STATISTICS IN MEDICINE
Statist. Med. **17**, 2685–2695 (1998)

ON THE CALCULATION OF A ROBUST *S*-ESTIMATOR OF A COVARIANCE MATRIX

N. A. CAMPBELL[1]*, H. P. LOPUHAÄ[2] AND P. J. ROUSSEEUW[3]

[1]*CSIRO Mathematical and Information Sciences, Wembley 6014, WA, Australia*
[2]*Delft University of Technology, 2628 BL Delft, The Netherlands*
[3]*UIA, Vesaliuslaan 24, B-2650 Edegem, Belgium*

SUMMARY

An *S*-estimator of multivariate location and scale minimizes the determinant of the covariance matrix, subject to a constraint on the magnitudes of the corresponding Mahalanobis distances. The relationship between *S*-estimators and *w*-estimators of multivariate location and scale can be used to calculate robust estimates of covariance matrices. Elemental subsets of observations are generated to derive initial estimates of means and covariances, and the *w*-estimator equations are then iterated until convergence to obtain the *S*-estimates. An example shows that converging to a (local) minimum from the initial estimates from the elemental subsets is an effective way of determining the overall minimum. None of the estimates gained from the elemental samples is close to the final solution. © 1998 John Wiley & Sons, Ltd.

INTRODUCTION

Robust estimators of covariance can be used to identify atypical observations, which are then examined more closely to identify the reasons why they are atypical. At the same time, robust estimators provide estimates of means, standard deviations and correlations for the main body of data.

One of the difficulties with calculating useful robust estimates for multivariate problems is that the final result can depend on the initial estimates. A common implementation, referred to as *w*-estimation, is: (i) calculate some initial estimates of means and covariances; (ii) determine the individual Mahalanobis distances of the observations from these means, relative to the covariance matrix; (iii) calculate weights which are related inversely to the magnitudes of the Mahalanobis distances; and (iv) calculate weighted means and covariances. There are two common forms of weight function: one in which the influence of an observation on the means increases linearly for an observation which belongs to the main body of data, and then remains constant; and one in which the influence of an observation is zero for very discrepant observations.

Robust *w*-estimators of covariance (see for example, Hampel *et al.*,[1] p. 283), are effective in detecting grossly atypical observations. However, these estimators will only be effective, in the

* Correspondence to: N. A. Campbell, CSIRO Mathematical and Information Sciences, Wembley 6014, WA, Australia

CCC 0277–6715/98/232685–11$17.50

sense that the final estimates will not be affected by the atypical observations and hence will not breakdown, if the proportion of such observations is less than or equal to $1/v$, where v denotes the number of variables (see for example, Hampel *et al.*,[1] p. 298).

An ideal approach is one in which the resulting means and covariances relate only to the main body of data, and are not affected by the atypical observations. Rousseeuw[2] introduced a high-breakdown estimator of covariance, based on minimizing the volume of the ellipsoid based on roughly half the data. His suggested numerical implementation is to take so-called elemental samples based on subsets of the observations. The means and covariance matrix are calculated for each subset, and the corresponding determinant is scaled by an appropriate quantile of the corresponding Mahalanobis distances. This is repeated for a large number of elemental samples. The means and covariance matrix corresponding to the minimum determinant over the elemental samples are taken as the robust estimates.

S-estimators have been proposed as a generalization of the minimum value ellipsoid procedure (see for example, Lopuhaä and Rousseeuw[3]). The determinant of the covariance matrix is again minimized, this time subject to a constraint on the magnitude of the corresponding Mahalanobis distances. Details are given in Section 2.

Section 3 presents a new numerical implementation for the calculation of robust *S*-estimators. The proposed implementation solves the *S*-estimator equations, based on iterative weighted estimator (*w*-estimator) calculations, using as starting values the means and covariances from the elemental samples. Section 4 gives a heuristic description of how *S*-estimators, and the implementation proposed here, work. The final estimates are shown by numerical example in Section 5 to perform better than those from the elemental sample calculations or from one-step of the *w*-estimator calculations.

2. *S*-ESTIMATORS

The *S*-estimator of multivariate location and covariance is defined as finding $(\hat{\mu}, \hat{\Sigma})$ to minimize $\det(\hat{\Sigma})$ subject to

$$n^{-1} \sum_{m=1}^{n} \rho[\{(x_m - \hat{\mu})^t \hat{\Sigma}^{-1} (x_m - \hat{\mu})\}^{1/2}] = b_0 \tag{1}$$

where $\rho(\cdot)$ is symmetric with continuous derivative, $\psi(\cdot)$, and $\rho(0) = 0$. The choice of b_0, and of ρ and hence ψ, is discussed in Section 3.

Lopuhaä[4] (Section 2.3) has shown that the 'score' equations derived from minimizing (1) can be reduced to the following equations:

$$\hat{\mu} = \sum_{m=1}^{n} w(d_m) x_m \Big/ \sum_{m=1}^{n} w(d_m)$$

and (2)

$$\hat{\Sigma} = \sum_{m=1}^{n} w(d_m)(x_m - \hat{\mu})(x_m - \hat{\mu})^t \Big/ v^{-1} \sum_{m=1}^{n} d_m \psi(d_m)$$

where

$$w(d_m) = \psi(d_m)/d_m$$

and

$$d_m^2 = (x_m - \hat{\mu})^t \hat{\Sigma}^{-1} (x_m - \hat{\mu}).$$

3. CALCULATION OF *S*-ESTIMATORS

Direct implementation of the *w*-estimator equations in (2), starting from the usual means and covariances, may or may not lead to the required minimum solution. The approach iteratively downweights those observations with the largest d_m^2 values. If these correspond to atypical values, then a robust solution will result. This is likely to be the case for gross outliers. However, *w*-estimation will not necessarily identify less marked atypical values (see the example in Section 5).

The approach proposed here is to combine the elemental sampling suggested for the minimum volume ellipsoid solution with the *w*-estimator equations in (2). The steps are:

S1. Take a sample (with or without replacement) of size $v + k$, where k is typically 1 or 2.
S2. Calculate the means and covariance matrix.
S3. Scale the covariance matrix so that the constraint on $\rho(d)$ in (1) is satisfied:
this is done by cycling through the calculation in (1), rescaling the covariance matrix (and hence the Mahalanobis distances), until equality is reached.
S4. Iterate through the *w*-estimator equations in (2) until convergence:
(i) at each stage, the covariance matrix is scaled to ensure that (1) is satisfied, and the determinant is calculated;
(ii) an option exists to stop the calculations if the determinant increases, irrespective of whether the means and covariances are sufficiently close to those from the previous iterations.

Various choices for the rho function in (1) have been proposed in the literature. The calculations reported here are based on the so-called biweight function proposed by Tukey, specifically

$$\begin{aligned} \rho(d_m) &= d_m^2/2 - d_m^4/2c_0^2 + d_m^6/6c_0^4 \qquad |d_m| \leqslant c_0 \\ &= c_0^2/6 \qquad\qquad\qquad\qquad\quad |d_m| > c_0. \end{aligned}$$

The derivative gives a redescending psi (ψ) function which depends only on the tuning constant value at which ψ becomes zero. The psi function is given by

$$\begin{aligned} \psi(d_m) &= d_m\{1 - (d_m/c_0)^2\}^2 \qquad |d_m| \leqslant c_0 \\ &= 0 \qquad\qquad\qquad\qquad\quad |d_m| > c_0. \end{aligned}$$

The value of the tuning constant value, c_0, is chosen here by specifying a cut-off constant, u_0, on the univariate scale as the number of standard deviations from the mean at which ψ becomes zero, and then converting u_0 to a value of c_0 on the chi-squared scale of d_m^2 by using the Wilson and Hilferty transformation (see for example Campbell,[5] p. 476). The constant b_0 is taken as the expected value of $\rho(d_m)$ assuming a multivariate Gaussian distribution. Specifically:

$$\begin{aligned} b_0 = {} & v\chi^2(v+2;c_0^2)/2 - v(v+2)\chi^2(v+4;c_0^2)/2c_0^2 + v(v+2)(v+4)\chi^2(v+6;c_0^2)/6c_0^4 \\ & + (c_0^2/6)\{1 - \chi^2(v;c_0^2)\} \end{aligned}$$

where $\chi^2(v;c_0^2)$ denotes the cumulative distribution for a χ^2 variable on v degrees of freedom.

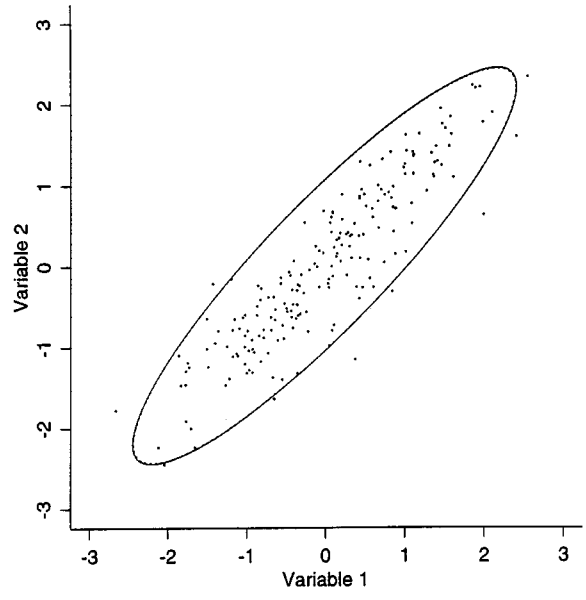

Figure 1. Covariance ellipse calculated from the means and covariance matrix resulting from steps S1–S4 for 200 observations generated from a bivariate Gaussian density with standard deviations of 1, and correlation coefficient of 0·90.

4. HOW *S*-ESTIMATORS WORK

A heuristic description of how and why the proposed *S*-estimator implementation is effective follows.

Figure 1 shows 200 observations generated from a bivariate Gaussian distribution with variances = 1·0 and correlation = 0·9, together with the 95 per cent covariance ellipse calculated from the means and covariance matrix resulting from the *S*-estimator implementation outlined in Section 3.

Figure 2 shows the results of steps S1–S3 for the initial elemental sample depicted by crosses ($\times$); the initial covariance ellipse has roughly the same orientation as that based on all the observations, though the corresponding correlation is somewhat lower. Given that this initial covariance ellipse has roughly the same orientation as the overall covariance ellipse, it would be expected that the majority of observations would fall within the final covariance ellipse resulting from repeated applications of step S3. (In effect, the covariance matrix is scaled so that the resulting Mahalanobis distances for the overall sample are consistent in magnitude with those that would be expected from a multivariate Gaussian sample.) For this example, ten iterations of step S3 are required before the covariance matrix stabilizes to (virtually) the final solution. The majority of observations fall within the final covariance ellipse. The corresponding standard deviations are similar to those based on the overall sample. However, the corresponding correlation coefficient is the same as that based on the initial elemental sample.

Figure 3 shows the results of repeated iterations of step S3 for the initial elemental sample depicted by circles ($\circ$). The initial covariance matrix defines an ellipse which is oriented differently from that based on all the data. Even when the covariance ellipse is inflated until the standard

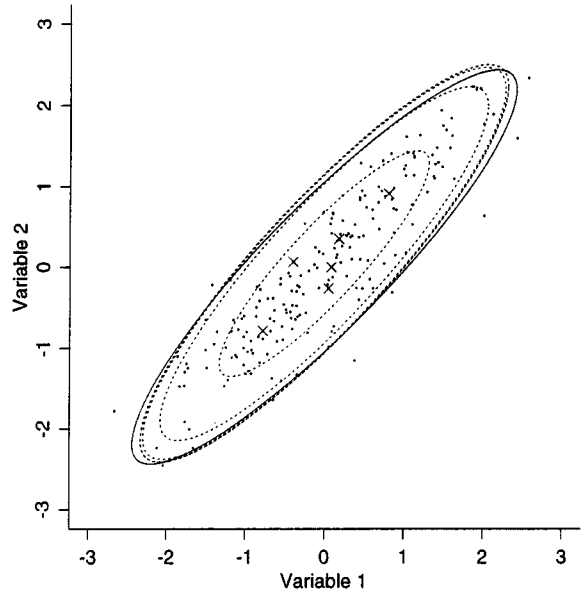

Figure 2. *S*-estimate from small good sample: sequence of covariance ellipses resulting from repeated iterations of step S3 based on the initial elemental sample depicted by crosses (×)
– – – *S*-estimator iterations
—— covariance ellipse for all the data

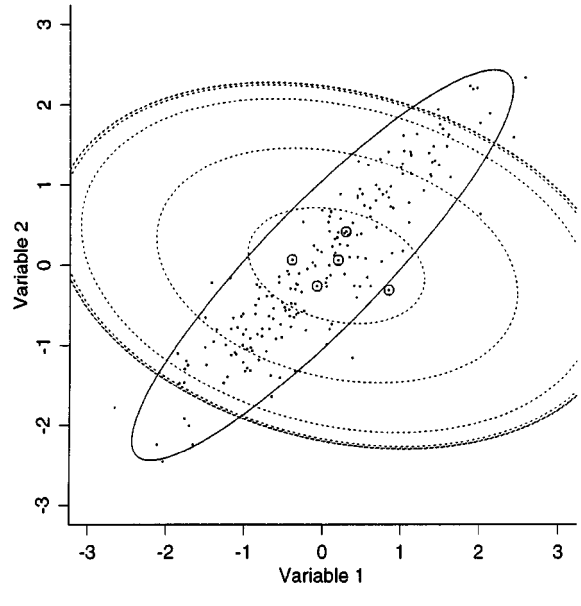

Figure 3. *S*-estimate from small bad sample: sequence of covariance ellipses resulting from repeated iterations of step S3 based on the initial elemental sample depicted by circles (○)
– – – *S*-estimator iterations
—— covariance ellipse for all the data

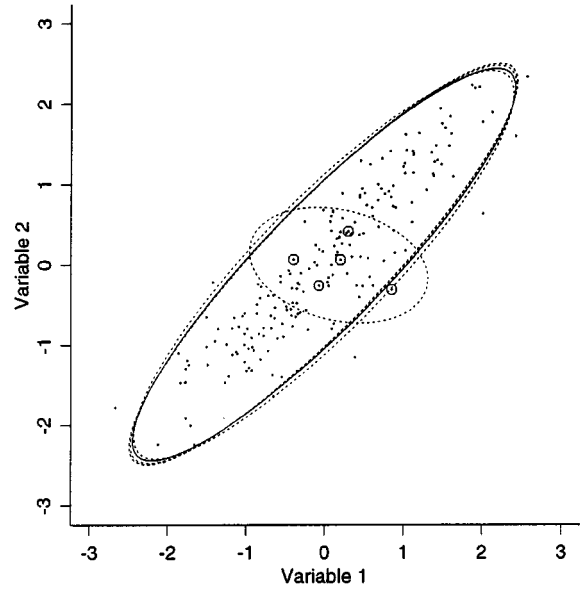

Figure 4. *w*-estimate from small bad sample: sequence of covariance ellipses resulting from repeated iterations of steps S3 and S4 based on the initial elemental sample depicted by circles (○)
– – – *w*-estimator iterations
—— covariance ellipse for all the data

deviations are similar to those based on the overall sample, a considerable number of observations still lie outside the covariance ellipse, and so the sum of the $\rho(d)$ values is still larger than that expected from a reasonable multivariate Gaussian sample. Hence it is necessary to further inflate the covariance matrix until sufficient of the observations lie within the current ellipse so that the sum of the observed $\rho(d)$ values agrees with the sum expected from a Gaussian sample. This results in a covariance matrix which has standard deviations considerably larger than those from the overall sample. The final estimate of correlation is the same as that based on the initial elemental sample.

Clearly, the first solution is to be preferred to the second, since the determinant of the final covariance matrix is smaller for the first solution than it is for the second (since the size of the covariance ellipse is smaller).

However, neither solution is comparable to that based on the overall sample. This is because for the usual implementation based on elemental samples (steps S1–S3), the initial sample effectively defines the magnitude of the final correlation coefficient, so that many elemental samples have to be generated to ensure that one of them has an initial covariance ellipse which is consistent in shape and relative size with that obtained from the covariance matrix for the majority of the observations.

Figure 4 shows the results of steps S1–S4 for the initial elemental sample depicted by circles in Figure 3 (the iterations of step S3 are not shown). After one iteration, the covariance ellipse has roughly the same orientation as that based on the overall sample. After three iterations, the procedure has almost converged to the final solution. Figure 5 shows the weights (in broad

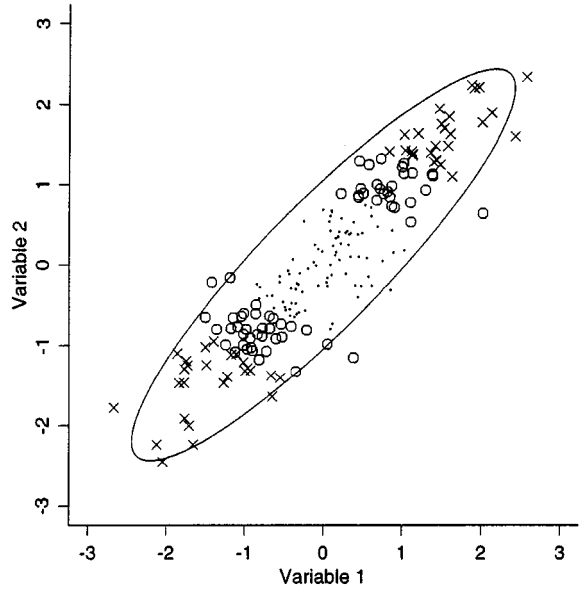

Figure 5. Weights (broad categories) after repeated iterations of step S3 for the elemental sample depicted by circles (○) in Figures 3 and 4

● weights in range 0·95–1·0
○ weights in range 0·85–0·95
× weights in range 0·0–0·85

categories) which result from the first set of repeated iterations of step S3; these provide the initial weights to be input into step S4. Even though the initial elemental sample defines a covariance ellipse with a markedly different orientation to that based on the overall sample, the resulting weights are roughly consistent with those resulting from the elemental sample in Figure 2.

The proposed implementation seems to be successful in determining robust estimates of means and covariances because all that is required is that atypical observations have small initial weights for one or more elemental samples, even though some reasonable observations will also have small weights initially. As the examples show, the weights for the reasonable observations rapidly increase, whereas those for the atypical observations remain small. This provides ready identification of the atypical observations and at the same time provides sensible estimates for the main body of data.

5. AN EXAMPLE

The data sets examined here relate to carapace measurements on the crab *Leptograpsus* (see Campbell and Mahon[6]). The basic data set consists of five variables (front lip, FL; rear width, RW; carapace length, CL; carapace width, CW; body depth, BD) on 50 crabs in each of four groups – males and females for a blue species and an orange species. The data are highly multivariate – correlations within groups are greater than 0·95 (some are as high as 0·99, see Table III).

Table I. Means for the blue males, for the augmenting orange data and for the augmented data sets

Species	n	FL	RW	CL	CW	BD
Blue	50	14·8	11·7	32·0	36·8	13·4
Orange (small)	10	11·9	9·4	23·5	26·1	10·6
Orange (middle)	10	16·6	12·1	33·3	36·5	15·1
Blue + small	60	14·4	11·3	30·6	35·0	12·9
Blue + middle	60	15·1	11·8	32·2	36·8	13·6
Robust						
Blue + small	60	14·3	11·3	30·4	34·8	12·8
Blue + middle	60	14·7	11·5	31·6	36·3	13·2

A canonical variate analysis shows clear separation between the blue and orange species. Huber (personal communication) has shown similar separation using projection pursuit on all 200 observations from 100 random starts.

However, more conventional clustering approaches (including a principal component analysis) do not reveal clusters relating to the species/sex differences. Given the very subtle nature of the differences between the species, two 'contaminated' data sets have been constructed, to see if robust estimation correctly identifies the atypical values.

For the two data sets that have been constructed, the 50 blue males have been augmented by 10 orange males. The first set adds 10 middle-sized specimens while the second adds 10 small specimens. For the 'middle-sized' data set, the augmenting observations lie to the side of the 'bulge' of the ellipsoid, while for the 'small-sized' data set, the augmenting observations lie near the bottom of the ellipsoid.

Table I summarizes the means for the various data sets. (The standard deviations are around 20 per cent of the mean.) Robust w-estimation using the usual means and covariances as starting values does not identify any atypical observations.

Table II summarizes the results of the S-estimator calculations for the 'middle-sized' data set, based on elemental samples of size 6, and a univariate cut-off of 3·5 (which corresponds to around 20 per cent breakdown on the univariate scale). The iterations converge to one of two solutions: that corresponding to the larger of the two final determinants partially downweights three of the augmented observations; that corresponding to the minimum determinant downweights (to zero) all 10 augmented observations.

As Table I shows, the robust means are little changed (and similarly for the standard deviations). The correlation matrix from the robust S-estimation is virtually identical to that for the blue males (since all the augmenting observations have weight zero). The correlations are higher than those for the augmented data set (see Table III).

None of the determinants calculated after step S3 in Section 3 is close to the final determinant, highlighting the difference between the solution gained by minimizing over elemental subsets and that from effective function minimization.

As the number of observations sampled in step S1 is increased, the frequency of the solution with the lower determinant decreases. With 6 observations, 11 of the 30 samples converged to the lower determinant; for 10 observations, 7 converged; for 15 observations, 3 converged; and for 20 observations, 1 converged to the lower determinant.

Table II. Results of the proposed *S*-estimator implementation for the 'middle-sized' data set. Each of the 30 elemental samples generated contains six observations. The univariate cut-off constant is 3·50. 'initial' denotes the determinant of the initial scaled covariance matrix after step S3. 'one-step' denotes the determinant after the first cycle of step S4

Number of iterations	Determinants		
	initial	one-step	final
9	0·170	0·073	0·070
7	0·138	0·074	0·070
8	0·106	0·092	0·091
10	0·158	0·095	0·091
9	0·118	0·093	0·091
12	0·201	0·121	0·091
8	0·102	0·092	0·091
9	0·100	0·071	0·070
11	0·186	0·098	0·091
12	0·132	0·097	0·091
10	0·137	0·094	0·091
8	0·140	0·096	0·091
12	0·337	0·139	0·091
10	0·098	0·092	0·091
8	0·085	0·070	0·070
11	0·129	0·095	0·091
8	0·111	0·093	0·091
9	0·117	0·089	0·070
11	0·171	0·081	0·070
6	0·115	0·071	0·070
8	0·123	0·073	0·070
13	0·220	0·114	0·091
10	0·150	0·106	0·091
6	0·091	0·071	0·070
12	0·122	0·094	0·091
10	0·182	0·098	0·091
9	0·122	0·095	0·091
9	0·126	0·085	0·070
9	0·140	0·095	0·091
7	0·111	0·079	0·070

For the robust calculations on the 'small-sized' data set, based on elemental samples of size 6 and a univariate cut-off of 3·5, a single solution results, with two of the augmenting observations being downweighted; the other eight small orange males tend to be indistinguishable from the blue males.

6. DISCUSSION

The proposed calculation of robust estimators leads to robust estimates of means and covariance matrix with high breakdown point. Though the calculations were motivated by the relationship

Table III. Correlation matrices for the sample of blue males and for the augmented 'middle-sized' data set

	FL	RW	CL	CW	BD
Blue males/robust S-estimates					
FL	1·000	0·969	0·996	0·995	0·993
RW		1·000	0·977	0·978	0·970
CL			1·000	0·999	0·995
CW				1·000	0·996
BD					1·000
Augmented 'middle-sized' data set					
FL	1·000	0·954	0·983	0·966	0·992
RW		1·000	0·976	0·973	0·957
CL			1·000	0·995	0·983
CW				1·000	0·968
BD					1·000

between *S*-estimators and a *w*-estimator solution, it is the use of the elemental sampling to provide initial estimates for the *w*-estimator calculations which is the key to the success of the approach.

Two aspects are essential. The elemental sample size should be small (see step S1), typically no more than one or two more than the number of variables. Increasing the size of the sub-samples decreases the probability of drawing a pure (outlier-free) trial (this probability goes down exponentially with the sub-sample size). This observation is supported by the calculations for the middle-sized example, where larger sub-samples were much less successful in identifying solutions which downweighted the atypical values. The corresponding covariance matrix must then be scaled so that the constraint in (1) is satisfied (see step S3); in the examples studied here, initial scaling factors of around 1·5 were common.

Step S4 also involves a similar scaling to ensure that (1) is satisfied. While this step was included in the calculations reported here, it made little difference after the first iteration. This aspect of the calculations could be speeded up for large sample sizes by calculating a scale equivariant statistic, computed only once, which is used to adjust the covariance matrix, $\hat{\Sigma}$, by the appropriate factor (see (1)). One possibility is to replace (1) by a one-step M-estimate of scale, as suggested by Rousseeuw and Leroy[8] (equation 2.12, p. 174). This could be done by first calculating a measure of scale, s_0, based on the median of the Mahalanobis distances, d_m, and then using a one-step estimate:

$$s_1 = s_0 \left[b_0^{-1} \left\{ n^{-1} \sum_{m=1}^{n} \rho(d_m/s_0) \right\} \right]^{1/2}.$$

Such an estimate is fast to calculate, and will provide a good approximation to the scaling in (1), with the same breakdown properties.

It is clear from the results in Table II (and from related tables for different elemental sample sizes and choices of cut-off constants) that the determinants of the initial scaled covariance matrices after step S3 are much worse than the final iterated results. Clearly, large numbers of elemental samples would need to be generated to achieve the solution obtained from the iterated calculations.

It is also clear from the results of this study that the solution from the first (or one-step) iteration is improved by iterating to the final solution. The initial and final solutions tend to show little relationship. While the smaller one-step solutions generally lead to the best final solution, the results suggest that it is necessary to carry out the local improvement (step S4) for each iteration. The results also indicate that it is adequate to carry out the iterated calculations for only relatively few elemental samples to obtain the optimal solution.

REFERENCES

1. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. *Robust Statistics. The Approach Based on Influence Functions*, Wiley, New York, 1986.
2. Rousseeuw, P. J. 'Multivariate estimation with high breakdown point', in Grossman, W., Pflug, G., Vince, I. and Wertz, W. (eds), *Mathematical Statistics and Applications*, Reidel Publishing Company, Dordrecht, 1986, pp. 283–297.
3. Lopuhaä, H. P. and Rousseeuw, P. J. 'Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices', *Annals of Statistics*, **19**, 229–248 (1991).
4. Lopuhaä, H. P. 'On the relation between S-estimators and M-estimators of multivariate location and covariance', *Annals of Statistics*, **17**, 1662–1683 (1989).
5. Campbell, N. A. 'Mixture models and atypical values', *Mathematical Geology*, **16**, 465–477 (1984).
6. Campbell, N. A. and Mahon, R. J. 'A multivariate study of variation in two species of rock crab of the genus *Leptograpsus*', *Australian Journal of Zoology*, **22**, 417–425 (1974).
7. Huber, P. J. *Robust Statistics*, Wiley, New York, 1981.
8. Rousseeuw, P. J. and Leroy, A. M. *Robust Regression and Outlier Detection*, Wiley, New York, 1987.

