ROBUST COVARIANCE MATRIX ESTIMATION AND MULTIVARIATE OUTLIER DETECTION

Daniel Peña and Francisco J. Prieto

97-08

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 97-08
Statistics and Econometrics Series 04
February 1997

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

ROBUST COVARIANCE MATRIX ESTIMATION AND MULTIVARIATE OUTLIER DETECTION

Daniel Peña and Francisco J. Prieto*

Abstract

A severe limitation for the application of robust position and scale estimators having a high breakdown point is a consequence of their high computational cost.

In this paper we present and analyze several inexpensive robust estimators for the covariance matrix, based on information obtained from projections onto certain sets of directions. The properties of these estimators (breakdown point, computational cost, bias) are analyzed and compared with those of the Stahel-Donoho estimator, through simulation studies. These studies show a clear improvement both on the computational requirements and the bias properties of the Stahel-Donoho estimator.

The same ideas are also applied to the construction of procedures to detect outliers in multivariate samples. Their performance is analyzed by applying them to a set of test cases.

Key Words

Kurtosis; multivariate statistics; breakdown point; linear projection.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. E-mail: fjp@est-econ.uc3m.es. This research was supported by CICYT grants PB93-0232 and PB94-0374.

Robust covariance matrix estimation and multivariate outlier detection

Daniel Peña[1]
Dept. Estadística y Econometría
Univ. Carlos III de Madrid, Spain
E-mail: dpena@est-econ.uc3m.es

Francisco J. Prieto[1]
Dept. Estadística y Econometría
Univ. Carlos III de Madrid, Spain
E-mail: fjp@est-econ.uc3m.es

ABSTRACT

A severe limitation for the application of robust position and scale estimators having a high breakdown point is a consequence of their high computational cost.

In this paper we present and analyze several inexpensive robust estimators for the covariance matrix, based on information obtained from projections onto certain sets of directions. The properties of these estimators (breakdown point, computational cost, bias) are analyzed and compared with those of the Stahel-Donoho estimator, through simulation studies. These studies show a clear improvement both on the computational requirements and the bias properties of the Stahel-Donoho estimator.

The same ideas are also applied to the construction of procedures to detect outliers in multivariate samples. Their performance is analyzed by applying them to a set of test cases.

Keywords: Kurtosis; Multivariate Statistics; Breakdown Point; Linear Projection

1 Introduction

Most usual multivariate analysis techniques depend on the assumption of normality in the data, and require the use of estimates for both the location and scale parameters of the distribution. The presence of outliers may distort arbitrarily the values of these estimators, and render meaningless the results of the application of these techniques.

To avoid this difficulty many different classes of robust location and scale estimators have been proposed in the literature, see Huber (1981), for example. A common measure of the sensitivity of these estimators to the presence of outliers in the data is the breakdown point of the estimator, ϵ^*, defined for a given estimator T and a sample X of size n as

$$\epsilon_n^*(T, X) = \frac{1}{n} \min \left\{ m : \sup_{X_m} T(X_m) < \infty \right\},$$

where X_m denotes a sample obtained after replacing m observations randomly chosen from X with arbitrary values.

A high breakdown point will imply that, in the limit, the corresponding estimator will not be greatly affected by the presence of outliers. Also, it is expected that this property will be preserved in some measure for finite contaminations. An estimator with a breakdown point close to 0.5 would then be protected against arbitrary distortions caused by the presence of any outliers in the sample.

A significant amount of effort has been devoted in recent years to the development of procedures for the computation of robust position and scale estimators with high breakdown point, see for instance Maronna (1976), Campbell (1980), Stahel, (1981), Donoho (1982), Rousseeuw (1985), Hampel at al. (1986), Rousseeuw and Leroy (1987), Davis (1987), Rousseeuw and van Zomeren (1990), Tyler (1991),

[1]This research was supported by CICYT grants PB93-0232 and PB94-0374.

Rocke and Woodruff (1993), Maronna and Yohai (1995) and Rocke and Woodruff (1996). As a consequence, a large number of alternative estimators having this property are available. All these alternatives share the property of being based on the maximization of certain non-concave and non-differentiable criteria. As a consequence, a common practice consists on constructing these estimators through a resampling process, where candidate solutions are generated randomly from a discrete set. This procedure is terminated when the number of subsamples generated is large enough to guarantee, with a certain probability, the computation of the optimizer.

Unfortunately, the number of candidate solutions (total number of subsamples of a given size), and the number of subsamples that guarantee the computation of a solution for the optimization problem with a given probability, grow exponentially with the size of the problem. As a consequence, the corresponding procedures become computationally very expensive for even moderately sized problems. Hadi (1992, 1994) and Atkinson (1994) have presented methods to compute approximations for these estimates requiring reasonable computation times.

The robust estimation of the location and covariance matrix is very closely related to the problem of the identification of multivariate outliers. In the multinormal case, the likelihood ratio test leads to identifying outliers as points with large Mahalanobis distances from the center of the data. This test requires the estimation of the location and covariance matrix from the data, and these values can be greatly affected by the presence of more than one outlier. Thus, the identification of outliers requires as a first step a reasonably good estimate of the location and covariance matrix.

In this paper we present and analyze several alternative procedures, based on the analysis of the projections of the sample points onto a certain set of directions, that work well in practice and can be implemented using very moderate computational resources, requiring a small computation time.

2 Description of the algorithms

In this section we present several variants of an algorithm for the computation of robust covariance matrix estimators and the detection of outliers in multivariate samples. This algorithm is based on the application of robust univariate estimators for position and scale to the projections of the sample points onto certain directions. These ideas are similar to those for the Stahel-Donoho estimators for position and scale (Donoho, 1982), except for the procedure to select the projection directions.

The Stahel-Donoho estimator proceeds by computing the maximum over all possible projections of the robust measure of distance for each sample point $\mathbf{x}_i$, that is, the weight assigned to each point is obtained from

$$r_i = \max_d \frac{|d^T x_i - \text{median}(d^T x_j)|}{\text{MAD}(d^T x_j)} \tag{1}$$

As these values are obtained from all possible projection directions, their computation requires the solution for each sample point of one global optimization problem with discontinuous derivatives. The computational cost involved has led to the development of alternative procedures, based on resampling schemes. For example, in Rousseuw (1993) directions are obtained by randomly selecting p observations from the original sample, and computing the direction orthogonal to the hyperplane defined by these observations. The maximum is then determined only for the (finite) set of directions obtained in this way. The resulting value for the "outlyingness" measure r_i is used to assign weights to the points for the computation of the position and scale estimators as the weighted sample mean and covariance.

Maronna and Yohai (1995) have shown that this estimator has the least asymptotic bias and the maximum efficiency among a set of affine equivariant estimators with high breakdown point.

2.1 Proposed scale estimator

Our proposed scale estimation algorithms separate from this scheme in the way the projection directions are selected, this being the step that poses the most significant computational demands.

In univariate normal data outliers have often been associated to large kurtosis values, and some well known tests of normality are based on the asymmetry and kurtosis coefficients. These ideas have also been used to test for multivariate normality (Malkovich and Afifi (1973)). Finally, some projection indices that have been applied in projection pursuit algorithms are related to the third and fourth moments (Jones and Sibson (1987), Posse (1995)).

Following these lines, to compute the measures r_i and the associated weights $w(r_i)$, our proposed schemes make use of a prespecified set of p directions, that can be obtained as the solution of a set of p simple smooth optimization problems, with limited computational effort. More specifically, we propose computing the outlyingness measure r_i from the directions that maximize some high moments of the projected data (kurtosis), or the coefficients corresponding to these moments.

The choice of the moments more adequate for our purpose (covariance estimation/outlier identification) is not immediately clear. What we propose is to explore several alternatives of the basic algorithm, to determine the best choice for the criterion to replace (1). We consider the following variants for this algorithm:

Algorithm 1. Maximization of the kurtosis. In this first algorithm, the projection directions are computed as those that maximize the kurtosis of the projected observations. The algorithm is constructed so that it retains some invariance properties for data transformations. Given the choice of criterion (kurtosis) and the need to achieve a high breakdown point, the algorithm will not be invariant to affine transformations, but it will be invariant to changes of center and scale.

1. The data is scaled and centered. Let x_j denote the column vector of observations corresponding to the $j-$th variable, the median md_j and the MAD ν_j for each of the variables. These values are computed, and the points are centered and scaled
$$y_{ij}^{(1)} \equiv y_{ij} = (x_{ij} - md_j)/\nu_j. \tag{2}$$
We set the iteration index $k = 1$.

2. The direction that maximizes the kurtosis for the scaled points is obtained as the solution of the problem
$$\begin{aligned} d_k = \quad & \arg\max_d \quad \sum_{i=1}^{n} (d^T y_i^{(k)})^4 \\ & \text{s.t.} \qquad d^T d = 1 \end{aligned} \tag{3}$$

3. The sample points are projected onto the subspace orthogonal to $(d_1, d_2, ..., d_k)$,
$$y_i^{(k+1)} = (I - d_k d_k^T) y_i^{(k)}, \tag{4}$$
and we set $k = k + 1$. If $k < n$, go to step 2.

4. From the set of orthogonal directions we obtain the weights for the robust estimators.

   For each sample point the outlyingness measure

$$r_i = \max_k \frac{\left| d_k^T y_i - \text{median}(d_k^T y_l) \right|}{\text{MAD}(d_k^T y_l)} \tag{5}$$

   is computed.

5. This measure is then used to estimate a robust center and/or a robust covariance matrix for the observations,

$$\begin{aligned} m &= \frac{\sum_1^n w_i x_i}{\sum_1^n w_i} \\ S &= \frac{\sum_1^n w_i (x_i - m)(x_i - m)^T}{\sum_1^n w_i} \end{aligned},$$

   where $w_i = w(r_i)$ is a function of the outlyingness measure r_i. For example, the Huber function can be used,

$$w(r) = I_{\{r \le c\}} + \frac{c^2}{r^2} I_{\{r > c\}},$$

   where $c = \sqrt{\chi^2_{p,0.95}}$.

Algorithm 2. Maximization of the kurtosis coefficient. In order to have an algorithm that is affine equivariant, we have replaced the kurtosis by its coefficient in the criterion used to determine the projection directions. The resulting algorithm is identical to Algorithm 1, except that Step 1 is no longer necessary, as the algorithm will be affine equivariant. Step 2 is replaced by changing the normalization of the projection direction, so that the kurtosis coefficient (as opposed to the fourth moment) is maximized. The new step is

2. The direction that maximizes the coefficient of kurtosis for the original points is obtained as the solution of the problem

$$\begin{aligned} d_k = \quad & \arg\max_d \quad \sum_{i=1}^n (d^T y_i^{(k)} - \frac{1}{n} \sum_{j=1}^n d^T y_j^{(k)})^4 \\ & \text{s.t.} \quad \frac{1}{n} \sum_{i=1}^n (d^T y_i^{(k)} - \frac{1}{n} \sum_{j=1}^n d^T y_j^{(k)})^2 = 1 \end{aligned} \tag{6}$$

The constraint in the preceding formulation is used to normalize the size of d by requiring the sample variance to be equal to one. As a consequence, under this condition the maximization of the kurtosis coefficient is reduced to the maximization of the fourth central moment.

Note that the property of orthogonality is not preserved by general affine transformations. As a consequence, Step 3 needs also to be modified to ensure that the resulting algorithm is affine equivariant. The projection is done to ensure that the resulting directions are orthogonal with respect to the sample covariance matrix, that is, we require that $d_i^T S d_j = 0$ for all $i \neq j$. The projection step (4) is replaced with

$$y_i^{(k+1)} = (I - \frac{1}{d^T S d} d d^T S) y_i^{(k)},$$

where S denotes the original sample covariance matrix.

Algorithm 3. Hybrid algorithm. While Algorithm 1 has a high breakdown point, Algorithm 2 is affine equivariant but its breakdown point is low (see Appendix A). It is possible to combine the procedures in both algorithms, to exploit the advantages of each one. This can be done by generating the set of $2p$ directions obtained from the application of both methods, that is, we apply Step 2 from both algorithms, and conduct Steps 4 and 5 using the full set of $2p$ directions, in order to obtain the outlyingness measures and the weights for the estimators.

Algorithm 4. Maximization of an absolute deviation measure. In Appendix A it is shown that using the kurtosis coefficient as an indication of outlyingness has some disadvantages with respect to the breakdown point of the resulting estimator. An alternative with better breakdown properties can be constructed by replacing the objective function of the optimization problem in Step 2 of Algorithm 2 with the following maximization problem

$$\begin{aligned} d_k = \quad \arg\max_d \quad & \sum_{i=1}^{n} |d^T y_i^{(k)} - \frac{1}{n}\sum_{j=1}^{n} d^T y_j^{(k)}|^3 \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n}(d^T y_i^{(k)} - \frac{1}{n}\sum_{j=1}^{n} d^T y_j^{(k)})^2 = 1 \end{aligned} \tag{7}$$

that is, the fourth moment has been replaced with the third moment for the absolute deviations.

2.2 Computation of the projection directions

The most relevant aspect of these algorithms is the computation of the projection directions (step 2 in the algorithms); we analyze it in greater detail. Consider first Algorithm 1; the optimality conditions for (3) are

$$\begin{aligned} 4\textstyle\sum_{i=1}^{n}(d^T y_i^{(k)})^3 y_i^{(k)} - 2\lambda d = 0 \\ d^T d = 1 \end{aligned}$$

and the first of these conditions can be rewritten as

$$\left(\textstyle\sum_{i=1}^{n} 2(d^T y_i^{(k)})^2 y_i^{(k)} y_i^{(k)T}\right) d = \lambda d. \tag{8}$$

Note that this equation indicates that the solution of the problem, d, should be an eigenvector of the matrix

$$M_k(d) \equiv \textstyle\sum_{i=1}^{n}(d^T y_i^{(k)})^2 y_i^{(k)} y_i^{(k)T},$$

that is, of a weighted covariance matrix for the sample, with positive weights (depending on d) given by $(d^T y_i^{(k)})^2$.

Also, if both sides of equation (8) are multiplied by d and the second condition is used, we obtain that λ is equal to the value of the objective function, and from (8) it follows that d should be a unit eigenvector of $M_k(d)$ corresponding to its largest eigenvalue (the principal component).

Given this result, we have used the following iterative procedure to compute the direction d:

1. Select a initial value d_0 such that $\|d_0\| = 1$.
2. Compute d_{k+1} as the (unit) eigenvector associated with the largest eigenvalue of $M(d_k)$. We choose the eigenvector having a positive first component.

3. Terminate whenever $\|d_{k+1} - d_k\| < \epsilon$.

This procedure converges very quickly for problems of reasonable size, and is much faster than the Stahel-Donoho resampling procedure.

It is easy to see that this algorithm will work as intended in extreme cases. Suppose that we add to a sample of $n(1-\alpha)$ observations x_i from a normal distribution $N(0,\Sigma)$ a number $n\alpha$ of identical outliers x_a. Note that this case of $n\alpha$ identical outliers is, according to Rocke and Woodruff (1996), the hardest to detect. For $k = 1$ (8) can be written as

$$\sum_i (d^T y_i)^3 y_i + n\alpha (d^T y_a)^3 y_a = \lambda d, \tag{9}$$

where the y_i, y_a have been defined according to (2). If we let $\|x_a\| \to \infty$, we must have that, with probability one, all y_i remain bounded and $\|y_a\| \to \infty$. As a consequence, if we let $y_a = \omega u$ with $\|u\| = 1$, condition (9) is equivalent to

$$\sum_i (d^T \frac{1}{\omega} y_i)^3 \frac{1}{\omega} y_i + n\alpha (d^T u)^3 u = \frac{\lambda}{\omega^4} d,$$

and after taking limits as $\omega \to \infty$, the two solutions of the equation that satisfy the constraint are

$$d^T u = 0, \quad \|d\| = 1, \quad \lambda = 0 \qquad \text{and} \qquad d = u, \quad \frac{\lambda}{\omega^4} = n\alpha.$$

The first choice corresponds to the minimizer, while the second one defines the maximizer. As a consequence, the procedure is able to detect the direction along which the outliers are going to infinity.

Consider now Algorithm 2, and assume that initially the points have been normalized to have mean equal to zero. This normalization poses no problem, as the algorithm is affine equivariant. From (6) the optimality conditions are

$$\begin{array}{r} 4\sum_{i=1}^n (d^T y_i^{(k)})^3 y_i^{(k)} - 2\lambda \sum_{i=1}^n d^T y_i^{(k)} y_i^{(k)} = 0 \\ \sum_{i=1}^n (d^T y_i^{(k)})^2 = n. \end{array}$$

We obtain the equation

$$\left(\sum_{i=1}^n \left((d^T y_i^{(k)})^2 - \sum_{j=1}^n (d^T y_j^{(k)})^4 \right) y_i^{(k)} y_i^{(k)T} \right) d = 0. \tag{10}$$

The solution of the problem, d, should be a vector in the null-space of the matrix

$$\tilde{M}_k(d) \equiv \sum_{i=1}^n \left((d^T y_i^{(k)})^2 - \sum_{j=1}^n (d^T y_j^{(k)})^4 \right) y_i^{(k)} y_i^{(k)T},$$

that is, of a weighted covariance matrix for the sample, with weights (possibly negative and depending on d) given by $(d^T y_i^{(k)})^2 - \sum_{j=1}^n (d^T y_j^{(k)})^4$.

The iterative procedure to compute the direction d is now:

1. Select a initial direction d_0, and scale it so that the constraint $\sum_{i=1}^n (d^T y_i^{(k)})^2 = n$ is satisfied.

2 Compute d_{k+1} as the eigenvector associated with the eigenvalue of $M(d_k)$ closest to zero. We choose the eigenvector having a positive first component.

3. Terminate whenever $\|d_{k+1} - d_k\| < \epsilon$.

This procedure also converges very quickly for problems of reasonable size.

For Algorithm 4, the procedure is similar to the one presented for Algorithm 2, with the only difference that the equation to be solved, for the case where the observations have been normalized to have mean equal to zero (this algorithm is also affine equivariant), is now

$$\left(\sum_{i=1}^{n}\left(|d^T y_i^{(k)}| - \sum_{j=1}^{n}|d^T y_j^{(k)}|^3\right) y_i^{(k)} y_i^{(k)T}\right) d = 0.$$

2.3 Multivariate outlier detection

As the preceding procedures are based on the assignment of weights related to the outlyingness of each observation, they should be easy to adapt to the detection of outliers in a multivariate sample; in this case we would be interested in assigning weights zero and one to each point in the sample.

To formalize this procedure, we propose the following algorithm:

1. Apply any of the preceding procedures for the estimation of scale to the original sample, and obtain the outlyingness measures for each sample point from 5.

2. Eliminate from the sample those observations having values $r_i > c_1$, and repeat step 1.

   In our computational experiments we have taken $c_1 = 3$.

3. Terminate the procedure either after all remaining observations have outlyingness measures smaller than c_1, or before the number of observations left in the sample becomes less than half the original number of observations.

4. The sample mean m and covariance matrix S are computed for the remaining observations, and the Mahalanobis distances for the whole sample are computed using these values,

$$d_i = (x_i - m)^T S^{-1}(x_i - m).$$

5. Those observations having distances such that

$$\gamma d_i > \chi^2_{p,v} \qquad v = \alpha/n,$$

   where α is the desired level of significance, are labelled as outliers.

   The value $\gamma < 1$ is a factor that attempts to correct the bias introduced in S given that some of the observations (those considered most likely to be outliers) have been removed prior to its computation. As the test to remove the observations depends on the outliers, it is not possible to compensate for this bias exactly. The value of γ that we have used in our tests has been obtained from simulation experiments for an uncontaminated multivariate normal sample, and is shown in the following table:

p	2	3	4	5	6	8	10	15	20
γ	0.72	0.69	0.65	0.63	0.60	0.55	0.51	0.41	0.33

Table 1: Correction factors

In order to justify the choices made in this algorithm note that, for an uncontaminated sample from a normal distribution, if we take $c_1 = 3$ the value of $P(r_i \leq 3)$ converges asymptotically to 0.956 ($\simeq 0.95$). Also, in the limit the values d_i will follow a χ^2_p distribution. The Bonferroni bound that leads to $\nu = \alpha/n$ is probably conservative.

3 Properties of the estimators

In this section we analyze the properties of the preceding estimators. We first concentrate on those properties that may be justified through the use of analytic tools, such as the affine equivariance of the estimators and their breakdown point. Some other properties (bias, identification errors) have been studied via simulation experiments, and these are covered in the latter part of the section.

An important property of the algorithms presented in this paper is the very reduced computational cost associated with its application. The optimization problems to be solved in Step 2 are much simpler than the corresponding optimization problems associated with the computation of the outlyingness measures in the Stahel-Donoho procedure. As a consequence, the computation time is significantly smaller for problems of a given size, even if a resampling procedure is used in the computation of the Stahel-Donoho estimator.

3.1 Affine equivariance and breakdown points

Algorithm 1. As a consequence of the transformations introduced in Step 1 of the procedure to estimate the sample scale, the estimator computed according to this procedure will be invariant to translations and changes of scale; nevertheless, it will not be invariant to general affine transformations.

Regarding the breakdown point, if we have some observations going to infinity, the kurtosis along the corresponding direction will go to infinity, and the projections for these points along the direction maximizing the kurtosis will also grow without bound. As a consequence, the outlyingness measure r_i for these observations will become arbitrarily large, ensuring a high breakdown point (see also the comments in Section 2.2 for the case of point-mass contamination).

Algorithm 2. The algorithm is affine equivariant. On the other hand, it is not possible to show that it has a high breakdown point, even if it works very well for contaminations due to outliers grouped in a single cluster. In Appendix A it is shown that when all outliers are concentrated at the same point (point-mass contamination), the projection direction obtained in Step 2 of the algorithm is always (that is, with independence of the proportion of outliers or the distance from the outliers to the center of the uncontaminated observations) orthogonal to the direction where the outliers are concentrated (measured from the center of the uncontaminated observations), except for the last computed direction (when $k = p$). As a consequence, this method may not be able to detect the outliers in the case of concentrated contaminations when the observations are slightly apart from each other and are far removed from the uncontaminated sample (the direction of the outliers may not be recognized in this case due to the condition that all directions should be orthogonal with respect to the sample covariance matrix), or when the outliers form several clusters.

Algorithm 3. As a combination of Algorithms 1 and 2, this algorithm will have a high breakdown point, but will not be affine equivariant. On the other hand (from Algorithm 2) some of the directions generated by the algorithm should be preserved

under affine transformations, and it should also present a satisfactory behavior for point-mass contaminations not far from the sample center.

Algorithm 4. This last algorithm shares with Algorithm 2 the property of being affine equivariant, but it has better breakdown point properties. The breakdown point of the algorithm is analyzed in Appendix B.

3.2 Simulation results. Scale estimation

We have analyzed the bias and variability of the procedures through an extensive set of simulation experiments. As we mentioned above, we have chosen to compare these results with those obtained for the Stahel-Donoho estimator with subsampling, as it is the estimator with the best asymptotic bias and efficiency, see Maronna and Yohai (1995).

In these simulations, and for a given contamination level α, we have generated a set of $(1-\alpha)n$ points from a $\mathrm{N}(0, I)$ distribution, and we have added αn additional points generated from a $\mathrm{N}(ce_1, 0.1I)$ distribution, where e_1 denotes the unit vector along the first coordinate direction and c is the distance from the origin. The choice of a concentrated contamination pattern can be justified from the difficulty associated with its detection; see Maronna and Yohai (1995) for some practical remarks on these contamination patterns, and Rocke and Woodruff (1996) for a theoretical analysis.

As Algorithms 1 and 3 are not affine equivariant, for each sample we have generated a random matrix A with singular values equal to $\sigma_i = 2^i$. The data for the experiments have been obtained as $y_i = Ax_i$, where x_i denotes the observations from the original sample.

The experiments have been repeated for different values of the sample space dimension ($p = 5$, 10, 15 and 20), contamination level ($\alpha = 0.2$ and 0.3) and distance of the outliers to the uncontaminated sample ($c = 2\sqrt{\chi^2_{p,0.95}}$ and $4\sqrt{\chi^2_{p,0.95}}$). The number of subsamples for the Stahel-Donoho estimator has been fixed for all cases at 5000. Finally, for each set of values 500 samples have been generated, and the Stahel-Donoho and proposed robust estimators of the covariance matrix have been computed. In order to compare the results, the condition numbers of the matrices generated by each of the procedures have been obtained, after rescaling the matrices using the transformation that makes the sample covariance of the uncontaminated sample equal to the identity matrix. As a result, a low bias using this measure would correspond to a value close to one. Table 2 shows the results of this comparison.

Note that all the proposed algorithms show improvements in the bias of the scale estimator with respect to the Stahel-Donoho algorithm with resampling. In particular, some of the algorithms are shown to be remarkably efficient, see for example the results for the hybrid algorithm (Algorithm 3).

3.3 Outlier detection

In order to test the detection of outliers in a meaningful way, we have chosen to compare the performance of our proposed algorithms on four datasets obtained from the literature. For each dataset a table with the corresponding results is shown, indicating the number of observations identified as outliers for each of the preceding algorithms, the list of the outliers from largest to smallest Mahalanobis distance, and the relative distance between the last observation identified as an outlier and the first observation considered "normal". The datasets used for this comparison are:

c	α	p	SD	Alg.1	Alg.2	Alg.3	Alg.4
$2\sqrt{\chi^2_{p,0.95}}$	0.2	5	11.54	10.00	7.13	6.64	6.78
		10	31.19	19.16	16.40	11.46	13.50
		15	68.87	27.08	36.74	17.88	25.81
		20	140.31	34.23	72.38	28.62	48.41
	0.3	5	54.93	26.99	29.49	26.32	27.19
		10	106.56	44.87	67.31	49.38	64.20
		15	153.11	60.24	112.62	83.19	119.83
		20	192.44	74.54	180.90	136.38	184.64
$4\sqrt{\chi^2_{p,0.95}}$	0.2	5	11.96	12.55	8.60	6.96	7.33
		10	36.05	34.92	29.87	11.75	18.67
		15	97.61	59.11	92.13	18.27	47.11
		20	273.40	87.03	237.14	29.39	119.77
	0.3	5	91.64	45.90	55.15	29.63	39.94
		10	275.18	118.12	238.69	58.90	160.97
		15	536.07	186.30	440.96	103.12	417.44
		20	768.36	247.84	711.80	188.14	729.60

Table 2: Outlyingness measures

- The Hawkins, Bradu, Kass set (Hawkins et al. (1984)), composed of 75 observations in dimension 3. The first 14 observations are outliers. Note that for

Algorithm	# outliers	outliers	rel. gap
Stahel-Donoho	14	14,12,13,11,4,5,9,3,10,7,2,6,8,1	0.99
Algorithm 1	14	14,12,11,13,4,5,9,3,10,7,6,2,8,1	0.98
Algorithm 2	14	14,12,11,13,4,5,9,3,10,7,6,2,8,1	0.99
Algorithm 3	14	14,12,13,11,4,5,9,3,10,7,2,6,8,1	0.99
Algorithm 4	14	14,12,13,11,4,5,9,3,10,7,2,6,8,1	0.98

Table 3: Hawkins-Bradu-Kass results

all algorithms the 14 outliers have been correctly identified, and their ordering is very similar in all cases.

- The bushfire dataset, composed of 38 observations in dimension 5. In Maronna and Yohai (1995) 13 observations were identified as outliers, from largest to smallest 35, 38, 33, 37, 34, 36, 32, 9, 8, 10, 11, 31 and 7. The results shown in Table 4 are markedly different for Algorithm 1 and the other methods, stressing the importance of using an affine equivariant procedure. Note that in all cases too many observations are identified as outliers, suggesting that the cutoff point proposed in Section 2.3 may be too conservative.

- The milk dataset, composed of 86 observations in dimension 8. Atkinson (1994) identified the observations 70, 2, 41, 1, 44, 74, 12, 13, 14, 3, 15, 47, 75, 17 and 16 as outliers, for a total of 15 outliers. Again, the results in Table 5 for the different methods are very similar, and as in the preceding case, they identify as outliers a larger number of observations than Atkinson.

- A synthetic data set with 34 observations, 30 of them generated from a normal multivariate distribution in dimension 6 with an ill-conditioned covariance

Algorithm	# outliers	outliers	rel. gap
Stahel-Donoho	16	33,35,38,34,37,36,32,9,8, 31,10,11,7,12,30,29	0.38
Algorithm 1	9	9,8,7,10,11,12,32,31,29	0.12
Algorithm 2	16	33,35,34,38,37,36,32,9,8, 31,10,11,7,30,29,12	0.57
Algorithm 3	16	33,35,34,38,37,36,32,9,8, 31,10,11,7,30,29,12	0.57
Algorithm 4	17	32,33,35,34,36,38,37,31,9, 8,7,10,11,30,29,12,28	0.71

Table 4: Bushfire results

Algorithm	# outliers	outliers	rel. gap
Stahel-Donoho	20	70,2,41,44,1,12,74,13,14,15, 47,3,75,16,11,20,27,17,77,18	0.22
Algorithm 1	19	70,2,41,1,74,12,44,13,75,3, 15,14,47,16,20,11,77,17,18	0.22
Algorithm 2	19	70,2,41,1,44,74,12,15,14,13, 47,75,3,16,11,20,17,77,18	0.02
Algorithm 3	20	70,2,41,1,44,74,3,12,75,13, 14,15,47,16,20,17,77,11,18,27	0.18
Algorithm 4	20	70,2,1,41,44,74,12,13,15,14, 16,75,3,47,20,11,17,18,77,27	0.06

Table 5: Milk results

matrix, and the last four generated as outliers from a different normal distribution with small covariance matrix, and located along the smallest principal component for the initial 30 observations. This dataset corresponds to the typically difficult case of a concentrated contamination. Again, Algorithm 1

Algorithm	# outliers	outliers	rel. gap
Stahel-Donoho	8	33,31,32,34,5,8,22,24	0.14
Algorithm 1	4	13,5,8,22	0.20
Algorithm 2	7	5,33,31,32,34,14,8	0.25
Algorithm 3	6	33,31,32,5,34,8	0.44
Algorithm 4	12	33,31,32,14,18,34,10,22,26,20,8	0.15

Table 6: Synthetic data results

does poorly, as could be expected from the properties of the algorithm. The remaining procedures perform reasonably well, but again identify too many observations as outliers.

From all these results, the main conclusion is that all algorithms, except for Algorithm 1, behave quite acceptably, but the cutoff point for the identification of outliers has been chosen as too pessimistic. Nevertheless, the selection of a better cutoff value is complicated by two considerations: firstly, a value having a probabilistic interpretation, such as for example identifying outliers with a given probability, would be affected by the presence and location of these outliers, and that would greatly complicate its determination, and the corresponding algorithm; secondly,

for real data it is not always clear which observations should be labeled as outliers. To illustrate this situation, consider Algorithm 3 when applied to the milk data. Figure 1 shows the ordered values of the logarithms of the Mahalanobis distances for the observations. The continuous line gives the cutoff point as described in the algorithm, and the dashed line gives a cutoff point that would identify the outliers as described in Atkinson, except for observation 20. The difference is not large, and if anything it seems that the proposed cutoff point gives a better separation between outliers and normal observations.

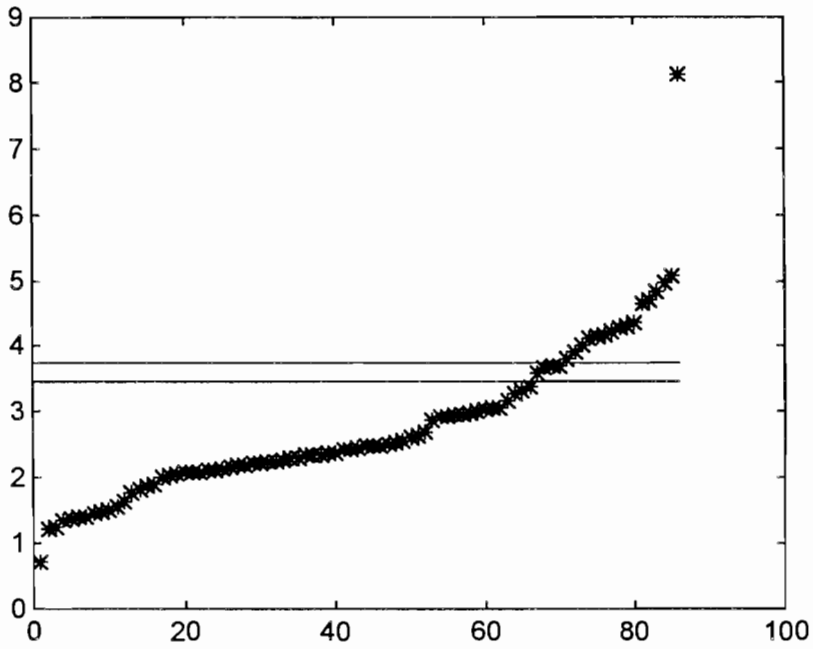

Figure 1: Milk data: ordered log of Mahalanobis distances and cutoff point using Algorithm 3

3.4 Running times

Finally, a most important property of the proposed algorithms is their limited computational requirements, both from the point of view of their ease of implementation, and of their reduced running times.

Table 7 illustrates these advantages by comparing the running times required by each one of the algorithms and the Stahel-Donoho procedure (for 5000 replications). Each value in the table corresponds to the running time to complete the estimation of the covariance matrix 100 times, for each one of the datasets described in this section. The running times are given in seconds, and have been obtained on an HP 735 workstation for Matlab versions of each algorithm.

Note the decrease by several orders of magnitude in the running times. This decrease is particularly marked when the number of observations in the sample is large with respect to the dimension (Hawkins, Bradu, Kass data set). Note also that as the dimension increases, the number of replications needed by Stahel-Donoho to guarantee a satisfactory result would increase exponentially (this number has been kept fixed for the preceding experiment), while the proposed algorithms would show only a polynomial increase in their running times.

	Stahel-Donoho	Alg. 1	Alg. 2	Alg. 3	Alg. 4
Hawkins, Bradu, Kass	3446.9	4.0	7.8	11.1	10.9
Bushfire	1530.9	5.4	10.6	15.5	21.3
Synthetic data	1432.3	10.9	15.6	26.0	25.7
Milk data	3593.3	22.9	63.9	87.5	120.3

Table 7: Running times

4 Conclusions

From the analysis of the proposed procedures, and although some of their theoretical properties may be improved upon, they seem to behave very well in practice, at least those that are affine equivariant or make use of information generated in an affine equivariant manner. In particular we recommend the hybrid algorithm, as it has shown a very good performance on all the tests we have conducted. On the other hand, they are much simpler to implement, and much faster than equivalent procedures with high breakdown point, such as the Stahel-Donoho estimator. The running times are shown to be several orders of magnitude smaller than those required by Stahel-Donoho.

In summary, the procedures presented in this paper for the estimation of robust covariance matrices and the identification of outliers seem to be good choices whenever these techniques must be applied and computational efficiency is an important consideration (large sample sizes or high data dimensions).

Appendix A

In this appendix we analyze the breakdown point properties of Algorithm 2 (the maximization of the kurtosis coefficient). In particular, we study the behavior of the method for a specific but representative contamination pattern, point-mass contamination. To model this case consider a sample $y_1, ..., y_m$ of $m \equiv n(1-\alpha)$ observations obtained from a $N(0, I)$ distribution, and a group of $n\alpha$ observations concentrated at a distance δ along the first coordinate direction, δe_1 (where e_1 denotes the first unit vector). To simplify the computations in what follows we will assume that the observations satisfy $\sum_{i=1}^{m} y_i = 0$ and $\sum_{i=1}^{m} y_i y_i^T = mI$. Note that Algorithm 2 is affine equivariant, implying no loss of generality due to the preceding assumptions.

Consider now an arbitrary projection direction ωu, where $\|u\| = 1$. The projected observations will be composed of a group of $n(1-\alpha)$ observations following a univariate $N(0, \omega^2)$ distribution, and a group of $n\alpha$ observations concentrated at $\delta\omega u_1$, where $u_1 = e_1^T u$ denotes the first component of u.

The value of ω will be determined by the satisfaction of the constraint (variance equal to one) in (6). Replacing the preceding values in this constraint, using the notation $x_i \equiv \omega u^T y_i$ and taking into account that $\sum_{i=1}^{m} x_i = 0$ and $\sum_{i=1}^{m} x_i^2 = m\omega^2 = n(1-\alpha)\omega^2$, we obtain

$$(1-\alpha)\omega^2 + \alpha\delta^2\omega^2 u_1^2 - (\alpha\delta\omega u_1)^2 = 1.$$

The value of ω that satisfies this constraint is

$$\omega^2 = \frac{1}{(1-\alpha)(1+\alpha\delta^2 u_1^2)}. \tag{11}$$

Consider now the expression for the kurtosis—the objective function in (6),

$$\psi(u_1) \equiv \sum_{i=1}^{m} (x_i - \alpha\delta\omega u_1)^4 + n\alpha\,(\delta\omega u_1 - \alpha\delta\omega u_1)^4 .$$

If the terms in this expression are expanded and expected values are taken, we have that on the average,

$$\varphi(u_1) = E\left[\psi(u_1)\right] = n(1-\alpha)\left(3\omega^4 + 6\left(\alpha\delta\omega u_1\right)^2\omega^2 + \left(\alpha\delta\omega u_1\right)^4\right) + n\alpha(1-\alpha)^4\left(\delta\omega u_1\right)^4 .$$

Replacing the value (11) and rearranging terms we obtain

$$\varphi(u_1) = \frac{n(3 + 6\alpha^2\delta^2 u_1^2 + (\alpha - 3\alpha^2 + 3\alpha^3)\delta^4 u_1^4)}{(1-\alpha)(1+\alpha\delta^2 u_1^2)^2}.$$

Observing that $0 \leq u_1^2 \leq 1$, the local extrema for this expression as a function of u_1 lie at 0, 1 and

$$u_1^2 = \frac{3}{\delta^2}\frac{1-\alpha}{3\alpha - 1}.$$

Comparing the values of φ at these points, we have that the global maximizer of $\varphi(u_1)$ is found at $u_1 = 0$, independently of the values of α and δ, that is, on a direction *orthogonal* to the direction where the outliers are located.

Note that this situation does not pose a difficulty for the proposed algorithm, as the direction where the outliers are located will nevertheless be selected as one (the last one) of the set of p orthogonal directions generated by the algorithm.

But if the outliers do not lie exactly in the same point or there is more than one cluster of outliers, the last computed direction need not be one along which the outliers will be revealed, and the algorithm may not be able to compute the correct projection directions. As a consequence it may have a low breakdown point.

Appendix B

Consider Algorithm 4, and a sample $x_1, ..., x_n$. Consider further that in this sample we have $n(1-\alpha)$ arbitrary sample points y_i, $i = 1, ..., n(1-\alpha)$, and we also have $n\alpha$ points (the outliers) at a set of arbitrary locations, v_i, $i = 1, ..., n\alpha$; some of the outliers are made to go to infinity, that is, for some i we have $\|v_i\| \to \infty$.

To analyze the breakdown point of this algorithm, note that the measure of outlyingness (5) is computed in terms of robust univariate location and scale measures. It is then sufficient to find the smallest value of α for which the direction d solution of the maximization problem (7), is one onto which the outliers have a bounded projection. In any other case, by applying iteratively the proposed algorithm and removing those observations with very high values of r_i, it would be possible to detect all outliers with a breakdown point equal to that of the univariate measures used in the computation of (5), that is, 50%.

Another important consideration is that, by requiring that the directions generated by the algorithm be orthogonal with respect to the sample covariance matrix, we have that the breakdown point will be defined by the projections onto the first direction obtained from (7). If this direction is not correctly identified, then there is no guarantee that any of the remaining directions will be able to identify the outliers. To further illustrate this point, from the affine equivariance of the algorithm it is equivalent to consider the case when the observations have been rescaled to have $S = I$. In this case, the defining property of the direction along which the (unscaled) outliers go to infinity is that the MAD of its (scaled) projections is

equal to zero (in the limit), and this property is quickly lost when small errors are introduced in the identification of this direction.

As a consequence, we will study the breakdown point of this algorithm by comparing the smallest value that the criterion

$$\varphi(d) \equiv \frac{\frac{1}{n}\sum_{i=1}^{n}\left|d^T x_i - \frac{1}{n}\sum_{j=1}^{n} d^T x_j\right|^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(d^T x_i - \frac{1}{n}\sum_{j=1}^{n} d^T x_j\right)^2\right)^{3/2}}, \tag{12}$$

may take along any direction d where the outliers go to infinity, with the largest value that this same criterion may take along any other direction where the outliers remain bounded.

For a given direction d we introduce the notation $z_i \equiv d^T y_i$ and $w_i \equiv d^T v_i$. Using this notation, the criterion (12) takes the form

$$\varphi(d) \equiv \frac{\sqrt{n}\left(\sum_{i=1}^{n(1-\alpha)}\left|z_i - \frac{1}{n}\left(\sum_{j=1}^{n(1-\alpha)} z_j + \sum_{j=1}^{n\alpha} w_j\right)\right|^3 + \sum_{i=1}^{n\alpha}\left|w_i - \frac{1}{n}\left(\sum_{j=1}^{n(1-\alpha)} z_j + \sum_{j=1}^{n\alpha} w_j\right)\right|^3\right)}{\left(\sum_{i=1}^{n(1-\alpha)}\left(z_i - \frac{1}{n}\left(\sum_{j=1}^{n(1-\alpha)} z_j + \sum_{j=1}^{n\alpha} w_j\right)\right)^2 + \sum_{i=1}^{n\alpha}\left(w_i - \frac{1}{n}\left(\sum_{j=1}^{n(1-\alpha)} z_j + \sum_{j=1}^{n\alpha} w_j\right)\right)^2\right)^{3/2}}.$$

For a direction along which the outliers are going to infinity, d_∞, we may write $w_i = \beta_i w$, where $w \to \infty$ and $|\beta_i| \leq 1$, with at least one $\beta_i = 1$. If we take limits when $w \to \infty$, we obtain

$$\varphi(d_\infty) = \sqrt{n}\frac{\sum_{i=1}^{n(1-\alpha)}\left|\frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right|^3 + \sum_{i=1}^{n\alpha}\left|\beta_i - \frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right|^3}{\left(\sum_{i=1}^{n(1-\alpha)}\left(\frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right)^2 + \sum_{i=1}^{n\alpha}\left(\beta_i - \frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right)^2\right)^{3/2}}. \tag{13}$$

We now derive a lower bound for this expression. From $|\beta_i| \leq 1$, the denominator satisfies

$$\sum_{i=1}^{n(1-\alpha)}\left(\frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right)^2 + \sum_{i=1}^{n\alpha}\left(\beta_i - \frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right)^2 = \sum_{i=1}^{n\alpha}\beta_i^2 - \frac{1}{n}\left(\sum_{j=1}^{n\alpha}\beta_j\right)^2 \leq n\alpha.$$

For the numerator, using $|u+v|^3 \leq |u|^3 + |v|^3$ we have

$$|\beta_i|^3 \leq \left|\beta_i - \frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right|^3 + \left|\frac{1}{n}\sum_{i=1}^{n\alpha}\beta_i\right|^3,$$

and as a consequence,

$$\begin{aligned}\sum_{i=1}^{n(1-\alpha)}\left|\frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right|^3 + \sum_{i=1}^{n\alpha}\left|\beta_i - \frac{1}{n}\sum_{j=1}^{n\alpha}\beta_j\right|^3 &\geq \frac{n(1-\alpha)}{n^3}\left|\sum_{i=1}^{n\alpha}\beta_i\right|^3 + \sum_{i=1}^{n\alpha}|\beta_i|^3 - \frac{n\alpha}{n^3}\left|\sum_{i=1}^{n\alpha}\beta_i\right|^3 \\ &= \frac{n(1-2\alpha)}{n^3}\left|\sum_{i=1}^{n\alpha}\beta_i\right|^3 + \sum_{i=1}^{n\alpha}|\beta_i|^3.\end{aligned}$$

As at least one of the β_i must be equal to one, a (rough) bound for this expression is given by

$$\sum_{i=1}^{n(1-\alpha)} \left| \frac{1}{n} \sum_{j=1}^{n\alpha} \beta_j \right|^3 + \sum_{i=1}^{n\alpha} \left| \beta_i - \frac{1}{n} \sum_{j=1}^{n\alpha} \beta_j \right|^3 \geq 1.$$

Finally, replacing both bounds in the criterion we obtain for d_∞,

$$\varphi(d_\infty) \geq \frac{1}{n} \frac{1}{\alpha^{3/2}}. \tag{14}$$

Note that the bound is not tight.

If we now analyze the projections onto a direction d_b where all observations take bounded values, there is no need to differentiate between outliers and the original observations, as all of them may take arbitrary, although bounded, values. We now denote by z_i the projections for all the observations, both the original ones and the outliers. Due to the affine equivariance of the procedure, we may assume without loss of generality that $\sum_{i=1}^{n} z_i = 0$ and $\sum_{i=1}^{n} z_i^2 = 1$. Under these conditions, the criterion (12) takes the value

$$\varphi(d_b) = \sqrt{n} \sum_{i=1}^{n} |z_i|^3, \tag{15}$$

and as under the preceding conditions $|z_i| \leq 1$, an upper bound would be

$$\varphi(d_b) \leq \sqrt{n}. \tag{16}$$

Combining (14) and (16), the values of α for which the proposed algorithm will be able to identify the outliers are those satisfying

$$n\alpha \leq 1. \tag{17}$$

This bound for the breakdown point does not depend on the dimension of the sample space, but behaves poorly with respect to n, the sample size.

Given the derivation procedure, and in particular that the result (17) has been obtained from bounds that are not tight, it is reasonable to prove the validity of this result by showing a particular contamination pattern that presents this kind of behavior.

Consider a point-mass contamination, and assume that we have $n(1-\alpha)$ observations x_i, the uncontaminated sample, such that $\sum_{i=1}^{n(1-\alpha)} x_i = 0$ and $\sum_{i=1}^{n(1-\alpha)} x_i x_i^t = I$ (again, note that the procedure is affine equivariant), and $n\alpha$ observations (the outliers) concentrated at δe_1, where e_1 denotes the first unit vector.

For the projections along a direction d_b orthogonal to e_1 the criterion to optimize will be given by

$$\varphi(d_b) = \sqrt{n} \sum_{i=1}^{n(1-\alpha)} |z_i|^3 \leq \sqrt{n},$$

with $z_i \equiv d_b^T x_i$. Note that although this bound is not tight, there are distributions for the uncontaminated sample that satisfy asymptotically the bound as $n \to \infty$, such as for example having one isolated sample point, and all other uncontaminated sample mass concentrated at a different point.

Consider now the projections onto e_1. The value of the criterion is

$$\varphi(e_1) = \sqrt{n} \frac{\sum_{i=1}^{n(1-\alpha)} |z_i - \alpha\delta|^3 + n\alpha\delta^3(1-\alpha)^3}{\left(\sum_{i=1}^{n(1-\alpha)} (z_i - \alpha\delta)^2 + n\alpha\delta^2(1-\alpha)^2 \right)^{3/2}}.$$

If we let $\delta \to \infty$ and simplify the resulting expression, we have

$$\varphi(e_1) = \frac{\alpha^2 + (1-\alpha)^2}{\sqrt{\alpha(1-\alpha)}}.$$

The condition under which the maximizing direction is e_1 would be $\varphi(e_1) \geq \varphi(d_b)$, and this condition will hold if

$$\frac{\alpha^2 + (1-\alpha)^2}{\sqrt{\alpha(1-\alpha)}} \geq \sqrt{n}.$$

From this bound we find again the relationship $\alpha = O(n^{-1})$ (although with a different constant). This result confirms that for this procedure the breakdown point presents the expected type of behavior with respect to the sample size.

Nevertheless, note that this behavior is associated with very unusual dispositions for the uncontaminated sample. For example, for the case of point-mass contamination, if the uncontaminated sample would have zero mean, variance equal to one and a third moment bounded with respect to n (let us say, bounded by k), then the breakdown point would be given by an expression of the form

$$\alpha \leq \left(\frac{n}{k} \frac{\alpha^2 + (1-\alpha)^2}{\sqrt{(1-\alpha)}} \right)^2.$$

References

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 E. Stahel, Basel: Birkhäuser.

Atkinson, A.C. (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.

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

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

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

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

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

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., 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.

Jones, M.C. and Sibson, R. (1987), "What is Projection Pursuit?," *Journal of the Royal Statistical Society A*, 150, 29-30.

Malkovich, J.F. and Afifi, A.A. (1973), "On Tests for Multivariate Normality," *Journal of the American Statistical Association*, 68, 176-179.

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.

Posse, C. (1995), "Tools for Two-Dimensional Exploratory Projection Pursuit," *Journal of Computational and Graphical Statistics*, 4, 83-100.

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

Rocke, D.M. and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047-1061.

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

Rousseeuw, P.J. (1993), "A Resampling Design for Computing High-Breakdown Point Regression," *Statistics and Probability Letters*, 18, 125–128.

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.

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.