ELSEVIER

Journal of Statistical Planning and Inference 57 (1997) 245–255

journal of statistical planning and inference

Robust estimation of multivariate location and shape

David M. Rocke*, David L. Woodruff

Graduate School of Management, University of California, Davis, CA 95616, USA

Received October 1994; revised May 1995

Abstract

In this paper, we describe an overall strategy for robust estimation of multivariate location and shape, and the consequent identification of outliers and leverage points. Parts of this strategy have been described in a series of previous papers (Rocke, *Ann. Statist.*, in press; Rocke and Woodruff, *Statist. Neerlandica* **47** (1993), 27–42, *J. Amer. Statist. Assoc.*, in press; Woodruff and Rocke, *J. Comput. Graphical Statist.* **2** (1993), 69–95; *J. Amer. Statist. Assoc.* **89** (1994), 888–896) but the overall structure is presented here for the first time. After describing the first-level architecture of a class of algorithms for this problem, we review available information about possible tactics for each major step in the process. The major steps that we have found to be necessary are as follows: (1) partition the data into groups of perhaps five times the dimension; (2) for each group, search for the best available solution to a combinatorial estimator such as the Minimum Covariance Determinant (MCD) — these are the preliminary estimates; (3) for each preliminary estimate, iterate to the solution of a smooth estimator chosen for robustness and outlier resistance; and (4) choose among the final iterates based on a robust criterion, such as minimum volume. Use of this algorithm architecture can enable reliable, fast, robust estimation of heavily contaminated multivariate data in high (>20) dimension even with large quantities of data. A computer program implementing the algorithm is available from the authors.

AMS classification: 62H12

Keywords: M-estimator; Minimum covariance determinant (MCD); Minimum volume ellipsoid (MVE); S-estimator

. Introduction

The estimation of multivariate location and shape is one of the most difficult problems in robust statistics. 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 outlier-free predictors (Huber, 1981). The multivariate location and shape problem is more difficult, since most known methods will break down if the fraction of outliers is larger than

* Corresponding author. Tel. 916 752 7368; Fax. +1 916 752 2924; e-mail: dmrocke@ucdavis.edu.

$1/(p+1)$, where p is the dimension of the data (Maronna, 1976; Donoho, 1982; 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* (Grübel and Rocke, 1990). A location estimator $\boldsymbol{t}_n \in \mathcal{R}^p$ is affine equivariant if and only if for any vector $\boldsymbol{b} \in \mathcal{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$,

$$\boldsymbol{t}_n(\boldsymbol{AX} + \boldsymbol{b}) \quad \boldsymbol{A}\boldsymbol{t}_n(\boldsymbol{X}) + \boldsymbol{b}.$$

A shape estimator $\boldsymbol{C}_n \in \mathrm{PDS}(p)$ is affine equivariant if and only if for any vector $\boldsymbol{b} \in \mathcal{R}^p$ and any non-singular $p \times p$ matrix $\boldsymbol{A}$,

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

This implies, for example, that stretching or rotating measurement scales will not change the estimates.

Computational methods have been reported in the literature for a number of approaches for finding robust estimates of multivariate location and shape (and therefore identifying outliers). Combinatorial estimators, such as the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators (Rousseeuw, 1985; Hampel et al., 1986; Rousseeuw and Leroy, 1987; Rousseeuw and van Zomeren, 1990, 1991) have been addressed with random search (Rousseeuw and Leroy, 1987), steepest descent with random restarts (Hawkins, 1993, 1994), and heuristic search optimization efforts (Rocke and Woodruff, 1993, 1995, 1996; Woodruff and Rocke, 1993, 1994). Iterative estimators such as maximum likelihood and M-estimators (Campbell, 1980, 1982; Huber, 1981; Kent and Tyler, 1991; Lopuhaä, 1992; 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). Sequential point addition estimators have been defined algorithmically by Atkinson (1993, 1995; Atkinson and Mulira, 1993) and Hadi (1992) working separately.

In this paper we review previous work on robust estimation and multivariate outlier detection. We describe an overall strategy for robust estimation of multivariate location and shape, and the consequent identification of outliers and leverage points. Parts of this strategy have been described in a series of previous papers (Rocke, 1996; Rocke and Woodruff, 1993, 1996; Woodruff and Rocke, 1993, 1994), but the overall structure is presented here for the first time.

2. A basic architecture for algorithms

In the following sections, we describe a basic structure for a class of methods of estimation for multivariate location and shape. The criteria on which this structure is based includes the following:

Affine equivariance: The solution should not depend on, for example, the measurement scales used for the variables.

Robustness: The method should be capable of obtaining a solution that reflects the majority of the data, even when a large fraction of the data are disparate.

Statistical efficiency: The solution should be reasonably statistically efficient, subject to robustness.

Computational practicality: The solution should be obtainable in a practical amount of computing time, even in high dimension or with large amounts of data.

Although these criteria are loosely defined, they limit the choice of methods severely, since many techniques fail one or more of the tests by any interpretation of the criteria.

Our research has indicated that several properties seem important for a method to be successful in estimation of multivariate location and shape, at least in higher dimension and with larger sample sizes (Rocke, 1996; Rocke and Woodruff, 1993, 1995, 1996; Woodruff and Rocke, 1993, 1994). These steps, which are explained in more detail in Sections 3–6, are as follows:

1. Partition the data into groups of a size that does not rise with the sample size. Our current practice is to use 5 times the dimension for the group size (Woodruff and Rocke, 1994).

2. Within each group, search for the best available solution to a combinatorial estimator such as the minimum covariance determinant (MCD). The collection of the best solutions found comprise the preliminary estimates for the next stage (Rocke and Woodruff, 1996; Woodruff and Rocke, 1994).

3. For each preliminary estimate, iterate in the entire data set to the solution of a smooth estimator chosen for robustness and outlier resistance. This means that the smooth estimator should have both a high breakdown and should downweigh points that are clear outliers, criteria that may seem to overlap but are not redundent (Rocke, 1996).

4. Choose among the final iterates based on a robust criterion.

Statistical efficiency considerations outlined later dictate that the final stage should be the solution of a smooth estimator such as an S- or M-estimator. Simple starting points for the iteration, such as the sample mean and sample covariance matrix, do not necessarily guarantee convergence to a high breakdown point solution to the estimating equations. 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 use of a preliminary combinatorial estimator.

We have examined both the methods of Ruppert (1992) and our own versions of direct search as a method of finding the good root for S- or M-estimation and have found that it seems superior to use a preliminary combinatorial estimator such as the MCD. The term *good root* here means that root of the estimating equations which has the minimum value of an auxiliary criterion (Clarke, 1983). The MCD, although the best of the preliminary estimates that we have tried, presents severe computational difficulties (Woodruff and Rocke, 1994).

3. Partitioning

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. That is, doubling the number of data points and doubling the amount of computation time can result in poorer estimates, unless partitioning is employed. 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 $5p$ (Rocke and Woodruff, 1994). This allowed the acquisition of good solutions with high probability with a computational time increasing only linearly with n (instead of exponentially).

Specifically, the proposal is that, in any preliminary combinatorial search, the data be partitioned into cells whose size does not grow with n, and the collection of solutions to these problems be used as starting points for further iteration. In our computational experience, attempting combinatorial search on the entire data set results in rapid degradation of performance with increasing sample size.

4. Combinatorial methods

Combinatorially defined estimators form an important part of the estimation strategy. Much earlier work was based on the minimum volume ellipsoid (MVE) estimator which is defined as the center and shape matrix of that ellipsoid from a given class of ellipsoids that has a minimum volume, subject to containing a fixed fraction of the data (Rousseeuw, 1985). We choose to refer to this estimate as combinatorial, since it is combinatorial in the sense that it is determined by the choice of the subset of points that are contained in the ellipsoid. However, the MVE is an $n^{-1/3}$ estimator (Davies, 1992), so has asymptotic efficiency zero. A more effective candidate is the minimum covariance determinant estimator (MCD), which is defined as the mean and covariance of that subset of the data of prescribed size such that the determinant of the covariance matrix is smallest. The MCD is an $n^{-1/2}$ estimator, so has better statistical properties than the MVE (Butler, et al., 1993). Recent work has shown that the MCD is much more effective in practice than the MVE, so our work has focused on this estimator (Woodruff and Rocke, 1994; Rocke and Woodruff, 1996). Fig. 1 shows the asymptotic location efficiencies for multivariate normal data for the MCD, a biweight S-estimator, and a translated biweight M-estimator (Rocke, 1996) as a function of the dimension. The MCD suffers some efficiency loss compared to the smooth estimators, but may still be acceptable under some cirumstances.

Fig. 2 shows the finite-sample location efficiencies of the MVE, MCD, the median, and the biweight S-estimator for normal data in dimension 1, where the estimators

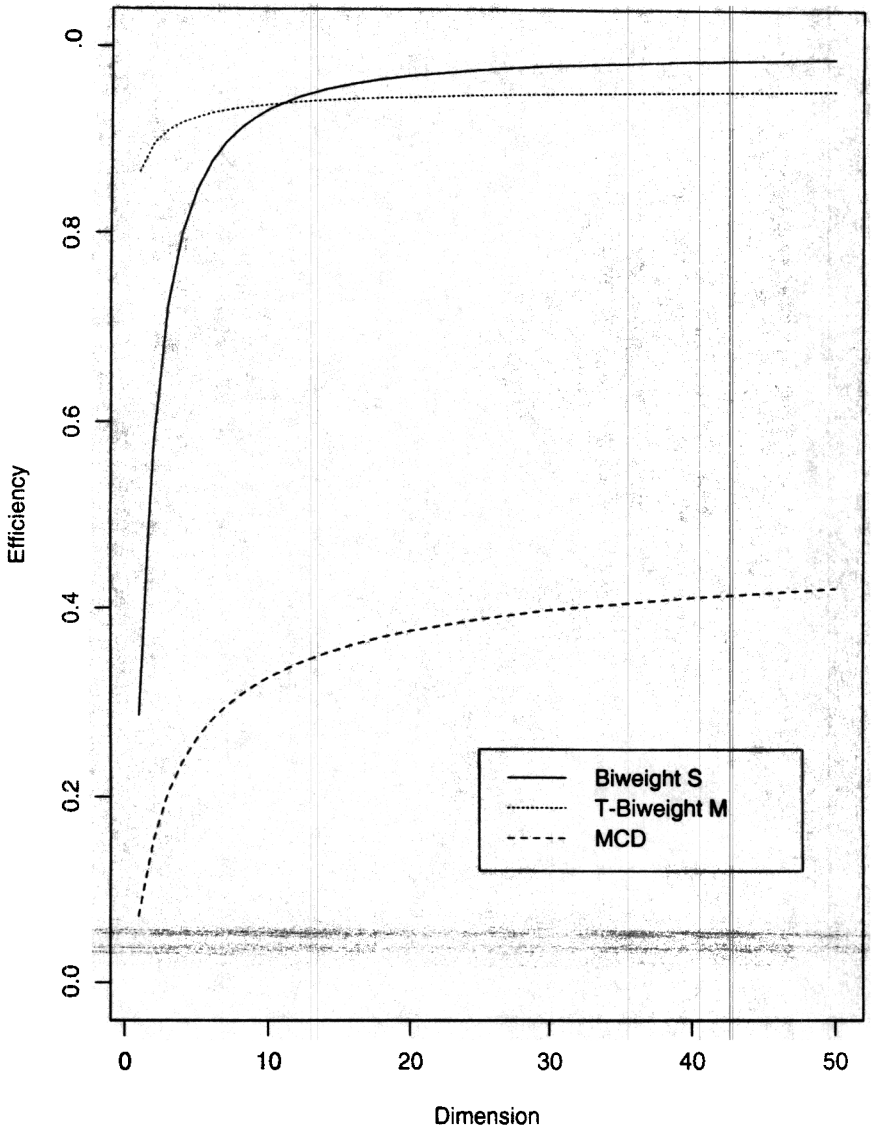

Fig. 1. Asymptotic efficiency for the location component of three estimators of multivariate location and shape as a function of the dimension.

can be computed exactly by enumerating intervals. The plot shows the log of the inefficiency (reciprocal efficiency) against the log of the sample size. An exact $n^{-1/2}$ estimator should have a level plot, whereas an exact $n^{-\eta}$ estimator, with $\eta < \frac{1}{2}$, should have a plot that is a straight line sloping up. The extreme inefficiency of the MVE can be seen, as well as the slow convergence of the MCD to its asymptotic behavior. On this standard, S- and M-estimators look attractive by comparison.

It should also be noted that the use of the MCD as a final estimator could not be recommended, even for outlier identification, for an important reason — computational methods are not known that can find good MCD estimates in contaminated data in high dimension and large sample sizes. Otherwise stated, the asymptotics of the *exact* MCD are only relevant if the exact MCD or something close to it can be computed, and there is no known method of accomplishing this objective in high dimension and/or large sample size.

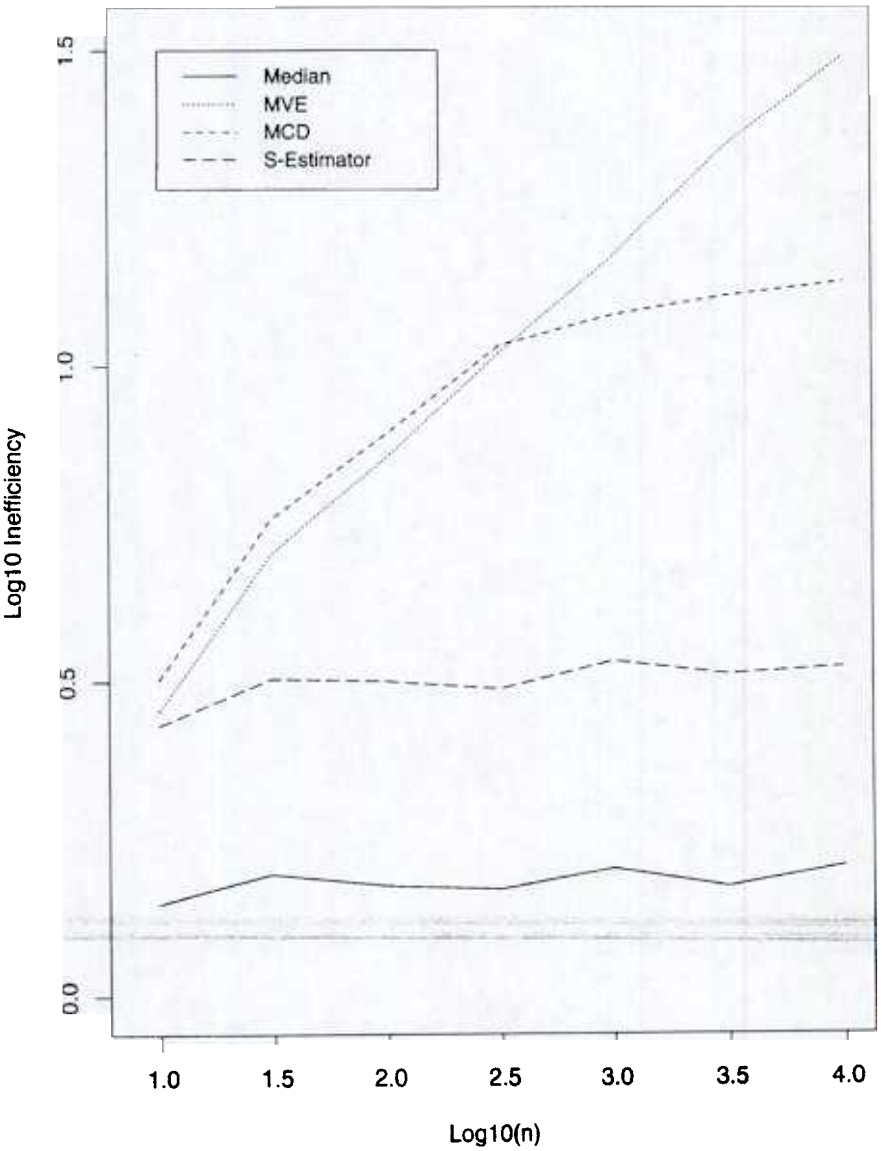

Fig. 2. Finite sample inefficiency of the location component of four estimators of location and shape in dimension 1 as a function of the sample size.

5. Smooth methods

The use of smooth methods is important for two reasons. First, the solution once obtained is much more statistically efficient. Second, a reasonable starting point will lead reliably to a good solution of the smooth estimator, whereas it is of little help in finding a better combinatorial estimator.

An S-estimate of multivariate location and shape is defined as that vector t and PDS matrix C which minimizes $|C|$ subject to

$$n \quad \sum \rho\left([(x_i - t)^{\mathrm{T}} C \quad (x_i - t)]^{1/2}\right) = b_0 \tag{1}$$

which we write as

$$n^{-1} \sum \rho(d_i) \quad b_0. \tag{2}$$

It has been shown by Lopuhaä (1989) that S-estimators are in the class of M-estimators with standardizing constraints with weight functions $v_1(d) = w(d)$, $v_2(d) = pw(d)$, $v_3(d) = v(d)$, where $\psi(d) = \rho'(d)$, $w(d) = \psi(d)/d$, $v(d) = \psi(d)d$, with constraint (2).

In Rocke (1996) it is shown that S-estimators in high dimension can be sensitive to outliers even if the breakdown point is set to be near 50%. We utilize the translated biweight (or t-biweight) M-estimation method defined in 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. Specifically,

$$w_t(d;c,M) = \begin{cases} & 0 \leqslant d < M, \\ -((d-M)/c)^2)^2, & M \leqslant d \leqslant M+c, \\ 0, & d > M+c, \end{cases}$$

$$\psi_t(d;c,M \quad \begin{matrix} d, & 0 \leqslant d < M, \\ d(1-((d-M)/c)^2)^2, & M \leqslant d \leqslant M+c, \\ 0, & d > M+c, \end{matrix}$$

$$\rho_t(d;c,M) \quad \begin{cases} d^2/2, & 0 \leqslant d < M, \\ M^2/2 - M^2(M^4 - 5M^2c^2 + 15c^4)/(30c^4) & \\ \quad + d^2(1/2 + M^4/(2c^4) - M^2/c^2) & \\ \quad + d^3(4M/(3c^2) - 4M^3/(3c^4)) & \\ \quad + d^4(3M^2/(2c^4) - 1/(2c^2)) & \\ \quad - 4Md^5/(5c^4) + d^6/(6c^4), & M \leqslant d \leqslant M+c, \\ M^2/2 + c(5c + 16M)/30, & d > M+c, \end{cases}$$

where we set M to the match the median of a χ^2_p to the median of the data, and $M+c$ is set to the 0.99 point of a χ^2_p. This estimator is shown in Rocke (1996) to be more outlier resistant in high dimension than typical S-estimators. The specifics of the iteration for both M- and S-estimators is given in Rocke and Woodruff (1996).

In accord with the theory in Rocke (1996), we have found that the use of the t-biweight M-estimator makes a large improvement in the performance of the hybrid algorithm compared to the use of biweight S-estimation. Some detailed evidence is given in Table 1 (Rocke and Woodruff, 1996). The situation here is that 20 replicates of shift outliers with indicated sample size, fraction of outliers, and computation time allowed (all computation times are CPU seconds on a DECStation 5000/200). 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.

Table 1
Comparison of biweight S-estimation with t-biweight M-estimation

n	α	Time (s)	Biweight (%)	t-biweight (%)
50	0.30	3	5	50
50	0.30	30	5	70
50	0.35	3	0	20
50	0.35	30	0	25
200	0.30	9	55	95
200	0.30	36	55	95
200	0.35	9	0	35
200	0.35	36	0	55

Another feature of this table 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. With 200 points in dimension 10, 30% outliers are quite managable, whereas 35% outliers are much more difficult. Seemingly, all robust multivariate estimation methods behave this way. There seems to be a critical level of contamination for any particular method and type of problem such that, when the contamination is much below that critical level, success is frequent, and when contamination is much above that level, success is rare.

6. Selection criterion

Given a set of solutions to the equations defining a smooth estimator, it remains to be decided which to use. For S-estimators, this is automatic — one uses the solution with the best value of the criterion function. For M-estimators, this is hardly more difficult. To avoid implosion, all such M-estimators require a constraint — we use the contraint of matching the median of the Mahalanobis square distances to the median of a χ^2_p. We can then use the candidate whose estimated 50% probability content ellipsoid has the smallest volume (the same one is chosen if any other content is used). This is equivalent to choosing the solution whose shape matrix determinant is smallest.

7. Overall performance

This section is devoted to the following question: for what dimensions, sample sizes, outlier distances, fractions of outliers, and computation times is the algorithm described in this paper effective? The theoretical results in Woodruff and Rocke (1994) demonstrate that any amount of contamination less than 50% can theoretically be handled with sufficient data and sufficient processing time. Here we ask a different question: what amount of contamination can be practically detected with an amount of data that is given and with practical processing times.

Table 2 shows some results (Rocke and Woodruff, 1996). For each indicated combination of dimension and outlier distance, a generalized linear model was fit to a

Table 2
Critical contamination level for 90% success with the hybrid algorithm

p	d	n	Time (s)	α
	2	50		0.27
	2	100		0.30
	2	200		0.33
	4	50		0.35
	4	100		0.37
	4	200		0.39
	2	200		0.26
	2	400		0.28
	2	800		0.30
	4	200		0.27
	4	400		0.32
	4	800		0.35

series of experiments. For each data set, the estimation method may succeed, so that the 'good root' was found and all outliers successfully identified. If not, the 'bad root' (obtainable by iterating from a starting value of the sample mean and covariance matrix) was found, and the process failed. Then the level of contamination 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, with 25–35% in dimension 20, and 20–25% in dimension 40 (not shown). Although these bounds are crude, it does give some feel for what problems are feasible.

A point that should not be overlooked is that advances in processor technology and parallel processing can have an important effect. For example, the DEC 3000/700 Alpha AXP workstation we used for the computations in this paper is about 10 times faster than the DECStation 5000/200 on which the simulations in Woodruff and Rocke (1994) were conducted, and multiple processor machines could also be used to multiply the effectiveness of the algorithm, which is parallelizable in a number of ways (Woodruff and Rocke, 1994). Further improvements in the algorithms, as well as the use of higher speed or parallel computers should extend the domain of performance of the algorithm to larger contamination and higher dimension.

Note on Software

A computer implementation (MULTPART) of the algorithms described in this paper is available from the authors, who may be contacted at dmrocke@ucdavis.edu and dlwoodruff@ucdavis.edu. The programs are written in the C programming language. Also available is a program to generate the instances used to test different features of the algorithm. These programs are also available from STATLIB.

Acknowledgements

The research reported in this paper was supported by grants from the National Science Foundation (DMS 93-01344, DMS 94-06193, DMS 95-10511) and the National Institute of Environmental Health Sciences, National Institutes of Health (P42 ES04699).

References

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

Atkinson, A.C. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers. In: S. Morgenthaler, E. Ronchetti, and W. Stahel, Eds., *Data Analysis and Robustness*, Birkhäuser, Basel.

Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers. *J. Amer. Statist. Assoc.* **89**, 1329–1339.

Atkinson, A.C. and H.-M. Mulira (1993). The Stalactite plot for the detection of multiple outliers. *Statist. Comput.* **3**, 27–35.

Butler, R.W., P.L. Davies and M. Jhun (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.* **21**, 1385–1400.

Campbell, N.A. (1980). Robust procedures in multivariate analysis I: robust covariance estimation. *Appl. Statist.* **29**, 231–237.

Campbell, N.A. (1982). "Robust procedures in multivariate analysis II: robust canonical variate analysis. *Appl. Statist.* **31**, 1–8.

Clarke, B.R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.* **11**, 1196–1205.

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

Davies, P.L. (1992). Asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20**, 1828–1843.

Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D. Thesis, Department of Statistics, Harvard University.

Grübel, R. and D.M. Rocke (1990). On the cumulants of affine equivariant estimators in elliptical families. *J. Multivariate Anal.* **35**, 203–222.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. JRSSB **54**, 761–771.

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

Hawkins, D.M. (1993). A feasible solution algorithm for the minimum volume ellipsoid estimator. *Comput. Statist.* **9**, 95–107.

Hawkins, D.M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Comput. Statist. Data Anal.* **17**, 197–210.

Hawkins, D.M., D. Bradu and G.V. Kass (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* **26**, 197–208.

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

Kent, J.T. and D.E. Tyler (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19**, 2102–2119.

Lopuhaä, H.P. (1989). On the Relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17**, 1662–1683.

Lopuhaä, H.P. (1992). Highly efficient estimators of multivariate location with high breakdown point. *Ann. Statist.* **20**, 398–413.

Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51–67.

Rocke, D.M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Ann. Statist.*, in press.

Rocke, D.M. and D.L. Woodruff (1993). Computation of robust estimates of multivariate location and shape. *Statist. Neerlandica* **47**, 27–42.

Rocke, D.M. and D.L. Woodruff (1996). Identification of outliers in multivariate data. *J. Amer. Statist. Assoc.*, in press.

Rocke, D.M. and D.L. Woodruff (1995). Multivariate outlier identification. *Proc. 26th Symp. on the Interface: Computing Science and Statistics* (in press).

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, T. Vincze and W. Werz, Eds., *Mathematical Statistics and Applications, Vol. B.* Reidel, Dordrecht.

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

Rousseeuw, P.J. and B.C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* **85**, 633–639.

Rousseeuw, P.J. and B.C. van Zomeren (1991). Robust distances: simulations and cutoff values. In: W. Stahel and S. Weisberg, Eds., *Directions in Robust Statistics and Diagnostics, Part II.* Springer, New York.

Ruppert, D. (1992). Computing S-estimators for regression and multivariate location/dispersion. *J. Comput. Graphical Statist.* **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.

Tyler, D.E. (1988). Some results on the existence, uniqueness, and computation of the M-estimates of multivariate location and scatter. *SIAM J. Scientific Statist. Comput.* **9**, 354–362.

Tyler, D.E. (1991). Some issues in the robust estimation of multivariate location and scatter. In: W. Stahel and S. Weisberg, Eds., *Directions in Robust Statistics and Diagnostics, Part II.* Springer, New York.

Woodruff, D.L. and D.M. Rocke (1993). Heuristic search algorithms for the minimum volume ellipsoid. *J. Comput. Graphical Statist.* **2**, 69–95.

Woodruff, D.L. and D.M. Rocke (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *J. Amer. Statist. Assoc.* **89**, 888–896.