b) is not shown separately, because it has been incorporated into the hybrid algorithm. Surreal was also tried, but its performance was not competitive with the others, and so it has been omitted from the summaries.

Given that some runs in high dimension may take up to an hour of CPU time, and that there are many conditions under which one should compare estimators, a fully comprehensive Monte Carlo study is impractical. The database used in the comparison study comprises more than 10,000 runs. The dimension p was 10, 20, and 40, with sample sizes of $n = 50$ to $n = 6{,}400$, with larger sample sizes used in higher dimensions. Several processing times t were tried for each case, varying from a few seconds to several hours in high-dimensional examples. The degree of contamination ε was varied from levels where the solution could almost always be found by most methods to levels where none of the methods could get them right.

To increase the utility of the number of runs that were practical to perform, a generalized linear model was fit to the outcomes of the experiments, each of which consisted of 20 trials at each case. The logit of the probability that a given estimator would succeed in identifying the outliers was taken to be a linear function of n, ε, and $\log(t/n)$. Different models were fit for each estimator, distance of outliers, and each dimension examined. We defined success to consist of identifying at least 90% of the outliers correctly as outliers. In almost every case, the hybrid algorithm had no errors if it succeeded at all, but we used the more liberal definition of success, because sometimes a method identified almost all of the outliers but missed a few, and it was thought to be unfair to call that "failure."

Figure 1 shows plots of the fitted probability of success as a function of the amount ε of contamination for three estimators in dimension 20, with $n = 400$ and $t = 1{,}600$ seconds. Figure 1a is for outliers set at a distance of $d = 2$; Figure 1b is for outliers at $d = 4$. The message is clear. The Forward algorithm is greatly superior to random search over elemental subsets at all levels of contamination. The hybrid algorithm in turn is noticeably more effective than Forward. Similar results obtain for other sample sizes, times, distances, and dimensions. In higher dimensions, limited trials suggest that the superiority of the hybrid algorithm is even greater. However, given the finite time available for computer simulations in high dimension, most of the runs were devoted to determining the envelope of feasible solution for the hybrid algorithm, rather than to documenting the exact degree of superiority over competing algorithms.

Another feature of the plots is worth noting. A small addition to the fraction of outliers converts a problem that is easy to solve into one that is quite difficult. For example,

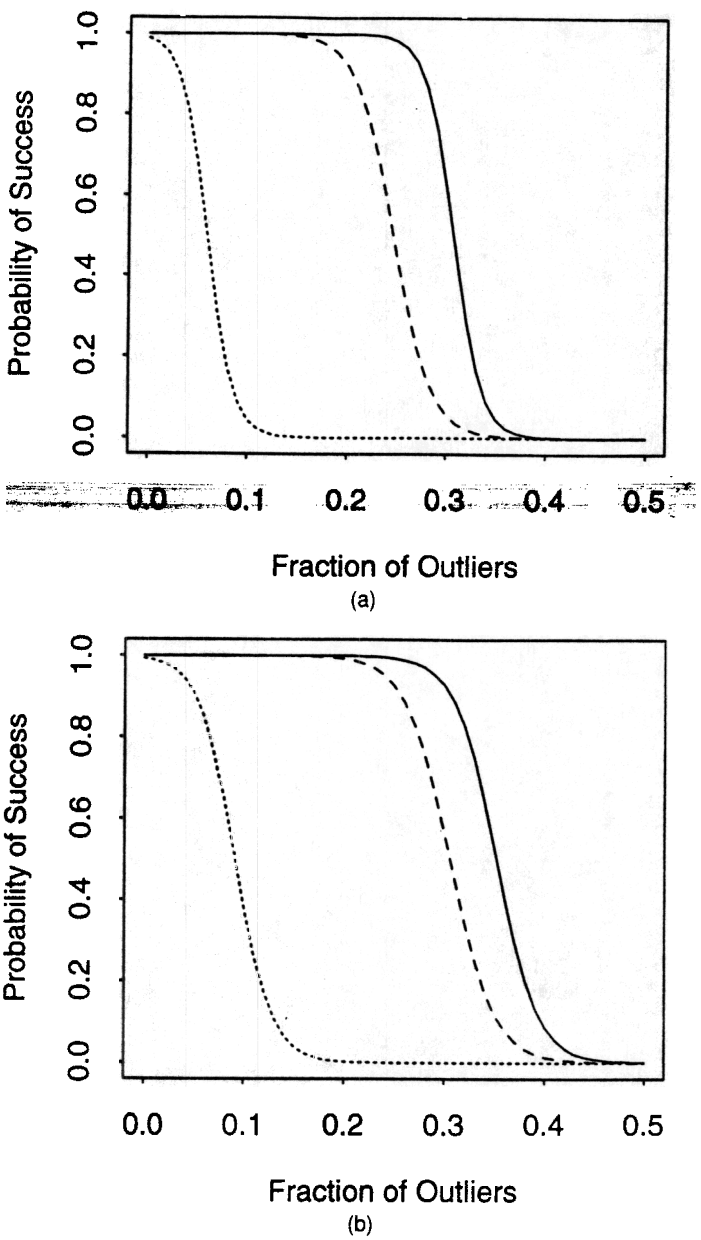

Figure 1. Predicted Performance of Outlier Detection Methods in Dimension 20 with n = 400. The dotted line represents MINVOL; the dashed line, FORWARD; the solid line, HYBRID. (a) Outliers at a distance of d = 2; (b) outliers at a distance of d = 4.

Table 3. Effect of Outliers in One or Multiple Clusters

d	Number of clusters	ε (percent)	Predicted success rate (percent)
2	1	30	55.6
2	2	30	98.6
2	4	30	99.9
2	radial	30	100.0
2	1	35	4.1
2	2	35	70.4
2	4	35	99.2
2	radial	35	100.0
2	1	40	0.1
2	2	40	7.5
2	4	40	81.9
2	radial	40	100.0
2	1	45	0.0
2	2	45	0.3
2	4	45	13.4
2	radial	45	100.0
4	1	30	92.9
4	2	30	99.9
4	4	30	99.9
4	radial	30	100.0
4	1	35	44.4
4	2	35	97.8
4	4	35	99.9
4	radial	35	100.0
4	1	40	4.6
4	2	40	72.8
4	4	40	99.3
4	radial	40	100.0
4	1	45	0.3
4	2	45	14.1
4	4	45	90.0
4	radial	45	100.0

NOTE: The experiments in this table were performed in dimension 20 with n = 400 and 1,600 seconds of processing time. The last column represents the fitted value from a generalized linear model fitted to the results of the experiment.

the hybrid algorithm presented with a data set of 800 points in dimension 20 with 30% shift outliers at a distance of $d = 2$ is predicted to succeed 85% of the time with 3,200 seconds of processing time. The success rate falls to 15% if the fraction of outliers rises to 35%.

4.3 Estimating the Envelope

This section is devoted to the following question: For what dimensions, sample sizes, outlier distances, fractions of outliers, type of outliers, and computation times is the hybrid algorithm effective? The theoretical results of Woodruff and Rocke (1994) demonstrated that any amount of contamination below 50% can be handled with sufficient data and sufficient processing time. Here we ask a different question: What amount of contamination can be practically detected with the amount of data given and with practical processing times? We also examine the effect of variables such as type of outlier, dimension, sample size, and computation time on the effectiveness of the algorithm. A series of tables show some results, which are based on simulations. To produce the predicted success rate, a generalized linear model was fit as described previously.

Table 4. Effect of Outlier Shrinkage

d	Time (sec)	Outlier type	Predicted success rate (percent)
			76
			56
			94
			89
			85
			82
			98
			97

NOTE: The experiments in this table were performed in dimension 20 with n = 800 and 30% outliers. The last column represents the fitted value from a generalized linear model fitted to the results of the experiment.

Table 5. Effect of Dimension

p	n	Time (sec)	ε (percent)	Predicted success rate (percent)
10	200			100.0
10	200			99.9
10	200			99.3
10	200			93.7
10	200			59.7
10	200			12.9
20	800			100.0
20	800			100.0
20	800			99.5
20	800			85.4
20	800			14.6
20	800			0.5
40	3,200			100.0
40	3,200			99.8
40	3,200			79.4
40	3,200			3.3
40	3,200			0.0
40	3,200			0.0

NOTE: The experiments in this table were performed with outliers at a distance of d = 2. The last column represents the fitted value from a generalized linear model fitted to the results of the experiment.

4.3.1 Type of Outliers. Most of our results are based on the use of shift outliers, in which the outlying values are generated from a distribution with the same covariance matrix and a shifted mean. Theorem 1 implies that this is the hardest shape that outliers can take. We tested this empirically by generating outliers that were in more than one

Table 6. Effect of Sample Size

n	Time (sec)	ε (percent)	Predicted success rate (percent)
200			96.8
400			98.3
800			99.5
200			46.8
400			62.3
800			85.4
200			2.5
400			4.6
800			14.6

NOTE: The experiments in this table were performed in dimension 20 with outliers at a distance of d = 2. The last column represents the fitted value from a generalized linear model fitted to the results of the experiment.

Table 7. Effect of Computation Time

ε (percent)	Predicted success rate (percent)
25	99.1
25	99.5
30	76.2
30	85.4
35	8.6
35	14.6

NOTE: The experiments in this table were performed in dimension 20 with $n = 800$ and outliers at a distance of $d = 2$. The last column represents the fitted value from a generalized linear model fitted to the results of the experiment.

cluster. Each cluster had a mean the same distance from the center as before ($d = 2$ or $d = 4$), but lying along one of the 2^p random diagonals $(\pm 1, \pm 1, \ldots, \pm 1)$. This provides a different overall outlier shape. The ultimate in this is to use a different random direction for each point, which amounts to radial outliers.

Table 3 shows some results of simulations with multiple clusters. These were done in dimension 20 with $n = 400$ and 1,600 seconds allowed for processing. It is apparent from these results that if the outliers lie in more than one cluster, then the process of identifying the good data becomes dramatically easier, culminating in radial outliers, for which we were successful in every trial of the experiment.

Another issue involves outliers that have a distribution with the same shape as the main data but of a different size. Theorem 2 implies that the smaller the size, the harder the problem. Point mass outliers have the smallest size but are easily detected with our pairwise comparison front end. A reasonable compromise as an alternative to shift outliers is what we call crossover outliers, in which the shrinkage is set just sufficiently large to make the mean distance of the outliers the same as the mean distance of the good data in the metric of the covariance of all the data; see Equation (4). Table 4 shows some example results for dimension 20 with $n = 800$ and 30% outliers. Shrinking the outliers always makes the search harder, but the effect is lessened if the search time is increased or if the outliers are relatively distant. Qualitatively, other comparisons remain the same whether shift or crossover outliers are used as the test bed, and so we have continued to use shift outliers in most instances.

Table 8. Critical Contamination Level for 90% Success With the Hybrid Algorithm

p	d	n	t (sec)	ε (percent)
10	2	50	100	
10	2	100	200	
10	2	200	400	
10	4	50	100	
10	4	100	200	
10	4	200	400	
20	2	200	800	
20	2	400	1,600	
20	2	800	3,200	
20	4	200	800	
20	4	400	1,600	
20	4	800	3,200	
40	2	800	3,200	
40	2	1,600	6,400	
40	2	3,200	12,800	
40	4	800	3,200	
40	4	1,600	6,400	
40	4	3,200	12,800	

NOTE: The last column shows the percentage of contamination at which a predicted 90% of the instances could be successfully completed. The predictions were from a generalized linear model fitted to the results of the experiment.

4.3.2 Dimension. Outlier problems in higher dimension are harder than those in lower dimension. More data usually will be needed, if only because the number of parameters to be estimated is higher, and more processing time will be needed. Even with some adjustments in this direction, the greater difficulty of problems in high dimensions shows up. In Table 5 we have let the sample size n grow with p^2 and allowed the processing time to be $4n$ seconds. Clearly, as the dimension rises, the amount of contamination that can be coped with falls, even after raising the amount of data and the computational time. Nevertheless, many problems in dimension 40 and higher will still be feasible, if the amount of contamination does not rise too high.

4.3.3 Sample Size. The results of Woodruff and Rocke (1994) show that for large enough sample sizes, even the hardest outlier problem can be tackled. Table 6 shows this effect in practice. These experiments use shift outliers at $d = 2$ in dimension 20. Whatever the amount of contamination, increasing the sample size (and the computational time proportionately) increases the success rate. However, to detect outliers at 35–40% in dimension 20 may require very large samples and long computation times.

4.3.4 Time. Often, increasing the available data is not a feasible option. Table 7 addresses the question of applying increasing amounts of time without increasing the sample size. These experiments involved shift outliers at a distance of $d = 2$ in dimension 20 with 800 points. Increasing the time does increase the success rate, although there is no assurance that the limit of the success rate as the time increases is 100%. An important avenue of future research is to make more effective use of large amounts of time with a fixed sample. One possible approach is to use multiple random partitions, with only a fixed amount of time allocated to each one. Thus as available CPU time increases, the number of partition cells examined will increase, rather than the intensiveness of the search in each cell.

4.3.5 The Current Envelope. By "the envelope" we mean the limits of the size and difficulty of problem that can be handled with current technology. This is dependent on the dimension, the sample size, the fraction and type of outliers, the distance of the outliers from the main data, and the available processing time.

Table 8 shows some results. For each indicated combination of dimension and outlier distance, a generalized linear model was fit as described earlier. Then the level of con-

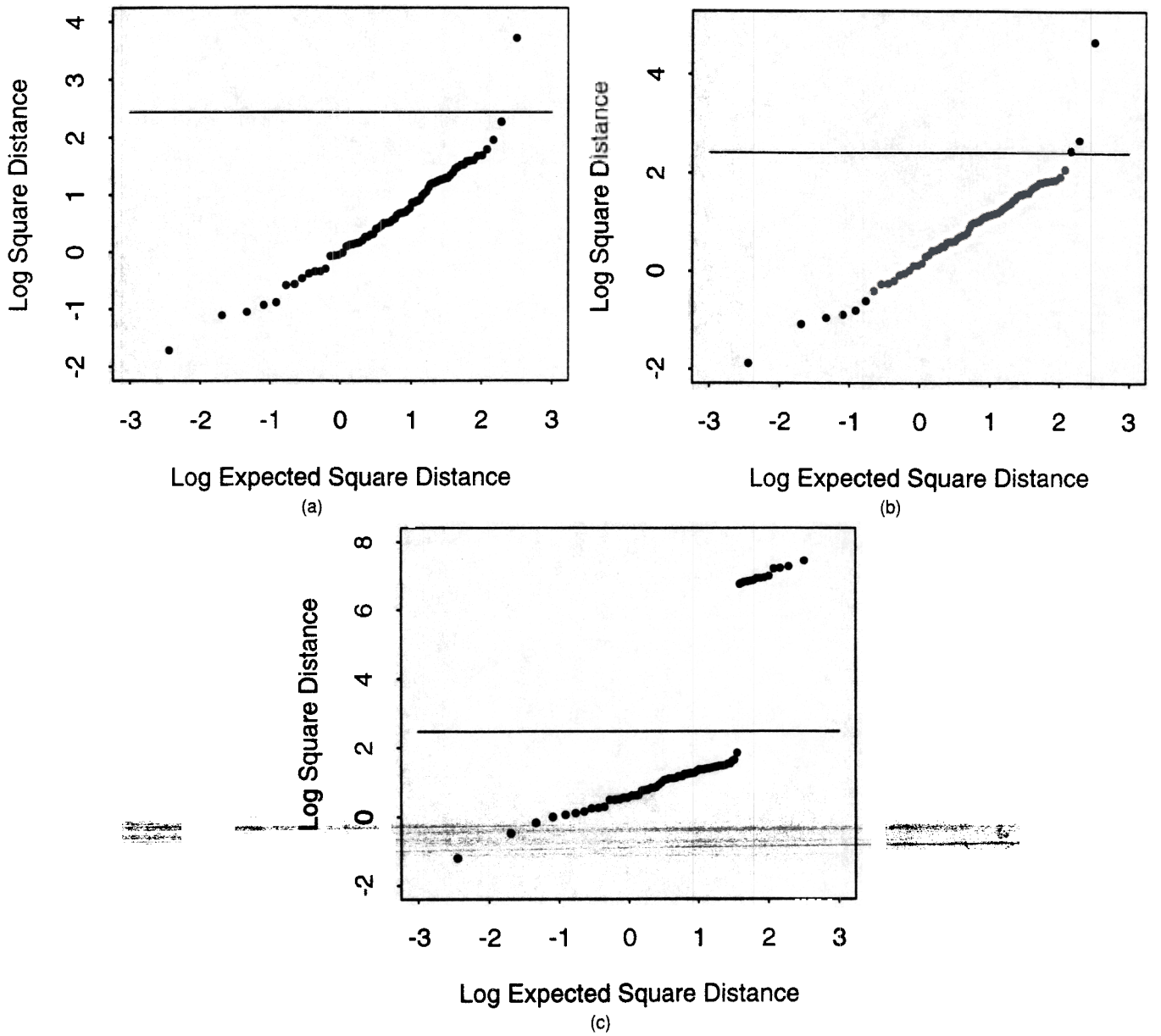

Figure 2. Mahalanobis Square Distances for the HBK Data Using Three Metrics Plotted Against χ^2_3 Expected Order Statistics on a Log-Log Scale. (a) The all-data metric. The horizontal line represents the 99% point of a χ^2_3 random variable. The three highest points are 14, 12, and 13, in that order. (b) The metric derived from the data after excluding points 12, 13, and 14. The horizontal line represents the 99% point of a χ^2_3 random variable. (c) The metric derived from the hybrid algorithm. The horizontal line represents a 99% cutoff as estimated by Phase II of the algorithm. The 14 points above the cutoff are the points introduced by Hawkins et al. as outliers (1–14).

tamination was found that allowed a predicted 90% of the data sets to be successfully completed. To avoid undue extrapolation, computation times and sample sizes were set to within the bounds of what were used for problems of that nature in our study.

The more data (and the more computation time), the greater the fraction of outliers that can be handled. Within our self-imposed bounds, we can say that outlier fractions in the 30–40% range can be reliably solved in dimension 10, 25–35% in dimension 20, and 20–25% in dimension 40. Although these bounds are crude, it does give some feel for what problems are feasible. It is likely that the sample sizes and processing times for dimension 40 are actually too small to be able to handle large amounts of contamination. For assured success with high contamination, substantially larger values of both than the ones that we used may be necessary. Multiple processor machines could also be used to increase the effectiveness of the algorithm, which is parallelizable in a number of ways (Woodruff and Rocke 1993).

5. EXAMPLES

In this section we examine the performance of the hybrid algorithm on several examples used in the literature. The first is the constructed data of Hawkins, Bradu, and Kass (1984), which consists of 75 points in dimension 3 plus a response. (We refer to these data as HBK for short.) We ex-

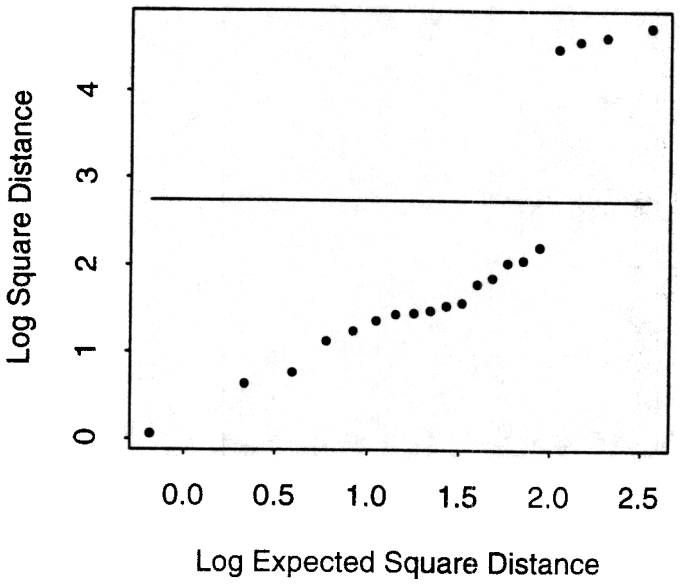

Figure 3. Mahalanobis Square Distances for the Modified Wood Gravity Data Using the Metric Derived From the Hybrid Algorithm. The horizontal line represents a 99% cutoff as estimated by Phase II of the algorithm. The four points above the cutoff are the points introduced by Rousseeuw as outliers; from largest to smallest, these are 19, 8, 6, and 4.)

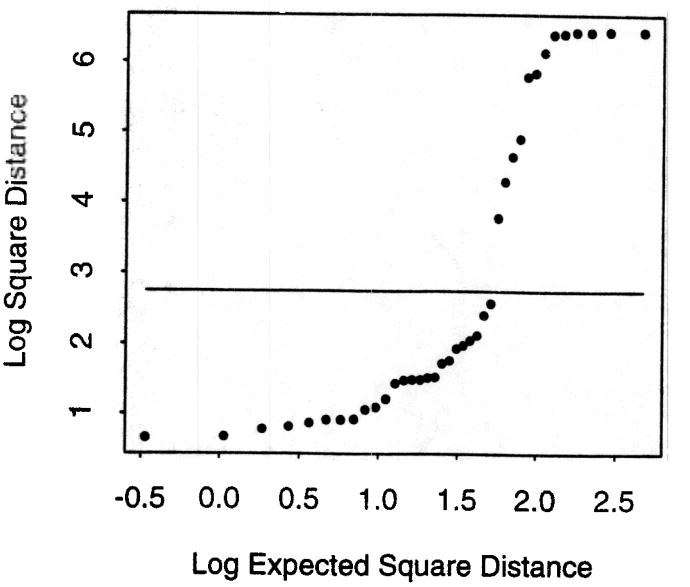

Figure 4. Mahalanobis Square Distances for the Bushfire Data Using the Metric Derived From the Hybrid Algorithm. The horizontal line represents a 99% cutoff as estimated by Phase II of the algorithm. The 13 points above the cutoff are the points identified by Maronna and Yohai as outliers using the Stahel–Donoho method; from largest to smallest, these are 35, 38, 33, 37, 34, 36, 32, 9, 8, 10, 11, 31, and 7.

amine just the three predictor variables. The first 14 points were designed to be leverage points, but only point 14 and possibly point 13 show up as such if ordinary Mahalanobis distances are used (Rousseeuw and Leroy 1987). Figure 2a shows the Mahalanobis square distances versus χ^2_3 order statistics with a horizontal line at the .99 percentage point of a χ^2_3 variable. Only point 14 appears out of line. If this point is omitted, then point 13 shows up as discrepant, and if point 13 is also omitted, then point 12 is close to the cutoff. Now no further points appear discrepant. Figure 2b shows the distances with points 12, 13, and 14 omitted; nothing appears out of the ordinary. Figure 2c shows the distances obtained from the hybrid algorithm. There is a clear separation of the 14 initial points from the remainder of the points—they are well above the cutoff derived by simulation in Phase II. The hybrid algorithm has identified the structure as specified by Hawkins et al. (1984). (Note that Hawkins et al. (1984), Rousseeuw and Leroy (1987), and others have also identified this structure.)

The second data set is the modified wood gravity data originally from Draper and Smith (1966), consisting of 20 observations of 5 explanatory variables and a response (wood specific gravity). We analyze only the explanatory variables. Rousseeuw (1984) and Rousseeuw and Leroy (1987) modified the data by replacing observations 4, 6, 8, and 19 by outliers. Figure 3 shows the distances from the hybrid algorithm. The four discrepant points are the observations that were perturbed by Rousseeuw. Again, these do not show up using the all-data metric, but the plots have been omitted because similar ones have appeared elsewhere.

The third example data set is the bushfire data used by Maronna and Yohai (1995) as an example. This consists of 38 observations (pixels) of satellite measurements on 5 frequency bands used to locate bushfire scars (Campbell 1989). Maronna and Yohai concluded that there were outliers in two groups: points 7–11, which are relatively easy to detect by various robust methods, and points 32–38, which are masked by the first group to estimators other than the Stahel–Donoho (SD) projection estimator as implemented by Maronna and Yohai. Figure 4 shows the distances by the hybrid algorithm. The most extreme group of points is 8, 9, and 32–38, whereas the next most extreme group of four is 7, 10, 11, and 31. These are almost the same points

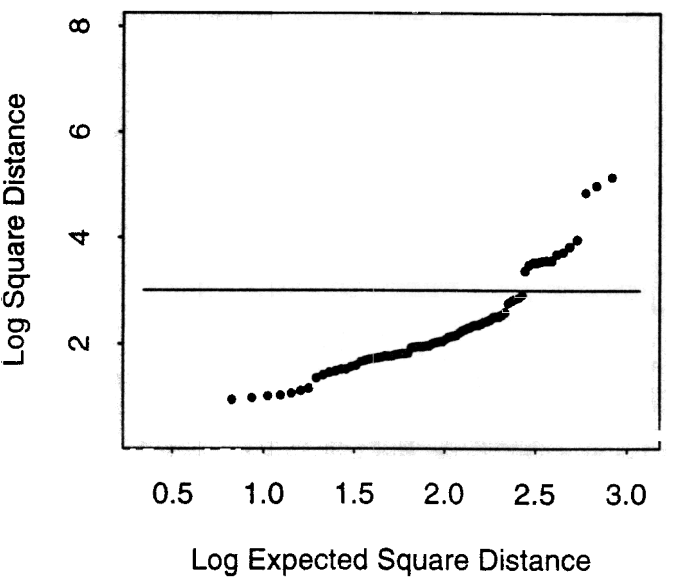

Figure 5. Mahalanobis Square Distances for the Milk Data Using the Metric Derived From the Hybrid Algorithm. The horizontal line represents a 99% cutoff as estimated by Phase II of the algorithm. The 15 points above the cutoff are the points identified by Atkinson as outliers using the Forward algorithm; from largest to smallest, these are 70, 2, 41, 1, 44, 74, 12, 13, 14, 3, 15, 47, 75, 17, and 16.

Table 9. Specifications of Example Data Sets

Data set	n	p	Number of outliers	Percentage of outliers
HBK data	75	3	14	21%
Modified wood gravity data	20	5	4	20%
Bushfire data	38	5	13	34%
Milk data	85	8	15	18%

NOTE: The HBK data are from Hawkins et al. (1984), the modified wood gravity data are from Rousseeuw and Leroy (1987), the bushfire data are from Maronna and Yohai (1995), and the milk data are from Daudin et al. (1988).

identified by the SD estimator (point 31 has been added) but sorted into slightly different groups. Further clarification of the clustering issue awaits development of methods for identifying clusters that may comprise less than half of the data.

The final data set consists of 8 measurements on 86 samples of milk (Daudin, Duby, and Trecourt 1988) analyzed by Atkinson (1994). Two of the points (63 and 64) are exact duplicates (which may explain the discrepancy between the 86 observations in Daudin et al.'s table 1 and the 85 observations said to be contained in it). We omitted the duplicate point but preserved the original numbering of the 86 points to facilitate comparison with Atkinson's results. Figure 5 shows the distances for this data set using the hybrid algorithm. The outlying points occur in three groups: point 70, points 1, 2, and 41; and the remainder. The small group of four lying just below the cutoff consists of points 11, 20, 27, and 77. The points lying above the cutoff coincide with the points identified as outliers or possible outliers by Atkinson. The appearance of clustering is even stronger here than in the bushfire data, suggesting that a more complex analysis might eventually be called for.

Table 9 summarizes the examples presented in this section. Two further features of the analysis of these data with the hybrid algorithm are worthy of note. First, each analysis was successfully accomplished with only a minimal amount of search time. Even though the fraction of contamination ranged up to 34%, in dimensions as low as 5–8, this does not fall within the problematic range of the method. Second, although there is a stochastic aspect to the MCD search, multiple tries on the same data set always yielded the same answer to many significant digits. The instability of some other methods does not seem to apply to the hybrid algorithm.

6. CONCLUSIONS

In this article we have investigated the nature of multivariate outliers and methods for their detection. We have shown that shift outliers provide a reasonable test bed for multivariate outlier detection, being difficult but not impossible to detect.

Using this test bed, we have shown a new hybrid algorithm to be superior to existing methods for this problem. Given sufficient data and processing time, this algorithm can deal with even heavily contaminated data in high dimensions. Roughly speaking, outlier fractions in the 30–40% range can be reliably solved in dimension 10, 25–35% in dimension 20, and 20–25% in dimension 40. The ability of this algorithm to accurately characterize the outliers in a multivariate distribution was also shown, using several examples from the literature.

Note: Programs in the C language to implement the hybrid algorithm and to produce the test data are available from the authors and from STATLIB. Please send requests by e-mail to dmrocke@ucdavis.edu or dlwoodruff@ucdavis.edu.

APPENDIX: PROOFS OF THEOREMS

Proof of Theorem 1

Because all of the statements are affine equivariant, we may without loss of generality take $\mu_0 = 0$ and $\Sigma_0 = \mathbf{I}$. Then, without loss of generality, we may rotate the coordinate until Ω is a diagonal matrix with diagonal elements $(\omega_1, \omega_2, \ldots, \omega_p)$. The constraint that the size (determinant) of Ω is fixed can be expressed as $\prod_{i=1}^{p} \omega_i = D$. Now $d^2_{\Sigma_0}(\mathbf{x}, \mathbf{O}) = \sum_{i=1}^{p} x_i^2$, whose expectation is $|\mu|^2 + \sum_{i=1}^{p} \omega_i$. Thus the theorem asserts that the minimum value of $\sum_{i=1}^{p} \omega_i$ subject to $\prod_{i=1}^{p} \omega_i = D$ is attained at $\omega_i = D^{1/p}$ for all $1 \le i \le p$. This can be shown by a straightforward application of the method of Lagrange multipliers.

Proof of Theorem 2

As before, because all operations are affine equivariant, we may set $\mu_0 = 0$ and $\Sigma_0 = \mathbf{I}$ without loss of generality. Also, without loss of generality, by affine equivariance, we may take μ to be a vector in which the first coordinate is η and the other coordinates are zero.

By Lemma 1, in the indicated coordinate system, Ω^{-1}, is a diagonal matrix with elements $((1-\varepsilon+\lambda\varepsilon+\varepsilon(1-\varepsilon)\eta^2)^{-1}, (1-\varepsilon+\lambda\varepsilon)^{-1}, \ldots, (1-\varepsilon+\lambda\varepsilon)^{-1})$. If the coordinates of a good point are $(x_1, x_2, \ldots, x_p)$, then

$$d^2_{\Sigma}(\mathbf{x}, \varepsilon\mu) = (x_1 - \varepsilon\eta)^2(1-\varepsilon+\lambda\varepsilon+\varepsilon(1-\varepsilon)\eta^2)^{-1} + \sum_{i=2}^{p} x_i^2(1-\varepsilon+\lambda\varepsilon)^{-1}. \quad \text{(A.1)}$$

Because the covariance matrix is $\mathbf{I}$, the expectation of this quantity is $(1+\varepsilon^2\eta^2)(1-\varepsilon+\lambda\varepsilon+\varepsilon(1-\varepsilon)\eta^2))^{-1} + (p-1)(1-\varepsilon+\lambda\varepsilon)^{-1}$, which for large η is $\varepsilon(1-\varepsilon)^{-1} + (p-1)(1-\varepsilon+\lambda\varepsilon)^{-1}$. Similarly, for a bad point,

$$E(d^2_{\Sigma}(\mathbf{x}, \varepsilon\mu)) = (\lambda^2 + (1-\varepsilon)^2\eta^2)(1-\varepsilon+\lambda\varepsilon+\varepsilon(1-\varepsilon)\eta^2)^{-1} + \lambda(p-1)(1-\varepsilon+\lambda\varepsilon)^{-1}, \quad \text{(A.2)}$$

which for large η is $(1-\varepsilon)\varepsilon^{-1} + \lambda(p-1)(1-\varepsilon+\lambda\varepsilon)^{-1}$.

For large η, the difference in $E(d^2_{\Sigma}(\mathbf{x}, \varepsilon\mu))$ between good and bad points is

$$\frac{1-\varepsilon}{\varepsilon} - \frac{\varepsilon}{1-\varepsilon} + \frac{(\lambda-1)(p-1)}{1-\varepsilon+\lambda\varepsilon} = \frac{1-2\varepsilon}{\varepsilon(1-\varepsilon)} + \frac{(\lambda-1)(p-1)}{1-\varepsilon+\lambda\varepsilon},$$

where a positive value indicates that the bad points have a larger expected Mahalanobis distance. With regard to varying λ, this is, up to a constant, $(p-1)\lambda/(1-\varepsilon+\lambda\varepsilon)$. This has derivative with respect to λ,

$$[(p-1)(1-\varepsilon+\lambda\varepsilon) - (p-1)\lambda\varepsilon]/(1-\varepsilon+\lambda\varepsilon)^2 = (p-1)(1-\varepsilon)/(1-\varepsilon+\lambda\varepsilon)^2 > 0.$$

The most difficult case is when this difference in distances is algebraically least, which is when $\lambda = 0$.

When $\lambda = 0$ (point mass contamination), the difference in the expected Mahalanobis distances is

$$\frac{1-2\varepsilon}{\varepsilon(1-\varepsilon)} - \frac{(p-1)}{}$$

which reaches zero when $\varepsilon = 1/(p+1)$. (Note that this is the breakdown point of many multivariate estimators.)

When $\lambda = 1$, the difference in the expected Mahalanobis distances is

$$\frac{1-2\varepsilon}{\varepsilon(1-\varepsilon)},$$

which is always positive but is a small fraction of the mean distance of either type of point for high dimension. Specifically, under normality, the variance of the Mahalanobis distance of a good point for large η is $2(p-1)$, so the difference between the expected Mahalanobis distances in units of the standard deviation of the Mahalanobis distance of a good point is

$$\frac{1-2\varepsilon}{\varepsilon(1-\varepsilon)\sqrt{2(p-1)}},$$

which goes to zero as the dimension rises (for fixed ε). Because both are asymptotically (in p) normal, the distributions converge.

The condition for the good and bad expected distances to be equal is (for large η)

$$\frac{1-2\varepsilon}{} + \frac{(\lambda-1)(p-1)}{1-\varepsilon+\lambda\varepsilon} = 0,$$

so

$$\lambda_0 \equiv \frac{(1-\varepsilon)(\varepsilon p - (1-\varepsilon))}{\varepsilon((1-\varepsilon)p - \varepsilon)}$$

[Received August 1993. Revised December 1995.]

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, NJ: Princeton University Press.

Atkinson, A. C. (1993), "Stalactite Plots and Robust Estimation for the Detection of Multivariate Outliers," in *Data Analysis and Robustness*, eds. S. Morgenthaler, E. Ronchetti, and W. Stahel, Basel: Birkhäuser.

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

Atkinson, A. C., and Mulira, H.-M. (1993), "The Stalactite Plot for the Detection of Multiple Outliers," *Statistics and Computing*, 3, 27–35.

Butler, R. W., Davies, P. L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, 21, 1385–1400.

Campbell, N. A. (1980), "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," *Applied Statistics*, 29, 231–237.

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

——— (1989), "Bushfire Mapping Using NOAA AVHRR Data," technical report, CSIRO.

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

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

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

Donoho, D. L. (1982), "Breakdown Properties of Multivariate Location Estimators," Ph.D. qualifying paper, Harvard University, Dept. of Statistics.

Draper, N. R., and Smith, H. (1966), *Applied Regression Analysis*, New York: John Wiley.

Glover, F. (1989), "Tabu Search—Part I," *ORSA Journal on Computing*, 1, 190–206.

——— (1990), "Tabu Search—Part II," *ORSA Journal on Computing*, 2, 4–32.

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

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

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

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

——— (1993a), "A Feasible Solution Algorithm for the Minimum Volume Ellipsoid Estimator," *Computational Statistics*, 9, 95–107.

——— (1993b), "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data," *Computational Statistics and Data Analysis*, 17, 197–210.

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

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

Kent, J. T., and Tyler, D. E. (1991), "Redescending M-Estimates of Multivariate Location and Scatter," *The Annals of Statistics*, 19, 2102–2119.

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

——— (1992), "Highly Efficient Estimators of Multivariate Location With High Breakdown Point," *The Annals of Statistics*, 20, 398–413.

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

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

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

Rocke, D. M. (1996), "Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension," *The Annals of Statistics*, in press.

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

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

——— (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications, Volume B*, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Werz, Dordrecht: Reidel.

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

Rousseeuw, P. J., and van Zomeren, B. C. (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 II*, eds. W. Stahel and S. Weisberg, New York: Springer-Verlag.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.

Stahel, W. A. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," Ph.D. thesis, ETH Zurich.

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

——— (1988), "Some Results on the Existence, Uniqueness, and Compu-

tation of the *M*-Estimates of Multivariate Location and Scatter," *SIAM Journal on Scientific and Statistical Computing*, 9, 354–362.

——— (1991), "Some Issues in the Robust Estimation of Multivariate Location and Scatter," in *Directions in Robust Statistics and Diagnostics Part II*, eds. W. Stahel and S. Weisberg, New York: Springer-Verlag.

Woodruff, D. L. (1995), "Ghost Image Processing for Minimum Covariance Determinants," *ORSA Journal on Computing*, 7-4, 468–473.

Woodruff, D. L., and Rocke, D. M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, 2, 69–95.

— - (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association*, 89, 888–896.